What "Exactly" Describes Planning in a Math Game-Based Assessment? Introducing *Fraction Ball: Exactly*

John Lopez
University of California, Irvine
johnll2@uci.edu

Katherine Rhodes
University of California, Irvine
ktrhodes@uci.edu

Kreshnik Begolli
University of California, Irvine
kbegolli@uci.edu

Andres Bustamante
University of California, Irvine
asbustam@uci.edu

June Ahn
University of California, Irvine
junea@uci.edu

Abstract

The digital assessment community aims to align assessment design and measurement selection with theory, but many digital assessments fail to establish these theoretical links. In this paper, we introduce Fraction Ball: Exactly (FBE), a digital game-based assessment for planning in a mathematics problem-solving context. Using gameplay learning analytics, we create distinct pilot measurement models for adult and child players. We demonstrate convergent validity of FBE, both with the Shallice Tower of London and with mathematical knowledge. We share our approaches to theoretically ground our assessment design, measure selection, and measurement model validation. By theoretically grounding each process, we outline ways in which digital assessment developers can integrate educational theory with assessment design and validation and theorize for themselves.

Keywords: assessment, learning analytics, game-based learning, game-based assessment, education research

1. Introduction

Although the digital assessment community aims to align assessment design and measurement selection with educational theory, many digital assessments fail to establish these theoretical links. To demonstrate a process of theorizing, we present a design and pilot validation study around the executive function (EF) skill of planning in a contextualized, mathematics game-based assessment.

Many existing EF assessments test EFs with tasks designed to target specific EFs. While these assessments show efficacy and reliability to train and test a variety of EFs, they ignore complementary skills present in contexts where people truly exercise their EFs, like mathematics (Gioia et al., 2010; McCoy, 2019). To design assessments around EFs embedded in contexts like mathematics, one must create a strong link between theory, assessment measures, and contextual elements. Unfortunately, many existing assessments weakly link theory to chosen assessment measures (Boorsboom, 2005). Thus, designing assessments for EF in contexts like mathematics requires bridging theory, measures, and context.

This paper seeks to support those who seek to narrow the gaps between theoretical embedding, assessment design, and measurement selection by creating theoretically-linked, contextualized EF assessments. We present how we developed *Fraction Ball: Exactly* (*FBE*), a game-based assessment for mathematics and the EF skill of planning. *FBE* (Figure 1) asks players to shoot basketballs on a court divided into fractions until they reach a "target" score. We hypothesized that *FBE* assesses key skills employed in planning within the context of mathematics. To explore this hypothesis, we designed *FBE* from theoretical principles of EF and planning, embedded theoretically-driven learning analytics into *FBE*, and psychometrically validated our learning analytics. We present distinct measurement models from pilot data collected with 80 adult players and of 58 child players. Our models establish convergent validity between *FBE*, the Shallice Tower of London (ToL) planning assessment (Shallice, 1982), and a mathematics assessment (Bustamante et al., 2022). We continue the process of theorizing to make sense of our initial results and guide next steps of iterative assessment design and validation. We make two contributions to digital educational assessment development:

URI: https://hdl.handle.net/10125/109434
978-0-9981331-8-8
(CC BY-NC-ND 4.0)

- We demonstrate a design and measurement validation study to measure planning in a mathematics context, shedding light on how to design EF assessments embedded within math contexts.
- We reflect on how this design and measurement validation process allows us to theorize planning in mathematics contexts beyond the design goals of the game.

2. Literature Review

2.1. Planning and Its Measurement

We characterize executive functions (EFs) as key cognitive processes involved with higher-order thinking skills like concentrating, organizing steps to execute a task, and adapting plans to environmental stimuli (Bagetta and Alexander, 2016; Diamond, 2013). Theories describing the exact nature of these higher-order thinking skills abound, and they do not necessarily converge into one, harmonious framework (Best and Miller, 2010; Blair et al., 2005, Diamond, 2013). Leading theories argue that EFs are both unified in one dimension and diversified into separable components (Diamond, 2013; Miyake et al., 2000). Moreover, they manifest differently at different stages of development (Best and Miller, 2010; Blair et al., 2005). The theoretical framework selected for conceptualizing EFs is largely guided by the research aims of one's discipline.

Planning is an EF skill that describes the ability to set, change, and accomplish goals when problem-solving (Bagetta and Alexander, 2016; Diamond, 2013). The construct of planning, in particular, is essential for accomplishing the multi-step, goal directed tasks in the context of mathematical problem solving. (Sikora et al., 2011). For example, students choose a problem-solving strategy among many possible approaches, commit to the desired strategy, and potentially change a strategy as they add new information. Prior research suggests that stronger planning skills predict mathematics performance in adolescents and children (Bull and Lee, 2014; Gerst et al., 2017; Sikora et al., 2011). Thus, planning is intrinsically connected to features of the contexts in which planning occurs.

The theory of planning that has dominated cognitive science for the past four decades draws from the Supervisory Attentional System (SAS) (Norman and Shallice, 1980). According to this theory, higher order thinking skills are controlled by two, complementary cognitive systems: a contention scheduling system that handles routine problems and an SAS that manages cognitive resources for non-routine problems. The contention scheduling system operates as a "cognitive manager", allowing attention to be malleable while we tackle routine problems in our daily lives, such as driving to a familiar place. Upon encountering an unfamiliar task, the SAS activates and draws attention to the task. Importantly, activation of the SAS involves a tradeoff – using the highly specialized resource of attentional control increases accuracy of completing the task but costs efficiency of how one completes the task.

It follows that cognitive and educational researchers have generally measured planning not only in terms of whether a desired goal has been accomplished (*accuracy*), but also in terms of how the solution was generated (i.e., with *speed* and *efficiency* of effort). In planning measures, *speed* is generally defined as the time spent after an examinee takes their first step on a problem until a problem is complete (Berg and Byrd, 2002; Unterrainer et al., 2004). Capturing this time is thought to capture planning because the more quickly an examinee can solve a problem after starting, the more likely the examinee would have planned an optimal solution (Berg and Byrd, 2002; Unterrainer et al., 2004).

Similarly to *speed*, *efficiency* measures how an examinee solves a problem in some optimal way. For example, if an examinee has effectively planned a solution, it is assumed they will have done so using the fewest number of actions (Berg and Byrd, 2002). Various planning assessments have operationalized *efficiency* through the number of actions taken to solve a problem (Berg and Byrd, 2002).

The domain of *accuracy* questions whether or not an examinee has successfully reached a goal state (Berg and Byrd, 2002). One interpretation of accuracy allows an examinee to make as many moves as they want to reach a target goal within a limited amount of time (Berg and Byrd, 2002). Another approach deems problems accurate only if an examinee achieved the target goal in the fewest possible steps (Unterrainer et al., 2004). Both scoring approaches require the ability to successfully execute steps toward a solution.

While a combination of *speed*, *efficiency*, and *accuracy* are thought to capture planning ability, the literature debates the exact nature of their measurement structure. For example, an exploratory factor analysis conducted by Unterrainer et al., 2004 found that planning measures loaded onto a single factor when compared to other EF measures. Similarly, a confirmatory factor analysis conducted by Debelak et al., 2015 found that a one-factor model represented planning better than models with multiple factors. Both studies point to a single set of measures representing

planning. Other studies, however, show that planning may encompass multiple latent structures. One study conducted an exploratory factor analysis with two planning measures and identified split loadings across two factors (Georgiou et al., 2017). Another study found that a three-factor model yielded a better psychometric fit than a two-factor model and a one-factor model (Trakoshis et al., 2022).

In sum, while existing measures are theoretically agreed to comprise planning, the latent structure of planning as constructed by these current measures is not clearly defined, requiring any new measures of planning to be taken under theoretical and psychometric scrutiny.

2.2. Assessing Planning

Based on the theoretical understanding of the SAS and planning, researchers identified specific problem-solving features that tend to activate the SAS. In general, for a task to involve the SAS, it must (1) be sufficiently difficult (2) be non-routine and non-familiar, meaning that there are no existing routines for solving the task, and (3) possibly, require corrections if the selected solution strategy fails or errors occur. Shallice (1982) utilized these key features to design the Tower of London assessment (Figure 2), a problem-solving task aimed to activate the SAS. The Shallice ToL requires problem-solvers to move three colored beads, one at a time, across a set of three pegs, into a target configuration, using the fewest number of moves. This sufficiently difficult, non-familiar, and reversible task became the standard assessment to generate new planning theory in cognitive and educational science.

Clinical psychologists originally designed and used assessments like the Tower of London to diagnose and treat challenges with EF, while cognitive scientists used these assessments to shed light on individual cognitive mechanisms which underlie EFs (Anderson et al., 1996). Reliability studies performed on many tools like the Tower of London have confirmed their validity for the purposes of evaluation (Berg and Byrd, 2002, McCoy, 2019). Digital versions of EF assessments like the Tower of London show the same reliability as physical assessments (Berg and Byrd, 2002). Thus, digital assessments led to increased scalability of EF assessments and allowed researchers to distribute them to a wide variety of populations, including children in schools. While these assessments are useful for measuring EFs in evaluative contexts, researchers question their efficacy beyond clinical applications (McCoy, 2019). Meeting the requirements of assessment ecological validity demands a balance between, (1) the relation between the targeted construct and the assessment, (2) demonstration of convergent validity with other, accepted assessments of the targeted construct, and (3) environmentally-based measures (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). To meet these requirements, a designer may have to iteratively a variety of measures from the environment before identifying those that truly represent the argued construct.

In the last decade, digital game-based assessments have set out to balance relationships between the targeted construct, argumentation for convergent validity, and designing environmentally-based measures (Mislevy et al., 2014; Rahimi and Shute, 2023). Over time, researchers developed design frameworks for game-based assessments, which may support the argumentation needed for ecological validity of new assessments (Kim et al., 2019; Mislevy et al., 2014). Evidence-centered game design (ECGD), for example, consists of the four components. First, the developer must *define assessment competencies from a non-game perspective*. To do this, a developer must understand what specific practices result in evidence of a given construct. From here, the developer must *create and link game mechanics to assessment competencies*. Game mechanics must tightly align to the defined argumentation structures and reasoning for competency. Game-based assessments often have some form of *feedback* to create awareness of the competency within the user or to dynamically adjust the assessment's difficulty over time. Lastly, developers must expect continual *iteration* of the game's design and its measures (Mislevy et al., 2014). Design processes like ECGD have produced games that train EF skills due to the intentional development of measures and contexts that elicit opportunities to draw measures from (Parong et al., 2017; Mejía et al., 2024). Many game-based assessment developers present strong argumentation for ecological validity, as assessment features draw from theoretical principles, grounded in diverse contexts (Rahimi and Shute, 2023). Game-based assessments provide opportunities for the broader assessment community to model how to more strongly align theory with assessment measurement and design. In the rest of this paper, we follow ECGD to produce a theoretically-linked game, measures, and further theories to guide our next steps of assessment development.

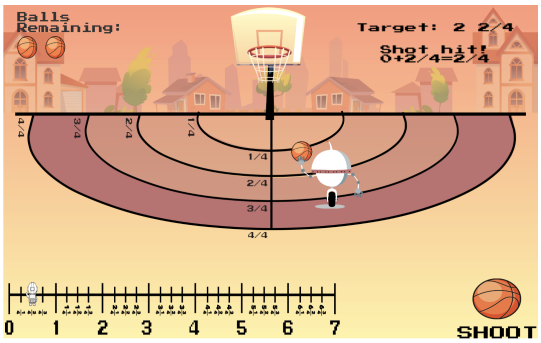

Figure 1. In Fraction Ball: Exactly (FBE), players must plan how each shot adds to the fraction sum!

3. Designing *Fraction Ball: Exactly*

We designed *Fraction Ball: Exactly* (*FBE*) to engage learners in planning and fraction mathematics skills simultaneously; the game assesses planning specifically in this rational-number environment. In each round of *FBE*, players shoot a basketball from a court divided into number units (as opposed to whole number units) they reach a "target" score. Players have no time limit to score. Players move their avatar to a desired position on the basketball court, press a "shoot" button, and watch their avatar shoot a basketball from that position. The interface shows the player's current score and any arithmetic changes to the player's prior score. Figure 1 displays the game's score, number line, and board labels in fraction notations (1/4, 2/4, etc). Each shot results in a possible arithmetic change to the player's summed total score. If their score surpasses the target score, or if they run out of basketballs to shoot with, they lose the round. Children must maintain a mental representation of a correct fraction sum while planning how to reach that sum from their current score. If this mathematical representation is inaccurate, children may plan their way to an incorrect target score. Thus, our pilot work for children evaluated *FBE* using both a planning and a fraction assessment.

We conjectured that players may exercise different planning processes depending upon the game situations that they play. To test this conjecture, we created a "Limited" condition where players received the exact minimum number of basketballs to achieve the target score. In this situation, we expected that players would plan their shots in specific ways to maximize *efficiency*. Conversely, we created an "Unlimited" condition where players could shoot as many basketballs as they wanted. To add an additional planning demand in this condition, players must consider that the probability of making each shot decreases the further their avatar is from the basket. In this condition, we expected more variation in player planning processes, as players were not constrained to the minimum number of balls required for efficiency.

Our design process followed steps from ECGD. (Mislevy et al., 2014). To *define our assessment competencies*, we broadly examined the theory of executive function (EF) and focused on a specific EF skill - planning. From here, we identified the subdomains of *speed*, *efficiency*, and *accuracy*. To transfer possible measures from thesecompetencies to *game mechanics*, we hypothesized the above game scenario that could capture a player's evidence of these three subdomains. Our design approach adds an additional layer of theoretical insight beyond theorizing the measures: we incorporate **design theory** from an existing, psychometrically reliable assessment, the Shallice ToL. Both assessments draw from design principles meant to elicit a participant's Supervisory Attention System (SAS). In *FBE*, participants must solve a problem using the fewest possible number of steps, for they must reach a target score using the fewest number of possible shots. Participants' goals change as they progress through the assessment, as each round offers different game rules or score notation. Lastly, participants must solve a problem in a unique, non-routine context, as *FBE* presents unique changes to a traditional basketball game. Thus, we design not only around theoretically-based measures, but we design around the theoretical underpinnings that inspired the measures to begin with.

We embed *iterative development* throughout the rest of this paper. Our next two sections describe how we designed our initial measures of *speed*, *efficiency*, and *accuracy* and used psychometric analyses to validate a measurement model based on these measures. In our discussion section, we integrate our data-driven insights with existing cognitive theory and hypothesize patterns of user engagement with with *FBE*. Our theories, then, guide our next steps for assessment and measurement design.

4. Method

4.1. Data Collection

To follow a psychometrically rigorous validation process, we performed two pilot studies, one with adults and children. For our adult pilot study, we recruited eighty English-speaking adults through Amazon Mechanical Turk. Participants played one round of each game variation in *Fraction Ball: Exactly* (*FBE*) (Figure 1) followed by a digital Shallice ToL (Figure 2). We did not collect major demographic

characteristics from the participants, such as sex, ethnicity, socioeconomic status, or prior mathematics skills.

For our child validation study, we presented *FBE* to two elementary schools in a major, urban, public school district in the Southwestern United States. Fifty-eight English-speaking children in grades 3-6 completed one round of each *FBE* variation, a digital Shallice ToL, and a 44-problem math assessment battery comprising decimal and fraction conversion, addition, and number line placement (Bustamante et al., 2022).

4.2. Exploratory Analytics Created

After characterizing the subdomains of *speed*, *efficiency*, and *accuracy*, we embedded exploratory analytics inspired by Shallice ToL measures with characteristics of our unique game setting. We represented *speed* as the number of milliseconds (*movement_time*) from the first action taken until a round ends, as some Shallice ToL studies actualize *speed* as *movement_time* (Anderson et al., 1996; Berg and Byrd, 2002; Trakoshis et al., 2022). We represented *efficiency* from the kinds of actions possible in *FBE*: the number of basketballs shot (*num_of_shots*), number of avatar movements on the court (*num_of_moves*), and the number of shots taken beyond the minimum number of shots needed to score (*excess_shots*). These representations parallel various *efficiency* measures in Shallice ToL studies (Berg and Byrd, 2002; Trakoshis et al., 2022). Lastly, we represented two forms of *accuracy* based on two common Shallice ToL scoring approaches of accuracy: whether a player had reached the target score (1 = true, 0 = false) (*simple_accuracy*) (Berg and Byrd, 2002), and whether a player scored using the fewest number of shots possible (1 = true, 0 = false) (*efficient_accuracy*) (Georgiou et al., 2017; Unterrainer et al., 2004).

4.3. Shallice ToL Implementation

We implemented a digital Shallice ToL (Figure 2) through Experiment Factory - an open-source, reproducible framework of digital measurement tasks (Sochat, 2018). We allow a player to make an unlimited number of moves to solve a problem, but they must solve the problem in under 20 seconds. All problems modeled those used by Shallice, 1982. We scored each player with *efficient_accuracy*: the number of problems the player solved correctly with the fewest number of moves possible (Unterrainer et al., 2004).

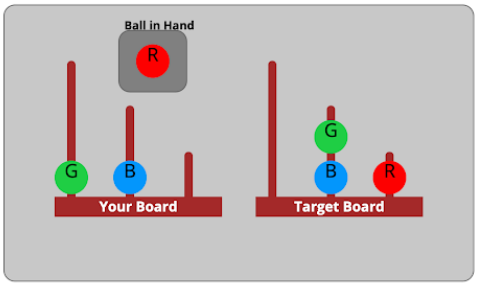

Figure 2. In the Shallice Tower of London (ToL), players arrange balls onto pegs to construct a desired arrangement in the fewest number of steps possible.

4.4. Descriptive Analysis For Adults - Narrowing Our Measures

Several observations of our adult descriptive data allowed us to select the variables for our measurement model. First, we did not find any significant correlations between *num_of_shots* to the Shallice ToL. We saw, however, that ToL correlated significantly and negatively, with Fraction-Unlimited *excess_shots* (-.29, $p < .05$) and Decimal-Unlimited *excess_shots* (-.27 $p < .05$). Thus, we initially chose *excess_shots* over *num_of_shots* to represent the number of basketballs each player shoots with. Next, we chose to use *efficient_accuracy* over *simple_accuracy* in our Unlimited conditions due to strong correlations to the Shallice ToL in the Fraction-Unlimited condition (.46 $p < .001$) and the Decimal-Unlimited condition (.35 $p < .001$). Because we conceptualized *efficient_accuracy* as a composite of *excess_shots* and *simple_accuracy*, we eliminated *excess_shots* as measures of *efficiency*, leaving *num_of_moves* as our only set of *efficiency* measures.

Upon selecting the analytics to incorporate in our measurement model, we organized five confirmatory factor analyses to demonstrate convergent validity between *FBE* and the Shallice ToL; we sought the extent to which our *FBE* measures are also measures of planning. Following these arranged, confirmatory analyses, we examined an additional, exploratory, post-hoc model based on a synthesis of model results from the planned testing. We conducted all analyses through MPlus Version 8.4 with a Weighted Least Squared Mean and Variance adjusted estimator, robust to violations of multivariate normality. We re-coded any item greater than 3 standard deviations or less than -3 standard deviations as missing and removed any extreme cases of outliers.

Table 1. Fraction Ball: Exactly (FBE) Variables

Domain	Variable Name	Decimal-Limited Adults *M (SD)*	Decimal-Limited Children *M (SD)*	Decimal-Unlimited Adults *M (SD)*	Decimal-Unlimited Children *M (SD)*	Fraction-Limited Adults *M (SD)*	Fraction-Limited Children *M (SD)*	Fraction-Unlimited Adults *M (SD)*	Fraction-Unlimited Children *M (SD)*
Speed	*movement_time*	12.41 (14.73)	15.68 (10.01)	34.69 (27.52)	46.11 (38.61)	12.32 (12.73)	15.48 (9.92)	35.18 (31.31)	53.54 (31.88)
Efficiency	*num_of_shots*	3.59 (1.26)	3.45 (1.14)	12.26 (6.3)	5.76 (3.22)	3.76 (1.12)	3.45 (1.17)	10.58 (5.82)	6.69 (3.88)
	excess_shots			2.75 (2.6)	2.51 (2.67)			2.5 (2.5)	3.31 (3.26)
	num_of_moves	1.97 (1.96)	2.92 (2.89)	5.69 (6.52)	7.75 (5.90)	2.32 (2.16)	2.75 (2.38)	5.5 (5.68)	7.73 (5.46)
Accuracy	*simple_accuracy*	.33 (.47)	.43 (.50)	.75 (.44)		.34 (.48)	.31 (.47)	.65 (.48)	
	efficient_accuracy			.21 (.41)	.22 (.42)			.19 (.40)	.16 (.37)

n=80 for adults, n=58 for children

5. Results

5.1. Validating our Adult Measurement Model

Our first model tested the extent to which all our measures were a single underlying skill (planning). Both exact and approximate fit statistics indicated that this model was not a good fit for the data ($\chi^2(54)$ = 145.36, p $<$ .001; RMSEA = .15; CFI = .69).

Our second model represented two factors in which - fractions or decimals - created differences in planning. Both exact and approximate fit statistics indicated that this model was not a good fit for the data ($\chi^2(53)$ = 144.09, p $<$; .001; RMSEA = .15; CFI = .70).

Our third model represented two factors in which game rules - Limited or Unlimited - indicated differences in planning. Both exact and approximate fit statistics indicated that this model was not a good fit for the data ($\chi^2(53)$ = 144.03, p $<$ .001; RMSEA = .15; CFI = .70).

Although we observed a lack of separation by game conditions and formatting, we attempted to run a four factor model, in which *Fraction Ball: Exactly* (*FBE*) indicated differences based on the individual trials (e.g. Fraction-Unlimited, Decimal-Limited , etc.). Our model estimation did not converge, so model our estimates were not trustworthy or interpretable. Based on results from models two and three, we are not surprised how this additional degree of model structure separation was not viable and could confirm that our variables cannot be separated by formatting or game condition.

Our fifth model represented three factors the domains of *Speed*, *Efficiency*, and *Accuracy*. Both exact and approximate fit statistics indicated that this model was not a good fit for the data, although better than models presented thus far ($\chi^2(51)$ = .10, p $<$ .001; RMSEA = 88.67; CFI = .87). All loadings were statistically significant (p $<$ .05). *Efficiency* and *Speed* correlated strongly at singularity (r = .77), hinting that these are not separate constructs. *Accuracy*, on the other hand, did not correlate strongly with *Speed* (r = .02) or *Efficiency* (r = .29) indicating these are separate constructs.

Following our five planned models, we examined a post hoc, exploratory model based on observations from prior models. First, we remove Decimal-Limited *movement_time*, Fraction-Unlimited *movement_time*, and Fraction-Unlimited *num_of_moves*, which seemed to not fit any of the presented models so far except for our final model. Second, we collapse *Efficiency* and *Speed* into a single *Efficiency* factor due to strong correlations in our fifth confirmatory model. Our sixth model for our adult sample comprised two factors in which *FBE* measures captured the domains of *Accuracy*and *Efficiency*. Figure 2 shows the full results of this model. Both exact and approximate fit statistics indicated that this model was a good fit for the data ($\chi^2(47)$ = 58.12, p = .12; RMSEA =.056; CFI = .97). Our *Accuracy* factor did not correlate strongly with *Efficiency*, (r = .25, p = .08) indicating these were separate constructs. Our *Accuracy* factor showed strong correlation with ToL (r = .78, p $<$ .001), but *Efficiency* did not (r = .23, p = .12). Thus, we showed convergent validity with the ToL through *Accuracy*.

5.2. Validating our Children's Measurement Model

After identifying a valid measurement model for the adults, we applied this model with factors in *Efficiency*

Table 2. Two-Factor Fraction Ball: Exactly (FBE) Model with Adult Sample

Factor	Indicators	Standardized Factor Loadings (SE)	Standardized Intercepts/ Thresholds (SE)	Standardized Residual Variances
Efficiency	Decimal-Unlimited *movement_time*	.46 (.11)***	1.297 (.26)***	.79 (.10)***
	Fraction-Limited *movement_time*	.42 (.09) ***	.91 (.25) ***	.83 (.07) ***
	Decimal-Limited *num_of_moves*	.41 (.10) ***	1.01 (.25) ***	.83 (09) ***
	Decimal-Unlimited *num_of_moves*	.68 (0.9) ***	.88 (.26) ***	.53 (.12) ***
	Fraction-Limited *num_of_moves*	.88 (.10) ***	1.07 (.20) **	.23 (.18)
Accuracy	Decimal-Limited *simple_accuracy*	.93 (.05) ***	.51 (.15) **	.12
	Decimal-Unlimited *efficient_accuracy*	.90 (.09) ***	.90 (.17) ***	.20
	Fraction-Limited *simple_accuracy*	.92 (.06) ***	.47 (.15) **	.16
	Fraction-Unlimited *efficient_accuracy*	.79 (.09) ***	.94 (.17) ***	.38

Note: * indicates significance at $p<.05$; ** indicates significance at $p<.01$; *** indicates significance at $p<.001$

and *Accuracy* to our children population. Surprisingly, all indices of model fit indicated significant issues ($\chi^2(47) = 85.36$, $p < .001$; RMSEA = .112; CFI = .76). Though all indicators of efficiency loaded significantly onto this factor, they were not salient indicators (i.e., they did not appear to be reliable measures of a the same underlying construct). Following this discovery, we considered each of the 5 confirmatory models described for adults on children, and unsurprisingly, all evidenced major problems with fit (full model results available upon request).

Due to the strong fit from our *Accuracy* variables in every model presented thus far, and the uncertainty of our measures of *Efficiency*, we present a unidimensional model of planning for our children's data, consisting solely of *Accuracy* measures. We present the full model results in Table 3. Both exact and approximate fit statistics indicated that this model was a good fit for the data ($\chi^2(11) = 10.84$, $p = .69$; RMSEA =.001; CFI $>$ .99). Thus, we converged on a model where children's *Accuracy* were the most informative indicators around the planning construct. We also found that this model had convergent validity with the ToL ($r = .50$ $p <$ .001). The model also showed convergent validity of the children's mathematics skills ($r = .61$ $p < .001$). Thus, our measures of *Accuracy* show evidence of planning in our mathematics game-based context.

6. Discussion

6.1. Intertwining Theory with Validation Results and Assessment Design

Our design and evaluation process demonstrates one approach to embed theory in different stages of game-based assessment design and validation. Although we followed steps outlined by ECGD, the ways we approach theory are not limited to ECGD. In Section 3, we outline how planning theory and design theory of the Shallice ToL influenced *Fraction Ball: Exactly* (*FBE*), particularly when *defining assessment competencies* and *designing assessment mechanics*. Theory also allowed us to validate and better understand our game's measures. We not only developed psychometrically valid analytics of *accuracy*, but we could and better theorize how planning occurred in our adult and student populations. Our adult measurement model supports theoretical hypotheses that measures of planning fall into different categories (Georgiou et al., 2017; Trakoshis et al., 2022). Interestingly, we show that a player's *accuracy* represents a distinct dimension from *efficiency*. While our measures of *accuracy* showed a relationship to planning in adults, our measures of *efficiency* did not. Contrary to our adult measurement model, our

Table 3. One-Factor Fraction Ball: Exactly (FBE) Model with Children Sample

Indicators	Standardized Factor Loadings (SE)	Standardized Intercepts / Thresholds (SE)	Standardized Residual Variances
Decimal-Limited *simple_accuracy*	.81 (.11) ***	.27 (.16) **	.34
Decimal-Unlimited *efficient_accuracy*	.65 (.17) ***	.90 (.17) ***	.58
Fraction-Limited *simple_accuracy*	.99 (.09) ***	.97 (.19) ***	.01
Fraction-Unlimited *efficient_accuracy*	.57 (.16) ***	.97 (.19) ***	.67

Note: * indicates significance at $p<.05$; ** indicates significance at $p<.01$; *** indicates significance at $p<.001$

child measurement model supports hypotheses that planning variables fall onto a single category (Debelak et al., 2015; Unterrainer et al., 2004). Though *accuracy* was evident and precisely measured for children, *efficiency* did not appear as a single identifiable construct for children. Despite theoretical and data-driven justifications for operationalizing *efficiency* as *movement_time* and *num_of_moves*, we could not validate *efficiency* as a component of planning. This does not mean that *efficiency* is unrelated to planning, theoretically, we simply could not justify this theoretical relationship using our chosen measures. Our model could only show the extent to which *accuracy* showed a relationship to planning and mathematical abilities. One possible explanation for realizing two measurement models across different ages stems from considering deliberate attention control and the SAS. When one performs a task that is sufficiently difficult, non-routine, and requires a changing goal, the brain activates the SAS and offers more attentional control. This attentional control exchange results in the person completing a task less efficiently. From this theory, we postulate that children in *FBE* may focus primarily on *accuracy* at the expense of *efficiency*. Our descriptive statistics in Table 1, support this hypothesis as well - while both samples achieved similar accuracy results, adults generally solved problems quicker than children and made more infrequent court moves in all problem scenarios. Thus, insights from our data allow us to develop the following theory of *FBE*: **'planning' tasks in *FBE* successfully challenge supervisory attentional control, and this attentional control may evidence different patterns of efficiency and accuracy depending on the developmental stages of *FBE* players**.

We note that our measurement models cannot provide scientific theories themselves. Measurement models function primarily as a blueprint of a theory (Boorsboom, 2006). However, we generated a theory of how students in our game applied attentional control. Moreover, rigorously validating models with both adults and children allowed us to identify this theory. **Our rigorous validation process not only strengthened a theoretically sound argument for convergent validity, but it allowed us to better theorize how our children interacted with the problems in ways different from adults**.

Intertwining validation results and theory helped us with *iterative improvement* to create stronger measurement models. One iterative change involved collapsing two theoretically similar, yet distinct enough, subdomains - *speed* and *efficiency* - into a single *Efficiency* factor for our adult measurement model. We can theorize that *speed* is theoretically linked to *efficiency* and we will not treat them as separate constructs moving forward. Defining them as distinct subdomains at first, however, gave us more confidence of their similarity as we progressed through analysis.

Beyond the scope of our current study, **theory guides how we might improve the future iterations of our game and measures**. How we grounded **design theory** of the Shallice ToL confirms the importance of ensuring our tasks are sufficiently difficult, non-routine, and require a changing goal. Our results provide a theoretical explanation for how these principles lead to a relationship with planning - *FBE* challenges a student's attentional control and challenges students to trade *efficiency* for *accuracy*. Future game scenarios ought to identify with tasks that challenge alternative hypotheses or dimensions of planning not addressed by the current study. Additionally, our theorized similarity between *efficiency* and *speed* pushes us to identify alternative *efficiency* measures beyond *movement_time* and *num_of_moves*. As we strengthen our understanding of planning in the literature and explore the theories we generate from this existing research, we hope to

further strengthen the way we understand planning in our unique mathematical context and work toward better assessments of planning and mathematics.

6.2. Limitations

Several limitations in this study create opportunities to further validate and iteratively improve *FBE*. First, our existing samples lack specific demographic data. Future research should investigate both adult and child samples, including demographics that may be important sources of individual differences among players. Predictive and discriminant validity will be particularly important to consider in future validation studies. Next, our current study considered *FBE* with regard to planning in a context of mathematics. Our strong correlation to rational number knowledge indicates a need to further consider this relationship and provides an opportunity to incorporate detailed analytics surrounding mathematics. We also propose identifying other EF skills in *FBE*. Next, both the adult and child samples in the current pilot study were smaller than what would be ideally included in a construct validation study. Approximate fit statistics that are less sensitive to sample size (i.e., RMSEA and CFI) all indicated approximate good fit for the models retained. Still, in future research, we hope to continue rigorously validating *FBE* with a significantly larger sample size, as our existing sample size limits claims to generalizability. Finally, we note that game-based assessments ought to continue to evolve and iterate. For example, we note that the present iteration of *FBE* does not incorporate feedback mechanisms - a key step in ECGD. Future designs may use our measurement model - or an improved version of the model - to dynamically adjust difficulty levels.

7. Conclusion

In this paper, we introduce *Fraction Ball: Exactly* (*FBE*) and outline how we embedded *FBE*'s assessment objectives, measurement validation process, and future directions in theory. Our pilot work highlights a tension in designing theoretically-driven assessments for children, as many theories - like those of planning - originate from adult behavior, not children. Through theorizing over the differences in adults and children uncovered via measurement design and configuration, we work toward improved understanding of theories created from adults. We leave the educational technology community with two considerations. First, we encourage assessment developers to continue building theory into *all* stages of the assessment design, validation, and iteration process. Second, we encourage assessment developers to *use* theory and data to generate new understanding of how users engage with a digital assessment, rather than relying on data alone. By linking data-driven findings to theory, we identified a deeper understanding of our assessment construct. We challenge educational technology developers to generate new research theories with the guidance of literature, careful research methodologies, and interdisciplinary collaboration with domain experts. Ultimately, through this exchange of scientific ideas generated from both developers and educational researchers, we can work to create stronger theories for science and better technologies for students.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Anderson, P., Anderson, V., & Lajoie, G. (1996). The tower of london test: Validation and standardization for pediatric populations. *Clinical Neuropsychologist*, *10*(1), 54–65. https://doi.org/https://doi.org/10.1080/13854049608406663

Bagetta, P., & Alexander, P. A. (2016). Conceptualization and operationalization of executive function. *Mind, Brain, and Education*, *10*(1), 10–33. https://doi.org/10.1111/mbe.12100

Berg, W., & Byrd, D. (2002). The tower of london spatial problem-solving task: Enhancing clinical and research implementation. *Journal of Clinical and Experimental Neuropsychology*, *24*(5), 586–604. https://doi.org/10.1076/jcen.24.5.586.1006

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, *81*(6), 1641–1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x

Blair, C., Zelazo, P. D., & Greenberg, M. T. (2005). The measurement of executive function in early childhood. *Developmental Neuropsychology*, *28*(2), 561–571. https://doi.org/10.1207/s15326942dn2802_1

Boorsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Boorsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*,

425–440. https://doi.org/10.1007/s11336-006-1447-6

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives*, *8*(1), 36–41. https://doi.org/10.1111/cdep.120592

Bustamante, A. S., Begolli, K. N., Alvarez-Vargas, D., Bailey, D. H., & Richland, L. E. (2022). Fraction ball: Playful and physically active fraction and decimal learning. *Journal of Educational Psychology*, *114*(6), 1307–1320. https://doi.org/10.1037/edu0000714

Debelak, R., Egle, J., Köstering, L., & Kaller, C. (2015). Assessment of planning ability: Psychometric analyses on the unidimensionality and construct validity of the tower of london task (tol-f). *Neuropsychology*, *30*(3), 346–360. https://doi.org/10.1037/neu0000238

Diamond, A. (2013). Executive functions. *Annual review of psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Georgiou, G. E., Li, J., & Das, J. (2017). Tower of london: What level of planning does it measure? *Psychological Studies*, *62*, 261–267. https://doi.org/10.1007/s12646-017-0416-8

Gerst, E., Cirino, P., Fletcher, J., & Yoshida, H. (2017). Cognitive and behavioral rating measures of executive function as predictors of academic outcomes in children. *Child Neuropsychology*, *23*(4), 381–407. https://doi.org/10.1080/09297049.2015.1120860

Gioia, G. A., Kenworthy, L., & Isquith, P. K. (2010). Executive function in the real world. *Journal of Head Trauma Rehabilitation*, *26*(6), 433–439. https://doi.org/10.1097/HTR.0b013e3181fbc272

Kim, Y. J., Ruipérez-Valiente, J. A., Tan, P., Rosenheck, L., & Klopfer, E. (2019). Towards a process to integrate learning analytics and evidencecentered design for game-based assessment. *Companion Proceedings 9 th International Conference on Learning Analytics Knowledge*, 204–205.

McCoy, D. C. (2019). Measuring young children's executive function and self-regulation in classrooms and other real-world settings. *Clinical Child Family Psychology Review*, *22*(1), 63–71. https://doi.org/10.1007/s10567-019-00285-1

Mejía, C., Herrera-Marmolejo, A., Rosero-Pérez, M., Quimbaya, J., & Cardona, J. F. (2024). Design of a video game for assessment of executive functions in deaf and hearing children. *Applied Neuropsychology: Child*, 1–8. https://doi.org/10.1080/21622965.2024.2311096

Mislevy, R., Oranje, A., Bauer, M., von Davier, A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K., & John, M. (2014). *Psychometric considerations in game-based assessment*. Glasslab.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howeter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100. https://doi.org/10.1006/cogp.1999.0734

Norman, D. A., & Shallice, T. (1980). *Attention to action: Willed and automatic control of behavior* (tech. rep.). Center for Human Information Processing.

Parong, J., Mayer, R. E., Fiorella, L., MacNamara, A., Homer, B. D., & Plass, J. L. (2017). Learning executive function skills by playing focused video games. *Contemporary Educational Psychology*, *51*, 141–151. https://doi.org/10.1016/j.cedpsych.2017.07.002

Rahimi, S., & Shute, V. J. (2023). *Educational technology research and development*. https://doi.org/10.1007/s11423-023-10232-1

Shallice, T. (1982). Specific impairments of planning. *Philosophical transactions of the Royal Society of London. Series B*, *298*(1089), 109–209. https://doi.org/10.1098/rstb.1982.0082

Sikora, M. D., Haley, P., Edwards, J., & Butler, R. W. (2011). Tower of london test performance in children with poor arithmetic skills. *Computer Games and Instruction*, *55*(2), 503–524. https://doi.org/10.1207/S15326942DN2103_2

Sochat, V. (2018). The experiment factory: Reproducible experiment containers. *The Journal of Open Source Software*, *3*(23), 521. https://doi.org/10.21105/joSs.00521

Trakoshis, S., Ioannou, M., & Fanti, K. (2022). The factorial structure of the tower test from the delis-kaplan executive function system: A confirmatory factor analysis study. *Assessment*, *29*(2), 317–331. https://doi.org/10.1177/1073191120960812

Unterrainer, J., Rahm, B., Kaller, C. P., Quiske, K., Hoppe-Seyler, K., Meier, C., Muller, C., Leonhart, R., & Halsband, U. (2004). Planning abilities and the tower of london: Is this task measuring a discrete cognitive function? *Journal of Clinical and Experimental Neuropsychology*, *26*(6), 846–856. https://doi.org/10.1080/13803390490509574