Behavior Research Methods & Instrumentation
1981, Vol. 13 (4), 430-433

Microcomputer-aided assessment using standard psychometric procedures

J. GRAHAM BEAUMONT
University of Leicester, Leicester LE1 7RH, England

This paper describes the opportunities offered by the application of commercially available, inexpensive microcomputers to computer-assisted administration of standard psychometric procedures. The rationale underlying this application and the requirements of software design are discussed. A project based upon the APPLE microcomputer is described, together with an example of the software being developed. Preliminary work and planned development and evaluation are reported.

The microprocessor revolution has brought rapid and significant change to psychology. Not only has the character of laboratory work in experimental psychology been transformed, a process that began with the introduction of minicomputers in the early 1970s, but other fields of psychological research have also begun to make use of the powerful new technologies available. Research is now in progress to see how these developments may contribute to the various aspects of applied psychology. This paper will explore one potential contribution, the use of inexpensive commercial microcomputers to assist in the administration of standardized psychometric procedures for the assessment of psychiatric and neurological states. The preliminary work of a project designed to evaluate this contribution is described.

There have been surprisingly few reports of the application of small computers to automated test administration. Undoubtedly, before the advent of the cheap and portable microcomputer, the relative cost of laboratory minicomputers and the problems of communicating with them over a distance discouraged the development of such applications. There have been, however, a few reports of the use of on-line computer facilities, besides those that have used off-line facilities for scoring and administration. Among the latter group, the application of a decision key approach to interpretation of the Halstead-Reitan Neuropsychological Battery (Russell, Neuringer, & Goldstein, 1970) was an important early development, and other systems for the interpretation of batteries of tests, including those for cognitive function, have been reported by Gilberstadt, Lushene, and Buegel (1976) and by Vincent (1980).

The major contribution to this field, although not centrally directed at cognitive assessment, has been the work of Johnson and Williams and their colleagues with the Psychiatric Assessment Unit (PAU). The system as first described used a large-scale computing facility and employed a specially designed terminal (Cole, Johnson, & Williams, 1975, 1976; Klingler, Miller, Johnson, & Williams, 1977), but it was later modified to operate around a PDP-11/V03 with standard CRT terminals (Johnson & Williams, 1978; Johnson, Giannetti, & Williams, 1978). Studies of the reliability and validity of this system (Johnson, Klingler, Giannetti, & Williams, 1980) are extremely encouraging.

Other on-line applications have included the Halstead Category Test (Beaumont, 1975), using a PDP LINC-8, and the Peabody Picture Vocabulary Test, using a special terminal linked to an IBM 1800 (Klinge & Rodziewicz, 1976). Both of these applications seem to be limited not by the inherent nature of the automated assessment, but by the relative cost and practical difficulties associated with the computer hardware.

The factor that has changed the picture so radically has been the advent of the inexpensive microcomputer, costing between $2,000 and $10,000, with a variety of inexpensive peripherals and a wealth of good software available. This development has been supported by the presence of the market for small business and professional users, as well as by the "hobbyist" owner of a "home computer." It is extremely fortunate that these machines, partly because of their adaptability, are well suited to the requirements of automated test administration. I am aware of only one report that describes the application of a small microcomputer to automated assessment, and that is in the area of interviewing rather than test administration (Bremser & Davidson, 1978), but there seem to be a very large number of small-scale projects currently in progress, which it is to be hoped will yield published reports in the near future. In the meantime, it is important that, rather than piecemeal development of software, there is a careful consideration of the problems associated with computer-assisted test administration, of the quality of the software that is to be developed, and a responsible approach to the introduction of good, reliable, and valid assessment procedures for distribution among clinical users.

This project was supported by the Department of Health and Social Security of the United Kingdom government.

0005-7878/81/040430-04$00.65/0

STANDARD TESTS OR NOVEL PROCEDURES?

Given that the principle of computer-aided assessment is accepted, the first question is whether it is sensible to attempt to automate the standard procedures already in use by clinicians or whether novel procedures more suitable to the system being employed should be developed. The introduction of new technology often demands the reappraisal of the overall design characteristics of a system. It has been said that the automobile has never recovered from the failure to reconsider the design of highway vehicles when the internal combustion engine was fitted to the horse-drawn carriage. While it would be mistaken to deny the validity of this principle (see, e.g., Cory, Rimland, & Bryson, 1977; Levander & Lagergren, 1973), it would be unreasonable not to explore the possibility of "converting" some standardized psychological tests to automated versions, along with the development of perhaps more appropriate new instruments.

If tests could be converted to a parallel automated form, there would be a considerable saving of time and resources, and these tests would be of utility in the period during which other tests might be developed. The automated form of a test still should be validated against the original form of the test, and its reliability should be independently assessed. Only if the automated version performs as if it is a parallel form of the original version of the test will the norms established for the original version be appropriate. If the forms cannot be considered parallel, then the norms are inapplicable, and the advantage that is conferred by starting with a standardized procedure is lost.

SOFTWARE DESIGN

There are a number of considerations that suggest that care should be exercised not only in the design of the actual software for a particular test, but also in the selection of a system and language for this application in general. Among these considerations are the appropriateness of the implementation, the ease of operation, the appearance of the system to the testee, and the stability of the system.

The appropriateness of the system refers to fairly obvious matters, such as whether the facilities required can be offered by the system. The implementation of a particular language must be reasonably efficient in terms of both speed and memory requirements, although this may not be the primary consideration. The system should offer facilities, both graphics and peripherals, appropriate to automated assessment. A system that cannot generate high-resolution graphics and does not offer the possibility of response modes other than keypresses is bound to be restricted in the tests that can be implemented. Not only must the system be capable of supporting the hardware for these facilities, but good software should also be available in order to control it.

Ease of operation is the second consideration. Software in this application must be designed not only for the testee but also for the test administrator, who may well be unfamiliar with microcomputers. The system should not be more difficult to operate than the average tape recorder. There are a number of ways in which this can be achieved, but given a language in which it is possible to write highly protected software (see, e.g., Kernighan & Plauger, 1976), the test administrator should be required to do no more than insert a floppy disk or cassette, turn the power on, and then respond to an initial dialogue with the machine to select a test and then establish the parameters for a particular test session. The software must guard against possible input/output errors, maintain control within the program, and instruct the administrator on action to be undertaken in the event of errors being detected.

The appearance of the system to the testee is also of considerable importance. Above all, the system should be "friendly." However, there is a difficult balance to be struck between a system that is cold and impersonal and one that is falsely human. Programs that constantly employ the testee's first name and offer encouragement such as "Right on there" or "Way to go" are not generally appropriate to most adult populations. The system should offer instructions and feedback in clear, simple language that is neither too involved nor too unnecessarily personal.

Another aspect of the appearance of the system to the testee is the form of response demanded. It is likely to be disconcerting for a testee either to feel uncertain of the outcome of actions in terms of the machine's response or to feel that actions may result in damage to the system itself. Whether the testee feels confident in interacting with the system and feels a degree of control over the progress of the test is highly dependent upon the software employed. This confidence can be encouraged by introducing the testee to the system and its behavior appropriately (through the initial dialogue) and by protecting the system against "incorrect" actions by the testee. The software must consider any conceivable action on the part of the testee and respond appropriately. Guidance should be available within the program to guide the testee's interaction with the system.

The stability of the system is implicit in much that has been stated above. That is, the programs must be designed so that they are failsafe. No input/output failure and no response or combination of responses by the testee must cause the program to crash. Given an appropriate language and well designed software, it is possible to achieve this without difficulty, and it confers advantages both in the reliability of the system and in the character of the interaction between the system and the test subject.

THE DHSS/LEICESTER PROJECT

The United Kingdom Department of Health and Social Security (DHSS) has recently funded a project at

the University of Leicester to investigate the feasibility of employing an inexpensive microcomputer to administer standardized psychometric procedures. The aims of the project are to : (1) investigate the appropriateness of the system selected, (2) develop software for a number of tests that will embody the principles outlined above, and (3) conduct a preliminary investigation of the performance of the automated as opposed to the conventional forms of the tests.

The hardware selected was the Apple II Plus microcomputer with 48K of memory, twin minifloppy-disk drives, the Apple language card, clock, graphics tablet, light pen, and Heuristics Speechlab. The system also contains the keyboard filter, PAL card (for European color televisions), and a serial interface board. This system was selected for a number of reasons. Good high-resolution (280 by 192) graphics, in color, supported by good software, are available, as well as a number of easily interfaced peripheral devices, again accompanied by good software. While the processor chip on which the system is based (6502) is slow and already seems rather outmoded, the performance of the system is adequate to this application and is more than compensated for by the amount of software available for the Apple. Most important, the Apple supports PASCAL, a good language for applications of this nature.

The reasons for using this system are probably quite clear. The system is highly portable and reliable. It can be easily transported and plugged into any color receiver. It is easy to operate and relatively inexpensive. Given that the system must be cheap enough for multiple machines to be available in each clinic, a total figure of around $5,000 for each system seems reasonable. This is especially true when the system can also act as a word processor, maintain records, undertake resource accounting, perform statistical tests, and assist in teaching and in clinical research.

We decided to avoid the use of specialized keyboards and response terminals. Response panels made in small numbers cannot match the reliability of the keyboard and other devices offered by most manufacturers. It seemed preferable to us to control and limit the use of keys on the keyboard by software and to investigate the use of other standard peripheral response devices, rather than to employ custom-built keyboards that are a feature of a number of other projects.

It is worth mentioning that although the project is at an early stage, we are delighted with the performance of the Apple in this context, although some questions remain about the practicability of using the PAL conversion card and any handy domestic color receiver. The color graphics quality is markedly inferior to that obtained with a color monitor or a receiver adapted for a direct RGB color drive, but it may still be adequate and is the subject of continuing evaluation. The Apple graphics tablet, although an aid to software development rather than a primary response medium, is truly excellent. The Heuristics Speechlab, although impressive, is probably not sufficiently reliable for clinical application. The light pen works better than might have been expected and is under consideration as an alternative response medium to the keyboard. The use of light pens by nonexperts is still an uncertain area and is worthy of further evaluation.

While there is a program of tests for which software is to be developed, of increasing sophistication in the graphics, responses, and scoring required, the first test completed will serve as an example of the principles to be embodied in each. This test is the Mill Hill Vocabulary Test (Set B). The program is written in PASCAL.

If the examiner inserts his floppy disk and turns on the power, the program runs automatically. Initial announcements are made, the administrator declares a password that will prevent the testee's gaining access to the test results, and test-item data are called from disk to set up the test. The dialogue is conducted through the television screen and the Apple keyboard. For each response (say, typing <space> to advance one page, or entering the six-character password), the program will only accept valid input from the keyboard. Typing of any kind of unexpected response results not in the program's failing, as in many languages, but in the Apple's bleeping and ignoring the input. The program is thus protected, and the respondent is guided into making the responses that the system expects. Similarly, while the program calls the test-item data from disk, input/output checking is suspended, so that if an error is detected (perhaps the disk has been inadvertently removed or replaced or the drive door has been opened), the system does not return to the system monitor but remains within the program, prints appropriate (and comprehensible) diagnostic messages to the administrator, and describes remedial action.

Once the test is set up, the testee is seated in front of the console for administration of the test. Instructions are given to the subject, close in form to the paper-and-pencil version, and, following guidance, the first problem is presented. This problem is repeated, with appropriate messages, until the testee gives the correct solution. The remaining 33 problems follow. For each, the problem is presented in a format parallel to the printed form, and the keyboard is "locked" to the numbers 1 to 6 (for possible responses) and the letters L and S. Pressing L allows the testee to leave or skip any item, and pressing S allows him to stop the test. Following presentation of the last unanswered item (Item 34 on first pass), the program presents again the items that have been skipped (in order of difficulty). This process is repeated until all items are answered or the test is stopped. The testee may then inspect all the responses made, change any response made, or add responses for items that were originally skipped. Opportunities to recheck the answers recur until the testee decides to stop checking. The program then bids the testee a polite "goodbye." By this approach, it is hoped to mirror the flexibility available in the conventional administration, to vary the order in

which items are attempted, and then to return to check over items and to make changes to previous responses. Preliminary tests indicate that the program is successful in achieving this, as well as being adequately stable.

At this point, the administrator is invited to enter the password (and given four chances to get it right), and the system responds with the test score, the total expected test score, the expected difficulty percentile level achieved, and, on entering the testee's age, the grade obtained, the verbal label associated with the grade, the percentile range in which the score falls, and the percentage of psychiatric patients scoring as highly.

DISCUSSION

I hope to continue developing software that fulfills the general requirements outlined above and intend to proceed not only to other verbal tests with multiple-choice responses, but also to tests that involve graphical stimuli (the Progressive Matrices Test, for example), as well as those that involve more complex responses. There are even plans to program the Wisconsin Card Sorting Test with response by the light pen, and the Block Design subtest of the WAIS, again using the light pen to manipulate images of the blocks. It may even be possible to develc an algorithm to undertake adequate scoring of the WAIS vocabulary subtest, but this is further in the future.

For the present, software development on less ambitious projects is continuing, and the essential empirical investigation of the performance of the automated tests with reference to standard administration is about to begin. This will initially be with normal subjects, the results of which should be available by the end of this year, and will later be extended to clinical groups.

We also hope to see increasing interest in the precise nature of the interaction between testees and automated systems and in the possibilities offered by computer-assisted test administration. These possibilities, which extend well beyond replacement of the human administrator, into such areas as test programs that can automatically adjust their norms and discrimination criteria on the continual receipt of feedback information, have hardly begun to be explored.

REFERENCES

BEAUMONT, J. G. The validity of the Category Test administered by on-line computer. *Journal of Clinical Psychology*, 1975, **31**, 458-462.

BREMSER, R. F., & DAVIDSON, R. S. Microprocessor-assisted assessment in the clinical research laboratory. *Behavior Research Methods & Instrumentation*, 1978, **10**, 582-584.

COLE, E. B., JOHNSON, J. H., & WILLIAMS, T. A. Design considerations for an on-line computer system for automated psychiatric assessment. *Behavior Research Methods & Instrumentation*, 1975, **7**, 195-198.

COLE, E. B., JOHNSON, J. H., & WILLIAMS, T. A. When psychiatric patients interact with computer terminals: Problems and solutions. *Behavior Research Methods & Instrumentation*, 1976, **8**, 92-94.

CORY, C. H., RIMLAND, B., & BRYSON, R. A. Using computerized tests to measure new dimensions of abilities: An exploratory study. *Applied Psychological Measurement*, 1977, **1**, 101-110.

GILBERSTADT, H., LUSHENE, R., & BUEGEL, B. Automated assessment of intelligence: the TAPAC test battery and computerized report writing. *Perceptual and Motor Skills*, 1976, **43**, 627-635.

JOHNSON, J. H., GIANNETTI, R. A., & WILLIAMS, T. A. A self-contained microcomputer system for psychological testing. *Behavior Research Methods & Instrumentation*, 1978, **10**, 579-581.

JOHNSON, J. H., KLINGLER, D. E., GIANNETTI, R. A., & WILLIAMS, T. A. The reliability of diagnoses by technician, computer and algorithm. *Journal of Clinical Psychology*, 1980, **36**, 447-451.

JOHNSON, J. H., & WILLIAMS, T. A. Clinical testing and assessment: Using a microcomputer for on-line psychiatric assessment. *Behavior Research Methods & Instrumentation*, 1978, **10**, 576-578.

KERNIGHAN, B. W., & PLAUGER, P. J. *Software tools*. Reading, Mass: Addison-Wesley, 1976.

KLINGE, V., & RODZIEWICZ, T. Automated and manual testing of the Peabody Vocabulary Test on a psychiatric adolescent population. *International Journal of Man-Machine Studies*, 1976, **8**, 243-246.

KLINGLER, D. E., MILLER, D. A., JOHNSON, J. H., & WILLIAMS, T. A. Process evaluation of an on-line computer assessment unit for intake assessment of mental health patients. *Behavior Research Methods & Instrumentation*, 1977, **9**, 110-116.

LEVANDER, S. E., & LAGERGREN, K. Four vigilance indicators for use with a minicomputer. *Reports of the Psychological Laboratories of the University of Stockholm*, No. 381, 1973.

RUSSELL, E. W., NEURINGER, C., & GOLDSTEIN, G. *Assessment of brain damage: A neuropsychological key approach*. New York: Wiley Interscience, 1970.

SPACE, L. G. A console for the interactive on-line administration of psychological tests. *Behavior Research Methods & Instrumentation*, 1975, **7**, 191-193.

VINCENT, K. R. Semi-automated full battery. *Journal of Clinical Psychology*, 1980, **36**, 437-446.

NOTE

1. This program is designed as a demonstration only, as permission has not been obtained from the test publishers.