Actions Speak Louder Than Words:

An exploration of game play behavior and results from traditional assessments of individual differences

Laura Levy, Rob Solomon, Maribeth Gandy
Georgia Institute of Technology
85 5th Street NW
Atlanta, Georgia 30308
+1-404-894-4195
[laura, rob, maribeth]@imtc.gatech.edu

Joann Moore, Jason Way, Ruitao Liu
ACT, Inc.
500 ACT Drive
Iowa City, Iowa 52243
+1-319-337-1499
[joann.moore, jason.way, ruitao.liu]@act.org

ABSTRACT

In this paper, we describe the results of an exploratory pilot study examining the use of a video game as an assessment tool. This is part of a larger project aimed at identifying the components and mechanics of accurate and engaging assessment games. A critical step towards this goal is to understand what player behaviors might naturally be associated with traditionally measured human variables (e.g. cognitive and non-cognitive individual differences). We use the custom game, *Food for Thought*, and its in-game logging capabilities to examine relationships between different play behaviors, like planning and level retrying, with results from traditionally measured psychosocial, multitasking, and personality variables. In a sample of 30 undergraduate students, we found 10 statistically significant correlations between in-game player behavior and measures from traditional assessments (ACT Engage, SynWin, and the BFI-44). These preliminary results suggest promise in understanding human individual differences through the exploration of player behaviors.

Categories and Subject Descriptors

K.8.0 [**Personal Computing**]: General – games.

J.4. [**Social and Behavioral Sciences**]: Psychology.

General Terms

Measurement, Design, Experimentation, Performance, Human Factors.

Keywords

Assessment, individual differences, cognitive psychology, personality, gaming

1. INTRODUCTION

The use of traditional scientific assessments to evaluate human performance and individual differences is prevalent in modern society. Traditional tests include familiar pen-and-paper exams, like the ACT college readiness exam and digital assessments, like the Graduate Record Examinations (GRE). Traditional assessments can be used to measure a number of human variables, such as knowledge, skills, performance, and self-reported attributes.

While traditional assessments often have the benefit of decades', if not a century's, worth of research into their validity and reliability, there are also drawbacks in using them. Traditional assessments can be time intensive for the test taker to complete and require special training for the test administrator to proctor and analyze. Test takers, knowing they are being assessed, might also experience increased stress that creates an opportunity for diminished performance. Additionally, a test taker might purposefully alter their selections on self-report assessments (e.g., personality) so that scores reflect more desirable qualities (such as appearing more agreeable). Finally, measurements of certain dimensions, like creativity, can be limited by the testing format itself.

Recently, there has been growing interest in the use of video games as assessment tools. Games can provide benefits over traditional tests, while still yielding accurate evaluations [1]. The complex environment of games is an interesting test environment that allows for a wide variety of actions and action sequences to be utilized by the player and analyzed by the researcher. By concealing the assessment within a compelling game, one can also create an engaging and less stressful testing experience. A properly instrumented game could provide an accurate, efficient, and engaging means of assessing a variety of human variables.

Currently, there is a need to further our understanding of the design and implementation of assessment games. Particularly, research is needed into game mechanics that might provoke naturally emerging play behaviors appropriate for statistical analysis. In this study we use the custom game, *Food for Thought*, and its in-game logging capabilities to examine relationships between game play behavior and results of traditionally measured participant attributes of cognitive and non-cognitive variables.

2. RELATED WORK

One popular method utilized for the design and validation of assessment games is through evidence-centered design (ECD) [2]. This conceptual framework guides how to both identify and evaluate variables of interest that might not be directly observable. For example, creativity is a challenging variable to assess. However, the task becomes easier after defining some parameters of creativity and then matching those to identified performance thresholds [3].

Therefore, one of the first steps to understanding the design of valid and appropriate assessment games is to evaluate what naturally occurring game play behaviors might be associated with inherent characteristics of the player. Human individual differences that drive observable behavior include cognitive and non-cognitive variables. Perhaps analysis of player strategy can

Proceedings of the 10th International Conference on the Foundations of Digital Games (FDG 2015), June 22-25, 2015, Pacific Grove, CA, USA.
ISBN 978-0-9913982-4-9. Copyright held by author(s).

reveal qualities about that person (like their personality) that would normally be measured with a traditional test. It may be possible that a player's behavior and reaction to in-game events can be just as informative and valid as any traditional assessment.

Using games as assessment tools presents an interesting opportunity to cleverly hide what is being evaluated, known as a "stealth assessment" [4, 5]. This can preserve feelings of engagement for the player, and also aid in self-report situations whereby a player might exaggerate desirable qualities. Companies have long been aware of personality test-takers that charm their way through the test to get the job. Concealing what is being assessed is useful in protecting it from those that would seek to "game the game".

One example of stealth assessment was demonstrated using the commercially available video game, *Elder Scrolls IV: Oblivion* [6]. Using an ECD-based approach, player actions were identified and then analyzed using a Bayesian model to reveal novel behaviors as measures for problem solving and creativity [5]. For example, a player in the game challenged with crossing a river might choose a number of actions. The player might choose a common but un-efficient river crossing method of simply swimming across. Fewer players might choose a more novel, but also inefficient, passage of burrowing under the river. Even fewer players might freeze the river and then slide across. This last action represents a rare, novel, and efficient player choice. Games that afford many different executable actions naturally allow for many different possible combinations of actions. This variety of behaviors provides a rich dataset to analyze for differences between and within players.

Exploratory correlation analyses have also proved useful for examining relationships between player game behaviors and traditional measures of individual differences, like personality [7, 8, 9]. This study utilizes this approach to examine player behavior and its relationship with traditionally assessed results.

3. STUDY DESCRIPTION

This paper describes current results of an on-going collaborative study with Georgia Tech and ACT, Inc., creator of the ACT college readiness assessment. The data comes from a pilot study of 30 college-aged (18-23 years) students (9 female, 21 male) that played the desktop computer game, *Food for Thought*.

Food for Thought is a single-player abstracted cooking game that emphasizes multitasking. Players strategically move ingredients through the game kitchen, taking care to prepare items properly in order to complete the recipe and earn a high score. Meanwhile, mini-games and distractions (such as flies and fires) add additional challenges to the gameplay. *Food for Thought* encourages planning ahead, multitasking, working efficiently, and reviewing end-level feedback.

A unique benefit of using *Food for Thought* as an assessment tool is the ability to examine player action and behavior through its automated data logging system. All player input actions are captured, time-stamped, and saved in a participant and session specific XML file. This allows for analysis of performance at a global, station, and step-wise level, in addition to fine measurement of durations spent in specific game stages and the ability to examine player strategy through mouse activity heat maps. Together, this provides a powerful tool in examining player strategies and the impact those might exert on player performance.

.

4. METHOD

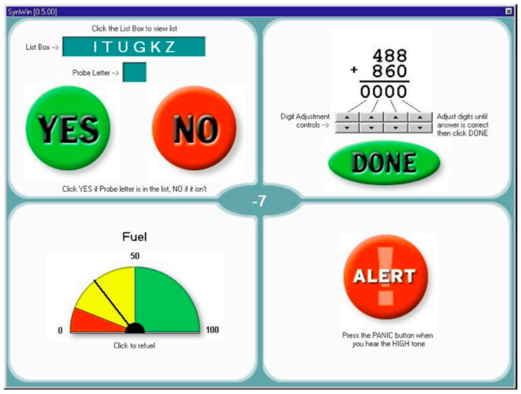
4.1 Materials

4.1.1 ACT Engage

ACT Engage College is an assessment of college students' academic behavior and measures 10 psychosocial factors (PSFs) that have been shown to reliably predict academic performance and persistence in college [10], even beyond traditional measures such as high school GPA and standardized achievement tests [11]. It contains 108 items that are measured on a 6-point Likert scale ranging from 1 (Strongly Disagree) to 6 (Strongly Agree). The 10 PSFs are grouped into three broad domains. *Motivation* consists of personal characteristics that help students to succeed academically by focusing and maintaining energies on goal-directed activities (i.e., Academic Discipline, General Determination, Goal Striving, Commitment to College, Study Skills, Communication Skills). *Social Engagement* consists of interpersonal factors that influence students' successful integration or adaptation into their environment (i.e., Social Connection, Social Activity). Finally, *Self-Regulation* consists of cognitive and affective processes used to monitor, regulate and control behavior related to learning (i.e., Academic Self-Confidence, Steadiness). These definitions, and more detail about the Engage College assessment, can be found in the Engage College User's Guide [12].

4.1.2 Multitasking in SynWin

SynWin (previously known as SYNWORKS1) is a customizable digitally administered multitasking environment that presents two cognitive (memory and arithmetic) and two perceptual tasks (visual and auditory monitoring) to the user (see Figure 1) [13, 14, 15]. Individual and global task scores are generated by the tool providing data on a participant's performance in parallelizing multiple tasks.



Click the List Box to view list

Click YES if Probe letter is in the list, NO if it isn't

Adjust digits until answer is correct then click DONE

Press the PANIC button when you hear the HIGH tone

Figure 1. The participant view of SynWin includes memory (upper left), arithmetic (upper right), visual monitoring (lower left), and auditory monitoring (lower right) tasks.

Participants learn to use SynWin first in a tutorial setting where they practice only one quadrant task at a time for one minute each. Once each task has been tried alone, all four are presented at once in a five minute timed trial. Points are earned for correct responses and lost for incorrect or missed ones. Scores are generated for each task, along with a global score, out of 100 points. Higher individual task and global scores indicate higher performance of simultaneous SynWin tasks.

SynWin has been shown to be valuable tool in measuring arousal-related variables and complex task performance [14]. SynWin has also been evaluated in studies investigating cognitive computer games [16], and concurrent task performance [17].

4.1.3 Big Five Inventory

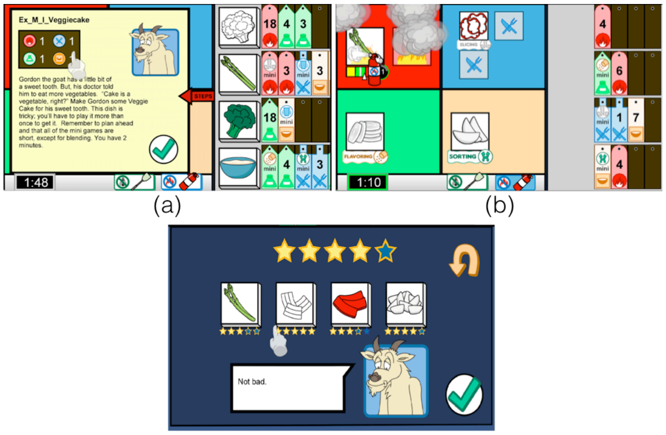
The Big Five Inventory – 44 item (BFI-44) is a self-report personality assessment scored across five personality dimensions: extraversion vs. introversion, agreeableness vs. antagonism, conscientiousness vs. carelessness, neuroticism vs. emotional stability, and openness vs. cautiousness [18]. The inventory contains 44 items and is scored on a 5-point Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree). Questions from the BFI-44 begin with the prompt "I see myself as someone who…" followed by a statement the participant rates with a Likert scale number (e.g., "I see myself as someone who is curious about many different things").

The BFI-44 has been proven reliable in multiple languages [19, 20], and to possess relationships with other variables such as academic success [21].

4.1.4 Food For Thought

Food for Thought was initially designed as a cognitive training game by researchers at Georgia Tech and North Carolina State University as a part of a four-year National Science Foundation funded cognitive gaming study.

Players engage in three main playing stages within the game: planning, playing, and reviewing (see Figure 2). A *planning* screen appears before the game play begins. This screen contains information on the recipe to be prepared, as well as the amount of time allotted to the recipe, and the first few steps of each ingredient. Players can spend time on this screen planning their strategy for when the level play begins.



(a) (b)

(c)

Figure 2. The three game stages of *Food for Thought*: planning (a), playing (b), and feedback review (c).

Once past this screen, players enter the second stage of *level game play*. The level begins and a level timer initiates counting down. Those that play quickly and efficiently should finish within the time allotted. Those that exceed the given level play time can continue to finish their recipe, but their performance score will suffer. The third game stage is end-level *feedback review* of the score. After the level concludes, players receive a star rating that reflects their performance on the dish they cooked; with five stars being the highest possible score. This screen allows players to get both high and low level feedback on their performance. Players can get further information on where things went wrong or right by examining their score by ingredient and station. A player that improperly cooks, mixes, chops, and stirs ingredients will incur a negative impact on their level score for that meal. Additionally, players must manage mini-games that occur in tandem with the main kitchen (see Figure 3).

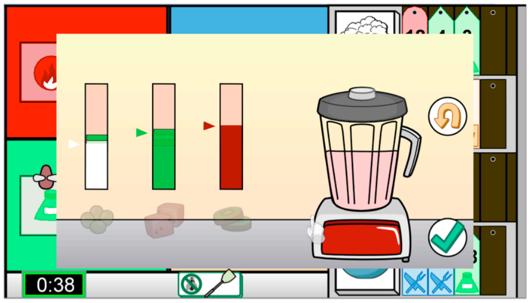


Figure 3. *Food for Thought* mini-games occur in parallel with the main kitchen game play, challenging a player to variably prioritize and multitask.

Players switch between the kitchen and mini-game screens as they attempt to properly prepare their kitchen ingredients while also tending to the visual search, spatial reasoning, and time estimation demands of the mini-games. Food items left on the counter too long will begin to lose "freshness", also impacting the score. Players that work quickly and efficiently, while managing multiple tasks at once will achieve maximum points.

Finally, players must select a fire extinguisher to put out station fires and wield a fly swatter against invading houseflies. A "Try Again" button icon on the feedback screen allows players to retry the level, whether that is for a higher score or simply to repeat the experience.

4.2 PROCEDURE

This study consisted of three days of participation. In the first session, participants spent approximately 60 minutes completing traditional questionnaires and instruments to evaluate participant psychosocial factors (ACT Engage, 30 minutes), multitasking performance (SynWin, 15 minutes), and personality (BFI-44, 15 minutes).

On sessions two and three, participants played *Food for Thought* for approximately one hour each day, with level difficulty increasing linearly. Players in *Food for Thought* encounter three main playing stages for each level: 1) a planning screen where a player sees the ingredients and required steps before the level begins, 2) the level itself where a player completes the recipe, and 3) the end-level feedback screen where a player sees the final score out of five stars and can examine the score justification in fine detail. From these playing stages, we analyzed the game play variables of overall level performance, total number of session levels played, number of player initiated level retries, average time spent on the planning screen, average time spent playing each level, and average time spent reviewing end-level feedback. A Pearson correlation analysis (*r*) was used to analyze relationships between in-game play behaviors and traditionally measured participant discrete variables.

5. OBSERVATIONS AND RESULTS

We found 10 statistically significant correlations between game play behaviors and scores from Engage, SynWin, and the BFI-44 (see Table 1).

Table 1.
ENGAGE, SynWin, and BFI-44 with game play variable correlations

	Planning	Playing	Reviewing	Retries
Steadiness	-0.10	**0.396***	**-0.376***	**-0.390***
Self-confidence	0.08	0.081	**-0.571****	-0.016
Communication Skills	-0.31	0.246	0.086	**-0.389***
General Determination	-0.36	0.224	-0.023	**-0.395***
Study Skills	-0.24	0.151	0.022	**-0.431***
Social Activity	-0.10	**0.527****	0.118	-0.268
SynWin Math	0.13	-0.344	-0.066	**0.453***
BFI Agreeableness	-0.26	0.196	0.024	**-0.490****

* = p < 0.05, ** = p < 0.01

5.1 ENGAGE

Eight statistically significant correlations were found between Engage psychosocial factors and *Food for Thought* game play behavior.

There was a negative relationship found between average time spent reviewing end-level feedback and two Engage measures of Steadiness (r = -0.376, n = 30, p = 0.05) and Self-Confidence (r = -0.571, n = 30, p = 0.01). Four negative correlations were found to exist between the number of retried levels a player initiated and the Engage measures of Steadiness (r = -0.390, n = 30, p = 0.05), Communication Skills (r = -0.389, n = 30, p = 0.05), General Determination (r = -0.395, n = 30, p = 0.05), and Study Skills (r = -0.431, n = 30, p = 0.05). A positive correlation was found between the average time spent playing game levels and measures of Steadiness (r = 0.396, n = 30, p = 0.05) and Social Activity (r = 0.527, n = 30, p = 0.01).

5.2 SYNWIN

One positive correlation was found between number of retried game levels and the individual SynWin arithmetic performance score (r = 0.453, n = 30, p = 0.05). No other correlations were found between game play behavior and overall, memory, or perceptual monitoring SynWin performance scores.

5.3 BFI-44

One negative correlation was found between number of retried game levels and the Agreeableness domain of the BFI-44 (r = -0.490, n = 30, p = 0.01)

5.4 GAME PLAY

In the two hours of game play sessions, participants completed an average of 76.8 levels with a range of 56 (minimum) to 96 (maximum) levels played. Players performed competently, averaging 4.71 out of 5 stars per level (min = 3.2, max = 5.0, SD = 0.02). The average time spent on the planning screen was 15.7 seconds (min = 6.84, max = 55.31, SD = 8.61), the average time spent in game level play was 65.86 seconds (min = 52.85, max = 77.07, SD = 6.56), and the average time spent reviewing feedback was 6.5 seconds (min = 2.01, max = 23.41, SD = 4.13). The majority of participants (66.7%) retried a level once (36.7%) or none (30%) of the time. The rest (33.3%) initiated between 5 and 9 level retries in the game.

6. DISCUSSION

The results presented here show some promise in using a video game, otherwise not intended as an assessment tool, to analyze relationships between game play behavior and results from traditional measures of individual differences.

One interesting result concerns two Engage measurements from the Self-Regulation domain with game play behavior. Higher Steadiness scores are correlated with more time spent in the game level play, less time spent in the score review feedback screen, and fewer number of retries. There was no relationship found between Steadiness scores and game performance or total number of levels played, however, suggesting an association between Steadiness high scorers and inefficient (longer) game play that does not necessarily result in poor performance or inspire motivation to retry for a higher score.

Self-Confidence was found to have a negative relationship with time spent in the feedback review screen. In other words, low self-confidence is correlated with more time spent in score review. Perhaps possessing low confidence in one's ability to perform or a proclivity to becoming easily overwhelmed or frustrated is related with the desire or need to fully inspect the details of one's performance.

We found level retries to be a game action correlated with each of the traditional assessments we used: Engage, SynWin, and the BFI-44. Three of these PSFs (Communication Skills, General Determination, and Study Skills) belong to the Motivation domain and one, the aforementioned Steadiness, belonging to the Self-Regulation domain. There appears to be a correlation with higher retries and Engage items associated with difficulty working in teams, lower levels of commitment, struggling with academic performance, and easily becoming overwhelmed and frustrated.

Finally, more frequent level retry initiations were correlated with lower agreeableness scores from the BFI-44. Facets of agreeableness include compliance, trust, and feeling sympathetic towards others. The opposite dimension to agreeableness is antagonism and is associated with stubbornness, vindictiveness, and indifference towards others. Perhaps agreeable individuals are more prone to giving the level another go and not taking their score personally.

Interestingly, participants that initiated the highest numbers of retries across both sessions (8 or 9 level retries) many times had already achieved 5 stars for the played level. Often, because their first performance had been high, their second attempt would result in a slightly lower score than the first. Other times, high retry initiators would dramatically improve their score (for example, from 3 to 5 stars) suggesting they were able to correct a mistake in understanding and/or execution in the game play.

7. CONCLUSION

The use of game-based assessment is an active and growing field of research. Assessment games present a unique opportunity to measure human individual differences in engaging and creative ways. The first step towards understanding the design and examination of assessment games is in identifying natural player behaviors that may be associated with individual differences. After all, these are the qualities that underlie that which produces observable behavior.

In this study, we found several statistically significant correlations between game play behavior and results from traditional psychosocial, multitasking, and personality assessments. These findings have informed our next steps in deploying a variety of other custom assessment game versions to a much larger sample size. Future development and research into game mechanics that are highly correlated with traditional measures of individual differences is necessary in developing accurate and compelling assessment games.

8. ACKNOWLEDGMENTS

This work is funded through grants from ACT, Inc and NSF Grant #0905127.

9. REFERENCES

[1] Grenhart, W., Sprufera, J. F., McLaughlin, A. C., and Allaire, J. C. 2013. Pilot: Customizing a commercially available digital game to assess cognitive function. Presentation at the *North Carolina Cognition Conference*. Raleigh, NC. Retrieved 12 February 2015 from http://www.gainsthroughgaming.org/wp-content/uploads/2012/07/NCCC2013.ppt

[2] Mislevy, R. J., Steinberg, L. S., and Almond, R. G. 2003. *Focus article:* On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, 1, 1, 3-62.

[3] Shute, V. J., Masduki, I., and Donmez, O. 2010. Conceptual framework for modeling, assessing, and supporting competencies within game environments*. Technology, Instruction, Cognition, and Learning*, 8, 2, 137-161.

[4] Shute, V. J. 2011. Stealth assessment in computer-based games to support learning. in Tobias, S. and Fletcher, J. D. eds. *Computer games and instruction*, Information Age Publishing, Charlotte, NC, 503–524.

[5] Shute, V.J., Ventura, M., Bauer, M. I., and Zapata-Rivera, D. 2009. Melding the power of serious games and embedded assessment to monitor and foster learning: flow and grow. In Ritterfeld, U., Cody, M., and Vorderer, P. eds. *Serious games: Mechanisms and effects*, Routledge, Taylor, and Francis, Mahwah, NJ, 295-321.

[6] Bethesda Softworks. 2006. *Elder Scrolls IV: Oblivion.* Retrieved 12 February 2015 from http://www.bethsoft.com/games/games_oblivion.html.

[7] Tekofsky, S., Spronck, P., Plaat, A., van den Herik, J., and Broersen, J. 2013. PsyOps: Personality Assessment Through Gaming Behavior. In *Proceedings of the 8th International Conference on the Foundations of Digital Games* (Crete, Greece, May 14 – 17, 2013). Retrieved 12 February 2015 from http://www.fdg2013.org/program/papers/paper22_tekofsky_etal.pdf.

[8] van Lankveld, G., Schreurs, S., and Spronck, P. Psychologically verified player modelling. In *Proceedings of the 10th International Conference on Intelligent Games and Simulation*, Breitlauch, L. ed., 12-19.

[9] van Lankveld, G., Spronck, P., van den Herik, J., and Arntz, A. 2011. Games as personality profiling tools. In *2011 IEEE Conference on Computational Intelligence and Games*, Preuss, M. ed., 197–202.

[10] Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., and Carlstrom, A. 2004. Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychol. Bull.,* 130, 2, 261–288.

[11] Robbins, S. B., Allen, J., Casillas, A., Peterson, C., and Le, H. 2006. Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *J. Educ. Psychol.,* 98, 3, 598–616.

[12] ACT. 2011. ENGAGE College user's guide. Retrieved 12 February 2015 from http://media.act.org/documents/engage_college_users_guide.pdf.

[13] Elsmore, T. F. 1994. SYNWORK1: A PC-based tool for assessment of performance in a simulated work environment. *Behav. Res. Meth. Ins. C.,* 26, 4, 421-426.

[14] Proctor, R. W., Wang, D. Y., and Pick, D. F. 1998. An empirical evaluation of the SYNWORK1 multiple-task work environment*. Behav. Res. Meth. Ins. C.*, 30, 2, 287-305.

[15] Wong, J. 2005. SynWin Version 1.2 Activity Research Services*. Ergonomics in Design: The Quarterly of Human Factors Applications*, 13, 4, 30-32.

[16] Kearney, P. R. 2005. *Cognitive calisthenics: Do FPS computer games enhance the player's cognitive abilities*? In the *Proceedings of DiGRA 2005 Conference* (Vancouver, Canada, June 15-20, 2005).

[17] Salthouse, T. A., Hambrick, D. Z., Lukas, K. E., and Dell, T. C. 1996. Determinants of adult age differences on synthetic work performance. *J. Exp. Psychol.-Appl.,* 2, 4, 305-329.

[18] Rammstedt, B., and John, O. P. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German*. J. Res. Pers*. 41, 1, 203-212.

[19] Benet-Martínez, V., and John, O. P. 1998. Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English*. J. Pers. Soc. Psychol*, 75, 3, 729-750.

[20] Fossati, A., Borroni, S., Marchione, D., and Maffei, C. 2011. The Big Five Inventory (BFI): Reliability and validity of its Italian translation in three independent nonclinical samples. *Eur. J. Psychol. Asess.*, 27, 1, 50-58.

[21] O'Connor, M. C., and Paunonen, S. V. 2007. Big Five personality predictors of post-secondary academic performance*. Pers. Indiv.*, 43, 5, 971-990.