Computerized adaptive testing in the Bundeswehr

AB-5 - Paper

Computerized adaptive testing in the Bundeswehr

Dr. Ernst G. Storm
Bundeswehr - Central Office
Bonn, Germany

Within the psychological service of the Bundeswehr about 250.000 selection and placement procedures were conducted per year (draftees, volunteers). In a technical concept of quality control in personnel psychology the minimum standards were defined for the introduction, the utilization, and the control of instruments and procedures.

One aspect of this technical concept is to optimize the system and to adapt it to future conditions by implementing adaptive tests, based on the item response theory (IRT), into the testing procedure. This aim has been pursued since many years under the cooperation of Prof. Dr. Lutz F. Hornke, University of Aachen (Germany), Department of Psychology.

Five subtests out of a test battery of psychological aptitude measurements were chosen for further adaptive testing, but only the first three will be of interest here:

- abstract logical reasoning (matrice type items)

- verbal reasoning (analogies)

- numerical reasoning (arithmetic items)

- verbal memory and serial learning (in progress)

In combination with adaptive testing, modern test construction ought to be based on a theory of the ability concept in question (e.g. Bejar & Yocom, 1991; Embretson, 1983; Hornke & Rettig, 1988; Storm, 1995) . Accordingly, it should be possible to derive a construction rationale from the concept. Based on these rule sets, items for the above-mentioned domains were designed with psychological and psychometric properties which are rooted in the theories specified.

For each of these aptitude domains the item parameters for several hundred items were estimated based on empirical data from several thousand examinees (draftees, partially volunteers, see Table 1). Logistic one - (1 PL) and two parameter (2 PL) model estimates were conducted with preference given to the 2 PL model (Baker, 1987; Lord & Novick, 1968; Rasch, 1980). Corresponding tests of model fit in respect to the latter model led to an item re-duction and the remaining items were assembled in three well-fitting item banks (see Table 1):

Table 1

Values concerning item bank building (item linking, -calibration, equating)

	item bank building (evaluation)			result
	testforms	items/form	examinees	items/bank
matrice test	76	12	29.728	456
analogy test	25	24	12.548	254
arithmetic test	32	18	17.008	288

In a further step the adaptive 2 PL-algorithm was evaluated in combination with the three item banks mentioned above. Hence 8000 examinees (draftees) worked on the usual conventional test battery. Additionally, they received one of the three tests which corresponded to one test in the test battery. The additional test was administered as an adaptive one.

The aim of adaptive testing in the testing situation is to gain maximum information with minimum effort. Hence the examinees should be given only those items that meet the person's level of functioning, so that the items will be solved in about fifty percent of the cases. If that holds, irrelevant items can be economized on, and, as one of various merits, it can be expected that adaptive testing, particularly the 2 PL, will need much less items as against conventional tests based on classical test theory. The conventional tests of our test battery consist of 20 items each with moderate reliability ($r_{tt} = 0.74$ to $r_{tt} = 0.88$). Figure 1 is based on the compound result of the three tests. It shows the joint distribution of the adaptive items needed to reach a reliability of $r_{tt} \gg 0.84$ for each person. This reliability is associated with a standard error of measurement of SEM = 0.39 for this sample. With the appearance of variance in the answer vector, the SEM is calculated whenever a person answers an administered item within the adaptive testing procedure. Before testing, the SEM value is fixed on a corresponding desired reliability level (in this case $r_{tt} = 0.84$). This cut-off value serves as stop criterion of the adaptive procedure as soon as the SEM falls below.

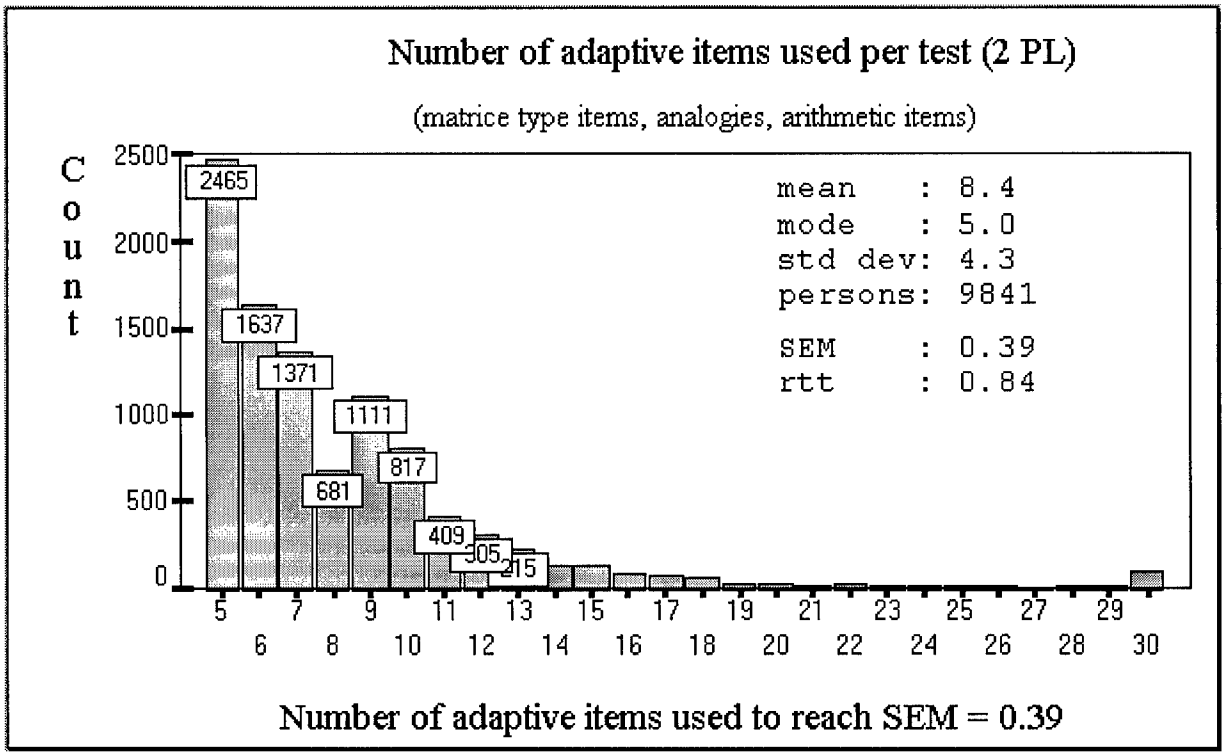

Figure 1. Distribution of the number of items needed for an adaptive testing situation

Fig. 1 illustrates that on an average 8.4 (matrice type items: 7.4; analogies: 9.9; arithmetic items: 7.8) items were needed per test to reach the SEM = 0.39 (r_{tt} = 0.84) with a mode of five items. This result clearly demonstrates the superiority of adaptive testing (2 PL) in respect of maximizing the output of item information. This leads to a drastically reduced number of items that must be administered to gain a superior amount of information about the examinee in comparison with conventional test procedures.

Because of the comparatively different test situation in adaptive testing, it must be taken into account whether the time spent per item is increased. This could be put down to the facts that items with difficulties that are too low are omitted or that any speed factor must be reduced to an acceptable level. Hence, a reduced number of items through adaptive testing must not necessarily lead to a reduced test taking time in comparison to conventional testing.

Table 2

Differences of test taking time in conventional and adaptive testing

	test taking time in minutes		
	matrice test	analogy test	arithmetic test
conventional testing	13.4 (n=8318)	4.3 (n=9822)	13.7 (n=9823)
adaptive testing	8.2 (n=3313)	3.7 (n=3457)	8.1 (n=3071)
saving	5.2	0.6	5.6

Table 2 shows test taking time for conventional and adaptive testing. It is obvious that the reduction of items in adaptive testing in this study leads to a clear reduction of test taking time, especially in demanding tests (matrice-, arithmetic test). As to about 250.000 testings a year a reduction of about 10 minutes per examinee with regard to three tests means a lot of savings for the organization. On the other hand the construction of an adaptive test with about 150 and more items can take years of development especially the first time. In the end, adaptive tests that are based on the item response theory with items generated with the help of an item construction rationale seem to be nearly in every aspect superior to conventional tests.

Another aspect of interest was the correlation of the adaptive and the corresponding conventional test. It seems reasonable that the corresponding tests should correlate higher than any other cross combination although the correlations in general can't be high because of the low reliabilities of the conventional tests except the arithmetic test ($r_{tt(matrices)} = 0.74$, $r_{tt(analogies)} = 0.75$, and $r_{tt(arithmetics)} = 0.88$).

Table 3

Correlation between conventional and adaptive tests

	conventional		
adaptive	matrice test	analogy test	arithmetic test
matrice test	r = 0.71 (n=2762)	r = 0.42 (n= 3309	r = 0.54 (n= 3308)
analogy test		r = 0.64 (n= 3452)	r = 0.46 (n= 3061)
arithmetic test			r = 0.69 (n= 3061)

Table 3 confirms the expectations of highest correlations between the corresponding tests in the diagonal and noticeable lower correlations between all other combinations. The magnitude of these coefficients follows the steps of the corresponding reliabilities. As to be further expected the correlation between the arithmetic and the matrice test is higher than the correlation between the analogy test and each of the others. Apart from the fact that the correlations are generally low, the correlation matrix itself reflects all

expectancies with regard to construct validation.

It is sometimes mentioned that adaptive test procedures are more demanding than conventional ones, because every item chosen by the adaptive algorithm reflects the ability level of the examinee. None the less it can be argued that the examinee feels tested in the right way because items that are too hard and too easy were omitted (e.g. Wainer, 1990). Within the testing procedure 60 examinees were given a questionnaire on the evaluation of the adaptive and the corresponding conventional test. The preliminary main results are as follows:

Table 4

Comparison between adaptive and conventional testing

	test is more difficult	test is more aggreeable
adaptive test	30 %	45 %
conventional test	50 %	40 %
neither	20 %	15 %

Regarding the small sample, Table 4 indicates that adaptive testing is not evaluated as being the more demanding test procedure. The adaptive test was rated even slightly better regarding difficulty, whereas both test procedures seem to be comparatively agreeable.

Conclusion

The first application of adaptive testing within the selection and placement procedures of the Bundeswehr in general is very promising and yields partially better results than expected. Corresponding to the reduced number of items needed to reach an appropriate reliability (see Fig. 1), the test taking time could be reduced by about 36 percent (see Table 2). This in fact is an enormous percentage in regard to 250.000 diagnostic procedures the year. The actual profit, however, is higher, because the adaptive procedure was terminated at a relatively high reliability level and not at the low levels of the conventional tests (except for the arithmetic test). Also was the minimum number of items fixed at five, because some examinees with 3 or 4 items mistrusted such a short test which is clearly shorter than its introduction. From this point of view it seems promising to build up an adaptive test battery with 10 and more tests (except for speed tests).

But even if there is the know-how of building an adaptive item bank, the costs at the moment are much higher as against conventional test construction. Reducing these starting costs is an aim which is pursued by our new test concepts of verbal memory and serial learning. On the one hand these tests have pleasant multi-media features and on the other they are constructed consistently, based on rule sets, so that the adjusted r^2 is

about 90 percent. When such rule sets are stringent enough, most of the items have a good chance to „survive" the fit tests of the logistic models. Furthermore, there is almost no problem to generate one or two hundred such items, which is a reasonable size for a well-designed item bank. One step further, such rule sets should allow to construct items on line, while testing. If that works, there will be no need for an explicit item bank, which, along the lines of procedural adaptive testing (Bejar, 1986), is replaced through an item-generating-algorithm derived from such a rule set.

References

Baker, F.B. (1987). Methodology Review: Item parameter estimation under the one-, two-, and three-parameter logistic models. Applied Psychological Measurement, 11 (2), 111-147.

Bejar, I.I. (1986). Final Report: Adaptive assessment of spatial abilities. Princeton, N.J.: ETS.

Bejar, I.I. & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden figure items. Applied Psychological Measurement, 15 (2), 129-137.

Embetson, Susan (1983). Construct Validity: Construct representation versus nomothetic span. Psychological Bulletin, 93 (1), 179-197.

Hornke, L.F. & Rettig, K. (1988). Regelgeleitete Itemkonstruktion unter Zuhilfenahme kognitionspsychologischer Überlegungen (Rule based item construction with help of cognitive psychology). In: K.D. Kubinger (Hrsg.), Moderne Testtheorie (140-162). Weinheim u.a.: Psychologie Verlags Union.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

Storm, E.G. (1995). Theoriegeleitete Testkonstruktion. Erfassung visueller Analyseleistungen anhand neuartiger, computergenerierter „Eingekleideter Figur-Aufgaben" (Theory based test construction. Assessing visual perception with new, computer generated embedded figure items). Arbeitsberichte Psychologischer Dienst der Bundeswehr. Bonn: Bundesminsterium der Verteidigung - P II 4.

Wainer, H. (1990). Computerized adaptive testing. A primer. Hillsdale, N.J.: LEA.

INTERNET DOCUMENT INFORMATION FORM

A . Report Title: Computerized Adaptive Testing in the Bundeswehr

B. DATE Report Downloaded From the Internet 4/21/99

C. Report's Point of Contact: (Name, Organization, Address, Office Symbol, & Ph #): **Navy Advancement Center**
Attn: Dr. Grover Diel (850) 452-1815
Pensacola, FL

D. Currently Applicable Classification Level: Unclassified

E. Distribution Statement A: Approved for Public Release

F. The foregoing information was compiled and provided by:
DTIC-OCA, Initials: VM_ **Preparation Date**: 4/21/99__

The foregoing information should exactly correspond to the Title, Report Number, and the Date on the accompanying report document. If there are mismatches, or other questions, contact the above OCA Representative for resolution.