Military Psychology
2013, Vol. 25, No. 4, 402–411

0899-5605/13/$12.00 DOI: 10.1037/mil0000006

The Predictive Role of Simulations in Assessing Military Performance

Eldad Rom
The College of Management, Rishon LeZion

Yael Kalderon
Israeli Navy, Haifa, Israel

The current study assessed the predictive validity of simulations to improve the military selection system. Four navy simulations were developed and their predictive validity was measured. The performance of 1007 Israeli navy soldiers was measured in a longitudinal study, which was carried on for almost two years. Participants' performance in four simulations (naval-navigation test, raft sailing, rubber boat [zodiac] mounting, and military tent assembly) was measured and used as a behavioral predictor for their performance at the end of their first year of active military service on combat ships. All but the raft sailing simulation predicted participants' performance.

Keywords: simulation, predictive validity, military performance

Making accurate predictions of individuals' future performance is a longstanding challenge that has captured the interest of personnel psychologists (Arthur & Villado, 2008). In pursuing the goal of effective personnel decision-making, researchers and practitioners have developed and employed a variety of assessment procedures, including cognitive ability tests (e.g., LePine, Colquitt, & Erez, 2000), personality measures (e.g., Hurtz & Donovan, 2000), and knowledge tests (e.g., Ones & Viswesvaran, 2007). Occasionally, personnel psychologists use simulations as predictor measures for achieving effective selection (e.g., Lievens & Patterson, 2011). During these simulations, applicants carry out a selected set of tasks similar to those performed on the actual job (Ployhart, Schneider, & Schmitt, 2010). The simulations are imitations of task operations and can vary from computerized flight simulators to vivid assessment centers. At their best, simulations maximize the correspondence between the content of the evaluation measure and that of the work domain, while leaning on the basic tenet of behavioral consistency—that past performance is the best predictor of future performance (Motowidlo, Dunnette, & Carter, 1990).

Simulation tests are embedded in the work sample testing domain, which is believed to be among the most valid predictors of job performance (Roth, Bobko & McFarland, 2005). Callinan and Robertson (2000) suggest that work sample testing is more than just a method or a procedure, but rather an approach toward assessing individuals where actual hands-on performance and a real work setting are presented to the applicant. Measures developed under this approach are designed to sample work behaviors and elicit signs of underlying predispositions (Motowidlo et al., 1990). Aside from simulations, which are probably the most straightforward and literal case of work sample tests, this approach also encompasses trainability tests, situational judgment tests, and job knowledge tests.

Work sample testing demonstrates high levels of predictive validity when compared with other selection methods (Lievens & Patterson, 2011). It enables us to evaluate candidates' readiness for complex environments and to sustain high levels of organizational effectiveness (Salas, Rosen, Held, & Weissmuller, 2009). In a review of the meta-analytic evidence available for the validity of selection methods, Schmidt and Hunter (1998) found work sample testing to have the highest reported validity for an indi-

Eldad Rom, School of Behavioral Sciences, The College of Management, Rishon Lezion, Israel; Yael Kalderon, Israeli Navy, Haifa, Israel.

Correspondence concerning this article should be addressed to Eldad Rom, Ph.D., Department of Behavioral Sciences, College of Management, Yitzhak Rabin Boulevard, Rishon LeZion, Israel. E-mail: eldadr@colman.ac.il

vidual method. Specifically, they calculated the incremental predictive validity of work sample testing, and found it to represent a 24% increase in validity over cognitive ability measures. Moreover, they argue that work sample testing taps performance-related factors that are unrelated to general intelligence and by doing so provides additional predictive validity (Callinan & Robertson, 2000). Although Roth et al. (2005) have recently updated and lowered the mean magnitude of the criterion-related validity of work sample testing, they still determine that this approach is valid for predicting job performance.

Aside from their predictive quality, simulations follow the recent focus of personnel selection research on the applicant perspective rather than the organizational perspective (Arthur & Villado, 2008). This line of research emphasizes candidates' perception of equality, face validity, test-taking motivation, test performance, and self-withdrawal from the selection process (e.g., Anderson, 2004). Moreover, empirical evidence demonstrates that work sample testing is associated with substantially lower levels of ethnic group bias than is cognitive ability and relatively lower adverse impacts (e.g., Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001). In fact, while cognitive ability tests administered to minority groups generate mean differences of approximately one standard deviation (Schmitt & Mills, 2001), work sample testing shows a considerable decrease in this effect. To illustrate, Schmidt, Clause, and Pulakos (1996) calculated a standard mean difference of .38 between Caucasian and African American participants on work sample tests in comparison to a standard deviation of .83 on traditional ability measures. Furthermore, work sample tests are positively viewed by applicants (e.g., Hausknecht, Day, & Thomas, 2004) and provide a preview of the actual job (e.g., Steiner & Gilliland, 1996); thus, they are perceived by candidates as more just than other selection procedures (Callinan & Robertson, 2000).

Despite these overt advantages that have been lauded by researchers for more than 50 years (Feinstein & Cannon, 2002), simulation testing has not yet challenged the superiority of mental ability assessment methods, and is still perceived as a supplementary measure (Schmidt & Hunter, 1998). Perhaps due to its inherent complexity and high cost (Lievens & Patterson, 2011), research in this domain remains relatively limited. Nonetheless, scholars have encouraged researchers to replicate the results obtained in laboratory studies in field situations (Schmitt & Mills, 2001), and to conduct predictive validity studies with simulations, as a means of assisting personnel decision-makers in improving their selection systems. Therefore, in the current study we followed this line of thinking and attempted to advance our knowledge further and to provide additional validity of simulation measurements in a military setting, which is characterized by high degrees of complexity and competitiveness (Salas et al., 2009). Specifically, we have conducted a longitudinal study with hundreds of Israeli navy soldiers and measured the predictive role of several simulations in assessing future military performance.

Research and implementation of simulations in military settings is not novel (e.g., Siegel & Bergman, 1975). The rationale for developing this procedure particularly in a military selection system is based on the above mentioned advantages. Of special relevance to the military setting is the tendency of work sample tests to reduce, if not eliminate, the adverse impact against minorities that is routinely manifested in cognitive ability measures (see Hough, Oswald, & Ployhart, 2001, for a comprehensive review). Since recruitment for military service in the Israeli army is compulsory by law, all the subgroups of the population are obliged to pass through the same selection procedure. This poses a challenge to developing an unbiased military selection system, and necessitates serious consideration of minority and equality issues. As mentioned above, simulations provide adequate rates of validity, offer a realistic job preview for candidates, and produce favorable applicant reactions as well as lower bias against minority candidates (Clevenger et al., 2001). Therefore, the development of these tests in the military selection system appears to be not only adequate, but crucial.

The Current Study

For the purposes of the current study, we created a set of hands-on performance tests that tapped particular behaviors relevant to navy soldiers' performance. We developed these simulations because they are believed to be better indicators of future job performance, according to the logic of behavioral consistency (Motowidlo et al., 1990). As previously demonstrated

(e.g., Feinstein & Cannon, 2002, 2003; Lievens & Patterson, 2011; Schmitt & Mills, 2001), simulations hold certain advantages over traditional ability tests. Specifically, they produce less adverse impact, higher positive reactions of examinees, and higher degrees of content validity. However, such simulations are probably the highest cost method in terms of time and resources, since they require the expensive development and implementation of necessary equipment (Motowidlo et al., 1990). Nevertheless, they enable us to maximize the potential predictive validity and to capture some unique behavioral features that cannot be obtained through traditional assessment procedures (Weekley & Ployhart, 2006).

In the current study's simulations we strived to mirror actual military tasks and contexts, while maintaining point-to-point correspondence with the navy soldier's actual role. In this sense, we provided candidates a platform in which they could translate their procedural knowledge into actual behavior (Lievens & Patterson, 2011). We assembled these task-specific simulations in a manner that did not require any previous experience, hence further reducing adverse and ethnic effects (Schmitt & Mills, 2001).

In order to develop the study's simulations, we followed Wenzler's (2009) practical perspective and focused our attention on recognizing the unique needs of the navy selection system and delivering value to personnel decision-makers. For the purpose of clearly defining the challenges of navy soldiers' performance, we implemented Feinstein and Cannon's (2003) bottom-up framework and initially assembled a panel of eight experts in the field of naval training. The panel was composed of senior navy instructors and navy personnel officers, and its mission was to provide significant behavioral requirements for navy soldiers' effective performance. This preliminary procedure was conducted in line with Cannon and Burns' (1999) recommendation first to evaluate the actual job performance requirements and then to design simulations to represent these behavioral requirements.

Members of the panel initially were instructed to individually generate a list of important behavioral competencies in navy soldiers' performance. Next, they were asked to combine their lists and to omit duplicate items. The first unified list consisted of seven items. Subsequently, members of the panel were consulted to address the relative importance of these items, thus further validating their list. During this process they decided to omit four items. Hence, the final list of central behavioral dimensions in navy soldiers' performance included intellectual capabilities (i.e., mental and cognitive skills that are relevant for learning and acquiring new information), social skills (i.e., interacting and communicating with others while promoting cooperation and collaboration), and sailing adjustment (i.e., overcoming seasickness and performing effectively while sailing on a craft). Based on these general behavioral dimensions, we constructed four simulations: (a) naval-navigation test simulation, (b) raft simulation, (c) rubber boat (zodiac) mounting simulation, and (d) military tent assembly simulation. In constructing these simulations, we followed the recommendations of Motowidlo et al. (1990) and used realistic materials and equipment to represent task situations fully. In this sense, we provided applicants with a fair opportunity to respond almost exactly as if they were in an actual military setting.

To examine the predictive validity of these four simulations, we conducted a longitudinal study with new recruits to the Israeli Defense Forces (IDF). The study was divided into three main sessions. In the first session, the new recruits' fitness for combat ship service was evaluated in a 2-day screening session. In the second session, several months later, participants were assigned to training courses (e.g., radar operator, navigator, mechanic, etc.), and their performance was evaluated upon completion of the courses. In the third session, which took place at the end of the first year of active military service, the participants' performance on combat ships was evaluated by their commanding officers.

To assess participants' potential performance in the upcoming training courses, we administered the naval-navigation test simulation. The scores in this simulation served as a predictor while the participants' training course final grade served as a criterion. During this simulation, all the participants were put together in a large room where they heard a lecture on the topic of naval navigation. Subsequently, we provided them with a brief written summary of the lecture and asked them to study the topic

and prepare for a knowledge test. In the upcoming day we administered this test. Through this simulation we imitated a normal routine in the naval training courses. During this routine trainees attend classes, take exams, read written materials, and need to manage their time and find effective learning opportunities. Hence, we hypothesized that:

> *H1:* Participants' final grade score in the training courses will be predicted by their score in the naval-navigation test simulation.

The aim of the raft simulation was to evaluate sailing adjustment and to assess participants' performance at sea. The scores in this simulation served as a predictor while the commander's evaluation of sailing adjustment served as a criterion. During this simulation we placed in the middle of the sea a 3-m inflatable life raft that is commonly used for emergency situations. The participants sailed to the raft in a speed-boat and were left there for an average of 90 minutes, accompanied by trained instructors. While being on the raft, participants had to complete a U.S. National Aeronautics and Space Administration moon-survival task. During this task, participants were instructed to analyze the written situation through group discussion, and to decide the rank order of the essential equipment. Hence, we hypothesized that:

> *H2:* Participants' adjustment for sailing aboard combat ships will be predicted by their performance during the raft simulation.

The third and fourth simulations were designed in order to evaluate participants' professional performance and social competence, respectively. These simulations were found effective for evaluating navy candidates' performance in a previous study we conducted (Rom & Mikulincer, 2003). During the zodiac simulation, the participants' mission was to mount a rubber boat, while following specific assembly instructions. The mounting mission is complicated and requires precision. This exercise simulates actual professional performance since it encompasses several essential skills, such as: analyzing and understanding technical information, reading figures, and construction abilities. In addition, we hypothesized that the score grade of the training course will also play a role in predicting professional performance. Hence, we used here the training course final grade score and the zodiac simulation score as predictors and the commander's rating of professional performance as a criterion. We hypothesized that:

> *H3:* Participants' professional performance aboard combat ships will be predicted by their performance during the zodiac mounting simulation as well as by their final training course score.

In the military tent simulation, the participants' mission was to construct a complex tent that requires high levels of coordination and cooperation for its construction. The tent consists of large sheets of rigid fabric that are attached to a number of poles and supporting ropes. On an average the tent assembly takes about 30 minutes for a small group of individuals. This simulation's purpose was to tap into social competencies, which commonly are divided (e.g., Rom & Mikulincer, 2003) into socioemotional functioning (i.e., the contribution to morale and cohesion of the team as well as conflict resolution among teammates) and instrumental functioning (i.e., the contribution to the successful completion of team tasks and the accomplishments of its goals). Hence, we used the tent assembly simulation score as predictor and the commander's rating of social adjustment as a criterion. We hypothesized that:

> *H4:* Participants' social adjustment aboard combat ships will be predicted by their performance during the military tent assembly simulation.

Method

Participants

A total of 1007 18-year-old men participated in the study. All of the participants had just begun their compulsory service in the IDF. Before the start of the screening session, all of the participants had undergone rigorous IDF tests and had been found to be suitable for serving in the army. All of the participants were single and had completed high school. Most of them resided in urban areas. Originally, the sample was composed of 1053 participants, but 46 were dropped because they failed to fill out all of the

research questionnaires or failed to complete the entire 2-day screening session.

Measures

A number of predictors and criterion variables comprised the current study's measures. The predictor variables comprised simulation scores and training performance (i.e., course completion grade), whereas the criterion variables comprised job performance evaluations (i.e., commander's ratings of professional performance, social adjustment, and sailing adjustment during active service aboard combat ship). Note that the final training course grade score served both as a criterion for the naval-navigation test simulation and as a predictor for commander's rating of professional performance.

To assess the participants' performance in the naval-navigation test simulation, a 15-item test was composed (e.g., "What is the difference between true north and grid north?"). Each test item had four alternative answers as well as a *don't know* response. Responses were scored 1 if correct or 0 if incorrect (M = 10.68, *SD* = 2.97). Cronbach's alpha for the naval-navigation test was .78. Also, in order to support the claim that this test measures intellectual capabilities, we calculated its correlation with the Wechsler Adult Intelligence Score (WAIS IV) which was administered to the participants at an earlier stage (M = 51.10, *SD* = 20.73). This analysis yielded a significant correlation of .49. To assess performance on the other three simulations (i.e., raft sailing, zodiac mounting, and military tent assembly), we used instructors' ratings. Namely, trained instructors were asked to evaluate participants' performance on a 9-point Likert-type scale ranging from 1 (*very poor performance*) to 9 (*excellent performance*). All simulations had behavioral anchors that described poor versus excellent performance. For example, in the raft sailing simulation, poor performance is manifested in a numb behavior where individuals tend to minimize communication with their surroundings, feel sick, or feel sleepy. Excellent performance, on the other hand, is manifested in a vital behavior where individuals tend to communicate effectively with their surroundings and are well focused. Raft sailing simulation mean rating was 5.81 (*SD* = 1.14), zodiac mounting simulation mean rating was 5.87 (*SD* = 1.20), and tent assembly simulation mean rating was 5.68 (*SD* = 1.29). Participants' performance during the training course was reflected in their final course grade. This grade was calculated by their course instructors as an average of all their exams during the training process. This score ranged from 0 to 100 (M = 90.09, *SD* = 5.72). Participants' performance on combat ships was evaluated by their direct commanding officers by the end of their first year on the vessel. Specifically, we asked the officers to assess participants' professional, social, and sailing functioning on a 5-point Likert-type scale ranging from 1 (*not at all*) to 5 (*very much*). Commanding officers' mean assessment of their professional performance (i.e., analyzing and understanding technical information, precision, and stress management), was 4.10 (*SD* = .90). Cronbach's alpha was .68. Their mean evaluation of participants' social adjustment (i.e., conflict resolution, contributing to morale and cohesion, and promoting collective tasks) was 4.09 (*SD* = 1.01). Cronbach's alpha was .68. And their mean score evaluation of participants' sailing adjustment (i.e., vital behavior in sea, not feeling seasickness, and maintaining focus while sailing) was 4.30 (*SD* = .79). Cronbach's alpha was .67.

Procedure

The study was conducted at an IDF base in the northern area of Israel. In the first session, the participants underwent a 2-day screening session in which their fitness to serve in combat units was evaluated. During this session we administered all the simulations. It is worth mentioning that at that stage the soldiers were still screened by original tests and interviews. The simulations were not taken into consideration during the selection process since they were still in a validation stage. Also, we assured all participants that the confidentiality of their responses would be maintained and that all data obtained would be used only for research purposes. On the first day of this session, we conducted the naval-navigation test simulation, which simulates performance in training courses. As mentioned above, participants heard a lecture regarding naval navigation and subsequently received a brief written summary of this

lecture. The following day, a naval-navigation test was administered to the participants.

On the second day of the screening session, all the participants were randomly assigned to small groups of 6–9 members wherein they completed the remaining three simulations, namely, the raft sailing simulation, zodiac mounting simulation, and military tent assembly simulation. The order of these three missions was randomized across groups. An instructor escorted and evaluated each group. All the instructors were navy soldier veterans that were trained for measuring participants' performance. The instructors' training process included reading reports, observing senior instructors' evaluations, and receiving on-the-job training. The instructors' role was to explain each mission and rate the participants' performance. During each mission, the instructor acted only as an observer and did not intervene in the deliberations, decisions, or performance of a group.

In the study's second session, which took place several months later, we evaluated participants' performance in the training courses. Throughout these courses the participants were trained for specific positions in combat ships, such as radar operators, navigators, mechanics, and radio operators. After completing the training courses, participants were assigned to a number of combat ships, where they commenced their active military service. By the end of their first year as soldiers on these vessels, we conducted the third session of the study. During this session, we approached the combat ships' commanding officers and asked them to evaluate the participants' performance throughout their first year of service on the ship. Each commanding officer rated a number of soldiers ($M = 5.7$, $SD = 1.11$) that were under his direct command for the past year.

Results

In order to test our predictions, we initially calculated the associations between the participants' performance scores in the four simulations, their final grade training course score, and the performance evaluations made by their commanding officers during their active service aboard battle ships (see Table 1).

As the table shows, all three group simulations (i.e., raft sailing, zodiac mounting, and tent assembly) yielded high degrees of significant intercorrelation (ranging from .77 to .82), significantly higher than the .11 average correlation between the naval-navigation test simulation and these three simulations. This pattern of correlations provides some support for the notion that the three group simulations have an incremental validity over the naval-navigation test simulation. This significant difference confirms our expectation that the three group simulations measure a different kind of knowledge from what the naval-navigation test simulation measures. Specifically, we speculated that the group simulations tap into knowledge, skills, and abilities that are relevant to teamwork, instrumental functioning, and social skills, while the naval-navigation test simulation taps into analytical and intellectual capabilities. As the table also shows, the training course final grade score correlated significantly with all four simulations, and its highest correlation was with the

Table 1
Zero-Order Correlations for Course Score, Simulations, and Performance Evaluations

	1	2	3	4	5	6	7
1. Course score							
Simulations:							
2. Navigation	.14**						
3. Raft sailing	.08*	.11**					
4. Zodiac	.07*	.11**	.82**				
5. Tent	.08*	.10**	.79**	.77**			
Commanding Officer's Evaluations:							
6. Professional	.27**	.02	.03	.30**	.08		
7. Sailing	.15**	.03	.17**	.17**	.16**	.43**	
8. Social	.09	.06	.08	.11	.15**	.41**	.62**

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

navigation test simulation. All commanding officers' evaluations (i.e., professional, social, sailing) yielded moderate degrees of significant intercorrelation (ranging from .41 to .62). However, carefully observing the associations between these evaluations and the study's predictors yielded a multifaceted pattern. Specifically, professional performance evaluation correlated significantly with course final grade score and zodiac mounting simulation score. These correlations support the claim that the zodiac simulation indeed taps into professional knowledge, skills, and abilities, and that achieving high scores in the training course is associated with effective professional functioning on the combat vessel. In addition, sailing adjustment yielded small degrees of significant correlation with all the predictors but the navigation test simulation. This pattern of findings suggests that the sailing adjustment is a central competency that is probably associated with a number of behavioral skills and abilities that are measured in the study's predictors. Also it further supports the notion that the three group simulations and the naval-test simulation measure different kinds of mental constructs. Finally, commanding officers' social performance evaluations correlated significantly only with tent assembly simulation. This association further supports our notion and previous research findings (Rom & Mikulincer, 2003) that the tent simulation taps into social knowledge, skills, and abilities. Given the correlation degrees among some of the variables, we conducted collinearity analyses in order to revoke multicollinearity issues. These analyses revealed low levels of multicollinearity (VIF lower than 3).

After conducting these preliminary analyses, we computed a series of regressions in an attempt to assess the predictive validity of our measures (see Table 2).

To test our first hypothesis, namely that participants' final grade score in the training courses would be predicted by the naval-navigation test simulation score, we conducted a simultaneous forward linear regression analysis. In this analysis, all four simulations were introduced as predictors. This regression analysis yielded the expected significant effect of naval-navigation test simulation. None of the other predictors' contributions was significant. Hence, our first hypothesis was supported.

Our second hypothesis was that participants' adjustment for sailing in combat ships would be predicted by their performance during the raft sailing simulation. To test this hypothesis, we conducted a simultaneous forward linear regression analysis, in which all four simulation scores as well as the training course final grade score were introduced as predictors. This regression analysis did not yield the expected effect, as the raft sailing simulation did not reveal a significant effect. No other effects were significant. Hence, our second hypothesis was not supported and the raft sailing simulation did not fulfill its predictive role.

To test our third hypothesis, namely that participants' professional performance in combat ships would be predicted by the training course final grade score as well as by the zodiac mounting simulation score, we conducted a simultaneous forward linear regression analysis, while introducing as predictors the same set of variables as in the previous analysis. This regression analysis yielded the expected significant effects for course final grade score and zodiac mounting simulation. None of the other predic-

Table 2
Summary of Regression Analyses for Variables Predicting Course Score and Performance Evaluations

	Course Score			Professional			Sailing			Social		
Variables	B	*SE* B	β	B	*SE* B	β	B	*SE* B	β	B	*SE* B	β
Course Score				.70	.2	.25**	.07	.18	.02	.17	.22	.06
Navigation	.33	.16	.13**	.2	.2	.08	.28	.25	.11	.14	.18	.07
Raft	.15	.19	.04	.07	.18	.02	.11	.12	.01	.15	.19	.07
Zodiac	.08	.19	.03	.76	.28	.28**	.12	.14	.03	.10	.11	.03
Tent	.20	.21	.08	.08	.19	.03	.14	18	.09	.48	024	.13*
R^2		.11			.19			.03			.07	
F		2.42**			5.48**			.36			4.66*	

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

tors' contributions was significant. Hence, our third hypothesis was supported.

Finally, to test our fourth hypothesis, namely that participants' social competency would be predicted by their performance during the military tent assembly simulation, we conducted a simultaneous forward linear regression analysis while introducing as predictors the same set of variables as in the previous analyses. This regression analysis yielded the expected significant effects, with the tent assembly simulation yielding a significant effect. None of the other predictors' contributions was significant. Hence, our fourth hypothesis was supported.

Discussion

Feinstein and Cannon (2003) claim that there is a traditional deficit of empirical findings on simulation validity. In focusing the current study on the validation process of simulations, we have attempted to contribute modestly to this domain by evaluating the predictive power of simulations. Throughout this process, we have also endeavored to advance the military selection system and promote its personnel decision making. In this sense, the main contribution of the current study is in advancing our knowledge concerning the potential role of simulations as predictors for job performance. By doing so, we also followed researchers' recommendations (e.g., Schmitt & Mills, 2001) to study simulations in field research rather than in the laboratory.

Overall, most of the current study's findings were in line with our hypotheses, thus providing further support for the predictive validity of simulations. More specifically, all but one simulation (raft sailing) successfully predicted the performance of navy soldiers in actual combat ship settings. These findings are aligned with previous studies (e.g., Lievens & Patterson, 2011) that have demonstrated the validity of simulations in predicting job performance.

We claim that the three other simulations (i.e., naval-navigation test, zodiac mounting, and tent assembly) possess some degree of external validity, since they have met the requirements made by scholars (e.g., Feinstein & Cannon, 2003) and were significantly associated with a real-world system (i.e., commanding officers' actual performance evaluations).

To compare the study's predictors, we use the basic logic of the multitrait-multimethod matrix (originally introduced by Campbell and Fiske [1959] in their classical framework). As stated above, Arthur and Villado (2008) highlight the importance of distinguishing between constructs and methods. Hence, we speculate that the high degrees of positive significant correlation between the three group simulations (raft sailing, zodiac mounting, and tent assembly) partially can be attributed to resemblance of method. In a similar vein, the relatively low, though significant, correlations between the naval-navigation test simulation and the three group simulations may further support this claim by representing the incremental validity of the simulations' methods. Nevertheless, this pattern of correlations may also tap a different hypothetical construct. In other words, the high degree of association between the three group simulations along with the low degree of association of these simulations with the naval-navigation test simulation possibly may be explained not only by the general resemblance of and differences between the simulation methods, but also by the type of knowledge, skills, and abilities required for these simulations. More specifically, while the group simulations presumably require procedural knowledge that is related to the combat ship experience of sailing, engaging in professional duties, and acting as a member of a team, the naval-navigation test simulation taps into intellectual capabilities that are more relevant to training course experience. Since we did not apply alternative measurement methods for assessing the same construct in the current study, we could not attain a conclusive answer and determine the degrees of variance that are attributable to method and construct separately.

Although we were very encouraged by the findings that most of our simulations significantly predicted navy soldiers' performance on combat ships, we strived to understand the reasons for the prediction failure of the raft sailing simulation. We speculate that this simulation did not fulfill its predictive role partially due to design and development issues. Specifically, in our endeavor to design the sailing simulation, we might have overlooked some crucial aspects of the actual combat ship sailing experience. Namely, the physical sensations of raft sailing are very different from those elicited during

combat ship sailing. Furthermore, the limited time of the simulation (90 minutes) also damaged its external validity, since it did not provide the time required to imitate a sailing experience. Thus, in the design of the raft sailing simulation we encountered a familiar problem in the simulation literature, where simulations appear to be so elegant and rational that it is hard to resist the temptation of assuming they fully and truthfully represent reality (Wenzler, 2009).

Before concluding, it is worth mentioning several boundaries that may limit the generalizability of the current study's findings. First, the participants in the study were all young male Israeli citizens; hence, the findings should be replicated using other age, gender, and cultural samples. Second, the study focused on a military setting, and should be replicated in other organizational contexts, such as public sector and business organizations. Third, all group simulation scores and job performance evaluations were done subjectively. Future research should attempt to employ more objective robust measurement methods for evaluating individuals' performances. In order to further develop the simulation literature, we urge future researchers to apply complex study designs in which diverse methods of measurement are applied along with simulations, thus making it possible to assess the convergent and discriminant validity of this procedure. Moreover, we encourage future researchers to develop complex simulations that will truly imitate reality and thus help to stem the burgeoning phenomenon of independent coaching firms that assist applicants to perform successfully in selection tests (Lievens & Patterson, 2011). We believe that simulations, in which candidates can act realistically in many aspects, may provide an accurate picture of genuine capabilities and authentic patterns of behavior, thus improving their predictive potential. Therefore, although designing and developing simulations tends to be a challenging and expensive process (Lievens & Patterson, 2011), we believe that its overall value is justified. In this sense we follow other scholars' view of simulation testing as a complex but effective selection method that motivates candidates and enhances their positive perception (e.g., Anderson, 2004).

In sum, although simulations have been praised by researchers for more than five decades, there is a substantial lack of empirical studies of this domain (Feinstein & Cannon, 2002). The relative lack of progress in simulation research frequently is attributed to inherent difficulties in creating acceptable evaluation methodologies. In this study we have attempted to overcome these difficulties. Despite its possible limitations, the current study emphasizes the relevance of simulations in predicting performance in complex contexts and contributes to the conceptual and empirical integration of the field of selection and personnel decision-making.

References

Anderson, N. (2004). The dark side of the moon: Applicant perspectives, negative psychological effects (NPEs), and candidate decision making in selection. *International Journal of Selection and Assessment, 12,* 1–8. doi:10.1111/j.0965-075X.2004.00259.x

Arthur, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93,* 435–442. doi:10.1037/0021-9010.93.2.435

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8,* 248–260. doi:10.1111/1468-2389.00154

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105. doi:10.1037/h0046016

Cannon, H. M., & Burns, A. C. (1999). A framework for assessing the competencies reflected in simulation performance. *Developments in Business Simulation and Experiential Exercises, 26,* 40–44.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Schmidt-Harvey, V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86,* 410–417. doi:10.1037/0021-9010.86.3.410

Feinstein, A. H., & Cannon, H. M. (2002). Constructs of simulation evaluation. *Simulation & Gaming, 33,* 425–440. doi:10.1177/1046878102238606

Feinstein, A. H., & Cannon, H. M. (2003). A hermeneutical approach to external validation of simulation models. *Simulation & Gaming, 34,* 186–197. doi:10.1177/1046878103034002002

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Per-*

sonnel Psychology, 57, 639–683. doi:10.1111/j.1744-6570.2004.00003.x

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9,* 152–194. doi:10.1111/1468-2389.00171

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85,* 869–879. doi:10.1037/0021-9010.85.6.869

LePine, J. A., Colquitt, J. A., & Erez, A. (2000). Adaptability to changing task contexts: Effects of general cognitive ability, conscientiousness, and openness to experience. *Personnel Psychology, 53,* 563–593. doi:10.1111/j.1744-6570.2000.tb00214.x

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology, 96,* 927–940. doi:10.1037/a0023496

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low fidelity simulation. *Journal of Applied Psychology, 75,* 640–647. doi:10.1037/0021-9010.75.6.640

Ones, D. S., & Viswesvaran, C. (2007). A research note on the incremental validity of job knowledge and integrity tests for predicting maximal performance. *Human Performance, 20,* 293–303. doi:10.1080/08959280701333461

Ployhart, R. E., Schneider, B., & Schmitt, N. (2010). *Staffing organizations: Contemporary practice and research*. Mahwah, NJ: Erlbaum.

Rom, E., & Mikulincer, M. (2003). Attachment theory and group processes: The association between attachment style and group-related representations, goals, memories, and functioning. *Journal of Personality and Social Psychology, 84,* 1220–1235. doi:10.1037/0022-3514.84.6.1220

Roth, P., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58,* 1009–1037. doi:10.1111/j.1744-6570.2005.00714.x

Salas, E., Rosen, M. A., Held, J. D., & Weissmuller, J. J. (2009). Performance measurement in simulation-based training: A review and best practices. *Simulation & Gaming, 40,* 328–376. doi:10.1177/1046878108326734

Schmidt, F. L., Clause, C. S., & Pulakos, E. (1996). Subgroup differences associated with different measures of some common job-relevant constructs. In C. L. Cooper & J. T. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 11, pp. 115–140). New York, NY: Wiley.

Schmidt, F. L., & Hunter, J. E. (1998). The validity of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274. doi:10.1037/0033-2909.124.2.262

Schmitt, N., & Mills, A. E. (2001). Traditional test and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology, 86,* 451–458. doi:10.1037/0021-9010.86.3.451

Siegel, A. I., & Bergman, B. A. (1975). A job learning approach for performance prediction. *Personnel Psychology, 28,* 325–339. doi:10.1111/j.1744-6570.1975.tb01540.x

Steiner, D. S., & Gilliland, S. W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *Journal of Applied Psychology, 81,* 134–141. doi:10.1037/0021-9010.81.2.134

Weekley, J. A., & Ployhart, R. E. (2006). An introduction to situational judgment testing. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 1–10). Mahwah, NJ: Erlbaum.

Wenzler, I. (2008). The ten commandments for translating simulation results into real-life performance. *Simulation & Gaming, 40,* 98–109. doi:10.1177/1046878107308077

Received July 8, 2013
Accepted July 9, 2013 ■