5-2010

Gathering behavioral samples through a computerized and standardized assessment center exercise yes, it is possible

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

Etienne VAN KEER
Hudson, Belgium

Ellen VOLCKAERT
Hudson, Belgium

Published in Journal of Personnel Psychology, 2010 January,
Volume 9, Issue 2, Pages 94-98
http://doi.org/10.1027/1866-5888/a000010

Gathering Behavioral Samples Through a Computerized and Standardized Assessment Center Exercise

Yes, It Is Possible

Filip Lievens,[1] Etienne Van Keer,[2] and Ellen Volckaert[2]

[1]Ghent University, Belgium, [2]Hudson, Belgium

Abstract. Although computerization and standardization might make assessment center (AC) exercises easier to administer and score, drawbacks are that most of such exercises have a static and multiple-choice format. This study reports on the development and initial validation of a computerized and standardized AC exercise that simulates key managerial tasks. This AC exercise capitalizes not only on the benefits of computerization and standardization (efficiency and cost savings) but at the same time aims to avoid their usual drawbacks (lower response fidelity and interactivity). The composite exercise score was significantly related to several criteria of interest and had incremental validity beyond cognitive ability. The exercise was also significantly related to candidates' people management competencies.

Keywords: assessment center, technology, selection

In the last decade, the increased use of technology and the growing computerization and standardization of assessment center (AC) exercises have provoked debate (Lievens & Thornton, 2005; Thornton & Rupp, 2006). On the one hand, computerization and standardization have advantages in terms of providing a consistent mode of delivery of exercise stimuli. They also streamline the time-consuming observation and rating process as there is no longer a need to use trained assessors, leading in turn to cost reductions. On the other hand, computerization and standardization have been criticized because they detract from one of the hallmarks of ACs, namely the observation of behavior (Thornton & Rupp, 2006). That is, computerized and standardized exercises typically have a multiple-choice format wherein participants choose one predetermined alternative of what they intend to do/say. Hence, they are primarily measures of knowledge (see also McDaniel, Hartman, Whetzel, & Grubb, 2007). All of this might decrease the response fidelity because in the workplace "life is not multiple choice" (Ryan & Greguras, 1998). Thornton and Rupp (2006) also argued that computerized and standardized exercises might be less interactive because candidates are no longer interacting with other candidates and/or role players. These drawbacks might turn AC exercises into low-fidelity simulations.

These advantages and disadvantages of computerizing and standardizing AC exercises are best illustrated with one of the most popular AC exercises: the in-basket. In paper-and-pencil in-baskets, candidates generate their reply to the materials and assessors score these responses. Conversely, in computerized in-baskets participants read emails (e.g., Lievens & Anseel, 2007) and pick a multiple-choice option. The scoring is done via an expert system built into the computer.

Our premise is that these commonly cited disadvantages of computerization and standardization are related to the particular approach of computerizing and standardizing AC exercises. Therefore, this study reports on the development and initial validation of an AC exercise that capitalizes not only on the benefits of computerization and standardization (consistent delivery, standardized scoring, and cost savings) but also aims to avoid their typical drawbacks (lower response fidelity and interactivity).

Method

Sample and Procedure

A first sample consisted of 106 students (72% males and 28% females; mean age = 23.6 years) who were pursuing a postgraduate master in General Management at a large

Belgian Business School. They had no work experience. A second sample consisted of 39 finalists (50% females and 50% males; mean age = 32.1 years) of a national contest "Employee of the Year." These candidates had several years of work experience and came from different organizations.

In both samples, the procedure was nearly identical. Candidates independently completed the computerized exercise (and other tests) in large PC rooms. Participants of the second sample also went through two role-play AC exercises. In Sample 1, candidates knew that their scores on the exercise would be used only for research. One week after the exercise they received brief feedback about their strengths and weaknesses.

Development of Exercise

The exercise (Electronic Assessment Simulation Exercise, EASE) to simulate three managerial tasks: information gathering/analyzing, decision-making, and strategic planning. The rationale behind its development was to construct a computerized and standardized exercise that was less close-ended and more interactive. Hence, participants could choose among a myriad of actions themselves (instead of choosing one of a limited number of multiple choices) and the PC provided a response depending on their action.

A company granted us permission to serve as model for the development of the exercise. First, we gathered a variety of contextual information (e.g., context, history, structure, financial data, products, personnel records, culture, and policies). Second, we held interviews to obtain a description of the strategic challenges the company was facing at the time. Third, we conducted interviews to inventory the approaches considered to face these challenges.

Description of Exercise

The exercise consisted of three main tasks. In the first task (information gathering/analyzing), participants had 88 min to collect information. In the exercise, this corresponded to 5.5 fictitious days (i.e., 1 day in the exercise lasted 16 min in reality). At the start, they had only two emails in their inbox. They could gather additional information via three approaches: (a) send emails, (b) retrieve archival documents, and (c) schedule meetings. In addition, participants had the freedom to decide how many questions to ask, which questions to ask, and to whom they wanted to address them. To this end, they used dropdown lists. There were 3,906 possible questions. In this first task, the computer responded interactively and realistically to the participants' actions. In case of archival requests, information was immediately sent to participants' inbox. With regard to email questions, participants received a reply within 24 h (16 min in reality). Finally, participants received meeting notes within 4 h after the end of the meeting.

In the second task (decision-making, 10 min), participants had to make decisions on the basis of Task 1. There were 8,430 possible decisions that could be taken: per decision participants had to decide on four aspects: domain (what is the decision about?), keyword (which topic is treated by the decision?), verb (what has to be done with it?), and person (who has to do it?).

In the third task (strategic planning, 30 min), participants had "to put it all together" and develop a strategic production plan on the basis of the previous tasks. Using an excel-like sheet, participants planned the start, stop, transfer, or integration of production lines at various time periods.

Scoring

In traditional AC exercises, assessors observe behavior and provide evaluations on the basis of their notes. Similar to this practice, the computerized AC exercise gathered descriptive and evaluative data. The *descriptive* information issued by the PC was represented by one score from Task 1. Specifically, the PC recorded whether participants relied on interpersonal (face-to-face) sources of information gathering (i.e., meeting requests vs. archive/email requests). To this end, a score from 0 (no reliance on interpersonal sources) to 10 (sole reliance to interpersonal sources) was issued by the PC. This score was labeled interpersonal inclination and indicated the proportion of meeting requests in all information requests (e.g., when the proportion of meeting requests was .36, this score was 3.6).

The *evaluation of the effectiveness* of participants' actions was determined on the basis of 30 experts (managing directors and senior managers; 23 males, 7 females, and mean age = 38.2 years). To score all possible questions in Task 1 beforehand, these questions were distributed across the experts. So, each expert scored over a 100 questions. Experts decided how many points a question deserved. For instance, asking a question regarding budgets to the financial manager received more points than asking this to the HR manager. Questions were given zero points when they were addressed to persons who could not give information (e.g., asking Indonesian managers about the transfer to Eastern Europe). Of the 3,906 possible questions, 1,397 were such nonsensical questions. Regarding Task 2, a similar approach was followed to assign beforehand points to decisions that could be taken. In Task 3, the scoring was done by checking whether participants' planning obeyed specific rules. For instance, negative points were given when the number of nonoperational lines was not kept to a minimum. Per task, all points were summed. In Tasks 1 and 2, this sum was divided by the number of items (either questions asked or decisions taken). This led to three effectiveness scores (one for each task). The distributions of these three scores approximated normal distributions. Given that the metric differed across the three tasks, z scores were used to express effectiveness scores. Next, a composite exercise score (the sum of the z scores on each task) was computed.

Other Predictor and Criterion Measures

Apart from the computerized exercise, participants of both samples also completed a *cognitive ability test* (Abstract Reasoning Ability Test (A-RAT); Bogaert, Trbovic, & Van Keer, 2005) that measured the ability to detect principles in abstract information and to apply them to new information. The test contained 40 items with five response alternatives (time limit = 20 min). The test manual mentioned a correlation of .52 with Raven's Progressive Matrices.

In Sample 1, archival criterion data were retrieved from students who completed the master in general management. A composite score (grade point average, GPA) was obtained by averaging scores on 18 courses. Most courses focused on economic and business-related topics.

In Sample 2, an *AC score* served as criterion measure. Given that AC exercises represent realistic work situations, they are often used as criterion measures (e.g., Atkins & Wood, 2002). Eight experienced and trained assessors (five females and three males) used 9-point behaviorally anchored rating scales to evaluate candidates on three dimensions in two role-plays. These dimensions were guiding people (ensuring that everyone behaves appropriately by giving clear instructions, monitoring results, and taking measures in accordance with the objectives/available means), cooperating (creating and encouraging team spirit by sharing one's own opinions/ideas and identifying oneself with the common goals), and interacting (getting to know persons via targeted questions, by paying attention to what drives them and by trying to fully understand the underlying message of their communication). The across-exercise dimension ratings were used. As these three dimensions correlated around .70, they were combined in a composite people management rating ($\alpha = .85$).

In Sample 2, we also used the *decision score* for rank ordering the candidates of the Employee of the Year competition. This decision score was a weighted sum of various tests (cognitive ability tests, a personality inventory, a creativity test, and a motivational drives questionnaire) completed by the candidates (*except* the computerized exercise and the role-plays). Assessors were not aware of candidates' computerized exercise scores when providing the final decision score.

Results

Table 1 presents the descriptive statistics and correlations in Sample 1. The composite exercise score was significantly correlated with cognitive ability. This was due to the information gathering/analyzing score. Both the composite exercise score and the strategic planning score were significant predictors of GPA. Next, we examined whether the computerized exercise explained incremental validity above cognitive ability. In a hierarchical regression, the cognitive ability test was entered first and the various scores next. As shown in Table 2, the descriptive/effectiveness scores added incremental validity beyond cognitive ability ($\Delta R^2 = .11$).

Table 2. Results of hierarchical regression analysis for predicting GPA ($N = 106$)

	β	R	R^2	Difference R^2
Step 1				
Cognitive ability	.38*	.30	.09*	
Step 2				
Interpersonal inclination	−.01			
Information gathering/analyzing	−.14			
Decision-making	−.06			
Strategic planning	.32*	.45	.20*	.11*

*$p < .01$.

Table 1. Means, *SD*s, and correlations of variables of sample 1 ($N = 106$)

	M	*SD*	1	2	3	4	5	6
Descriptive score								
1. Interpersonal inclination[a]	3.7	2.1						
Effectiveness scores								
2. Information gathering/analyzing (Task 1)[b]	152.76	26.16	.27**					
3. Decision-making (Task 2)[b]	4.01	1.27	.16	.07				
4. Strategic planning (Task 3)[b]	−806.50	803.00	.08	.02	.23*			
5. Composite exercise score[b]	.08	1.84	.23*	.56**	.65**	.68**		
Other measures								
6. Cognitive ability	18.66	5.55	.11	.45**	.07	.02	.29**	
7. GPA[c]	14.12	.64	.01	.03	.02	.30**	.21*	.30**

*$p < .05$. **$p < .01$.

[a]Expressed as a score from 0 to 10, with 10 meaning that participants exclusively searched for information from interpersonal sources (meetings).

[b]Although these scores were expressed as *z* scores, the table provides the *M* and *SD*s of the unstandardized variables of Tasks 1–3.

[c]GPA varied between 0 and 20.

Table 3. Means, *SD*s, and correlations of variables of sample 2 ($N = 39$)

	M	*SD*	1	2	3	4	5	6	7
Descriptive score									
1. Interpersonal inclination[a]	3.5	1.8							
Effectiveness scores									
2. Information gathering/analyzing (Task 1)[b]	158.38	21.36	.30						
3. Decision-making (Task 2)[b]	4.01	1.22	.14	.28					
4. Strategic planning (Task 3)[b]	−570.17	822.65	.01	.42*	.17				
5. Composite exercise score[b]	.15	2.08	.17	.74**	.65**	.75**			
Other measures									
6. Cognitive ability	23.00	5.50	−.01	.28	.15	.22	.26		
7. People management (role-plays)	4.63	1.36	.39*	.11	.05	.20	.10	−.23	
8. Final decision	59.60	2.85	.29	.17	.14	.35*	.33*	.43**	−.19

*$p < .05$. **$p < .01$. For information on superscripts, see Table 1.

Table 3 presents the descriptive statistics and correlations in Sample 2. The composite exercise score was a significant predictor of the final decision score. In line with Sample 1, the strategic planning score was the only significant predictor among the effectiveness scores. Again, a hierarchical regression showed that the exercise scores had incremental validity beyond cognitive ability.

A different picture emerged when an overall people management rating served as criterion as the descriptive score then mattered. Candidates' interpersonal inclination (i.e., the use of more interpersonal sources to search for information) was positively correlated with the people management rating.

Discussion

This study reported on the development and initial validation of a computerized and standardized AC exercise that simulated key managerial tasks. This exercise is distinctive because it includes the efficiency benefits of computerization and standardization (administration and scoring ease), while at the same time avoiding their common pitfalls (lower response fidelity and lower interactivity). So, this study illuminates that there exist possibilities of computerizing and standardizing AC exercises other than the traditional formats employed.

Although it is only initial validity evidence, this study's results are promising. First, in both samples, the composite exercise score predicted several criteria of interest. So, gathering behavioral "samples" of candidates' information gathering/analyzing, decision-making, and especially strategic planning via this exercise enables to gather useful information. Second, apart from the information gathering/analyzing task, the scores do not exhibit substantial relationships with cognitive ability, providing the exercise with opportunity to go beyond cognitive ability. However, this should be interpreted with caution because of possible range restriction (samples of candidates high on cognitive ability). Third, the descriptive score of candidates' interpersonal inclination was related to their people management competencies as rated in AC exercises. This exemplifies that the computerized and standardized exercise is related not only to task-related performance but also to interpersonal performance (see also Meriac, Hoffman, Woehr, & Fleisher, 2008), which is remarkable given its individual character.

One limitation of our approach is that it might be especially relevant for individual exercises. This study also included GPA and role-play performance as criteria. Future research is needed to build on this work by using job performance. Another intriguing issue consists of comparing the validity of this new measure to extant multiple-choice measures. In addition, we need to examine the adverse impact of the exercise. Finally, we should scrutinize candidates' perceptions to examine how managerial candidates react to it.

References

Atkins, P. W. B., & Wood, R. E. (2002). Self-versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55*, 871–904.

Bogaert, J., Trbovic, N., & Van Keer, E. (2005). *Ability test suite – Level III – Manual*. Hudson: Ghent.

Lievens, F., & Anseel, F. (2007). Creating alternate in-basket forms through cloning: Some preliminary results. *International Journal of Selection and Assessment, 15*, 405–411.

Lievens, F., & Thornton, G. C. III. (2005). Assessment centers: Recent developments in practice and research. In A. Evers, O. Smit-Voskuijl, & N. Anderson (Eds.), *The Blackwell handbook of personnel selection* (pp. 243–264). Malden, MA: Blackwell Publishing.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. III. (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60*, 63–91.

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042–1052.

Ryan, A. M., & Greguras, G. J. (1998). Life is not multiple choice: Reactions to the alternatives. In M. Hakel (Ed.), *Alternatives to traditional testing* (pp. 183–202). Mahwah, NJ: Erlbaum.

Thornton, G. C. III., & Rupp, D. (2006). *Assessment centers in human resource management: Strategies for prediction diagnosis, and development*. Mahwah, NJ: Erlbaum.

Filip Lievens

Department of Personnel Management, Work and Organizational Psychology
Ghent University
Henri Dunantlaan 2
9000 Ghent
Belgium
Tel. +32 9 264 6453
Fax +32 9 264 6494
E-mail filip.lievens@ugent.be