Shapes and patterns of adaptive game-based learning: an experiment

Josine Verhagen, Kidaptive
David Hatfield, Kidaptive
Julie Watson, Kidaptive
Solomon Liu, Kidaptive
Dylan Arena, Kidaptive

Abstract: It is often claimed that adaptive educational games keep the learner more engaged and maximize the learning taking place in the game. We explored these two claims by evaluating adaptive and non-adaptive forms of a pattern- and shape-recognition game for preschoolers. We used a Bayesian IRT model to make this game adapt in real time to the learner's performance. Results indicate that adaptivity led to higher engagement, and we found some evidence of greater learning. We also note some important prerequisites for the success of adaptive games.

Introduction

Adaptive learning, game-based learning, and early learning are all hot topics in educational policy circles: Adaptive learning promises to help educators support the wide variety of learning needs and goals in our current educational system; game-based learning can improve engagement; and early learning sets the stage for future academic success or failure. Adaptive, game-based early learning is also a promising area of research because (a) young children learn naturally through play and (b) young children's abilities develop rapidly, making an adaptive-learning context especially appropriate.

There is ample evidence that game-based learning improves engagement (Steinkuehler, Squire, and Barab, 2012) and that personalization improves learning (e.g., Connor, Morrison, Fishman, Schatschneider, and Underwood, 2007). The few studies of real-time adaptivity in games have found mixed results for learning and engagement (e.g., Núñez Castellar, All, and Van Looy, 2014; Orvis, Horn, and Belavich, 2008; Sampayo-Vargas, Cope, He, and Byrne, 2013); the implementations, curricula, durations, and age groups (no preschoolers) in these studies were vastly different.

Here we describe the design of an iPad-based game designed to help preschoolers learn basic concepts about shapes and patterns, and we compare learning and engagement outcomes for the adaptive and non-adaptive versions of this game.

Making a game real-time adaptive through Bayesian IRT

An educational game can be adaptive in different ways. Learners may be assessed initially and then assigned to a fixed learning path, or their path may change after each "level" of the game. At the extreme is real-time adaptivity, where the game adapts every time a learner completes a challenge. This high level of adaptivity is intended to keep the game at a "Goldilocks" level of difficulty for the learner (neither too easy nor too hard) to optimize engagement and learning.

To achieve the continuous assessment required for real-time adaptivity, we used item response theory (IRT) models (e.g., Embretson & Reise, 2000) from computerized adaptive testing (e.g., van der Linden & Glas, 2010; Wainer et al., 2000). IRT provides the tools to estimate item (i.e., challenge) difficulty and person (i.e., learner) ability on a single scale (see Figure 1). For example, in a basic model, when an item's difficulty and a person's estimated ability are equal, the person has a 50% probability of answering the item correctly; when the person's ability is higher/lower than the item's difficulty, the person is more/less likely to get the answer correct.

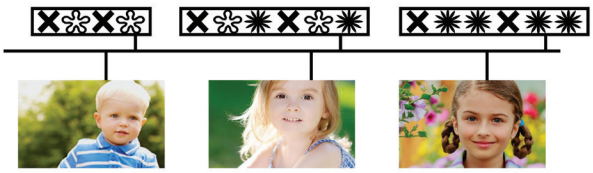


Figure 1: Increasing pattern complexity and learner ability on a single scale.

But games differ in important ways from tests. In tests, people typically answer many items at once, providing a

lot of data about both people's performance and items' difficulties. Educational games typically consist of multiple short interactions with a few responses each. And whereas tests attempt to capture a highly accurate snapshot measurement of a person's ability at a specific moment in time, educational games encourage learning over time, with the result that ability estimates change frequently. These features cause trouble for traditional computerized adaptive testing models: There is too little information to yield an accurate assessment of a learner's ability within one gameplay session, but taking all measurements together to estimate ability leaves no room to monitor growth.

Researchers have developed numerous approaches to address these concerns in computerized adaptive learning (e.g., Eggen, 2012; Klinkenberg, Straatemeier & van der Maas, 2011; Wauters, Desmet & Van Den Noortgate, 2010; 2011). Our approach is to use Bayesian IRT (for more on Bayesian IRT models, see Fox, 2010 and van der Linden & Glas, 2010). In Bayesian IRT, prior knowledge about a person's ability is considered in the estimation process, so previous gameplay results can be taken into account when learners begin a subsequent play session. To monitor change in performance over time, we weighted this prior information less heavily than if the sessions had been played in a single assessment situation. In this way, we could counter the effect of short test length by incorporating additional information, while still allowing measures of ability to change over time.

Study Description

We conducted a field experiment to measure engagement and learning in adaptive and non-adaptive versions of a learning game for preschool-aged children; we also measured learning in a control group not playing either version of the game. We recruited families in the US with a single child between the ages of 2.5 to 4.5 and an iPad 2 or later with wireless Internet access at home. Household income, ethnicity, and parent education varied widely in the sample. Participants and their parents used an iPad app created for this study (described further below) at home to complete pre- and posttests and for all game play.

All participants began by completing a pretest with nine questions about shapes and nine about patterns. Participants were then randomly assigned to the adaptive (n = 44), non-adaptive (n = 47) or control (n = 48) condition, with condition assignment stratified by pretest score. During the following six weeks, participants in the adaptive and non-adaptive conditions were asked to use the iPad app to play games designed to assess and teach shape identification/manipulation and pattern recognition. The six weeks were divided into 18 lessons, with each lesson lasting two or three days. During each lesson, participants were asked to play each of four games at least once, after which they could replay games as often as they wished. After six weeks, all participants were asked to complete an 18-question posttest (a parallel form of the pretest), which resulted in 36 (adaptive), 39 (non-adaptive), and 40 (control) completed posttests. Participants in the control group were given access to the game after completing the posttest.

For both the adaptive and non-adaptive conditions, each game was designed to continuously measure participant ability. For only the adaptive condition, participants were then presented challenges (i.e., items) with an expected 70% probability of correct response. For participants in the non-adaptive condition, challenge difficulty was increased at the beginning of the third and then every other lesson (i.e., every four to five days), regardless of the participant's ability or whether the participant had played once, multiple times, or not at all during that lesson.

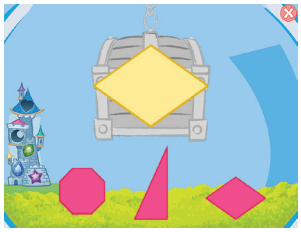
Gameplay

An iPad app was created to advance shape understanding and to teach pattern recognition and extension to preschoolers using two shape games and two pattern games, along with short educational video clips. Within each of these two domains, gameplay was largely the same.

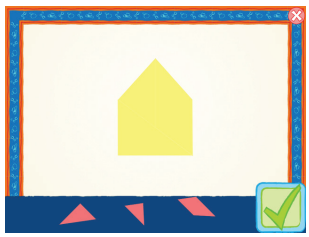
In the two shape games, participants worked on shape identification and manipulation (translation, rotation, scaling, and composition). Participants were shown a set of shapes at the bottom of the screen, varying from simple shapes like a circle or square to less familiar shapes like a pentagon or irregular octagon. In the easiest levels, participants had to match a target shape presented at the top of the screen. In more difficult levels, participants had to identify shapes by name only (e.g., "Tap the square") or rotate shapes to fit an outline. In the hardest levels, participants had to compose an outlined shape "puzzle" from multiple pieces, dragging them to the target area and rotating them to fit.

In the two pattern games, participants were shown a sequence of objects (such as ABAB, ABCABC, or ABBABB) and had to choose the correct object(s) to continue the pattern. At the easiest level, participants saw objects "A" and "B" in an ABAB pattern and asked, "What comes next?" with the choice of another "A" item or an unrelated "C" item. Higher levels had more difficult distractor objects (e.g., both "A" and B" objects in the example above) and asked for multiple pattern elements rather than simply the next element. In the highest levels, participants first

defined their own sequence of objects and then repeated that sequence.

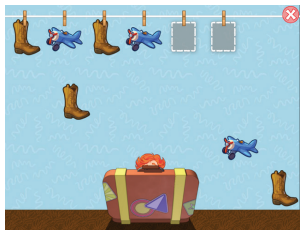


Shapes Game A Level 2
Task: Identify Advanced Shape

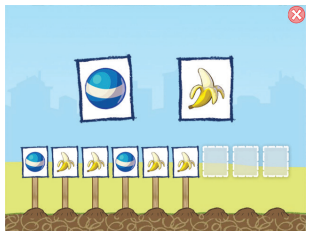


Shapes Game B Level 8
Task: Compose Shape: 3 parts

Figure 2: Screenshots of the two shapes games



Patterns Game C Level 4
Task: Extend AB(_)(_)



Patterns Game D Level 7
Task: Extend ABB(_)(_)(_)

Figure 3: Screenshots of the two patterns games

In all four games, participants were given corrective feedback and hints after incorrect responses, with multiple chances to provide the right answer (although only the first response counted towards ability estimation). In addition, participants were shown short video clips between the games to reinforce the concepts they had just been working on.

Results

Engagement

The first measure of engagement we looked at was the duration of play sessions, defined as the time between opening and closing the app that contained the games. Figure 4 shows the distribution of the duration per play session in minutes for the adaptive and non-adaptive conditions. Because the distributions are skewed (as is often the case with measures of time or duration), we took the log of the durations and assessed whether there was a difference in log(duration) between conditions with a linear mixed model with random effects for the participants. The difference was significant (= 6.10, *df* = 1, *p* = .01). The average duration of play sessions was 10.4 minutes in the adaptive condition and 8.7 minutes in the non-adaptive condition.

Next we looked at the retention of participants over lessons (the 2-3 day periods in which the participants were supposed to play all four games at least once). Because we expected usage to decrease (or decay) exponentially over time, we used log(lesson number) as a predictor, as well as condition and the interaction between log(lesson number) and condition. First, we fit a logistic regression model predicting whether participants in the two conditions played in a certain lesson period. We found a significant decrease in the probability of playing over time (*b* = -.89, *z* = - 9.39, *p* < .001) but no significant negative time-by-condition interaction.

Another way to look at retention is to consider the number of times participants played within each lesson period. Because this is a count (which is naturally very skewed) and because there were many zero counts, we performed a negative binomial regression analysis, with the same predictors as our previous analysis: log(lesson number), condition, and their interaction. There was a significant decrease in number of playthroughs per lesson for the

adaptive condition (b = -.68, z = -7.03, p < .001), but this decrease was steeper for the non-adaptive condition, as indicated by a significant interaction effect (b = -.30, z = -3.393, p < .001). Figure 5 illustrates these results. The dotted lines represent the observed average number of playthroughs in each lesson, and the solid lines represent the number of playthroughs predicted by the negative binomial regression model. Retention is significantly lower in the non-adaptive than in the adaptive condition.

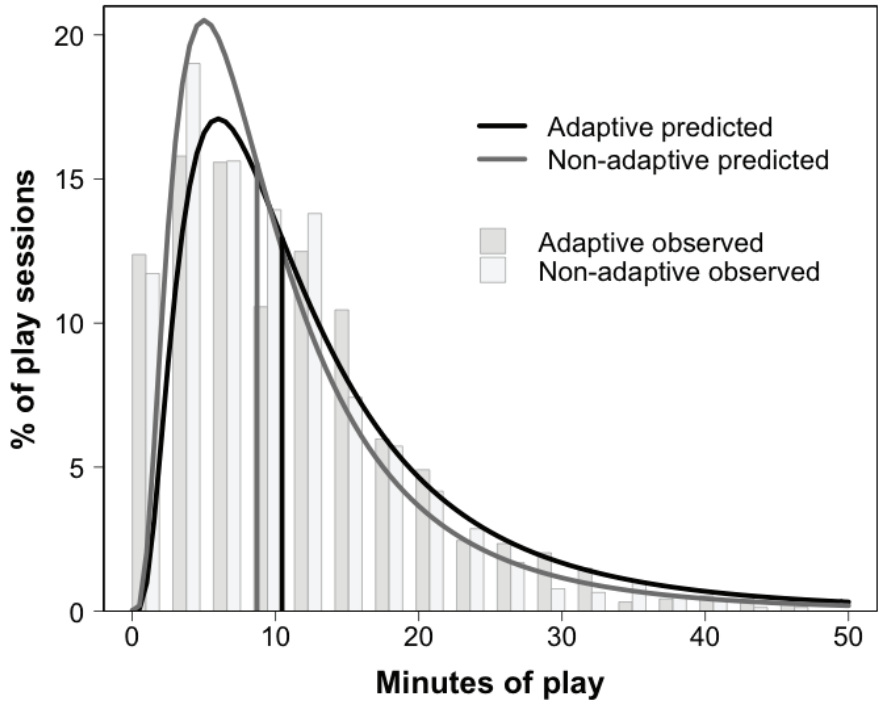


Figure 4: Distribution of play session duration in the adaptive and non-adaptive condition

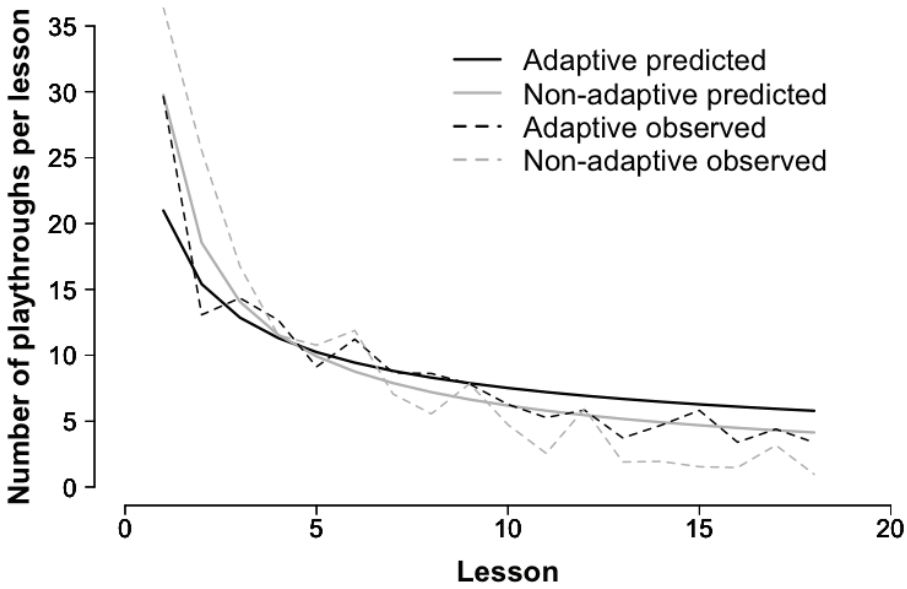


Figure 5: Retention of learners over time: Number of playthroughs in each of the lesson periods.

Learning

First we looked at pre-/posttest score changes, but there were no differences among the adaptive, non-adaptive and control conditions. (We will elaborate more on these results in the discussion.)

Next we looked at in-game performance measures, using ability estimates calculated throughout the study to see how ability estimates at the end of each playthrough changed over time. For both domains (shapes and patterns), we ran a linear mixed model with random effects for the participants, and fixed effects for lesson, condition, and the lesson-by-condition interaction.

Figure 6 presents results for the shapes and patterns domains. In the shapes games, the adaptive and non-adaptive conditions started out at equal ability (b = -.04, t = -.29). There was no increase in ability for the non-adaptive condition (b = .01, t = 2.22), but a positive interaction effect (b = .02 t = 3.521) shows a significant increase in ability for the adaptive condition over lessons (= 12.38, df = 1, p < .001). In the patterns games, the non-adaptive group started out at a slightly but not significantly (= 2.05, df = 1, p = .15) lower level after the first lesson than the adaptive group (b = -.23, t = -1.38). The ability in both conditions increased (= 48.46, df = 1, p < .001) but the interaction effect indicated that the increase in ability over lessons was not different for the *non-adaptive* than for the adaptive condition. We elaborate on possible causes for these mixed results below.

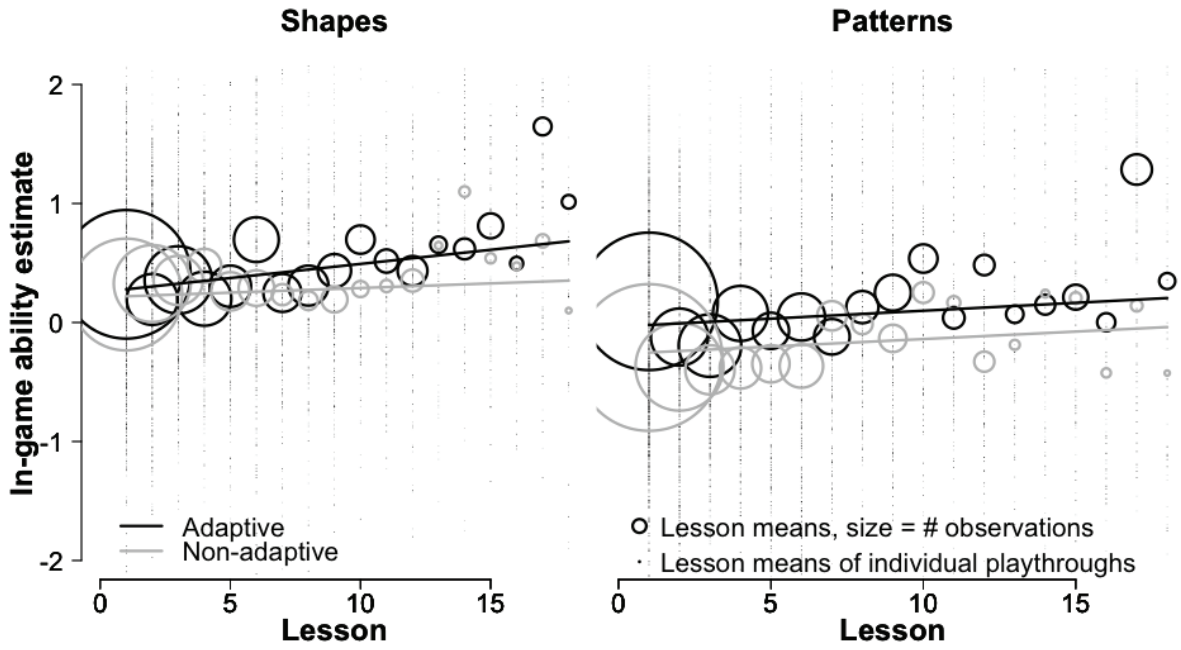


Figure 6: In-game ability estimates as a function of lesson and condition.

Adaptivity

One way to check the adaptive mechanism is to evaluate the percentage of items per playthrough answered correctly during the game and how this changed over lesson periods. In the adaptive condition, this percentage should stay at 70% (our target percentage when matching items to ability estimates). For the non-adaptive condition, the percentage correct should be high in the first lessons and decrease over time, because challenges were designed to increase in difficulty regardless of whether or how well the participant played. The results of a logistic mixed regression confirmed these expectations: The average percentage correct in the adaptive condition was 67% in the patterns game and 74% in the shapes game, and it did not change significantly over lessons. A significant interaction indicated a decrease in percentage correct for the non-adaptive condition in both the patterns (b = -.1,z = -10.92, p < .01) and shapes (b = -.07 z = -7.72, p < .01) domains.

Another way to check the adaptive mechanism is to evaluate whether the adaptive version of the game was able to adapt to the level of the participant adequately. To evaluate this, we looked at the in-game participant-ability estimates and the difficulties of the items offered to the participants in the adaptive condition. Because participants were given items that they were expected to answer correctly with 70% probability, the threshold they needed to reach for moving up in item difficulty was equal to the item's difficulty + 0.5. Figure 7 shows that item-difficulty thresholds (bold lines) and in-game participant-ability estimates (dotted lines) matched quite well for the shapes games. For the patterns games, however, the adaptivity of the games was not optimal: The range of item difficulties for this domain was too narrow, overlapping only with a small percentage of the actual participant abilities. Therefore, participants with very low or high ability got "stuck" on one level of the game.

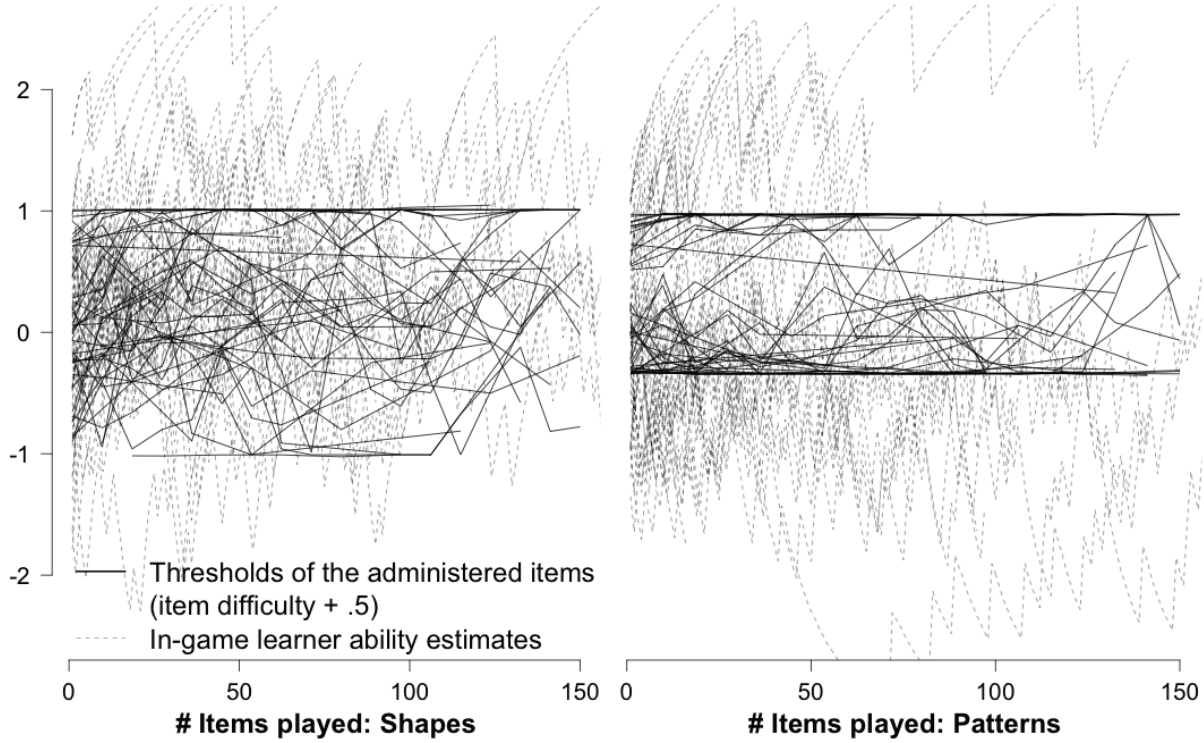


Figure 7: In-game learner ability estimates and item difficulties over time

Discussion

Our results clearly support the claim that adaptivity in educational games leads to more engagement. The participants in the adaptive condition played longer sessions than the participants in the non-adaptive condition, and the number of playthroughs per lesson period decreased more over time for the non-adaptive than for the adaptive condition.

The results are more mixed for the claim that adaptivity leads to greater learning. In the shapes domain, we found evidence of improvements in learning performance in the adaptive condition relative to the non-adaptive condition, but the pre-/posttest measures showed no condition differences, and the patterns domain showed no differences in learning for the two conditions.

The failure of pre-/posttest measures to show any differences is quite possibly due to the low reliability of the instruments with our participants (Cronbach's α ranged from .45 to .69 on the subscales). It is also possible that the instruments were not sensitive enough to overcome the effects of parental "support" in the tests: In a post-study survey, 75% of parents reported having assisted at least a little with either the pre- or posttest, even though they were explicitly instructed not to. Another possible explanation could be that staying focused and completing an 18-item test was simply too ambitious for our preschoolers (e.g., Jones, Rothbart & Posner, 2003).

The learning results of the patterns domain might be explained by the inferior functioning of the adaptive mechanism in the patterns games. Specifically, the range of difficulty levels for the games' challenges did not match the actual abilities of the participants playing the games, so the majority of adaptive-condition participants got "stuck" in either the easiest or hardest game challenges. In the non-adaptive condition, participants received challenges of increasing difficulty regardless of their performance, which could explain why these learners showed equal learning over time.

In conclusion, our study provided evidence of increased engagement in adaptive games and mixed evidence for increased learning. One important lesson from this experiment for designers of adaptive learning games is to include challenges (i.e., items) spanning a wide enough range of difficulties to match the full range of learners' abilities. This requires some form of item calibration (testing challenges with diverse learners to assess their difficulties) during the design of the game so that gaps in challenge difficulty can be filled with appropriately difficult new challenges. Future work on adaptive games could include investigating more efficient ways to calibrate item difficulties in adaptive games, to more quickly achieve the necessary range of challenges without increasing the already substantial work of recruiting learners solely for this process. Another important lesson from this experi-

ment is that adaptive learning games of this kind can be effective even with children as young as three years old. This is particularly important given that children vary tremendously in their levels of skill as well as their rates of development in this age range, and it suggests that more work could be done to develop effective learning games to provide more tailored learning experiences in this key developmental period.

References

Connor, C.M., Morrison, F.J., Fishman, B.J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science*, 315, 464–465.

Eggen, T. J. (2012). Computerized adaptive testing item selection in computerized adaptive learning systems. *Psychometrics in practice at RCEC.* Retrieved from http://doc.utwente.nl/80211/1/Chapter_2.pdf

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Jones, L. B., Rothbart, M. K., & Posner, M. I. (2003). Development of executive attention in preschool children. *Developmental Science*, 6, 498–504.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education,* 57, 1813–1824.

Núñez Castellar, E. P., All, A., & Van Looy, J. (2014). Games as adaptive learning tools: an effectiveness study. In *64rd International Communication Association (ICA) Annual Conference, Pre-Conference: Beyond the Pixels: A Look at Digital Games*.

Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human Behavior,* 24, 2415–2433.

Sampayo-Vargas, S., Cope, C. J., He, Z., & Byrne, G. J. (2013). The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education*, 69, 452–462.

Steinkuehler, C., Squire, K., & Barab, S. A. (2012). *Games, learning, and society: Learning and meaning in the digital age*. New York: Cambridge University Press.

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York, NY: Springer

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., & Steinberg, L. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Erlbaum.

Wauters, K., Desmet, P., & van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning*, 26, 549–562.

Wauters, K., Desmet, P., & van den Noortgate, W. (2011). Acquiring item difficulty estimates: a collaborative effort of data and judgment. *Paper presented at the 4th International Conference of Educational Data Mining.*