A Personality Testing Program in the U.S. Navy

Walter C. Borman
Janis S. Houston
Robert J. Schneider
Kerri L. Ferstl
Personnel Decisions Research Institutes, Inc.

William L. Farmer
Ronald M. Bearden
Navy Personnel Research, Studies, and Technology (NPRST)
Navy Personnel Command

This paper describes the background, development, and results of a pilot study of the Enlisted Computerized Adaptive Personality Scales (ENCAPS), an instrument designed to measure non-cognitive attributes for the purpose of selecting and classifying recruits into enlisted positions in the United States Navy. ENCAPS utilizes an adaptive testing framework to, hopefully, measure personality in a more precise way than has been possible to date.

Background

The notion of adaptive testing has had an application primarily in the cognitive ability domain. Briefly, the strategy is to first present to test-takers item(s) at an intermediate level of difficulty. If a test-taker answers the item(s) correctly, the computerized adaptive testing (CAT) algorithm will identify within an item pool an item that is somewhat more difficult. If the initial item(s) are answered incorrectly, then an item of somewhat less difficulty is presented. In this manner, the CAT administers items to test-takers tailored to their ability level. As an example, in the first case, the person taking the test will be administered more and more difficult items until he/she answers an item incorrectly. This provides an initial estimate of the person's ability level, and subsequent administration of items around that difficulty level fine-tune that ability estimate. With this approach, the test-taker's ability level is estimated with more precision and greater efficiency (i.e., less testing time), compared to that provided by a typical ability test.

Recently, this CAT concept was adapted for use in the domain of performance assessment. Borman, Buck, Hanson, Motowidlo, Drasgow, & Stark (2001), developed a system that presents pairs of behavioral statements scaled according to effectiveness and asks the rater to choose which of the two statements is more descriptive of the ratee. The effectiveness value of that statement then becomes the initial estimated performance level for the ratee. Next, an algorithm is triggered that identifies two more statements with effectiveness values bracketing the initial estimated performance level in such a way that the information provided by the next choice of the more descriptive statement is maximized in an item response theory (IRT) sense. This next choice of a statement also adjusts the estimated performance level for the ratee. The assessment process is iterative, with pairs of statements being presented using the same algorithm, for several additional statement pairs.

In a laboratory study with videotaped ratees performing at pre-scripted levels, business people raters used this computerized adaptive rating scale (CARS) and either a behaviorally anchored rating format or a numerical rating scale to evaluate the videotaped ratees. Results showed that the CARS ratings had a considerably lower standard error or measurement, higher correlational validity, and higher accuracy compared to the other formats (Borman et al., 2001).

Introduction

The purpose of the present research is to transport the CAT and CARS concepts to personality testing. Specifically, we wanted to use the CARS algorithm in a self-report personality inventory. The notion is to present initially a pair or personality items with different levels of the target trait represented, and ask the test-taker to select which of the two items is more descriptive of him/her. This choice establishes an initial estimated value on the trait, and the CARS algorithm selects two additional items with trait level values bracketing the initial estimated value, again, in such a way that the personality trait information on the test-taker is maximized in an IRT sense. Thus, the iterative CARS algorithm was applied to the measurement of personality.

The Project

The first step in developing this ENCAPS inventory was to select three constructs highly likely important for success in Navy ratings (i.e., jobs), and as a "proof of concept," write items to measure those pilot constructs. These initial constructs were Achievement, Social Orientation, and Stress Tolerance.

Four members of the project team wrote a total of 233 items targeted toward the three constructs. Table 1 presents three items for Achievement.

Table 1
Sample ENCAPS Items Targeting Various Levels of Achievement

Sample Achievement Item	Target Trait Level
Obstacles energize me because they give me a chance to show what I can really do.	High
When I encounter obstacles, I like to make a good effort to overcome them.	Medium
I don't waste a lot of time trying to overcome obstacles that I didn't create.	Low

Next, 25 researchers familiar with the personality domain rated each of the items with respect to the level of the target trait represented (from 7 = extremely high to 1 = extremely low). The 1-rater interrater reliability of these ratings was .87; the reliability for the mean of all raters was .99. Further, most levels of the traits were well represented (i.e., from 1 to 7). A second round of item writing filled in the gaps that remained.

The final 280 ENCAPS items were entered into a database accessed by the ENCAPS program that runs on the Word 2000 operating system. As mentioned, an algorithm similar to that used for

CARS is used to select pairs of items that maximize item information. As respondents choose one of the pairs presented, the program uses that information (the trait level value) to determine the next pair to present, a pair that will again maximize trait information. For the time being, a limit of 10 pairs has been designated as the number of pairs to present to any respondent. Previous research by Stark and others (2003) suggests that this is a sufficient number of items to arrive at a stable estimate of the respondent's standing on the trait level. Although some computer adaptive formats present pairs of items where the items are from two different constructs, ENCAPS always presents pairs of items from the same construct, while alternating constructs from one pair of items to the next.

A pilot test of ENCAPS for the 3 constructs was next conducted. We included in the pilot test: (1) ENCAPS; (2) items from well established "marker" measures of these constructs (e.g., NEO and CPI); and (3) a representative set of ENCAPS items in a typical personality inventory format. These instruments were administered to 194 ROTC and other college students.

Results showed, first, that the spread of the scores for ENCAPS was satisfactory. Second, correlations between ENCAPS scores and scores from each of the other two inventories were computed, and for the ENCAPS and traditional format ENCAPS the convergent validity correlations were .60, .61, and .68, respectively, for Achievement, Social Orientation, and Stress Tolerance. With the marker scales, correlations were somewhat lower (.48, .58., and .67). In all cases, the diagonal correlations (convergent validities) were higher than the off-diagonal correlations (discriminant validities).

Regarding length of time to complete the three inventories, ENCAPS appears to have a considerable advantage. On average, ENCAPS took 6.1 minutes to complete. Making some assumptions about what would be a "comparable length" for the other formats, they would have taken about 14 minutes to complete. This advantage is important operationally, because the applicant testing time is limited.

Currently, the project team is identifying additional personality constructs to target for ENCAPS. We anticipate measuring a total of 9-10 constructs.

Future Work

Most important is the question of validity for predicting important criteria. We will be conducting comparative validity studies with the ENCAPS and the other two inventories. In particular, Navy enlistees in "boot camp" will be administered the three inventories, and follow-up data will be gathered for these same Sailors regarding attrition, training performance, and early career job performance. Our hypothesis is that the additional precision of measurement provided by ENCAPS (as with the CARS compared to the other rating formats) will result in higher validities for this method.

The longer term vision is to use ENCAPS, in combination with the ASVAB and JOIN (the Navy's vocational interest instrument) to screen out individuals at the bottom of the distribution *and* to help classify new Sailors into ratings (jobs) where they are more likely to performance effectively and remain in the service.

References

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86,* 965-973.

Stark, S. & Chernyshenko, O. (2003, April). *Using IRT methods to construct and score personality measures that are fake-resistant*. Paper presented at the 18th Annual Conference of the Society of Industrial and Organizational Psychology, Orlando, FL.