Validating Stress Tolerance Exercise With Game-Based Assessment and Stress Checklists

Johann Wium

VALIDATING STRESS TOLERANCE EXERCISE WITH GAME-BASED ASSESSMENT AND STRESS CHECKLISTS

Jóhann Wium

Isavia Air Navigation Services

Iceland

Stress tolerance is an important attribute for air traffic controllers. Assessing an individual's stress tolerance should therefore be considered when selecting student air traffic controllers. Unfortunately, measures of stress tolerance are often self-reports and as such are subject to "faking good". This paper details the validation of an observation-based stress tolerance exercise and its convergent validity with behavioral signs of stress (from stress checklists) and results of game-based assessments of Emotional Stability and Performance Under Pressure. The resulting data suggests that ratings from the exercise are valid and that game-based assessments can be used to predict ratings made from observable behavior in candidates.

A high attrition rate is a persistent problem in air traffic controller (ATCO) training (Broach, 2017). Previous job analyses of the role of ATCOs emphasize being able to perform in high stress environments (Goeters et al, 2004; Nickles et at 1995; Suresh et al, 2012) so stress tolerance is an issue in ATCO selection. Stress tolerance is defined as being capable of reacting with a problem-solving approach rather than an emotional approach when faced with adversity and remaining calm, even-tempered, and composed in stressful situations (Nickles et at, 1995).

Some studies indicated that stress tolerance can affect student ATCO success (see e.g. Chapelle et al, 2015; Roe et al, 2012; Collins et al, 1989) while others found no significant connection (Geven et al, 2008; Luuk et al, 2009; Oakes et, 2001; Schroeder et al 1993). One possible explanation is that these studies used self-report questionnaires when measuring stress tolerance and they can be susceptible to "faking-good". Griffith and Converse (2012) estimate that 30% (+/-10%) of job applicants will give overly positive answers about their capabilities and that "faking good" is more likely in 'high stakes' selection (Ellingsen, 2012; Griffith et al, 2007).

To counter this bias, an assessment center exercise was developed to assess stress tolerance under pressure. The exercise was validated by comparing it to behavioral observations by two raters and scores from a game-based assessment (GBA). The behavioral observations were in the form of a checklist score of observed behaviors connected to stress (i.e. stress checklist). The game-based assessment is a novel method to assess various applicant characteristics, such as performance under pressure and emotional stability. Applicants had completed the GBA one month earlier as part of a multiple hurdle student ATCO selection process. While extensive, that process did not specifically select for stress tolerance or similar traits until the assessment center.

Inter-rater reliability between the two assessors was also calculated to ensure consistency of ratings.

Game-based assessments

Convergent validity for GBA's had been demonstrated with cognitive tests before in Icelandic student ATCO selection (Boardman, 2017), but this is the first study to use behavioral assessments. A GBA uses methods of psychological testing embedded in a gamified interface, collecting game performance data, both overtly (e.g. candidate choices in the game) and covertly (e.g. reaction times), to assess cognitive and personality factors (Arctic Shores, 2017). GBA's are also resistant to "faking-good" (Armstrong et al, 2016) making them ideal for this comparison.

GBA's combine research from I/O, neuropsychology (Ferreira-Brito et al, 2019), behavioral economics, and education (Reiners & Woods, 2015) to assess applicants. Reaction time, for example, can assess Neuroticism (Robinson & Tamar, 2005) or self-confidence (Wichmann et al, 2016), time on task can assess persistence (Ventura & Shute, 2013), time logs can assess collaboration (Mislevy at al, 2015), and so on.

The GBA used was Skyrise City from Arctic Shores. It measures several psychological constructs including Emotional Stability (ES), consistency of affect, i.e., stressful situations are dealt with in a calm and even-tempered manner, and Performance under Pressure (PP), maintaining goal-oriented behavior while subjected to negative stressors (Arctic Shores, 2017).

Stress tolerance assessment exercise

An applicant entered a room with two assessors and sat at a table with a countdown clock, 12 puzzle pieces, and a rectangular frame. A standardized briefing was given stating that the exercise would be 10 minutes long (counted down on the clock) and that the goal was to fit all puzzle pieces inside the rectangular frame. The briefing stated that warnings would be given when there were 5 minutes and 1 minute left. The applicant was not explicitly told that stress tolerance was being assessed, and, if asked, the standard answer was that assessment was based on task performance.

The puzzle was a reproduction of an assembly puzzle called Calibron-12 created by Theodore Edison in 1933. While it is theoretically possible to finish within the allotted time, it is exceptionally unlikely (estimated minimum time is about 4 hours). The puzzle, however, looks like it should be easily solved within the allotted time (Creative Crafthouse, 2020). Because an applicant is unlikely to solve the puzzle, the performance measure was stress tolerance, not completion time.

The stress in this exercise is created by the fact that the exercise is part of a high-stakes selection process and occurs under time pressure while the candidate is being observed by two assessors. Additionally, the candidate is faced with a frustrating task, which may engender a sense of impending failure. The capacity to continue to perform, while remaining unaffected by these stressors, is an indication of stress tolerance.

Once 10 minutes had passed and the applicant had left, each rater gave an independent rating (IR) of estimated stress tolerance. The score ranged from 1 (Very poor stress tolerance) to 6 (Very good stress tolerance). Scores from the raters was then averaged to give a final stress tolerance (ST) score for the exercise.

Second, each rater completed a checklist for observed signs of stress and tallied the number of signs to give a checklist score. The score ranged from 0-9 and the signs included were: shaking, fidgeting, stiffness, defensive behavior, avoiding behavior, inadvertent sounds, flushed skin, forceful movements and slow movements. Assessors received detailed descriptions of stress signs as part of their training.

Methods and results

Participants

Thirty participants were assessed in an assessment center for Icelandic student ATCO applicants. Internal data protection policy dictates that identifiable personal information is removed from research datasets so information on age and gender is not available.

Assessor measures

The independent scores (IR) for stress tolerance as given by the assessors, demonstrated an inter-rater reliability of $r(30) = 0.853$, $p>0.001$. For the checklist of observed signs of stress, assessors had a 68% agreement rate.

The total number of stress signs observed by each assessor was counted from the checklists. The correlation between the total number of stress signs and independent rating of stress tolerance (IR) was $r(60) = -0.582$, $p>0.001$ (with a lower IR meant that more signs were observed and vice versa). This supports the hypothesis that IR and the checklist score (observable signs of stress) demonstrate convergent validity.

Game-based assessment

The correlation between final stress tolerance score (ST) and Emotional Stability (ES) was $r(30) = 0.444$, $p>0.05$. The correlation between final stress tolerance score (ST) and Performance Under Pressure (PP) was $r(30) = 0.380$, $p>0.05$. The GBA measures also reached significance when compared to the independent scores of the raters (IR). Between IR and ES the correlation was $r(60) = 0.348$, $p>0.01$, and between IR and PP it was $r(60) = 0.295$, $p>0.05$. Correlation between checklist scores and ES or PP did not reach significance.

Discussion

The results suggest that the exercise is valid. It has reliability as evidenced by high inter-rater reliability of the independent ratings. The independent ratings and checklist scores demonstrate convergent validity as there is a significant negative correlation between

observable stress signs and stress tolerance. The assessment also demonstrates convergent validity in that the GBA measures of Emotional Stability (ES) and Performance Under Pressure (PP) both correlate significantly with stress tolerance (ST).

While this study suffers from a potential lack of generalizability due to its small sample size it provides two useful insights for practitioners. First, the exercise described shows an observational method to rate stress tolerance free from applicants "faking good". Second, as game-based assessments are a new approach in selection it is important to note that they can show significant correlations to more traditional ratings made by behavioral observations.

References

Arctic Shores (2017). Skyrise City: Technical manual. Game version 1.26, Scoring version 1.5.

Armstrong, M., Landers, R., & Collmus, A. (2016). Gamifying Recruitment, Selection, Training, and Performance Management: Game-Thinking in Human Resource Management. In Gangadharbatla, H., and Davis, D.Z. (Eds.), *Emerging Research and Trends in Gamification* (pp. 140-165). doi: 10.4018/978-1-4666-8651-9.ch007.

Boardman, E.C. (2017). Convergent and Predictive Validity of a Game-based Assessment in Air Traffic Control Recruitment. United Kingdoms: Arctic Shores

Broach, D. (2017). Attrition in U.S. Air Traffic Control Specialist (ATCS) Training: A Review of 50 Years of Data. 19th International Symposium on Aviation Psychology, 542-547.

Chappelle, W., Thompson, W., Goodman, T., Bryan, C., & Reardon, L. (2015). The Utility of Testing Noncognitive Aptitudes as Additional Predictors of Graduation From US Air Force Air Traffic Controller Training. *Aviation Psychology and Applied Human Factors. 5*. 93-103. doi: 10.1027/2192-0923/a000082

Collins, W. C., Schroeder, D. J., & Nye, L. G. (1989). Relationship of anxiety scores to academy and field training performance of air traffic control specialists. DOT-FAA-AM-89-7. Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Creative Crafthouse (2020, October 10). Calibron 12 – A Very Difficult Wood Puzzle Brain Teaser. Retrieved from https://www.creativecrafthouse.com/calibron-12-very-difficult-design-from-1933-now-with-cover.html

Ellingson, J. E. (2012). People fake only when they need to fake. In Ziegler, M., MacCann, C., and Roberts, R. (Eds.), *New Perspectives on Faking in Personality Assessment*, (pp. 19-33). New York: Oxford University Press.

Ferreira-Brito, F., Fialho, M., Virgolino, A., Neves, I., Miranda, A., Sousa-Santos, N., Caneiras, C., Carriço, L., Verdelho, A., & Santos, O. (2019). Game-based interventions for neuropsychological assessment, training and rehabilitation: Which game-elements to use? A systematic review. *Journal of Biomedical Informatics. 98*. doi: 103287. 10.1016/j.jbi.2019. 103287.

Geven, E.., Oprins, E., Veldhuijzen, E., and Roe, R.A. (2008). Using personality scales in ATC selection: Predictive validities for ATC task performance in Initial Training. Proceedings of the 28th European Association for Aviation Psychology Conference.

Goeters, K.M., Maschke, P., and Eißfeldt, H. (2004). Ability Requirements in Core Aviation Professions: Job Analyses of Airline Pilots and Air Traffic Controllers. In K. M. Goeters (Ed.), *Aviation psychology: Practice and Research* (pp. 135-140). Aldershot: Ashgate Publishing Limited.

Griffith, R. L., Chmielowski, T., and Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behaviour. *Personnel Review, 36*(3), 341-355. doi: 10.1108/00483480710731310

Griffith, R. L., and Converse, P. D. (2012). The rules of evidence and the prevalence of applicant faking. In Ziegler, M., MacCann, C., and Roberts, R. (Eds.), *New Perspectives on Faking in Personality Assessment* (pp. 34-52). New York: Oxford University Press.

King, R.E., Retzlaff, P.D., Detwiler, C.A., Schroeder, D.J., Broach, D. (2003). Use of personality assessment measures in the selection of air traffic control specialists. DOT/FAA/AM-03/20. Washington, DC: Federal Aviation Administration Office of Aviation Medicine

Luuk, K., Luuk, A., & Aluoja, A. (2009). Predicting professional success of air traffic control personnel from their personality profile at admission to Ab Initio training. *The International Journal of Aviation Psychology, 19(3),* 235–251. doi:10.1080/10508410902983896

Mislevy, R.J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M.I., von Davier, A., & John, M. (2015). Psychometrics and Game-Based Assessment, in F. Drasgow (Ed.), *Technology and Testing: Improved Educational and Psychological Measurement*. United Kingdoms: Routledge.

Nickles, B.J., Bobko, M.D., Sands, W.A., & Tartak, E.L. (1995). Separation and Control Hiring Assessment (SACHA) final job analysis report, Bethesda, MD: University Research Corporation

Reiners, T., & Wood, L.C. (2015). Gamification in Education and Business. Springer International Publishing. doi: 10.1007/978-3-319-10208-5

Robinson, M., & Tamir, M. (2005). Neuroticism as mental noise: A relation between neuroticism and reaction time standard deviations. *Journal of Personality and Social Psychology, 89(1),* 107– 114. doi: 10.1037/0022-3514.89.1.107

Roe, R.A., Oprins, E., and Geven, E. (2012). The predictive validity of personality tests in air traffic controller selection. Proceedings of the 30th European Association for Aviation Psychology Conference.

Schroeder, D. J., Broach, D., & Young, W. C. (1993). Contribution of personality to the prediction of success in initial air traffic control specialist training. DOT/FAA/AM93/4. Washington, DC: FAA Office of Aviation Medicine.

Suresh, A., Ramachandran, K., & Srivastava, A. (2012). Personality Based Job Analysis of Air Traffic Controller. *Indian Journal of Aerospace Medicine. 56.* 21-31.

Ventura, M. & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior. 29.* 2568-2572. doi: 10.1016/j.chb.2013.06.033.

Wichmann, T., Buchheim, A., Menning, H., Schenk, I., George, C., & Pokorny, D. (2016). A reaction time experiment on adult attachment: The development of a measure of neuro-psychological settings. *Frontiers in Human Neuroscience.* doi: 10.3389/ fnhum.2016. 00548