COMPUTERIZED TESTING IN THE GERMAN FEDERAL ARMED FORCES (FAF): EMPIRICAL APPROACHES

WOLFGANG WILDGRUBE
PSYCHOLOGICAL SERVICE OF THE FAF

In April 1982, the Federal Armed Forces initiated the first empirical pilot project in the area of computerized adaptive testing (CAT) to collect data for psychological testing. Four microprocessor-based stand-alone computers called TEST 2000 (manufactured by the German firm Zak, which specializes in psychological and physiological measurement) were installed at the recruiting center for volunteers in Munich. This was the first installation for computerized testing in Germany and, according to present information, in Europe. TEST 2000 had previously been installed at some locations in Western Europe only for physiological measurement and in connection with an optical mark reader for preprocessing of data and simple analyses.

The first empirical phase started with a conventional testing strategy and the standard entrance test battery to collect data (1) for the examination of technical aspects (e.g., at the moment there are some problems with the keyboard for the examinees, and some trouble with the memory, which has become overheated during a period of hot weather); (2) for organizational aspects (computerized testing carried out during the entire selection process) and attitudinal change (e.g., reducing the psychologists' "computer anxiety"); and (3) for "human factor" aspects (e.g., motivation of volunteers, problems in handling the test computer). Collection of these first empirical data also presented an opportunity to compare paper-and-pencil testing versus computerized testing.

Configuration

Hardware

The testing station in Munich consists of four TEST 2000 computers. Each microprocessor-based stand-alone computer has a screen, processor, two floppy disks (a system disk and a data disk for approximately 80 persons), a special keyboard for the examinee (digits 1 to 9 and 0; four buttons: "correct," "incorrect," "I don't know," and "next item"), and a typewriter keyboard for the psychologist/proctor. It is possible to present the items in black on a green background or in green on a black background. One testing computer is connected with a line printer for the output of the testing results (The data disks from the three other computers are transferred to this computer, which is connected with the printer subsequent to the testing session).

TEST 2000 is based on a highly integrated unit of the ZILOG Z80A-family.

The stand-alone version offers the following:
1. 64 KByte RAM, 8KByte used by the operating system,
2. 8 KByte PROM containing integer- and floating-point arithmetic and system utility routines,
3. 16 KByte video refresh RAM, which stores high resolution graphic display (130,000 dots) or 8 pages of alphanumeric information,
4. 2 RS-232 serial interfaces, with reprogrammable drivers included in the operating system,
5. 2 8-bit parallel general purpose interfaces,
6. 4 programmable timer/counter channels,
7. 2 integrated minidisk drives with 220Kbyte (formatted) each, with double density/double sided versions optional,
8. 9" industry quality standard monitor (monochromatic),
9. Separate operator keyboard (7-bit ASCII code),
10. Separate examinee keyboard, and
11. System expansion bus and slots for 2 S100-bus or 8 europe-single-format cards.

Psychological Section

Whole batteries of tests or inventories are presented automatically, including test instructions and training sessions. There is a selectable mode of presentation for each test. The proctor/psychologist may choose between power, speed, or power-with-time-limit tests. Selection of items to be presented follows a sequential strategy or another type of adaptive strategy. There is a selectable alphanumeric and/or high resolution graphic display of test items on the screen.

Examinee's answers are input by a simple, easy-to-handle keyboard and are automatically checked for correct, incorrect, or illegal response. Response times for each item are recorded and integrated into the set of test data. There is selectable feedback of correct/incorrect responses and elapsed item or test time.

Any latency between the presentation of successive items is eliminated by the "advanced processing concept." All jobs of computation, item selection, and picture processing for the next item are performed during the presentation of the actual item. (There is a delay of 15 seconds maximally for the preprocessing of graphical items, when the button "next item" is pressed continually without reading the item and solving the problem).

Apparatus tests--like reaction time measurement, vigilance tests, or tracking tests--may be integrated into the test battery and controlled by the computer system (real-time execution). The test data are stored on disk after each subtest during the test session, thus avoiding loss of data due to power outage. Test data are integrated into a structured test data bank after each session for further test or item analysis procedures.

Physiological Section

In connection with the Zak-A/D interface, the TEST 2000 system allows on-

line recording and analysis of up to six different physiological input channels. Long time off-line recording and analysis of psychophysiological data are accomplished by the pocket-sized, accu-driven microprocessor system BIOPORT, which interfaces directly to the TEST 2000 system. The BIOPORT system is especially suited for concurrent measurement of psychophysiological variables during test or training situations, thus revealing the examinee's reactions and resistance to stress-inducing factors. At present, this section is not implemented in the FAF; however, acquisition is planned for pilot selection and evaluation procedures.

Problems

There are currently problems in two areas. The first problem concerns the data transfer from the microcomputer to a large sized computer for further calculations. At present the data (scores and latencies/time for each subtest) are printed and then manually input to a larger computer. This problem will be solved in the fall of 1983 by using a tape drive and magnetic tape for the data transfer from the microcomputer at Munich to the large computer in Bonn. The second problem is that graphic items require a great deal of effort for programmers, since they are written in Assembler language in connection with a graphic preprocessor. In the future, video disks will be used to store the graphic items; these items will then be monitored by a microcomputer.

Test Material and Software

Aptitude Classification Battery

The first empirical phase started with the Aptitude Classification Battery (EVT), which is the standard test battery for entrance into the Federal Armed Forces and is quite similar to the ASVAB. At present six subtests of the EVT (the subtests without graphic items) are on the computer in two parallel forms (see Table 1). Further, four subtests that are not on the computer (e.g., radio test, test for reaction rate) measure special aspects and are speeded tests.

All six tests on the computer have time limits; within these limits, all items without an answer are presented again. Testing time varies between 45 and 75 minutes, since some examinees need more time for the sample items, whereas others do not use the entire time allocated for a subtest. It is practical to have, on an average, eight persons each day at the test station.

In the first empirical phase, volunteers who have had experience with the EVT battery have been tested. These persons took the EVT as draftees at the recruiting center. When they become volunteer-carriers, they take the EVT once more using a computer-administered parallel form. The next samples of testees will consist of draftees and persons without experience with the EVT and will be grouped according to education level. Furthermore, a small sample will be tested with the two ways of item presentation on the screen (green vs. black).

Software

The software is stored on the system disk. The disk contains, for example,

Table 1
The Six EVT Subtests Used for the FAF Pilot Project

Test	No. of Items/Alternatives	Time
Word Analogy Test (WAT)	20 items/5 alternatives	4 1/2 minutes
Figure Reasoning Test (FDT)	20 items/8 alternatives	10 minutes
Arithmetic Reasoning Test (RT)	20 items/input of the results	14 minutes (paper-and-pencil for notes)
Spelling (Orthographical Test; RST)	50 items/correct-incorrect	3 minutes
Mechanical Ability Test (MT)	20 items/5 alternatives	13 minutes
Electrotechnical Comprehension Test (EKT)	20 items/4 alternatives	10 minutes

the following programs (with content requested by a system command):

ZAKDOS: Operating system of TEST 2000.
EDIT: Editor for input and data management.
BASIC: Interpreter for BASIC (FORTRAN will soon be available).
GRAPHIC: Preprocessor for the graphic items (in Assembler).
SEQUEN: Management of the items with item text and alternatives (e.g., storage on disk and deletion of the blanks).
EVTA1, EVTA2, etc.: Item text and alternatives for the EVT.
EVTB1, EVTB3, etc.: The nongraphic subtests in two parallel forms.
EVT: Management of the EVT test session (creation of the data file, common instructions for EVT, item presentation--conventional/sequential). The item material is stored separately from the presentation procedure, so the system is very flexible.
TESBAS: Users' program in BASIC for handling the data/results, e.g., output on screen or printer, with results for each item (answer, scoring, latency/response time) or for the six subtests (scores, time used).

Results

Empirical Results

Table 2 shows the results calculated from the data of the first 208 examinees. These persons had first taken the EVT by paper and pencil (some weeks or months earlier) and then had taken the EVT by computer at the recruiting center for volunteers in Munich.

The comparison using the t test for dependent samples shows significant mean differences except for Arithmetic. The verbal subtests Word Analogy and Spelling had higher scores using paper and pencil and the subtests with figural items (Figure Reasoning, Mechanical, Electrotechnical) had higher scores using

Table 2
Scores from the Six Subtests of the EVT
for the Computerized Adaptive Test
and Paper-and-Pencil Test

Subtest	CAT	Paper and Pencil	t	p	r
Word Analogy	14.56	15.13	-2.86	.005	.606
Figure Reasoning	16.50	15.65	4.81	.000	.703
Arithmetic	12.01	11.70	1.59	.114	.785
Mechanical	13.37	12.58	4.45	.000	.734
Spelling	31.39	34.00	-5.47	.000	.772
Electrotechnical	7.19	6.14	4.11	.000	.682

computerized testing. The correlation coefficients varied between .60 and .79. Further analysis of these data and collection of data are in progress.

Other Results

There have been no problems handling the testing station by the psychologist, subsequent to a one-day training session by Zak. For example, the psychologist, after a one-day training session by Zak, starts the test session prepares the EVT allocating a data file (for data security, code numbers for the examinees are stored on the disk only), and prints the results. An additional aspect of the training was to reduce anxiety/reservation towards the computer and to familiarize the psychologists with the fact that the computer gives some assistance and support.

The examinees have not had any problems handling the test computer using the special keyboard. Before starting the session there is a motivation phase during which the psychologist gives an introduction to the computerized testing session. Then, the examinee is seated in front of the CRT, reads all further instructions on the screen, and uses the keyboard (paper and pencil for the Arithmetic test only for calculations). Questions are rare, except concerning the brightness of the screen.

The volunteers respond positively to this individualized way of testing. The difference between the fastest and slowest examinee is about 30 minutes. There is enough time allocated so that examinee can reread the instructions until they have understood how to solve the items. On the other hand, some examinees do not need the entire time allocated for a subtest and go on to the next subtest when they have answered all items. It is therefore possible to save some testing time with this individualized testing procedure.

There is positive response to feedback from the computer. After each item, the examinee presses the button "next item," and the computer reacts and presents the next item. There is also positive response to the special testing

situation. During the paper-and-pencil session, about 50 persons are seated in one room, and the tests are carried out in groups. During the computerized testing session there are only four examinees at the test computer, with the psychologist seated nearby for motivational purposes at the beginning of the testing session and for answering questions during the course of the test.

Future Developments

Hardware

At the end of October 1982 a major change in the hardware will occur. There will be an integrated testing station, which will consist of a microcomputer TEST 2000 for monitoring the testing station, with floppy disks, tape drive, and line printer, and about five testing terminals with a special screen for graphic items and moving pictures (especially developed for tracking items), 64 KByte RAM, and the special keyboard for the examinee. Two testing places will have amber screens, so that it will be possible to compare different forms of item presentation. The system is planned to connect eight testing stations via an interface with TEST 2000. The host computer loads a subtest to the RAM of the testing location, and the examinee then takes this subtest on his/her terminal. Following a subtest the results are stored on the disk, and the next subtest is loaded.

Second Empirical Phase

After the first empirical phase is terminated, the second empirical phase is scheduled for November 1982. Following the conventional testing phase, adaptive testing with fixed branching strategies (pyramidal testing) will commence. Three tests have thus been prepared: Word Analogy, Number Series, and Spelling, each consisting of 66 items and 11 stages. The items were prepared and selected in a pilot study, with the branching procedures developed using classical testing theory and, additionally, Mokken's (1971) probabilistic approach for the check of unidimensionality/homogeneity (Nauels & Wildgrube, 1981). There will thus be an opportunity to compare conventional results (collected either by paper and pencil or by computer) with data collected either by computerized adaptive strategies for the Word Analogy and Spelling tests.

Third Empirical Phase

In the third empirical phase, evaluation of variable branching strategies in adaptive testing are planned. It is still in an uncertain stage, however, because the FAF is searching for testing procedures that can be used in the daily selection process and that, furthermore, can be understood by the test users.

Parallel with starting the second phase of pyramidal testing, the first empirical approach (collection of data) to use the response time of each item (latency) as an additional ability parameter will be carried out (Birke, 1981).

Other Research

In June 1982 research was begun with Lutz Hornke, University of

Duesseldorf. Small projects are being undertaken (1) to investigate the "human factor" aspects of computerized testing (e.g., man-machine interaction, ergonomic aspects such as different screen colors, influence of different instructions for the test procedure) and (2) to compare these results with conventional paper-and-pencil testing. The item pool will be prepared for figural items (similar to the Figure Reasoning test) using logical rules for item construction to vary item difficulty based on Embretson's (1983) work. Collection of data for these items is planned in conventional paper-and-pencil format at the beginning of 1983.

REFERENCES

Birke, W. Item response time as a basis for ability and difficulty measures. Paper presented at the 23rd annual conference of the Military Testing Association, Arlington VA, October 1981.

Embretson, S. E. Component latent trait models for test design. In D. J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory/Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1983.

Mokken, R. J. A theory and procedure of scale analysis with applications in political research. New York: deGruijter/Berlin: Mouton, 1971.

Nauels, H.-U., & Wildgrube, W. Probleme des Itempools beim adaptiven Testen—Pilotstudie zum CAT. Zeitschrift f$r Differentielle und Diagnostische Psychologie, 1981, 4, 303-323.

Wildgrube, W. Computerized testing in the Federal Armed Forces. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, Psychometric Methods, 1980.