Journal of Applied Testing Technology, *Vol 20(S1), 2019*

Designing and Validating a Stealth Assessment for Calculus Competencies

Ginny Smith[1], Valerie Shute[2*] and Amber Muenzenberger[3]

[1]Doctoral Candidate at Florida State University
[2]Mack and Effie Campbell Tyner Endowed Professor of Education, Florida State University, 1114 W. Call Street, Tallahassee, Florida 32306-4453; vshute@fsu.edu
[3]Chief Learning Officer at Triseum

Abstract

The following research focuses on Game-Based Learning (GBL) for assessment through the lens of designing and validating a stealth assessment for the calculus game Variant: Limits™. The process of using Evidence-Centered Design (ECD) to create a valid assessment is highlighted through the development of a competency model and scoring rubrics. A sample of college students enrolled in a Calculus 1 course played the game outside of class during a 2-week period. Comparing the results of students' in-game and external assessments, researchers examined the validity of the stealth assessment measure. The stealth assessment was significantly correlated with the external measures and found to be a valid assessment of students' overall calculus knowledge as defined by the competency model.

Keywords: Assessment, Calculus, Game-Based Learning, Innovative Instruction, Limits, Stealth Assessment

1. Introduction

According to a report from the President's Council of Advisors on Science and Technology (PCAST) (2012), the United States needs more graduates with STEM degrees, as more than 60% of college students intending to enter the STEM fields do not graduate with STEM degrees. Strategies to increase retention in STEM degree programs is a major focus of the report with one main recommendation to improve introductory (i.e., occurring within the first two years of college) STEM courses (PCAST, 2012). Of these introductory courses, Calculus could arguably be the most important. Most STEM degrees require at least one semester of Calculus; therefore, it is considered a gateway course to the STEM fields (Bressoud, Mesa, & Rasmussen, 2015). However, according to the Conference Board of Mathematical Sciences and the American Mathematical Association, the national college failure rates of Calculus 1 are as high as 38% each year (Blair, Kirkman, & Maxwell, 2013). In the 2015-2016 school year, 421,239 students who were enrolled in one of the two AP Calculus courses in the United States (i.e., Calculus AB or Calculus BC) took the AP Calculus exam (College Board, 2016). Of the students enrolled in AP Calculus AB course, 40.6% did not make a high enough score to earn college credit. Of the students enrolled in an AP Calculus BC course, 18.5% scored below the level needed to earn college credit. New methods are needed to engage, teach, and retain Calculus students (Bressoud et al., 2015).

The National Academies of Science, Engineering, and Medicine recently examined undergraduate STEM education, and one of three overarching goals presented in their report for improvement in undergraduate STEM courses was the inclusion of evidence-based learning experiences (2018). A national study by the Mathematical Association America (MAA) and the National Science Foundation (NSF) suggests the importance of promoting higher-order thinking in calculus courses (Bressoud et al., 2015). Game-based learning is an example of a learning experience backed by educational research. Learning games offer an effective educational vehicle for a variety of outcomes, in a variety of content areas, including STEM education (Clark, Tanner-Smith, & Killings worth, 2016; Federation of American Scientists [FAS], 2006; Vogel,

*Author for correspondence

J., Vogel, D., Cannon-Bowers, Bowers, Muse, & Wright, 2006; Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013).

Along with the need for new learning experiences comes the need for new forms of assessment that align to the experiences. However, resistance to change can be a barrier to implementation. Teachers reported fears of increased grading time due to the incorporation of higher-order thinking assessment items and activities (Bressoud et al., 2015). Game-based learning can help bridge this divide as research has demonstrated its usefulness for both increasing (FAS, 2006; Wenglinsky, 1998) and assessing (Shute, Wang, Greiff, Zhao, & Moore, 2016) higher-order thinking skills. The use of game-based learning and game-based assessment has the potential to meet an important need within STEM education.

2. Stealth Assessment

Games are ubiquitous. Approximately, 155 million Americans play video games (Entertainment Software Association [ESA], 2015). Good games are engaging and require the application of various competencies to succeed (Kim, Almond, & Shute, 2016; Shute, Rieber, & Van Eck, 2011). So, how can assessment be placed within the game environment and not ruin the fun factor or disrupt flow, while accurately measuring students' competencies? One answer is through the combination of stealth assessment (Shute, 2011) and evidence-centered design (Mislevy, Steinberg, & Almond, 2003).

When students engage in gameplay, they generate copious amounts of data (i.e., time spent in-game, attempts, game actions). Stealth assessment takes advantage of this stream of information, allowing students to be assessed on a set of competencies without interrupting their gameplay. Burying the assessment in the game makes it invisible, to the point where learning and assessment become blurred (Wang, Shute, & Moore, 2015).

Stealth assessment is intended to be formative, supporting the development of competencies instead of judging final mastery. Learning games that incorporate stealth assessment can provide real-time assessment of students' learning and progress to both the students and instructor using gameplay data from log files (Shute & Ventura, 2013). These student models are personalized versions of the competency model developed for the game through evidence-centered design.

2.1 Evidence-Centered Design

Evidence-centered design (Mislevy et al., 2003) provides the theoretical framework for building stealth assessments in digital games. The Evidence-Centered Design (ECD) process starts with developing the competency model. The competency model is the representation of the theoretical concepts being assessed. Developing the competency model involves identifying and structuring the relevant variables into meaningful relationships. The competency variables comprise the knowledge set or skills to be measured by the assessment. The next step in the ECD process is to consider the evidence necessary to make claims about student competency. The evidence model establishes the specific relationships between the competency variables and their associated metrics. The evidence model is the link between the competencies and the tasks students perform within gameplay. The task model defines the specific features of tasks that will elicit the necessary evidence. The use of evidence-centered design as the frame for building the stealth assessment models aligns the learning activity and assessment by linking student actions to competency variables (Mislevy et al., 2003).

2.2 Crafting the Stealth Assessment for Variant: Limits™

The current study aims to validate the stealth assessment designed, developed, and used in conjunction with the game Variant: Limits™ (v1.0.1; 2017). Building the stealth assessment was one component of a larger research project examining the effectiveness of Variant: Limits™, a learning game designed to complement introductory college calculus courses by enhancing conceptual understanding through gameplay. The game is linear and organized into four *Zones* (i.e., game levels), each centered around specific learning objectives.

The research team together with subject matter experts set the scope for the game's educational content and developed the competency model for the stealth assessment. The competency model organized the overall educational content of the game into three main concepts and their associated sub-concepts (i.e., competencies). The learning tasks within the game were then linked to the competencies to ensure alignment between the assessment and gameplay. Figure 1 contains a representation of the competency model created for the stealth assessment.

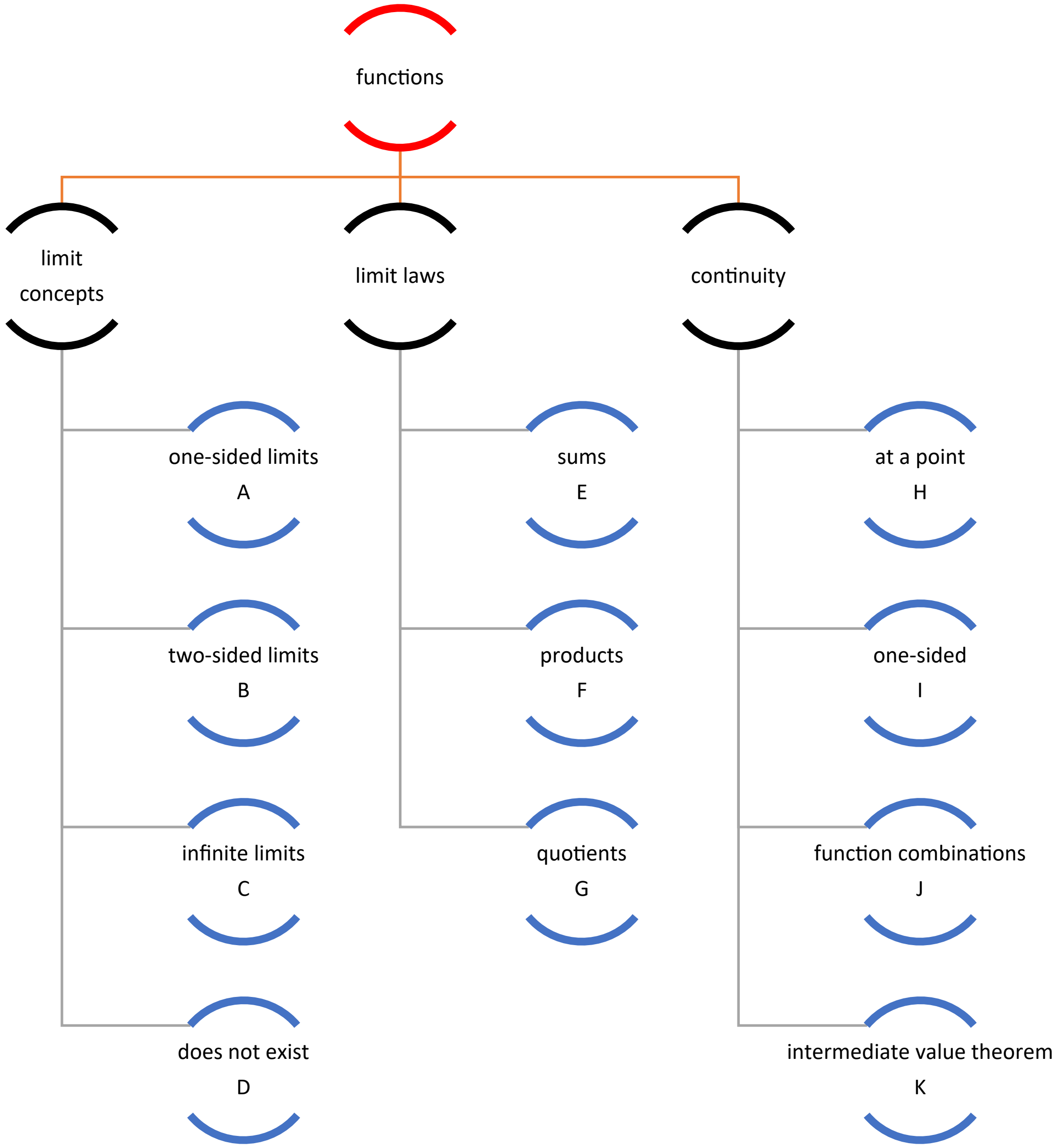

Figure 1. Variant: Limits competency model. The competency model is comprised of the educational content to be assessed. It also illustrates the relationships among the competencies.

After finalizing the competency model, the research team examined each puzzle (learning task) in the game to determine its connection to the competency model and document the associated game play actions required for the solution. The competencies were introduced in the game as students progressed. For example, puzzles in the first *Zone* are linked to only four competencies (i.e., Competencies A, B, H, and I). Competencies J and K are not covered until *Zone 3* and competency C is only addressed in *Zone 4* (i.e., the final level of the game). Therefore, as more gameplay occurs, more evidence of learning accumulates, and more competencies can be assessed for the student.

Analyzing learner actions and puzzle content led researchers to the development of scoring rubrics. The scoring rubrics connect the gameplay data, collected in the logfile, to estimates of the learners' mastery level of specific competencies (i.e., competency model). The scoring rubrics coupled with the evidence accumulation process comprise the evidence model of the stealth assessment. The validity of the evidence model will be assessed through a comparison of the stealth assessment outputs and the external calculus assessment results.

As another form of measuring calculus knowledge relevant to the game, two parallel external calculus assessments were developed by subject matter experts. The research team also evaluated the external calculus measures to ensure each of the competencies were covered on the external test by at least two items. Each external measure was comprised of 17 questions. The questions were all multiple-choice questions with four answer choices. Figure 2 shows the coverage between the external test questions and the competencies.

3. The Study

3.1 Research Question

Our overarching question in this validation study was as follows: Does the stealth assessment accurately measure students' conceptual knowledge of calculus compared to the results from an external assessment?

3.1.1 Methods

The research study of Variant: Limits^{TM} (v1.0.1; 2017) took place during the fall semester at a R1, Research University. Participants were college students enrolled in a Calculus 1 course (i.e., Math 151). Students were recruited through mathematics instructors, digital communication, and posters in public locations at the research site. Participation in the project was incentivized. Students who completed all tests and surveys, along with 4 hours of gameplay or game completion received extra credit on a test. At the time of the study, all participants had received prior instruction on the calculus content covered in the game.

Students enrolled in the study by completing the pretest and demographic survey. Once enrolled, students were randomly assigned to either the experimental (i.e., gameplay) or the control group. Students in the experimental group were given two weeks to complete four hours of gameplay or finish the game. They played the game outside of class on their own computers. During that initial two-week period, students in the control group did not play the game or complete any alternative intervention. Students in both the experimental and control groups were given a posttest at the end of the two weeks. After completing the posttest, students in the control group were given access to the game for an additional two-week period. In all, 481 students enrolled in the study by completing the pretest. Of the students who enrolled, 382 completed the posttest.

A sample from the study was selected to validate the stealth assessment. To evaluate the stealth assessment,

	CM_A	CM_B	CM_C	CM_D	CM_E	CM_F	CM_G	CM_H	CM_I	CM_J	CM_K
Q1								X	X		
Q2					X				X	X	
Q3											X
Q4	X	X						X			
Q5		X						X			
Q6					X	X	X				
Q7		X						X			
Q8								X		X	
Q9			X								
Q10			X								
Q11			X								
Q12					X						
Q13				X							
Q14				X		X					
Q15				X			X				
Q16	X										
Q17											X

Figure 2. External calculus test item and CM facet matrix. The matrix shows the test questions matched with each competency measured in the stealth assessment.

logfiles from gameplay were run through the stealth assessment machinery after the study was completed. The sample for the validation study was restricted to students in the experimental group who completed at least half of the game (n = 148) to ensure enough gameplay data for the competency and evidence models to assess. The demographics of the restricted population resemble the demographics of the larger sample. Table 1 contains selected demographic information of participants for both the game study and the validation study.

Total scores per competency from the evidence model of the stealth assessment were tallied and recorded for each student. Stealth assessment ratios (i.e., points earned/points possible) were calculated based on the final puzzle each student attempted in the game. Ratios were calculated to inform the competency model for each student: 1) The whole model is assessed through the variable total ratio, and 2) each competency is assessed through its individual variables (e.g., Aratio, Bratio, etc.). The amount of evidence accumulated for each competency is based on the amount of gameplay.

3.2 Results

The validation of the stealth assessment through its correlation with an external measure required an analysis of the external measure to which it would be compared. Item analysis of the external calculus pretest and posttest revealed moderate to low reliability with Cronbach's alphas ranging from .41 to .62. Table 2 provides the results of the item analysis for both forms of the pretest and posttest assessments. Means in the table are based on the proportion of correct responses to the questions. Most of the item difficulties revealed the questions were either too easy (M>.75) or too hard (M<.25), for a four-option multiple-choice item. All matched items performed similarly except for question 16, which was much harder for students on Form A than it was on Form B. Overall, the pretest questions were easy for students, demonstrated by the high student success rate on most of the pretest questions. For ten of the seventeen matched pretest items, the average proportion correct was greater than 75%. Somewhat high pretest scores were expected since the participants had received instruction on the calculus content prior to the start of the study. However, low item difficulty on the pretest can lead to ceiling effects (i.e., leaving little room for improvement).

After completing the analysis on the external assessments, the stealth assessment output was analyzed. The current study utilized a reduced sample for validation of the stealth assessment. Table 3 contains basic descriptive statistics for the variables used for analysis. Ratios from the stealth assessment were compared to their matched scores from the external assessments to evaluate construct validity. Bivariate correlations were calculated between the stealth assessment output and the external scores. As expected, the pretest and posttest scores were significantly correlated with the stealth assessment total ratio (Pretest: r = .19, p = .02; Posttest: r = .21, p = .02). Comparing the ratio for each of the individual competencies to the average of the matched external items also revealed significant correlations. The ratio for competency C was significantly correlated with its matched items on the pretest (r = .51, p = .002) and posttest (r = .45, p = .008), and the ratio for competency G was significantly correlated with its matched items on the posttest (r = .21, p = .014).

Given the established correlations between the two competencies (Cratio and Gratio) and the external tests, regression analysis was performed using Cratio and Gratio as predictors for posttest scores. The linear regression revealed the combined competencies accounted for 28% of the variation in posttest scores (r = .53, p = 006). The stealth assessment ratios for competency C (b^*_{Cratio}= .40, p = .013) and competency G (b^*_{Gratio}= .33, p = .040) were both significant predictors of posttest scores.

Due to the significant correlation between pretest and posttest scores (r = .48, p< .001), an additional regression

Table 1. Selected demographics of participants in the game study and validation study

	Gender		Race[a]		Avg. Weekly Gameplay	
	Male	Female	White	Asian	≤ 2 hrs	> 2 hrs
Game Study	60.0%	39.8%	72.8%	15.2%	60.3%	39.7%
Validation Study	63.5%	36.5%	74.3%	12.2%	53.4%	46.6%

[a] *race categories included were selected by at least 10% of the sample*

Note. Statistics were given in percentages of each sample for easier comparison.

Table 2. Sample size, reliability, and item analysis for the external calculus assessments

Test Items	Form A – Pretest n = 236 α = .62		Form B – Pretest n = 245 α = .57		Form A – Posttest n = 192 α = .41		Form B – Posttest n = 190 α = .43	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	.91	.29	.90	.30	.94	.24	.96	.19
2	.58	.49	.53	.50	.70	.46	.69	.46
3	.89	.31	.92	.27	.96	.19	.97	.16
4	.84	.37	.78	.42	.93	.26	.91	.29
5	.96	.19	.96	.20	.99	.10	.97	.16
6	.93	.25	.91	.29	.97	.17	.87	.33
7	.98	.14	.97	.18	.99	.10	.98	.14
8	.67	.47	.68	.47	.71	.46	.69	.46
9	1.00	.07	1.00	.06	[a]	[a]	.99	.07
10	.84	.37	.81	.40	.97	.16	.98	.14
11	.91	.29	.82	.39	.95	.21	.93	.25
12	.78	.42	.69	.46	.82	.38	.84	.37
13	.79	.41	.88	.33	.83	.38	.92	.27
14	.16	.36	.32	.47	.19	.39	.39	.49
15	.29	.45	.30	.46	.32	.47	.34	.47
16	.04	.20	.78	.42	.05	.21	.89	.31
17	.52	.50	.60	.49	.74	.44	.64	.48

[a] Item removed due to no variability

Table 3. Descriptive statistics for the variables in the validation study

Variable	***n***	***M***	***SD***	**Min**	**Max**
totalpre	148	12.66	2.40	5	16
totalpost	136	13.60	1.70	8	17
totalratio	148	.62	.08	.44	0.82
Aratio	148	.62	.09	.41	0.88
Bratio	148	.85	.11	.50	1.00
Cratio	33	.91	.09	.69	1.00
Dratio	148	.25	.18	.00	0.71
Eratio	148	.80	.14	.33	1.00
Fratio	148	.63	.22	.00	1.00
Gratio	148	.63	.14	.32	1.00
Hratio	148	.49	.17	.14	1.00
Iratio	148	.60	.16	.33	1.00
Jratio	141	.53	.18	.25	1.00
Kratio	52	.84	.22	.00	1.00

analysis was conducted with Cratio, Gratio, and pretest as predictors. For predicted posttest scores, Cratio ($b^* = .33$, $p = .024$) and pretest ($b^* = .43$, $p = .006$) were significant predictors, while the contribution of the Gratio was no longer significant ($b^* = .21$, $p = .140$). The new model accounted for 45% of the variation in posttest scores ($r = .67$, $p < .001$). However, analysis of the nested models revealed that the model containing the predictors, Cratio and pretest, was the most parsimonious ($R^2_{change} = .24$, $F(1, 30) = 12.29$, $p = .001$).

These results meet expectations based on the limitations of the external measures and suggest that the overall stealth assessment is valid. However, further validation of each of the individual competencies measured by the stealth assessment would strengthen this conclusion.

Gender disparity still exists in the STEM fields (National Science Foundation, 2017). The need for more diversity and equity in the STEM education is the second of the three main goals identified by the National Academies of Science, Engineering, and Medicine (2018). Therefore, the research team also examined the stealth assessment overall output (i.e., total ratio) for differences based on gender. Regression analysis revealed no significant gender difference on the stealth assessment (Male$_{totalratio}$: M= .63, Female$_{totalratio}$ M = .62, p = .52). Similarly, an analysis of covariance (ANCOVA) revealed no main effect for gender on total ratio when holding pretest scores constant, F(1, 145) = .05, p = .82. Also for this sample, there was no significant main effect for gender relative to posttest scores, when holding pretest scores constant, F(1,133) = 1.64, p = .20.

A final set of analyses was conducted to determine if the sample used to validate the stealth assessment (i.e., subset of the experimental group) showed any significant change in calculus understanding (i.e., gameplay). When compared to the control group from the study, the research team found no significant main effect of gameplay on posttest scores, holding pretest constant, F(1, 309) = .11, p = .74. However, the degree to which a student progressed in the game (i.e., highest *Zone* completed) was significantly correlated with posttest scores (r = .18, p = .02), showing students who played more of the game performed better on the posttest than those who played less of the game (as calculated by the amount of game successfully completed, not time spent on gameplay).

3.2.1 Discussion

The stealth assessment performed as expected and provided valid assessment of students' calculus understanding as defined by the competency model. However, the research team expected to find more correlations between the individual competencies and their matched items on the external calculus measures. Several factors limited the analysis. Poor discrimination of the assessment items in the external measures was a major limitation to the validation of the stealth assessment. Item analysis revealed only five (out of 17) questions on the pretest had a success rate less than 75%. The success rate on the pretest is evidence of elevated levels of prior knowledge on the educational content being assessed. As previously stated the students had recently covered limits (i.e., the main calculus concept in the game) in their calculus course.

Another limitation of the validation study was the difficulty of the game. Previous testing of the game indicated it would take approximately four hours of gameplay to complete the game. Therefore, the study design instructed the experimental group to either finish the game or complete four hours of gameplay during the two-week period. However, very few students finished the game and only a small number made it into the final *Zone* of the game (i.e., *Zone* 4). Table 4 shows the highest game level completed by students. Of the students in the gameplay condition the majority (84%) only made it through the first two levels of the game during the two-week period. Therefore, the majority of students accumulated limited or no evidence for six of the eleven competencies in the competency model. In fact, only 3% of students completed the game and thus accumulated all possible logfile data from gameplay for evaluation by the stealth assessment machinery.

A final limitation is the stealth assessment was not embedded in the game during the game study; all outcome ratios were calculated posthoc. One advantage of stealth assessment is that it can offer immediate, precise feedback as students are playing. This advantage of stealth assessment was not leveraged in this study. Students in the gameplay (i.e., experimental) condition were not given any feedback from the stealth assessment on their progress (i.e., current mastery of competencies) during gameplay. Task level feedback can influence and improve learning (Shute, 2008). Therefore, the research team recommends future validation studies for stealth assessments are conducted after the assessment is embedded directly into the game.

The current research highlights some of the possibilities for STEM education afforded through the use of game-based learning and assessment. Further research on the effectiveness of game-based learning and assessment in the STEM fields is warranted as evidence-

Table 4. Number and percent of students by highest zone completed

Zone	**< 1**	**1**	**2**	**3**	**4**
n	24	52	113	28	7
%	11%	23%	50%	13%	3%

Note. Students are only reported by the highest zone they completed, not for every Zone completed.

based approaches and programs are needed to help increase student success in STEM courses (The National Academies of Science, Engineering, and Medicine, 2018).

4. References

Bressoud, D., Mesa, V., & Rasmussen, C. (2015). *Insights and Recommendations from the MAA National Study of College Calculus*. Washington, DC: The Mathematical Association of America.

Blair, R. M., Kirkman, E. E., & Maxwell, J. M. (2013). Statistical abstract of undergraduate programs in the mathematical sciences in the United States: Fall 2010 CBMS survey. Providence RI: American Mathematical Society. Retrieved from: www.ams.org/profession/data/cbms-survey/cbms2010.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research*, *86(1),* 79-122. doi:10.3102/0034654315582065

College Board. (2016). 2016 AP Exam score distributions. Retrieved from: https://www.totalregistration.net/AP-Exam-Registration-Service/2016-AP-Exam-Score-Distributions.php.

Entertainment Software Association. (2015). *Essential facts about the computer and video game industry: 2015 sales,demographic and usage data*. Retrieved from: http://www.theesa.com/wp-content/uploads/2015/04/ESA-Essential-Facts-2015.pdf

Federation of American Scientists. (2006). Summit on educational games: Harnessing the power of video games for learning. Washington D.C. Retrieved from: https://fas.org/programs/ltp/policy_and_publications/summit/Summit%20on%20Educational%20Games.pdf

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying Evidence-Centered Design for the development of game-based assessments in Physics Playground. *International Journal of Testing*, *16(2),* 142-163.doi:10.1080/15305058.2015.1108322

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1(1),* 3-62.

National Academies of Sciences, Engineering, and Medicine. (2018). *Indicators for Monitoring Undergraduate STEM Education*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24943

President's Council of Advisors on Science and Technology. (2012, February). *Report to the President: Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics.*

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. doi:10.3102/0034654307313795

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

Shute, V. J., Rieber, L., & Van Eck, R. (2011). Games . . . and . . . Learning. In R. Reiser & J. Dempsey (Eds.), *Trends and issues in instructional design and technology* (3rd ed.) (pp. 321-332). Upper Saddle River, NJ: Pearson Education, Inc.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106-117. doi:10.1016/j.chb.2016.05.047

Variant: Limits [computer software]. (2017). Bryan, TX: Triseum.

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34(3),* 229-243.

Wang, L., Shute, V. J., & Moore, G. (2015). Lessons learned and best practices of stealth assessment. *International Journal of Gaming and Computer Mediated Simulations, 74(4),* 66-87. doi:10.4018/IJGCMS.2015100104

Wenglinsky, H. (1998). *Does it compute? The relationship between educational technology and student achievement in mathematics*. Educational Testing Service. Policy Information Center. Retrieved from: www.ets.org/research/pic

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105(2),* 249-265. doi:10.1037/a0031311