Computational Intelligence and Cognitive Performance Assessment Games

Christoffer Holmgård
Game Innovation Lab
NYU Tandon School of Engineering
Brooklyn, New York 11021
christoffer@holmgard.org

Julian Togelius
Game Innovation Lab
NYU Tandon School of Engineering
Brooklyn, New York 11021
julian@togelius.com

Lars Henriksen
Apex Group ApS
Pilestræde 43
1112 Copenhagen K, Denmark
lars@apex.dk

***Abstract*—In this paper, we present the idea that game design, player modeling, and procedural content generation may offer new methods for modern psychological assessment, allowing for daily cognitive assessment in ways previously unseen. We suggest that games often share properties with psychological tests and that the overlap between the two domains might allow for creating games that contain assessment elements and provide examples from the literature that already show this. While approaches like these are typically seen as adding noise to a particular instrument in a psychometric context, research in player modeling demonstrates that it is possible to extract reliable measures corresponding to psychological constructs from in-game behavior and performance. Given these observations, we suggest that the combination of game design, player modeling, and procedural content generation offers new opportunities for conducting psychometric testing with a higher frequency and a higher degree of personalization than has previously been possible. Finally, we describe how we are currently implementing the first version of this vision in the form of an application for mobile devices that will soon be used in upcoming user studies.**

I. Introduction

Psychological testing and assessment has been practiced since at least 2200 B.C. in China [1], while its modern Western form can trace its roots to the end of the 19th century [2]. Since then, the basic practice of testing has followed roughly the same template: individuals are tested before some liminal decision point, e.g. when a person is eligible for moving into a different grade in school, is applying for a new position, or when the effects of a training program need to be assessed. However, we believe that psychological assessment could be used to greater effect it were possibly to apply it not only at these liminal points, but more often, for instance on a daily basis. Based on the existing literature which shows that games, digital or otherwise, share many properties with psychological tests and may be capable of capturing much of the same information about their users, we intuit that games could be useful enablers for this vision.

If games can be used for psychological assessment, this may open up for new application areas where games or game-like systems can contribute with novel properties - for instance their ability to sustain user engagement and motivate recurrent use through intrinsically and extrinsically motivating features [3], [4].

One example of a novel use case could be for workers who currently undergo uncomfortable drug screening tests based on e.g. biological samples. Workers could avoid these, instead documenting their readiness for work simply by performing a short, daily cognitive test. In effect, a daily cognitive test could be a "breathalyzer for the brain", testing a person's fitness for a particular job or function immediately before starting it. Here, we suggest a number of principles to enable this, centered around using digital games to enable longitudinal, frequent, within-subject measurement of cognitive performance characteristics.

The field of psychometrics has, over the course of more than a century, developed practices for dealing with the specific challenges of testing human skills and capabilities at specific points in time. When developing new instruments and protocols for applying them, three challenges tend to recur in psychometrics: ensuring that the subjects taking tests are motivated and performing at a level representative of their best or typical behavior or abilities, ensuring that tests are valid, and ensuring that they are reliable [1].

Psychometrics already has a number of established and documented approaches to address these challenges, including different conceptions of validity, different ways of measuring reliability, and approaches that attempt to attune themselves to extract the maximal amount of information about the subject from a single test, such as Item Response Theory [5] based tests and Computerized Adaptive Testing [6].

In this paper, we argue that digital game design coupled with player modeling and procedural content generation can offer psychometrics a new set of approaches that may lead to new testing paradigms. Some of these ideas have already been explored in the games literature, extensively even, while others, to the best of our knowledge, are novel and so far unexplored.

We suggest three main design principles drawn from game design and computational intelligence for games and proceed to describe a currently ongoing research project that attempts to realize these ideas, including a description of the current prototype and planned user study. The three approaches from games and computational intelligence we believe could enable longitudinal, frequent, within-subject measurement of psychological performance characteristics are: 1) Using (digital)

game design to motivate subjects to participate in frequent testing. 2) Using player modeling to ensure that the maximal amount of information is learned about the subject from each play through. 3) Using personalized procedural content generation to mitigate and control learning effects, also known as test-retest effects [1], ensuring that tests remain valid and reliable in spite of frequent test administration.

The rest of this paper is structured into four main parts. First, we visit the state-of-the-art in cognitive assessment identifying where games or game-like applications may provide value and novelty. Second, we describe related work in using games to measure psychological characteristics. Third, we describe in detail how the three approaches listed above could be used to realize a specific, novel form of psychological assessment. Fourth, we describe how these ideas drive the design, implementation, and refinement of a prototype application called *SkillShow*. The purpose of SkillShow is to provide a platform for daily testing of performance indicators for simultaneous capacity [7], [8] and inductive reasoning intelligence [9].

II. Related Work

In this section, we start by describing the use of computational intelligence in *Work and Organizational Psychology* (WOP), with a focus on identifying challenges in assessment and selection. We then move on to describing related work in the use of games for identifying psychological characteristics and individual differences in players.

A. Work and Organizational Psychology

The field of WOP has been engaged in addressing the problem of personnel selection and training for more than a century. A typical approach has been to combine fundamental psychometric measurements such as cognitive tests and personality tests with samples of work and interviews for assessment [10]. The use of computerized adaptive testing is also well known in psychology and psychometrics [6], but WOP has only recently started to apply methods from the field of computational intelligence. A nascent movement in WOP is poised to embrace the applications of methods from the artificial and computational intelligence communities [11]. This may bring new levels of specificity and falsifiability to sub-disciplines such as job and task analysis, work behavior measurement, motivation modeling, performance management, personnel assessment and selection, and the modeling of individual differences [11]. The purpose of the research described here is to leverage computational game intelligence to embrace this opportunity.

B. Games for Assessment and Measuring Psychological Characteristics

Often, training simulations/games (such as those used in military or corporate settings) and educational games (such as those used in schools) have been used as measures of the learners'/players' abilities in the subject matter, providing a form of assessment. Sometimes, these assessments have been focused on documenting performance in a narrow field or curriculum or they have been focused on assessing not the students, but the transfer capability of the simulation or game [12].

However, games and assessment at the more general level have been related within the field of simulation and gaming for decades. As early as 1978, Spitz argued that performance in games such as Mancala and Three-in-a-row are related to intelligence [13]. Jones at al. [14] and Jones [15] described how video games might be used for performance assessment. Work in this vein continued throughout the 1980's and 1990's with a focus on assessing intelligence through video games [16], [17].

Using commercial games to quantitatively assess other psychological characteristics and individual differences is a newer idea. Recently, significant work has been done in measuring personality, such as the work by Van Lankveld et al. [18], [19] who specifically suggested games as personality profiling tools [20]. Yee et al. [21] have conducted work in the same vein while Canossa et al. [22] focused on assessing differences in life motivations and later also personality [23]. Tekofsky et al. have shown how play style varies with age [24] in *Battlefield 3* [25] and how performance and speed in the game decrease with age [26]. Finally, Boot el al. recently released an overview of the use of video games as tool for investigating cognitive processes [27].

Meanwhile, the notion of modeling the player for reasons directed at the game experience itself or for impacting the player, such as adaptively changing game difficulty or other parameters to control e.g. player engagement, has been explored for a number of years [28]–[31]. Particularly within applied games, such as educational games or games for health treatment, this has been a focus of significant research into both affective responses and in-game behavior [32]–[34]. Work by Schute et al. [35], [36] has shown how assessment can be built into (educational) games in a way that allows for automatic assessment of student mastery of curricular topics.

Neuroscientifically derived tasks have started appearing in gamification-based frameworks targeting talent identification for recruitment and organizational placement purposes [37]. These approaches to characterizing individuals through games and gamified activities make their way into the general field of WOP, as exemplified by companies such as *Knelf*, *Knack*, *pymetrics*, and *owiwi*[1], contributing to a general change toward a higher reliance on objective data, in the sense of Yannakakis et al. [31][2], and computational intelligence, supporting or supplanting traditional expert-based assessment practices [11].

Altogether, this growing body of research and products shows that activities that take place in games share characteristics with activities included in psychological assessment instruments, from small arcade games to expansive role playing games: The interactions with NPCs and the environment in a game like *Fallout 3* [38] may provide information about

[1] qa.knelf.com; knack.it; pymetrics.com, owiwi.gr

[2] We still consider models built on objective data subjective in the sense that any model's methods and data set are designed, selected, and deployed by individuals or groups thereof.

the same characteristics as the verbal indications given in response to a personality test [23]. Or the performance in a real time strategy game such as *StarCraft 2* [39] may be indicative of cognitive motor performance, a cognitive characteristic normally measured with a specialized test [40]. At the other end of this spectrum, simple tasks drawn from neuroscience, such as the Go/No-Go task, a task developed for assessing attention and inhibitory control/disinhibition, may be put in a gamified context [37] and can be perceived as a game activity by the test subject [41]. This shows that the lines between psychological assessment instruments and games in some instances can be blurry and that the two might be combined for some purposes. In the next section, we show why we believe daily cognitive assessment would be a suitable use case for such a combination.

III. Challenges for Daily Cognitive Assessment

In this section, we identify some of the specific challenges that prevent the use of daily cognitive assessment in WOP today and suggest how game design, computational intelligence, and procedural content generation together might address these. We start by reviewing how game design may alleviate typical issues in test fatigue and retain motivation for daily testing. Afterwards, we suggest that player modeling may drive personalized procedural content generation which in turn can support test-retest reliability while improving the subject/player experience. Building on this idea, we propose that existing research in active player modeling might be used to configure assessment tasks to gather more relevant information about individual players. Then, we propose that assessment games may also provide logistical benefits to frequent cognitive assessment by leveraging existing technologies and principles from game telemetry and game analytics. Finally, we expand upon what we believe is the potential of procedural content generation for psychological assessment.

A. Motivation for Tests

When candidates take psychological assessment tests today, we typically assume that they will be performing at the best possible level they are capable of at that specific moment in time. One reason for this could be that the test is acting as a gatekeeper between the individual and some desired outcome, so we assume the candidate is striving to perform well. Another reason could be that the test is assessing a person's performance levels as part of monitoring during e.g. rehabilitation, where we assume that the patient's goals are aligned with yielding an indicative assessment. However, we could risk that subjects were nervous test takers or under the influence of stereotype threat [42] and therefore performing below their actual optimum. In some instances we may suspect that individuals' goals are not aligned with identifying their maximal performance, such as when insurance claims are involved or e.g. during evaluation for conscription into armed forces. Still, even if subjects might be stressed or malinger we would assume that they took an active and engaged stance toward the test and we are less concerned about the subject's engagement with the test. Extrinsic motivation, other contextual factors, or simply the novelty of the test typically motivate the subject to engage [1].

In the case of frequent testing with the same test(s), however, the circumstances may be radically different. Exposure to the same test or similar tests over and over again is known to cause test or survey fatigue in humans. Simply put, it becomes boring. Subjects who are bored cannot be expected to produce data that is as indicative and valid as data from subjects who are highly motivated for and/or engaged with the test [43]. The amount of noise in the collected data should be expected to increase.

For the use case envisioned here, where we want to assess the same individual on a daily basis, principles drawn from game design seem like suitable approaches to enriching an assessment test with motivating elements reducing this undesirable noise. Identifying gratifying core loops for games that simultaneously work as cognitive performance indicators would be one approach to ensure that the subject remains attentive to and engaged with the test.

Appealing and motivating game designs often incorporate strong elements of feedback to the player, in order to communicate the game's evaluation of her performance and to guide her to play in certain ways or take certain actions [44], [45]. Typically, psychometric test construction avoids these kinds of feedback loops in order to minimize the amount of noise introduced into the test situation and to keep tests comparable. By keeping the context static it becomes easier to assess the individual and compare individuals [1]. In contrast, feedback and performance communication is considered integral to game design [44].

As noted above, the literature on game based testing shows that it is possible to create motivating games rich in feedback that still have acceptable validity as assessment tools. This indicates that it should also be possible to motivate the player through interesting, varying gameplay while still accurately evaluating player performance through player models that take the game as context into consideration [46]–[48] and by extension we should most likely also be able to evaluate their cognitive characteristics.

A problem related to dealing with the context of the game is dealing with the developing expertise of the player, as she becomes more proficient through experience. This, in turn, is related to the problem of test-retest reliability in psychological testing, which we approach in the following section.

B. Test-retest Reliability

Game design and player modeling excel at evaluating individuals in manners perceived as fair across contexts that are comparable in general, but vary in their specific configuration. Scores obtained in individual play sessions of e.g. *Tetris* [49] are generally considered comparable to one another even though the specific sequence of tetrominoes encountered may have been different. The balancing of the game is assumed to provide guarantees that even though specific game instances are unique, their difficulties are roughly equal over time and

part of the gameplay is managing this uncertainty [50]. If games feature progression, a player generally expects a well-planned challenge curve that matches her development of skill over time, or even adapts to it [28].

In psychological assessment, the development of expertise with regard to the test itself is generally viewed as undesirable, as this is assumed to obscure actual underlying performance characteristics of the subject which are applied to the test as a task, but not trained by or developed from it [1]. In games, and in particular games that leverage computational intelligence for difficulty adjustment, this development is generally leveraged as an asset. Knowing the player's development of expertise allows the game designer or the game artificial intelligence to configure the game to an appropriate difficulty level. Importantly, a game may keep a history of the player's skill development and may store the context for the exhibited performances too [51], [52]. This shows that player modeling already contains the necessary frameworks to track and calibrate for the player's development of expertise over time by adjusting the challenge/difficulty [31]. For the case of daily cognitive assessment using games, this becomes particularly important, as each subject/player may follow an individual learning curve. We engage with this topic in the following section.

C. Individualized Test Sensitivity

Classical psychological assessment would typically provide all subjects with the same instrument and use principles such as e.g. Item Response Theory, time limits, or progress measures through instrument items to gain sufficient information about each subject [1]. More recently, Computerized Adaptive Testing has started providing methods for making individualized test configurations that adapt during testing by selecting appropriate items from item pools [6]. This approach has a natural counterpart within games where personalized [53], experience driven content generation [54] is capable of generating content that is appropriate to e.g. a player's skill level or emotional state. Additionally, research has shown how player models can be used to drive not only the generation of content that matches the player's desired experience, but also content that will reveal the maximal amount of information about the player. As such, the combination of player modeling and procedural content generation provides methods for games that may extend current practices within psychological testing.

D. Testing Costs

A typical concern that may limit the application of psychological assessment today is the cost of deploying tests. Even though many psychological tests are now available in digital versions, ported from their original paper version and scored automatically by local or remote software, they still typically assume controlled environments and administration by professionals. Deploying psychological assessment solutions in the form of games on mobile devices will be able to tap into existing, standardized distribution platforms and may leverage existing data collection, aggregation, and analysis frameworks [51]. Taking this into account, games applied as assessment tools might be able to extend current psychological testing practice through the infrastructure that games (and mobile games in particular) bring.

In the following section we go into deeper detail about how player modeling and procedural content generation may offer new methods for psychological testing.

IV. Procedural Content Generation in Assessment Games

The promise of procedural content generation is to automatically create new games, new game variants or simply new game content each time a test is taken. By now, the general problem of generating content for games is fairly well understood, and a number of effective methods exist for e.g. generating levels, textures, puzzles, characters, and vegetation for games ranging from platformers to puzzle games to open-world adventures [55]. In the context of cognitive performance assessment games the role of procedural content generation would be to introduce variety, which serves both to avoid learning effects and to ensure continued player engagement.

Procedural generation can be applied to different levels in a game—in more constrained settings it might be a question of changing a few parameters, in less constrained settings a matter of generating new structural content such as a level, and in the least constrained setting procedural generation can be applied to the very rules of a game, creating new variants, but with similar underlying challenges.

We hypothesize that the less constrained the procedural generation is, the better learning effects can be avoided. If the test consists of the same game at each occasion with only minor variations in the parameters or the level, learning effects will likely only be partially mitigated—there will still be significant training benefits from having played another (similar) version of the same game. If instead an entirely new game is generated for each test, learning effects are likely to be negligible, as the time and effort spent taking previous tests is not likely to improve the individual's performance on the new test. On the other hand, the more the game changes between tests, the less comparable the results will be and reliability may suffer as a result and any prior established validity may be invalidated. This is a trade-off which will need to be explored in more depth in future research. Particularly investigating how much change along one or more dimensions impacts validity, as measured using e.g. well-known psychological tests as criteria, will be a significant challenge.

An initial strategy toward addressing this could be simulation-based testing of the generated games. This approach should be able to ascertain the level at which they challenge a particular cognitive skill, or in other words the performance/behavior in a game relative to a given latent construct. To do this, we envision developing computational agents that can learn to play any of these games with the same skill and playing style as a particular human player. We refer to these player-imitating agents as *procedural personas* [56], adaptive agents that learn to reproduce human play

skill, preferences, and implicitly playing styles in a single game [56]. This is done by identifying common patterns of skills and preferences in games, and biasing existing game-playing algorithms with these patterns. Such agents would be trained on a player's testing/playing history and become increasingly representative over time. Creating computational agents that can learn to play the testing games in a similar way to the human player/test-taker, and with the same performance, is not a trivial task, both because the performance needs to carry over from the game variant the agent was trained on to the new game variant, and because most computational game-playing agents tend to play in distinctly non-human-like ways. When a new game, or a new variation of a game, has been generated, it can first be played by the procedural persona to set a baseline for the player's expected performance on that game. This moves the problem of reliability from the particular instance of the test/game and instead places it on the persona as a user/player model. As long as this is representative of the player it should enable a procedural content generation system to choose configurations with an appropriate discriminatory ability. Prior work has demonstrated that it is possible to represent skill in game playing agents [57] and to bias game playing agents towards exhibiting more human-like play styles [58], [59].

Modeling players' performance and playing style can be taken even further in order to more accurately assess the player. The paradigm of Active Player Modeling uses active learning in player modeling [60]: the space of content is searched for the content areas where the model is least certain about the player's performance. In other words, the game content generator probes the content space to maximally improve its knowledge about the player. This approach could be very effective for exploratory cognitive assessment.

Active player modeling can be usefully compared to Computerized Adaptive Testing, where test items are chosen from a pool in order to provide the most appropriate test questions for a given test-taker in order to maximize the information gained from each item [6]. The difference to our envisioned system is that each configuration of the test/game, comparable to an item or an item set, is selected or generated and then configured based on simulations in response to the player model, which is continuously updated. This could allow for accurate generation of tests for individual test-takers, taking into account variation among multiple dimensions of cognitive performance.

Given our assumption that games applied as assessment instruments might bring new methods to the assessment of cognitive characteristics within motivation, reliability, sensitivity, and cost, we have developed a prototype test platform, which we describe in the following section.

V. The SkillShow Prototype

In this section we describe the work-in-progress SkillShow prototype. The prototype is built in accordance with the three principles outlined above in Section I: game design should be applied to build motivation, player modeling in context should be applicable through game playing agents, and procedural content generation should be supported to ensure novelty and maximal individual information gain from each test session. In this application the three principles are applied in moderation and the application design still leans heavily on existing psychological tests. SkillShow is designed to measure two different latent constructs: *simultaneous capacity* [8] and *inductive reasoning intelligence* [9] Each construct will be measured through a test/game that combines and adapts existing game designs and existing psychological tests. Given that we can successfully demonstrate that the suggested approaches work for these tests/games, we envision moving on to more complex tests/games later. Below, we describe each of the two tasks that we currently plan to include in our user studies.

TABLE I
Overview of the Design Characteristics of the *Interrupted SET* Test/Game.

Game genre/template	Card game: SET
Game mechanics	Set identification, sorting, search
Test inspiration	SIMKAP
Latent construct	Simultaneous capacity

A. Interrupted SET

The first mode of SkillShow is titled *Interrupted SET*. The fundamental psychological task is drawn from a well-known assessment instrument called *SIMKAP* [7], [8]. It is typically used in the assessment and selection of personnel for critical functions such as ship captains, fighter pilots, or air traffic controllers. SIMKAP measures a construct called simultaneous capacity: an individual's ability to perform several mental operations simultaneously and switching between these tasks based on outside demands. In order to fuse the properties of the SIMKAP test with a motivating game design that supports player modeling and adaptive procedural content generation, we borrow and adapt the rules of the game *SET* [61]. The game requires players to analyze a selection of 18 cards in order to create sets of three cards that are either all identical or all unique on four dimensions: count, color, shape, and fill type. The player must identify as many sets as possible within a time limit. While the player is solving this task we intermittently interrupt the player by presenting distracting, overlaid tasks that require the player to either sort integers or conduct visual search for characters. The resulting test/game is characterized in Table I and the two game modes are displayed in Figure 1. We expect the challenge to be configurable through the number of sets present in the each initial card spread and the frequency and complexity of the distracting tasks.

B. Simple MULTIFLUX

The second mode of SkillShow is titled *Simple MULTIFLUX* and is adapted from the work of Kröner et al. who show how interactive computer simulations can be used to assess *inductive reasoning intelligence* [9]. They develop a simulation-based task where the subject must go through three stages of understanding a dynamic system: identifying rules,

(a) Identifying sets. (b) One kind of interruption.

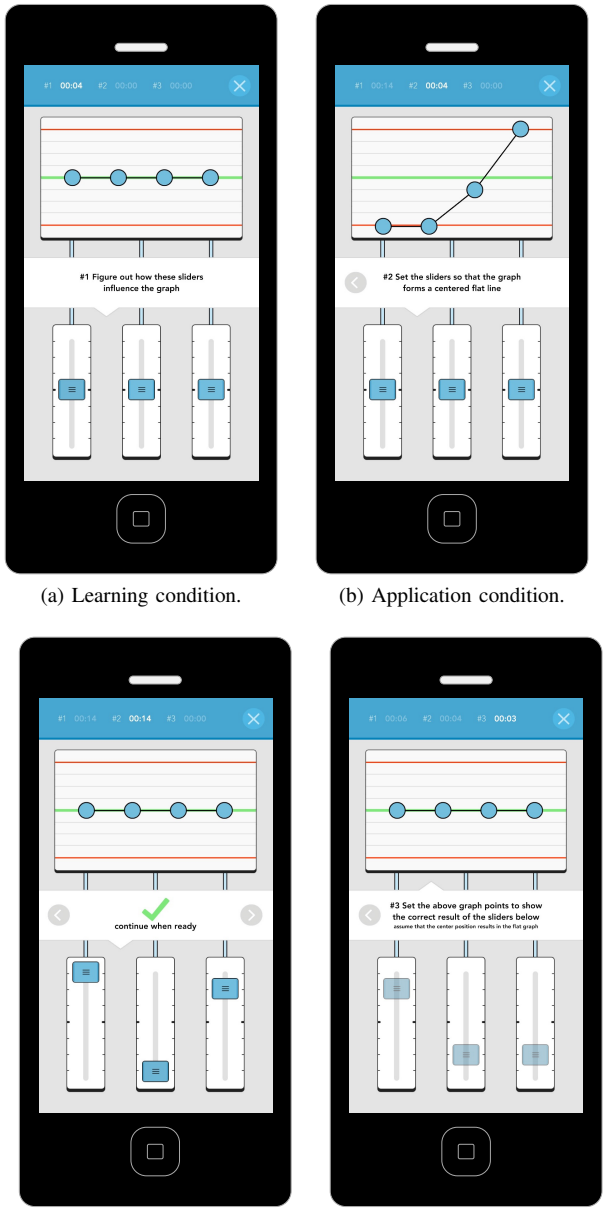
Fig. 1. Two modes of Interrupted SET.

TABLE II
OVERVIEW OF THE DESIGN CHARACTERISTICS OF THE *Simple MULTIFLUX* TEST/GAME.

Game genre/template	Puzzle games
Game mechanics	Dynamic systems learning/manipulation
Test inspiration	MULTIFLUX
Latent construct	Inductive reasoning intelligence

applying rules, and demonstrating understanding of rules. First, the subject is given time to analyze the relations between a number of control inputs and the outputs of an abstract, simulated machine. Secondly, the subject is given an instance of this abstract machine and asked to bring it into a particular goal state using the acquired information. Thirdly, the subject is given a pre-configured machine and asked to determine how the controls must be arranged to cause this output. This test is reminiscent of many abstract puzzle games and well suited to the mobile format. A design overview of this test/game is shown in Table II. In order to increase the appeal of the task, we designed a compelling user interface, simplifying the display and adding visual, auditive, and animation feedback. We did not change the core rules of the test as this was deemed appropriate and engaging in its original form. The resulting design is displayed in Figure 2.

The original MULTIFLUX test has four inputs and four outputs and gave subjects 7 minutes and 30 seconds to solve the tasks. Through informal experimentation, we deemed this to be too complex, and reduced the problem to our *Simple MULTIFLUX* variant with only three inputs and three outputs and a 5 minute time limit. We expect the difficulty of this test/game to be configurable via the complexity of the function that maps inputs to outputs in the simulated machine.



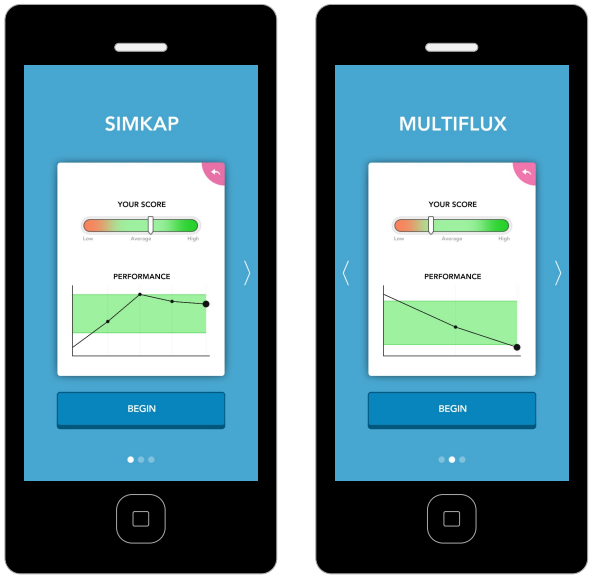
(a) Learning condition. (b) Application condition.

(c) Solved application condition. (d) Demonstration condition.

Fig. 2. Four states of Simple MULTIFLUX.

C. Feedback and Evaluation Screens

In order to support player motivation and adherence both test/games in SkillShow are built with rich visual and auditive feedback. The tests/games are built to a production value that would seem familiar to a player used to playing premium puzzle games on their mobile phone. Additionally, the application includes immediate feedback after each play session. The application rates the player's performance in the latest play in relation to previous performances and displays a graph with recent sessions, shown in Figure 3.



(a) SIMKAP evaluation (b) MULTIFLUX evaluation

Fig. 3. Two evaluation screens from SkillShow providing individualized feedback.

VI. Evaluating the SkillShow Application

The next step for evaluating the SkillShow application's usefulness for daily cognitive assessment will be to conduct a pilot user study with the two games. At the present time, two user study configurations are planned.

First, since the tests/games are based on existing psychological assessment instruments, a longitudinal study where subjects/players play the two games once a day will be run. Intermittently, during this period of time, the players will be asked to complete the original tasks under lab conditions. In addition to conducting the original, validated tasks that the tests/games are based upon, the subjects/players will also be asked to complete a battery of other, validated tests measuring well-known psychological constructs, such as attention and intelligence. Additionally, once a baseline is established for the players, a number of interventions will be staged, where participants will be split into a treatment group and a control group. The treatment group will undergo a procedure known to reduce cognitive performance; currently we are contemplating a moderate amount of sleep deprivation. We expect the treatment group to exhibit significantly worse performance in both games on the occasion when they are sleep deprived, relative to their baseline, and we expect them to return to this baseline when they are no longer sleep deprived. We expect this effect to be present for both the reference psychological tests, and the SkillShow versions. For the control group, we expect no effect other than performance increases due to learning effects.

Once this pilot study is complete we intend to build individual player models, modeling skill and style over time. This, in turn, will allow us to construct procedural difficulty adjustment systems for each task, that may be used in a subsequent second pilot study. If learning and development of skill is observed in the first study, the purpose of this second study will be to enable the SkillShow application to keep challenge steady and maximize information gain by adjusting the difficulty of the tasks to match the players' skill development.

VII. Conclusion

In this paper, we outlined a number of reasons, based on the literature, to assume that games applied as cognitive tests may be able to facilitate daily cognitive assessment through motivating, reliable, personalized, and cost effective tests. We described four key ways in which combining this approach with player modeling and procedural content generation might bring novel methods to cognitive testing in general. Additionally, we described a prototype for a first pilot user study testing the efficacy of games as daily assessment tools. The tasks included in the prototype are relatively conservative interpretations of existing psychological tests, but fused with game mechanics drawn from existing games, and game design principles such as such rich interfaces and feedback. If the pilot-studies show the tests/games can provide valid and reliable assessment we will expand the prototype application with more elaborate games, player modeling, and procedural content generation.

The overarching vision described in this paper is to explore the possibilities in combining key elements from game design and computational intelligence in games, specifically player modeling and procedural content generation, with psychological testing. The SkillShow prototype represents our first step in exploring these possibilities.

Acknowledgment

The SkillShow application is developed and owned by Apex Group ApS and Informantics ApS.

References

[1] R. Gregory, *Psychological Testing: History, Principles, and Applications*. Pearson, 2011.
[2] J. M. Cattell, "Mental tests and measurements with remarks by F. Galton," *Mind*, vol. 15, no. 59, pp. 373–381, 1890.
[3] M. R. Lepper and T. W. Malone, "Intrinsic motivation and instructional effectiveness in computer-based education," *Aptitude, learning, and instruction*, vol. 3, pp. 255–286, 1987.
[4] C. Klimmt and T. Hartmann, "Effectance, self-efficacy, and the motivation to play video games." 2006.
[5] F. B. Baker, *The basics of item response theory*. ERIC, 2001.
[6] H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, and R. J. Mislevy, *Computerized adaptive testing: A primer*. Routledge, 2000.
[7] B. Rosmark, "Validering av ett simultankapacitetstests prediktionsförmåga av framgång i utbildningen av båtchefer (stridsbåtförare)," *Stockholm, Sweden: Pliktverket Regionkontor Stockholm*, 2001.
[8] O. Bratfisch and E. Hagman, "Simultankapazität/multi-tasking (simkap) version 24.00: Handanweisung (simultaneous capacity/multi-tasking (simkap) release 24.00: Manual)," *Mödling, Austria: Schuhfried*, 2003.
[9] S. Kröner, J. Plass, and D. Leutner, "Intelligence assessment with computer simulations," *Intelligence*, vol. 33, no. 4, pp. 347–368, 2005.
[10] N. Chmiel, *An introduction to work and organizational psychology: a European perspective*. John Wiley & Sons, 2008.
[11] J. M. Weinhardt and J. B. Vancouver, "Computational models and organizational psychology: Opportunities abound," *Organizational Psychology Review*, vol. 2, no. 4, pp. 267–292, 2012.
[12] J. Chin, R. Dukes, and W. Gamson, "Assessment in simulation and gaming a review of the last 40 years," *Simulation & Gaming*, vol. 40, no. 4, pp. 553–568, 2009.

[13] H. H. Spitz, "The universal nature of human intelligence: Evidence from games," *Intelligence*, vol. 2, no. 4, pp. 371–379, 1978.

[14] M. B. Jones, R. S. Kennedy, and A. C. Bittner Jr, "A video game for performance testing," *The American Journal of Psychology*, pp. 143–152, 1981.

[15] M. B. Jones, "Video games as psychological tests." *Simulation & Games*, 1984.

[16] P. Rabbitt, N. Banerji, and A. Szymanski, "Space fortress as an iq test? predictions of learning and of practised performance in a complex interactive video-game," *Acta Psychologica*, vol. 71, no. 1, pp. 243–257, 1989.

[17] E. Donchin, "Video games as research tools: The space fortress game," *Behavior Research Methods, Instruments, & Computers*, vol. 27, no. 2, pp. 217–223, 1995.

[18] G. Van Lankveld, S. Schreurs, and P. Spronck, "Psychologically verified player modelling." in *GAMEON*, 2009, pp. 12–19.

[19] G. Van Lankveld, S. Schreurs, P. Spronck, and J. Van Den Herik, "Extraversion in games," in *International Conference on Computers and Games*. Springer, 2010, pp. 263–275.

[20] G. van Lankveld, P. Spronck, J. Van den Herik, and A. Arntz, "Games as personality profiling tools," in *Conference on Computational Intelligence and Games*. IEEE, 2011, pp. 197–202.

[21] N. Yee, N. Ducheneaut, L. Nelson, and P. Likarish, "Introverted elves & conscientious gnomes: The expression of personality in world of warcraft," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 753–762.

[22] A. Canossa, J. B. Martinez, and J. Togelius, "Give me a reason to dig: Minecraft and psychology of motivation," in *Conference on Computational Intelligence and Games*. IEEE, 2013, pp. 1–8.

[23] A. Canossa, J. B. Badler, M. S. El-Nasr, S. Tignor, and C. R. Colvin, "In your face(t). impact of personality and context on gameplay behavior," in *Foundations of Digital Games*, 2015.

[24] S. Tekofsky, P. Spronck, A. Plaat, J. Van den Herik, and J. Broersen, "Psyops: Personality assessment through gaming behavior," in *BNAIC 2013: Proceedings of the 25th Benelux Conference on Artificial Intelligence*, 2013.

[25] EA DICE, *Battlefield 3*. Electronic Arts, 2011.

[26] S. Tekofsky, P. Spronck, M. Goudbeek, A. Plaat, and J. van den Herik, "Past our prime: A study of age and play style development in battlefield 3," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 7, no. 3, pp. 292–303, 2015.

[27] W. R. Boot, *Video games as tools to achieve insight into cognitive processes*. Frontiers Media SA, 2015.

[28] R. Hunicke, "The case for dynamic difficulty adjustment in games," in *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*. ACM, 2005, pp. 429–433.

[29] R. Dias and C. Martinho, "Adapting content presentation and control to player personality in videogames," in *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*. ACM, 2011, p. 18.

[30] S. C. Bakkes, P. H. Spronck, and G. van Lankveld, "Player behavioural modelling for video games," *Entertainment Computing*, vol. 3, no. 3, pp. 71–79, 2012.

[31] G. N. Yannakakis, P. Spronck, D. Loiacono, and E. André, "Player Modeling," in *Artificial and Computational Intelligence in Games*. Saarbrücken/Wadern: Dagstuhl Publishing, 2013, pp. 45–55.

[32] J. Robison, S. McQuiggan, and J. Lester, "Evaluating the consequences of affective feedback in intelligent tutoring systems," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–6.

[33] M. D. Kickmeier-Rust and D. Albert, "Educationally adaptive: Balancing serious games." *International Journal of Computer Science in Sport*, vol. 11, no. 1, 2012.

[34] H. Wang, H.-T. Yang, and C.-T. Sun, "Thinking style and team competition game performance and enjoyment," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 7, no. 3, pp. 243–254, 2015.

[35] V. J. Shute and R. Glaser, "A large-scale evaluation of an intelligent discovery world: Smithtown," *Interactive Learning Environments*, vol. 1, no. 1, pp. 51–77, 1990.

[36] V. J. Shute, "Stealth assessment in computer-based games to support learning," *Computer games and Instruction*, vol. 55, no. 2, pp. 503–524, 2011.

[37] A. B. Collmus, M. B. Armstrong, and R. N. Landers, "Game-thinking within social media to recruit and select job candidates," in *Social Media in Employee Selection and Recruitment*. Springer, 2016, pp. 103–124.

[38] Bethesda Game Studios, *Fallout 3*. Bethesda Softworks, 2008.

[39] Blizzard Entertainment, *StarCraft II: Wings of Liberty*. Blizzard Entertainment, 2010.

[40] J. J. Thompson, M. R. Blair, and A. J. Henrey, "Over the hill at 24: persistent age-related cognitive-motor decline in reaction times in an ecologically valid video game task begins in early adulthood," *PloS one*, vol. 9, no. 4, p. e94215, 2014.

[41] A. Lieberoth, "Shallow gamification. testing psychological effects of framing an activity as a game," *Games and Culture*, vol. 10, no. 3, pp. 229–248, 2015.

[42] C. M. Steele and J. Aronson, "Stereotype Threat and the Intellectual Test Performance of African Americans," *Journal of Personality and Social Psychology*, vol. 69, no. 5, p. 797, 1995.

[43] J. E. Beck, "Engagement tracing: using response times to model student disengagement," in *Artificial Intelligence in Education*. IOS Press, 2005, pp. 88–95.

[44] T. Fullerton, C. Swain, and S. Hoffman, *Game Design Workshop: Designing, Prototyping, and Playtesting Games*. Focal Press, 2004.

[45] K. Salen and E. Zimmerman, *Rules of Play: Game design fundamentals*. MIT press, 2004.

[46] P. H. M. Spronck *et al.*, *Adaptive Game AI*. UPM, Universitaire Pers Maastricht, 2005.

[47] A. Zook, S. Lee-Urban, M. R. Drinkwater, and M. O. Riedl, "Skill-based mission generation: A data-driven temporal player modeling approach," in *Proceedings of the Third Workshop on Procedural Content Generation in Games*. ACM, 2012, p. 6.

[48] G. N. Yannakakis and J. Togelius, "A Panorama of Artificial and Computational Intelligence in Games," *IEEE Transactions on Computational Intelligence and AI in Games*, no. 99, p. 1, 2014.

[49] A. Pajitnov and V. Pokhilko, *Tetris*. Alexey Pajitnov, 1984.

[50] G. S. Elias, R. Garfield, and K. R. Gutschera, *Characteristics of Games*. MIT Press, 2012.

[51] M. S. El-Nasr, A. Drachen, and A. Canossa, *Game Analytics: Maximizing the Value of Player Data*. Springer Science & Business Media, 2013.

[52] J. J. Thompson, M. R. Blair, L. Chen, and A. J. Henrey, "Video game telemetry as a critical tool in the study of complex skill learning," *PloS one*, vol. 8, no. 9, p. e75129, 2013.

[53] N. Shaker, G. Yannakakis, J. Togelius, M. Nicolau, and M. O'Neill, "Evolving personalized content for super mario bros using grammatical evolution," in *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2012.

[54] G. N. Yannakakis and J. Togelius, "Experience-driven procedural content generation," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 147–161, 2011.

[55] N. Shaker, J. Togelius, and M. J. Nelson, *Procedural Content Generation in Games: A Textbook and an Overview of Current Research*. Springer, 2016.

[56] C. Holmgård, A. Liapis, J. Togelius, and G. N. Yannakakis, "Evolving personas for player decision modeling," in *Conference on Computational Intelligence and Games*. IEEE, 2014.

[57] A. Zook, B. Harrison, and M. O. Riedl, "Monte-carlo tree search for simulation-based strategy analysis," in *Foundations of Digital Games*, 2015.

[58] D. Whitehouse, P. I. Cowling, E. J. Powley, and J. Rollason, "Integrating monte carlo tree search with knowledge-based methods to create engaging play in a commercial mobile game." in *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2013.

[59] A. Khalifa, A. Isaksen, J. Togelius, and A. Nealen, "Modifying mcts for human-like general video game playing," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016.

[60] J. Togelius, N. Shaker, and G. N. Yannakakis, "Active player modelling," in *Foundations of Digital Games*, 2014.

[61] M. J. Falco, J. Langdon, and F. Vohwinkel, *SET*. 999 Games, 1988.