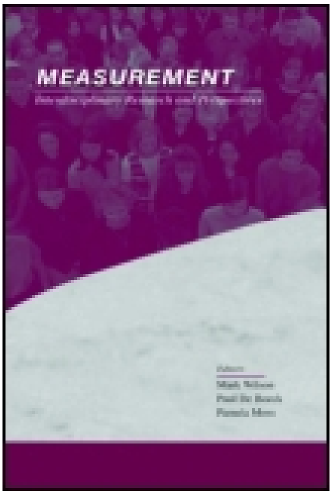

Measurement: Interdisciplinary Research and Perspectives

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmes20

Rejoinder to Comments on Task Features in Simulations and Games

Russell G. Almond[a], Yoon Jeon Kim[a], Gertrudes Velasquez[a] & Valerie J. Shute[a]

[a] Educational Psychology and Learning Systems, Florida State University

Published online: 22 Aug 2014.

To cite this article: Russell G. Almond, Yoon Jeon Kim, Gertrudes Velasquez & Valerie J. Shute (2014) Rejoinder to Comments on Task Features in Simulations and Games, Measurement: Interdisciplinary Research and Perspectives, 12:3, 118-124, DOI: 10.1080/15366367.2014.939528

To link to this article: http://dx.doi.org/10.1080/15366367.2014.939528

Measurement, 12: 118–124, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366367.2014.939528

Routledge
Taylor & Francis Group

REJOINDER TO ISSUE 12(1–2)

Rejoinder to Comments on Task Features in Simulations and Games

Russell G. Almond, Yoon Jeon Kim, Gertrudes Velasquez, and Valerie J. Shute
Educational Psychology and Learning Systems, Florida State University

First, we would like to thank all of the people who took the time to write commentaries on our article. There is a rather famous map of the United States drawn by Daniel K. Wallingford from a New Yorker's perspective and a similar one drawn by Mark Storm from a Texan's perspective; each has the home state much larger than the rest of the country. The humor in these maps comes from the natural tendency of people to know things close to them in great detail but to only have sketchy knowledge of things that are far away; that is, our natural view of any topic is like a fish-eye lens with things in the center magnified and things at a distance compressed. The fish-eye lens also applies to assessment design: Psychometricians, test developers, domain experts, computer programmers, game designers, test users all have their unique view of the test design process. Evidence-centered assessment design (ECD) looks different from these different perspectives, and it is hard to put the whole picture together without looking at ECD from many directions. Therefore, we are truly grateful to have such a wide array of perspectives on our article.

One thing that was obviously not clear in the original article was the purpose of including both the mathematics word problem and the game examples. Some of the earliest ECD presentations by Mislevy, Steinberg, and Almond[1] included a triangle. The 3 vertexes were *purpose* (formative versus summative), *score reporting* model (unidimensional versus multidimensional) and *task type* (multiple choice versus simulation based). The key idea was that changes in any one vertex would necessitate changes in the others as well. In particular, at the time (late 1990s) both simulation-based assessment and cognitively diagnostic assessment were becoming more popular and would induce changes in one (and as a consequence all) of the vertexes. ECD was

[1] Personal recollection, Russell Almond. These were mostly internal ETS presentations and did not have a formal publication date.

This is a rejoinder to the commentaries on "How Task Features Impact Evidence From Assessments Embedded in Simulations and Games" from issue 12(1–2) 2014 of *Measurement: Interdisciplinary Research & Perspectives*.

Correspondence should be addressed to Russell G. Almond, Educational Psychology and Learning Systems, Florida State University, 3204-J Stone Building, 1114 W. Call St., Tallahassee, FL 32306-4453. E-mail: ralmond@fsu.edu

intended to be a language for describing and realizing what those changes might be. Thus, the thesis of our article is properly that the same design principles that are used to construct conventional assessments can also be applied to game-based assessments, modified as appropriate to suit the new purpose.

Another thing that was unclear in the original article was the state of the work on *Newton's Playground* when the article was written. The initial draft of the article was written after the completion of the design process. Only some very preliminary piloting had been completed at that point, so we could not yet test our theories. One purpose of this article is to document some of the ways in which we applied ECD to the design process for this game. Between the time of the initial submission and publication, we had time to complete a field trial. Some of those results are discussed in this article or our rejoinder, but most will need to wait for a future publication.

An additional motivation for writing this article came from our (particularly Russell's) experience using Mislevy, Steinberg, and Almond (2002) in the classroom. The class was a practical class on assessment design. After completing the reading, the students were asked to pick a task (related to their final project) and list one variable that fills each of these roles. This task proved to be quite challenging. This article was intended to be more introductory than the Mislevy, Steinberg, and Almond (2002, 2003) references to help students with that task. (Despite its difficulty, we still recommend this activity as a way to get a design team up to speed on ECD).

DISCUSSION OF COMMENTARIES

DiCerbo's Comments

In her commentary, Kristen E. DiCerbo highlights the importance of having "a game designer, a learning designer, and an assessment designer, all involved in task design." To this list, we would add a psychometrician and a computer programmer. All of these roles were present during the design process for *Newton's Playground* (sometimes by the same team member wearing different hats). We hope that one contribution of this article is to provide these people with a framework in which to talk about task characteristics in a productive way.

One place in which our perspective differs from hers is in the need for automaticity in the modeling process. This likely stems from differences in the kinds of assessments we have been building. For example, the initial version of the *NetPASS* simulator had 9 tasks. Each task had an individual task model and a unique evidence model: a custom built Bayesian network (Levy & Mislevy, 2004). *Newton's Playground,* in contrast, had 75 levels. Although each level was hand drawn, it was necessary to find similarity among the levels—that is, to create task models—in order to structure the work of both task design and evidence model creation. The same ECD principles are present, but they look different due to the different natures of the assessments.

We agree that game-based assessments provide an opportunity to harvest some of the riches of the digital ocean. However, this has proved difficult to achieve in practice. The final version of *Newton's Playground* was much less ambitious than our original draft in large part because of the difficulty of producing operational definitions for key observables. It became clear that identifying which agents of motion the player was attempting to use was a necessary part of many of the observations we wanted to make. The design for that code took several iterations and several experiments before we got it right. DiCerbo and Behrens (2014) is an important article in that it sets a lofty goal for the assessment community, but it will still take a lot of hard work to reach that goal.

Graf's Comments

We would like to thank Aurora Graf for the very comprehensive and useful history of automatic item-generation; it makes a nice complement to our article. Although automatic item-generation is not the only reason for thinking about task model variables, experiences with automatic item-generation are informative for both automatic and manual item-generation in future efforts.

Among the work she cites are examples of both item models being used retrospectively (items from existing tests are coded as to their features and the relationship between the features and item parameters are studied) and item models being used prospectively (items are built from models and then field tested). Both of these are important: Retrospective studies help form theories of item development and prospective studies test those theories. These theories of item development can only predict some of the evidential properties of items. Michael Timms picks up on the difficulty of telling whether a task model variable is truly radical or incidental; the body of work cited by Graf includes many cases in which the task designers were surprised by the difference between their theory and the data. These studies are a good start, but more work is needed in this area.

Creativity proved to be the most difficult of the 3 skills targeted to measure by *Newton's Playground*. As Graf stated, open-ended items, which allow multiple solution paths, are best for measuring creativity; but those items are the weakest for measuring physics. This produced a challenge for the assembly model, particularly as students could choose which levels to address. Some of the problem was in the framing of the game (the delivery model). During our internal testing, the framing was often to explore creative solutions; so the development team excelled at coming up with creative solutions to delight our fellow teammates. During the data collection, students were told that the person who completed the most levels would get an extra gift card. This encouraged efficient rather than creative solutions. Finally, we will note that creativity poses the greatest challenge for both automated scoring and automatic item generation.

Markus's Comments

Keith A. Markus attempts to more clearly state the thesis for our article, but we would rather state it as, "The principles of ECD can be extended from traditional assessment to game-based assessment." Similarly, he poses the question of "whether evidence-centered design is intended to codify traditional best practices or is intended as an alternative to traditional practices." The answer (for the present article and the original Mislevy, Steinberg, and Almond [2002, 2003] articles) is both. In particular, ECD is designed to be a language for talking about what is the same and what is different and, hence, what can be reused from earlier designs and what must be changed.

Markus's comments on content-construct confusion are extremely interesting because this is an issue that we have found frequently arises during design conversations. In particular, it is easy to fall into a mode in which one designer is talking about construct and the other is talking about content and the two are talking past each other. We address the issue only tangentially in the current article because the focus is on the task rather than the proficiency model. This does seem like an area where the language of ECD could be improved.

Markus also accuses us of occasionally treating content-related evidence as construct validation rather than just one line of argument for construct validation. We agree with the thrust of this criticism and recognize that while content-related evidence is necessary for construct validation, it is not sufficient. Hopefully, his comments have helped clarify some of our unclear writing.

Markus asks about the effect that user control over task selection will have on construct completeness. We are interested in this issue ourselves, and Kim (2014) offers an initial exploration of this topic. We have speculated that it might be possible to dynamically alter the reward structure of the game (e.g., offer more points for game levels that use underrepresented parts of the construct) to encourage players to provide more information, but to our knowledge nobody has of yet experimented with this possibility.

Oliveri and Khan's Comments

The discussion of María Elena Oliveri and Saad Khan on noncognitive constructs is interesting. Although in principle, ECD can be applied to both cognitive and noncognitive constructs, we have found that students have a harder time applying ECD to noncognitive constructs. Consider 2 items from the Beck depression inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961): (1) "I get tired more easily than I used to" and (2) "I would like to kill myself." The second item is more difficult than the first,[2] where here psychometric difficulty means the patient must be more depressed before he or she will endorse the item. When working with noncognitive constructs we have found that the distance between the meaning of psychometric difficulty for that construct and the lay meaning of difficulty causes test designers (particularly students working on classroom projects) to think about game difficulty rather than psychometric difficulty. In general, the language of ECD assumes that the construct of interest is cognitive and often confuses test designers when the construct in noncognitve. An extension or clarification of ECD for noncognitive constructs would be welcome.

Oliveri and Khan's other comments focus on testing validity both internal to and external to the assessment. They rightly identify the agent identification and assignment subsystem as a potential weak link in the chain. This is mostly an issue of timing: After the initial article was drafted, we did extensive testing of this system and found very high agreement between the automatic system and the human raters (more than 95% agreement). More problematic is the rule that identifies a solution with an agent according to the last agent used before the goal is reached. There may be a way to improve on that rule, but the improved rule must be both customizable to the level and general enough to promote reuse.

Oliveri and Khan state, "Validation is a key concept that should be . . . thought of alongside all other aspects that are considered during game [or any assessment] design." We agree with this principle and regret that there is not enough room to discuss all aspects of assessment design

[2] There are also some differences in the other aspects of evidentiary focus. In particular, there are other causes of fatigue besides depression, giving the first item weak discrimination. Furthermore, depression can be present with our without suicidal thoughts, indicating that the second item may be tapping a different "proficiency" of depression.

within a single article. We note in passing that task models that are not used because they are too difficult to implement or to score are potentially suitable for validity studies, which generally involve smaller samples. The external validity measures proved to be both a strength and a weakness of *Newton's Playground.* One of the best results to come out of the project was a novel measure of persistence independent of the game (Ventura & Shute, 2013; Ventura, Shute, & Zhao, 2012). The external physics measure, on the other hand, was a custom-designed instrument that turned out to have low reliability. Consequently, it provided little validity evidence to support the initial field trial.

Timms's Comments

Michael Timms asks where game design fits into the models of the CAF. The answer from Mislevy, Steinberg, and Almond (2003) is the presentation model; indeed, in Mislevy, Steinberg, and Almond (2002) this was called the "simulator model." Ten years later, we think that earlier answer might be a bit naïve. Game design aspects touch upon all of the models of the CAF, and this is likely the point that DiCerbo makes in her commentary about including game designers in the design team. Certainly we have anecdotal evidence from our early field trials that the delivery model—how the game is framed to the players—makes a difference in the kind and quality of evidence delivered.

Timms also talks about the difficulty of developing evidence rules when the outcome space for the task (the work product) is open ended. One of the key lessons of the simulator-based assessment *HyDrive* (Mislevy & Gitomer, 1996) was the importance of identifying observables that separated good performances from bad. If the simulator is then designed to make determining the values of those observables easy, then you get better assessment (and instruction). It is slightly more difficult in game-based assessments because the key observables must also make sense within the narrative of the game.

Finally, Timms notes that the relationship between some observables and the proficiencies might be nonmonotonic. Time spent on a task tends to be especially problematic in this way. Often there is a contrast between increasing effort on the task (which improves the probability of success) and fluency with the skills involved (which decrease the time required). Most psychometric models assume a monotonic relationship between proficiency and outcome and hence are not suited for this purpose. This is even more problematic for game-based assessment as the amount of effort players are willing to spend on overcoming challenges differs and may confound the measurement of the targeted proficiencies.

Walker and Engelhard's Comments

We found the comments by A. Adrienne Walker and George Engelhard Jr. on the nomothetic and idiographic perspectives very interesting. This has frequently been a point of confusion among design teams when constructing the proficiency model. In our own work, we have often wavered between these perspectives, adding to the confusion. Some of the problem may lie in the distinction between formative and summative assessment. The goal of a formative assessment is to help a student grow to meet internal and/or external goals; consequently, the idiographic perspective is probably more helpful. The goal of a summative assessment is to compare a student to a set of standards or a reference population; here the nomothetic perspective is more helpful. A second problem is the grain size of the proficiency model. An idiographic proficiency model would be more detailed than a nomothetic one. This means there will be less evidence gathered per proficiency during a fixed time period in the idiographic assessment. This is problematic for high-stakes assessments.

Walker and Engelhard express concern about using "stealth assessments" for high-stakes purposes. In an era in which data privacy is becoming an important political issue, we share those concerns. We agree with the general principle that no authority should gather data about any person for the purpose of high-stakes assessment without their explicit knowledge. However, the term "stealth assessment" was never intended to apply to high-stakes assessments or to refer to gathering data without consent. Rather it was meant to invoke the idea that just as a good teacher continually uses formative assessment to adapt instruction to the needs of the students, a game-based assessment system should continually monitor the ability of the students and adapt itself accordingly. There is some evidence that commercially successful games already do this (Gee, 2010). Moreover, if games and simulations are used (with the examinees' full knowledge) in high-stakes assessments and if during the assessment the examinees lose their anxiety about being tested in the flow of the activity, this will surely produce better outcomes.

Bond's Comments

Lloyd Bond emphasizes the importance of ECD in both prospective and retrospective design tasks. We would like to add that these frequently happen in the process of the same assessment design. Indeed, *Newton's Playground* is based on an earlier commercial game, *Crayon Physics Deluxe.* The earliest stages of the design process were mostly the design team playing *Crayon Physics* and discussing how that game should be modeled. When it became clear that the existing game would not support the evidence identification we needed, we switched from retrospective to prospective design. Our design efforts were stronger because of the earlier retrospective stages.

Bond talks about the difficulty of both learning and applying ECD. It appears this article has only partially met its goal in providing a simpler introduction to ECD. One problem is that there are many nuances in assessment design and it is easy to get lost in the details and miss the bigger picture. While we have made much progress in talking more formally about assessment design over the past 10 years, we have not made nearly as much progress in developing a pedagogy of assessment design.

A second problem with learning and applying ECD is that so much of it is dependent on the purpose of the assessment. Parts of the ECD theory become important only in the context of particular applications. Consider the variable role of "which proficiency"; this role is central when the proficiency model is multidimensional but only of peripheral interest when the proficiency model is largely unidimensional. In our experience, design teams applying ECD often need to first build custom forms and tables that capture the parts of the ECD process that will be important in this particular application (Almond, 2010). We hope that the broader language of ECD, as well as the examples we provide in this article, help those design teams decide what needs to go into those custom forms.

CONTINUING THE CONVERSATION

We would like to once again thank the commenters for opening a discussion about many aspects of assessment design. We do not view our answers in this rejoinder as a final response but rather merely the second word in a longer conversation. We think this conversation should grow to include all of the assessment community (using the broadest possible definition, including test developers, psychometricians, instructional designers, game designers, test users, and people involved with the sales and marketing of assessment systems). Only when we view the problems of assessment design from many different perspectives will a true theory of assessment design emerge.

We would like to offer the ECD Wiki (http://ecd.ralmond.net/ecdwiki/ECD/ECD/) as one way to continue this conversation. The (minimal) content already on the wiki, including the start of an ECD glossary, can be freely viewed. If you would like to add to or discuss the content on the wiki, e-mail the first author to request a password. We hope that this and other opportunities will continue this important conversation.

REFERENCES

Almond, R. G. (2010). I can name that Bayesian Network in two matrixes. *International Journal of Approximate Reasoning*, *51*, 167–178.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571.

DiCerbo, K. E. & Behrens, J. T. (2014). *The impact of the digital ocean on education*. [white paper] London, UK: Pearson. Retrieved from http://research.pearson.com/digitalocean

Gee, J. P. (2010). Human action and social groups as the natural home of assessment: Thoughts on 21st century learning and assessment. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century: Supporting educational needs* (pp. 13–40). New York, NY: Springer.

Kim, Y. J. (2014). Search for the optimal balance among learning, psychometric qualities, and enjoyment in game-based assessment. Unpublished PhD dissertation. Florida State University.

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a simulation-based assessment. *International Journal of Measurement*, *4*, 333–369.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction*, *5*, 253–282.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. H. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (pp. 97–128). Mahwah, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, *1*(1), 3–62.

Ventura, M., & Shute, V. J. (2013). The validity of a game-based assessment of persistence. *Computers and Human Behavior*, *29*, 2568–2572.

Ventura, M., Shute, V. J., & Zhao, W. (2012). The relationship between video game use and a performance-based measure of persistence. *Computers and Education*, *60*, 52–58.