DOCUMENT RESUME

ED 346 662 EC 301 274

AUTHOR Engdahl, Brian
TITLE Computerized Adaptive Assessment of Cognitive Abilities among Disabled Adults.
PUB DATE 17 Aug 91
NOTE 19p.; Paper presented at the Annual Meeting of the American Psychological Association (San Francisco, CA, August 16-20, 1991).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Adults; Aptitude Tests; *Cognitive Tests; *Computer Assisted Testing; Head Injuries; Language Skills; Males; *Mental Disorders; Neurological Impairments; Participant Satisfaction; Spatial Ability; Substance Abuse; *Test Format; Testing

ABSTRACT

This study examined computerized adaptive testing and cognitive ability testing of adults with cognitive disabilities. Adult subjects (N=250) were given computerized tests on language usage and space relations in one of three administration conditions: paper and pencil, fixed length computer adaptive, and variable length computer adaptive. Subjects were classified into primary disability categories: medical, mentally ill, chemically dependent, brain injury, and no disability. Forty percent of subjects had multiple diagnoses, half of them with both chemical dependency and mental illness. Only three percent were female. Ages ranged from 20 to 76 years. Subjects taking the computerized forms perceived the tests to be easier, faster, more easily read, and more enjoyable than those taking the paper and pencil tests. Test time was shortest under the variable length condition. The mentally ill subjects took longer to complete computerized testing than other subjects. There were no differences in subject satisfaction with the test as a function of ability. Two factors emerged from factor analysis, the first comprising verbal abilities, math and language skills, recent and remote memory, and freedom from distractibility, and the second comprising perceptual abilities, abilities to process nonverbal materials, and psychomotor skills. Findings suggested that subjects performed somewhat better on the computerized version and that clinicians were less accepting of computerized assessment than were patients. (DB)

Computerized adaptive assessment of cognitive abilities among disabled adults

Brian Engdahl, PhD
Psychology Service
VA Medical Center
Minneapolis, MN 55417
612-725-2073

Division 5 Session Title: Introduction to computer-based adaptive testing: Applications to psychological assessment.

American Psychological Association Annual Convention

San Francisco, August 17, 1991

This project was funded by the U.S. Department of Veterans Affairs Health Services Research and Development Service under Project IR #88-114.B. Raina Eberly, PhD of the Minneapolis VA Medical Center was the co-principal investigator. We wish to thank Dr. James McBride of the Psychological Corporation, Dr. David Weiss of the University of Minnesota, Sara Madden, Sheri Locken, Cheryl Winch, John Bielinski, John Kline, and our subjects for their invaluable assistance.

Large numbers of disabled unemployed adults complete cognitive ability tests every year. More than 2000 were tested during 1990 at the Minneapolis VA alone. These assessments themselves consume substantial health care dollars and in turn, they play key roles in the planning and delivery of costly medical and vocational rehabilitation services.

Current cognitive ability testing takes two forms: 1. Brief intellectual screening to estimate IQ, that is typified by the Shipley Hartford Institute of Living Scale. It is given by clerical staff, often to groups of patients in a paper-and-pencil format, and requires about 30 minutes.

2. Extended cognitive ability assessment is exemplified by the Wechsler Adult Intelligence Scale, requires a individual administration by a trained clinician, and requires about 1 1/2 hours per patient. Multifactor aptitude tests such as the General Aptitude Test Battery (or GATB), and the Differential Aptitude Tests (or DAT) also are used to assess cognitive abilities.

The emerging third form of ability testing, Computerized Adaptive Testing (or CAT) was just described by Dr. Green. Research has shown that CAT tests perform satisfactorily and lead to high subject satisfaction among high school and college students, and military recruits. Other research has shown computerized tests to be <u>feasible</u> with deteriorated geriatric patients, although the performance and acceptability of such tests is not known.

The Psychological Corporation's DAT is an effective tool in the assessment of disabled adults. Its subtests index a range of abilities that predict performance in rehabilitation programs. The CAT version of the DAT (or CAT-DAT) is the first commercially available adaptive ability test. It has been shown to be highly similar to the DAT among high school students.

Dr. Raina Eberly and I examined CAT-based assessment's performance relative to currently used tests, and CAT's acceptability to patients and clinicians. We were particularly concerned about these issues because our population differs from others reported in the CAT literature in its high proportions of disabled, elderly, medically ill, and psychiatrically impaired patients.

Most of our subjects completed two DAT subtests in a counterbalanced order: Language Usage and Space Relations, selected because of their high correlations with our standard IQ estimating test, the Shipley, with its Verbal and Abstract scores. Subjects were randomly assigned to one of three administration conditions: Paper & Pencil, Fixed Length computer adaptive, and Variable Length computer adaptive.

Tests with variable item numbers (rather than the standard fixed item number) have not received much study. In principle, most termination criteria used in variable length tests should result in shorter tests that are as accurate as fixed length tests. For this project, an option for the Variable Length condition was programmed allowing the specification of a termination rule for each of the CAT-DAT's seven power subtests. To estimate an individual's "true score" to an accuracy of approximately $\pm$ 1 standard error of measurement, we specified posterior variance values of .06 for Space Relations and .10 for Language Usage. We derived these figures by subtracting (from 1.00) previously reported DAT subtest reliabilities. Relative to the commercial CAT-DAT, our CAT-DAT subtests incorporated expanded item pools and a three-parameter logistic model for estimating ability levels rather than a one-parameter model. Our CAT-DATs matched the commercial CAT-DAT in item format, introductory and support materials, and test result printouts.

The CAT-DATs were given at an IBM AT-compatible computer equipped with a high resolution 14" green monochrome monitor. Typing skills were not required, as the test administrator entered the subject's name and background information, and the subject needed only to enter responses on one of two keys (the

"Y" or "N" keys). Paper & Pencil DATs were given at desks under standard time limits (maximum allowable time = 45 minutes). Although we expected some would have difficulties completing the tests, all subjects successfully completed them, despite clear indications from their MMPIs that they were experiencing significant emotional turmoil at their time of testing.

Subjects were classified into primary disability categories: medical, mentally ill, chemically dependent, brain injury, and no disability. Nearly 40% had multiple presenting diagnoses, half of them with a combination of chemical dependency and mental illness, and another 30% with mental illness and chronic medical problems. Three percent were female. Our subjects ranged from 20 to 76 years of age, with a mean of 42.3. They had completed 7 to 18 years of education, with a mean of 12.9 years. A group of 302 non-volunteers was compared to the experimental subjects and found to be similar in demographics, diagnoses, and other test scores.

Subject satisfaction questionnaires were based on previous research and allowed direct comparisons across the three DAT modes. As Table 1 shows, both CAT-DAT modes were characterized by greater overall subject satisfaction than the paper & pencil mode. There was a trend for psychiatric patients to be less satisfied with all three modes of testing relative to the other groups. Satisfaction and age correlated .24 (older subjects reporting more satisfaction), satisfaction and education correlated-.10 (less educated subjects tending toward less satisfaction).

CAT-DAT subjects perceived their tests to be easier, faster, more easily read, and more enjoyable than paper & pencil subjects. They perceived little time pressure and would strongly prefer a computer-administered test over a paper & pencil test. Mixed findings included a perception by CAT-DAT subjects that corrections were somewhat difficult to make and being somewhat bothered by this. Frequencies of complaints about reading the test questions were low. One to four percent thought there was

Not enough space between the lines.	(1.3%)
Difficulty in reading the type of lettering.	(3.9%)
Layout of the questions on the page.	(3.5%)
Size of letters was too small.	(2.6%)
Too much glare on the screen/pages.	(3.9%)

Nearly 10% commented that the pictures in the test booklet or the screen for the three subtests with graphics were hard to understand. Spontaneous comments were made about the relatively coarse resolution of the CAT-DAT's graphics. We were informed that these are being improved through an upgrade in resolution level in the pending CAT-DAT revision. The overall rates of these complaints (including the open-ended response option) were relatively low, and they did not vary significantly across type of test.

Table 1

Overall Subject Satisfaction by Test Type and Medical Diagnosis

Test Type	Medical Diagnosis: Acute Medic.	Chron. Medic.	Mental Ill	Chem. Dep.	Brain Injury	None	Row Mean
Variable Length	(0)	15.25 (4)	10.33 (52)	12.55 (53)	(0)	8.00 (1)	11.55 a
Fixed Length	(0)	12.00 (1)	11.85 (13)	12.04 (25)	10.00 (1)	10.00 (1)	11.88 b
Paper and Pencil	(0)	5.25 (4)	9.06 (50)	9.60 (42)	5.50 (2)	18.00 (1)	9.15 ab
Column Mean		10.44	9.95	11.41	7.00	12.00	(n=250)

"a" and "b" denote pairs of groups significantly different at $p \leq .05$. N's are shown in parentheses.

All 22 Psychology staff received reports based on their patient's test performance and all were asked to rate their satisfaction with the information provided by the testing. Their overall satisfaction (defined as Items 1 + 2 + 3 + 4 - 5; in Table 2) did not vary significantly across type of test, as the information yield from the three modes was equivalent, and clinicians seldom receive feedback from their patients about testing conditions.

Table 2

Clinician Satisfaction with DAT Test Results at the Item Level

	Agree 1	2	3	4	Disagree 5	sd
1. I found this information useful in my report preparation.		----2.91---				1.43
2. I found this information useful in team discussions.	1	2	3	4	5	
			---3.09---			1.40
3. It complements the information provided by other tests.	1	2	3	4	5	
		---2.45---				1.30
4. I would consider requesting that other patients complete the DAT.	1	2	3	4	5	
		---2.60---				1.48
5. It provides no new information.	1	2	3	4	5	
			---3.47---			1.33

n = 223 responses. The number flanked by dashes is the mean response, and the dashes indicate ± 1 sd.

Total test time actually required for all DAT modes was recorded. Table 3 shows the effects of test type and medical diagnosis on test time. Test time was shortest under the variable length condition, intermediate under the paper & pencil condition, and longest under the fixed length condition. The mentally ill subjects, on average, took longer to complete computerized testing than subjects in the other diagnostic categories. Nearly all subjects used the maximum allowable time within the paper & pencil mode. The interaction of test condition and medical diagnosis was non-significant.

Table 3
Subject Testing Time by Test Type and Medical Diagnosis

	Medical Diagnosis						
Test Type	Acute Medic.	Chron. Medic.	Mental Ill	Chem. Dep.	Brain Injury	None	Row Mean
Variable Length	(0)	33.00 (4)	38.54 (59)	33.70 (66)	(0)	30.00 (1)	35.85ab (130)
Fixed Length	(0)	40.00 (1)	59.28 (18)	45.44 (27)	42.00 (1)	42.00 (1)	50.38ac (48)
Paper and Pencil	(0)	45.00 (4)	44.89 (63)	42.00 (53)	41.00 (2)	45.00 (1)	43.60bc (123)
Mean	-- (0)	39.70 (9)	44.06a (140)	38.88a (146)	41.33 (3)	39.00 (3)	(N=301)

(a,b,c) denote pairs of groups significantly different at $p \leq .05$.

As one might expect, there was a modest negative correlation between Language Usage and Administration Time of -.21; lower ability subjects require a bit more help from the test administrator. Table 4 shows some scatter among average subtest scores, raising some concern over the equivalence of the test forms. Supplementary ANCOVA and regression analyses were conducted examining effects of Shipley score, years of education, age, previous computer experience, and anxiety level upon DAT scores. The few effects found were small, and suggested that the paper and pencil format and higher anxiety levels contributed to lower test scores, independent of Shipley estimated IQ, education, and age.

The average number of items attempted under the three test conditions varied, with variable length involving the fewest, fixed length an intermediate (and by definition fixed) number, and paper & pencil involving the most items, roughly double the number required under fixed length CAT conditions. Variable length subtests presented one-third fewer items than fixed length subtests. Additional analyses (not shown) contrasted the two orders of administration conditions (SR/LU vs. LU/SR) and revealed no significant order effects on subtest score, test time, or subject satisfaction.

Table 4
Study 1 Subtest Scores and Items Completed by Test Type

	Test Type				
	Variable Length	Fixed Length	Paper & Pencil	F	p
Space Relations Score	49.2	44.8	42.0	2.15	.119
Space Rel. Items attempted	20.5	30	46.1	238.69	<.001
Language Usage Score	64.8	60.7	56.0	3.16	.069
Lang. Usage Items attempted	18.3	25	47.3	1100.40	<.001
N:	120	48	127		

Since CAT presents subjects of lower ability with items that are not beyond their ability level, do they express more satisfaction with CAT than those of higher ability, as has been suggested? We looked for interactions between ability level and satisfaction. The sample was divided into thirds on Language Usage scores, creating a factorial design with type of test as a second factor, general satisfaction being the dependent variable. Simple inspection of the means failed to support this hypothesis, although our sample was short on lower scoring subjects. They were equally or somewhat less satisfied with their testing when compared to medium and higher scoring subjects.

Table 5

Subject Satisfaction by Ability Level and Mode of Administration

	Language Usage Level			
	Low	Medium	High	Total
Test Type				
Variable Length	11.54 (28)	11.60 (45)	11.20 (41)	11.44a (114)
Fixed Length	10.62 (13)	12.56 (18)	12.30 (10)	11.88b (41)
Paper and Pencil	7.38 (42)	10.28 (25)	10.65 (34)	9.20ab (101)
Total	9.29 (83)	11.42 (88)	11.11 (85)	(256)

(a,b) denotes pairs of groups significantly different at $p \leq .05$.

Because item characteristics make ability estimation more demanding at a distribution's extremes, we also expected that the

Variable Length program would administer more items to high ability subjects relative to those in the middle. Again, relative to DAT development samples, our group has relatively few low ability subjects. The top third of the Language Usage distribution completed significantly more items on average than those in the middle and lower thirds, completing approximately 21 items versus 17. This 3.5 item difference appears practically insignificant, and did not increase test completion time, nor decrease subjects' satisfaction. It might influence test time and satisfaction among very low scoring subjects.

We examined the convergent and discriminant validities of the CAT-DAT relative to the MMPI, Shipley, WAIS-R, and GATB. Those expressing more distress via their MMPI F score had lower DAT scores, and those scoring higher on the Masculinity-Femininity scale had higher DAT scores. As the Shipley often is used to estimate Full Scale WAIS-R IQ, it is interesting to note that its overall correlation with WAIS-R IQ was .65. The combination of DAT Language Usage and Space Relations correlated noticeably better (about .80) with Full Scale WAIS-R IQ.

Two factors emerged from a CAT-DAT/WAIS-R factor analysis: the first defined by verbal abilities, math and language skills, recent and remote memory, and freedom from distractibility. The second factor is defined by perceptual abilities, abilities to process non-verbal materials, and psychomotor skills.

Intercorrelations were high within three pairs of CAT-DAT and GATB subtests that share the same names: Verbal, Numeric, and Spatial (Space). A disappointingly low correlation (.53) was found between the two Clerical Speed and Accuracy subtests, which are virtually identical in content and format. This is consistent with clinical impressions we formed in reviewing test results with patients, where performance on the CAT-DAT Clerical test was observed to be highly variable relative to GATB Clerical performance. If a substantially higher score was observed, it was invariably on the GATB. We concluded that the conditions of CAT-DAT administration do not facilitate maximal performance from all subjects because it is the only timed test in the midst of seven pure power tests. Other research has suggested there are substantial differences between printed and computerized versions of the Clerical Speed and Accuracy test.

The CAT-DAT shares little variance with the eye-hand performance GATB subtests (Motor Speed, Finger Dexterity, and Manual Dexterity). These GATB subtests often provide unique variance to the prediction of vocational and training performance. This information has not been tapped by Paper and Pencil tests, although may be tapped by computerized testing currently under development.

Discussion:

CAT's implementation in the VA's more than 170 medical centers where psychological testing via computer is already established should pose no major problems from a computing resources standpoint. The computational demands of the item selection and presentation process is minimal (it is, for practical purposes, unnoticeable on an IBM-PC). From a software development standpoint, products suitable for presentation on personal computers are already available, and preparation of mainframe versions for inclusion in the VA's Mental Health Package could be accomplished without great effort. Staff time required to administer a few CAT tests will be about the same as that required to administer a few Paper & Pencil type tests. For complete multi-aptitude batteries, CAT's staff administrative time is less, and when CAT tests can be substituted for individually administered tests, such as the WAIS-R, the staff time saved can be substantial.

The user interface is a concern. As I noted earlier, the graphic images' quality should be improved. The sheer amount of time subjects spend at the keyboard could tie up computer resources to a significant degree in some settings. We use seven CRT terminals for patient testing, and we would need one more to provide CAT test to our referrals. Most Study 1 patients required 30 minutes or more to complete their two subtests. Testing time could be decreased, perhaps at the risk of decreasing user satisfaction, through greater emphasis on "working as quickly as possible", the incorporation of a time clock as part of the screen display, etc. This might also enhance the validity of the Clerical Speed subtest by promoting maximal performance. Both of these possibilities could be examined in future research.

CAT tests are increasingly commercially available in prepackaged or "build it yourself" systems. They are typically administered on an Apple or IBM-compatible personal computer and require access to a dot matrix printer to print copies of test results. Earmarking a PC for use by patients remains a concern for many sites. A two-floppy drive PC (or a single floppy drive PC with a hard disk) plus monochrome graphics capability is required. Such new equipment may cost less than $500, and used equipment of this type may be available at many sites, being steadily displaced by more advanced equipment. Scored results may be saved on a floppy disk and printed from another PC, or an Epson FX-compatible dot matrix printer with cable may be purchased for under $150 and dedicated to printing test results. As a bonus, such a system would allow the ever-increasing number of non-adaptive psychological tests to be administered and scored on site. The direct cost per CAT-DAT administration is $2.50 to $3.00, depending on the quantity of scorings purchased at one time. This fee is fixed, whether one subtest or all portions of the CAT-DAT are administered. A more sophisticated test-use metering system that would permit administration of partial batteries and reduce administration costs would be desirable.

Under the conditions of this study, CAT produced limited reductions in testing times, although subject satisfaction with the testing process was high. The near-zero rate of incomplete tests also suggests that CAT is a less frustrating method of testing. Further analyses of response patterns are planned to estimate what proportion of subjects exhibited near-random response patterns. We expect this proportion will be small, further bolstering the assertion that CAT is acceptable (and valid) to VA patients. We predict patients would be more accepting of additional testing, allowing test batteries that contain CAT tests to be longer. This would yield more information that is clinically useful than we now obtain.

Our subjects did not exhibit test time savings comparable to those shown by high school students and Dr. Green's military recruits. Much of the reason may be due to the varying contexts of test administration. It appears that students and recruits worked their way quickly through the tests, typically in groups taking the tests simultaneously, which may enhance their performance speed. They were in better physical and psychological health than our subjects. They likely perceived their test outcomes as in some way affecting their futures.

In contrast, our subjects were introduced to the tests without a suggestion that the results would affect decisions made about them, only that they would be helping us to improve methods of testing. They did not begin their tests with others, and in fact most often worked in solitude. Only the Paper & Pencil condition subjects were told they were working under a time limit. A future study that compares CAT testers under time limits to those free from time limits would clarify this uncertainty.

Our unsuccessful post-hoc attempts to account for the higher scores obtained under CAT conditions, combined with our clinical impressions leave us with a concern that our subjects perform somewhat "better" on the CAT-DAT relative to their performance on standard measures. The initial equating process used may be appropriate for high school students, but other equating methods, or additional developmental research may be needed when CAT-based tests are used in clinical settings such as ours with adults tested one at a time.

We concur with Butcher's assertion that those least accepting of computerized assessment are clinicians, not their patients. A few clinicians expressed puzzlement over the meaning and validity of DAT test results throughout our project's course, despite much effort on our part. They appear to be most comfortable with the tests they learned about in graduate school. The majority of our staff, however, were quite receptive and used this new information in their clinical work. This bodes well for acceptance of CAT in other settings. It is clear that patients readily accept computerized adaptive testing.

Computerized adaptive assessment of cognitive abilities among disabled adults

Brian Engdahl, PhD
Psychology Service
VA Medical Center
Minneapolis, MN 55417
612-725-2073

Division 5 Session Title: Introduction to computer-based adaptive testing: Applications to psychological assessment.

American Psychological Association Annual Convention

San Francisco, August 17, 1991

This project was funded by the U.S. Department of Veterans Affairs Health Services Research and Development Service under Project IR #88-114.B. Raina Eberly, PhD of the Minneapolis VA Medical Center was the co-principal investigator. We wish to thank Dr. James McBride of the Psychological Corporation, Dr. David Weiss of the University of Minnesota, Sara Madden, Sheri Locken, Cheryl Winch, John Bielinski, John Kline, and our subjects for their invaluable assistance.

Large numbers of disabled unemployed adults complete cognitive ability tests every year. More than 2000 were tested during 1990 at the Minneapolis VA alone. These assessments themselves consume substantial health care dollars and in turn, they play key roles in the planning and delivery of costly medical and vocational rehabilitation services.

Current cognitive ability testing takes two forms: 1. Brief intellectual screening to estimate IQ, that is typified by the Shipley Hartford Institute of Living Scale. It is given by clerical staff, often to groups of patients in a paper-and-pencil format, and requires about 30 minutes.

2. Extended cognitive ability assessment is exemplified by the Wechsler Adult Intelligence Scale, requires a individual administration by a trained clinician, and requires about 1 1/2 hours per patient. Multifactor aptitude tests such as the General Aptitude Test Battery (or GATB), and the Differential Aptitude Tests (or DAT) also are used to assess cognitive abilities.

The emerging third form of ability testing, Computerized Adaptive Testing (or CAT) was just described by Dr. Green. Research has shown that CAT tests perform satisfactorily and lead to high subject satisfaction among high school and college students, and military recruits. Other research has shown computerized tests to be feasible with deteriorated geriatric patients, although the performance and acceptability of such tests is not known.

The Psychological Corporation's DAT is an effective tool in the assessment of disabled adults. Its subtests index a range of abilities that predict performance in rehabilitation programs. The CAT version of the DAT (or CAT-DAT) is the first commercially available adaptive ability test. It has been shown to be highly similar to the DAT among high school students.

Dr. Raina Eberly and I examined CAT-based assessment's performance relative to currently used tests, and CAT's acceptability to patients and clinicians. We were particularly concerned about these issues because our population differs from others reported in the CAT literature in its high proportions of disabled, elderly, medically ill, and psychiatrically impaired patients.

Most of our subjects completed two DAT subtests in a counterbalanced order: Language Usage and Space Relations, selected because of their high correlations with our standard IQ estimating test, the Shipley, with its Verbal and Abstract scores. Subjects were randomly assigned to one of three administration conditions: Paper & Pencil, Fixed Length computer adaptive, and Variable Length computer adaptive.

Tests with variable item numbers (rather than the standard fixed item number) have not received much study. In principle, most termination criteria used in variable length tests should result in shorter tests that are as accurate as fixed length tests. For this project, an option for the Variable Length condition was programmed allowing the specification of a termination rule for each of the CAT-DAT's seven power subtests. To estimate an individual's "true score" to an accuracy of approximately $\pm$ 1 standard error of measurement, we specified posterior variance values of .06 for Space Relations and .10 for Language Usage. We derived these figures by subtracting (from 1.00) previously reported DAT subtest reliabilities. Relative to the commercial CAT-DAT, our CAT-DAT subtests incorporated expanded item pools and a three-parameter logistic model for estimating ability levels rather than a one-parameter model. Our CAT-DATs matched the commercial CAT-DAT in item format, introductory and support materials, and test result printouts.

The CAT-DATs were given at an IBM AT-compatible computer equipped with a high resolution 14" green monochrome monitor. Typing skills were not required, as the test administrator entered the subject's name and background information, and the subject needed only to enter responses on one of two keys (the

"Y" or "N" keys). Paper & Pencil DATs were given at desks under standard time limits (maximum allowable time = 45 minutes). Although we expected some would have difficulties completing the tests, all subjects successfully completed them, despite clear indications from their MMPIs that they were experiencing significant emotional turmoil at their time of testing.

ubjects were classified into primary disability categories: medical, mentally ill, chemically dependent, brain injury, and no disability. Nearly 40% had multiple presenting diagnoses, half of them with a combination of chemical dependency and mental illness, and another 30% with mental illness and chronic medical problems. Three percent were female. Our subjects ranged from 20 to 76 years of age, with a mean of 42.3. They had completed 7 to 18 years of education, with a mean of 12.9 years. A group of 302 non-volunteers was compared to the experimental subjects and found to be similar in demographics, diagnoses, and other test scores.

Subject satisfaction questionnaires were based on previous research and allowed direct comparisons across the three DAT modes. As Table 1 shows, both CAT-DAT modes were characterized by greater overall subject satisfaction than the paper & pencil mode. There was a trend for psychiatric patients to be less satisfied with all three modes of testing relative to the other groups. Satisfaction and age correlated .24 (older subjects reporting more satisfaction), satisfaction and education correlated-.10 (less educated subjects tending toward less satisfaction).

CAT-DAT subjects perceived their tests to be easier, faster, more easily read, and more enjoyable than paper & pencil subjects. They perceived little time pressure and would strongly prefer a computer-administered test over a paper & pencil test. Mixed findings included a perception by CAT-DAT subjects that corrections were somewhat difficult to make and being somewhat bothered by this. Frequencies of complaints about reading the test questions were low. One to four percent thought there was

Not enough space between the lines.	(1.3%)
Difficulty in reading the type of lettering.	(3.9%)
Layout of the questions on the page.	(3.5%)
Size of letters was too small.	(2.6%)
Too much glare on the screen/pages.	(3.9%)

Nearly 10% commented that the pictures in the test booklet or the screen for the three subtests with graphics were hard to understand. Spontaneous comments were made about the relatively coarse resolution of the CAT-DAT's graphics. We were informed that these are being improved through an upgrade in resolution level in the pending CAT-DAT revision. The overall rates of these complaints (including the open-ended response option) were relatively low, and they did not vary significantly across type of test.

Table 1

Overall Subject Satisfaction by Test Type and Medical Diagnosis

Test Type	Medical Diagnosis: Acute Medic.	Chron. Medic.	Mental Ill	Chem. Dep.	Brain Injury	None	Row Mean
Variable Length	(0)	15.25 (4)	10.33 (52)	12.55 (53)	(0)	8.00 (1)	11.55 a
Fixed Length	(0)	12.00 (1)	11.85 (13)	12.04 (25)	10.00 (1)	10.00 (1)	11.88 b
Paper and Pencil	(0)	5.25 (4)	9.06 (50)	9.60 (42)	5.50 (2)	18.00 (1)	9.15 ab
Column Mean		10.44	9.95	11.41	7.00	12.00	(n=250)

"a" and "b" denote pairs of groups significantly different at $p \leq .05$. N's are shown in parentheses.

All 22 Psychology staff received reports based on their patient's test performance and all were asked to rate their satisfaction with the information provided by the testing. Their overall satisfaction (defined as Items 1 + 2 + 3 + 4 - 5; in Table 2) did not vary significantly across type of test, as the information yield from the three modes was equivalent, and clinicians seldom receive feedback from their patients about testing conditions.

Table 2

Clinician Satisfaction with DAT Test Results at the Item Level

	Agree 1	2	3	4	Disagree 5	sd
1. I found this information useful in my report preparation.		----2.91---				1.43
2. I found this information useful in team discussions.			---3.09---			1.40
3. It complements the information provided by other tests.		---2.45---				1.30
4. I would consider requesting that other patients complete the DAT.		---2.60---				1.48
5. It provides no new information.			---3.47---			1.33

n = 223 responses. The number flanked by dashes is the mean response, and the dashes indicate ± 1 sd.

Total test time actually required for all DAT modes was recorded. Table 3 shows the effects of test type and medical diagnosis on test time. Test time was shortest under the variable length condition, intermediate under the paper & pencil condition, and longest under the fixed length condition. The mentally ill subjects, on average, took longer to complete computerized testing than subjects in the other diagnostic categories. Nearly all subjects used the maximum allowable time within the paper & pencil mode. The interaction of test condition and medical diagnosis was non-significant.

Table 3
Subject Testing Time by Test Type and Medical Diagnosis

Test Type	Medical Diagnosis: Acute Medic.	Chron. Medic.	Mental Ill	Chem. Dep.	Brain Injury	None	Row Mean
Variable Length	(0)	33.00 (4)	38.54 (59)	33.70 (66)	(0)	30.00 (1)	35.85ab (130)
Fixed Length	(0)	40.00 (1)	59.28 (18)	45.44 (27)	42.00 (1)	42.00 (1)	50.38ac (48)
Paper and Pencil	(0)	45.00 (4)	44.89 (63)	42.00 (53)	41.00 (2)	45.00 (1)	43.60bc (123)
Mean	-- (0)	39.70 (9)	44.06a (140)	38.88a (146)	41.33 (3)	39.00 (3)	(N=301)

(a,b,c) denote pairs of groups significantly different at $p \leq .05$.

As one might expect, there was a modest negative correlation between Language Usage and Administration Time of -.21; lower ability subjects require a bit more help from the test administrator. Table 4 shows some scatter among average subtest scores, raising some concern over the equivalence of the test forms. Supplementary ANCOVA and regression analyses were conducted examining effects of Shipley score, years of education, age, previous computer experience, and anxiety level upon DAT scores. The few effects found were small, and suggested that the paper and pencil format and higher anxiety levels contributed to lower test scores, independent of Shipley estimated IQ, education, and age.

The average number of items attempted under the three test conditions varied, with variable length involving the fewest, fixed length an intermediate (and by definition fixed) number, and paper & pencil involving the most items, roughly double the number required under fixed length CAT conditions. Variable length subtests presented one-third fewer items than fixed length subtests. Additional analyses (not shown) contrasted the two orders of administration conditions (SR/LU vs. LU/SR) and revealed no significant order effects on subtest score, test time, or subject satisfaction.

Table 4
Study 1 Subtest Scores and Items Completed by Test Type

	Test Type				
	Variable Length	Fixed Length	Paper & Pencil	F	p
Space Relations Score	49.2	44.8	42.0	2.15	.119
Space Rel. Items attempted	20.5	30	46.1	238.69	<.001
Language Usage Score	64.8	60.7	56.0	3.16	.069
Lang. Usage Items attempted	18.3	25	47.3	1100.40	<.001
N:	120	48	127		

Since CAT presents subjects of lower ability with items that are not beyond their ability level, do they express more satisfaction with CAT than those of higher ability, as has been suggested? We looked for interactions between ability level and satisfaction. The sample was divided into thirds on Language Usage scores, creating a factorial design with type of test as a second factor, general satisfaction being the dependent variable. Simple inspection of the means failed to support this hypothesis, although our sample was short on lower scoring subjects. They were equally or somewhat less satisfied with their testing when compared to medium and higher scoring subjects.

Table 5

Subject Satisfaction by Ability Level and Mode of Administration

	Language Usage Level			
	Low	Medium	High	Total
Test Type				
Variable Length	11.54 (28)	11.60 (45)	11.20 (41)	11.44a (114)
Fixed Length	10.62 (13)	12.56 (18)	12.30 (10)	11.88b (41)
Paper and Pencil	7.38 (42)	10.28 (25)	10.65 (34)	9.20ab (101)
Total	9.29 (83)	11.42 (88)	11.11 (85)	(256)

(a,b) denotes pairs of groups significantly different at $p \leq .05$.

Because item characteristics make ability estimation more demanding at a distribution's extremes, we also expected that the

Variable Length program would administer more items to high ability subjects relative to those in the middle. Again, relative to DAT development samples, our group has relatively few low ability subjects. The top third of the Language Usage distribution completed significantly more items on average than those in the middle and lower thirds, completing approximately 21 items versus 17. This 3.5 item difference appears practically insignificant, and did not increase test completion time, nor decrease subjects' satisfaction. It might influence test time and satisfaction among very low scoring subjects.

We examined the convergent and discriminant validities of the CAT-DAT relative to the MMPI, Shipley, WAIS-R, and GATB. Those expressing more distress via their MMPI F score had lower DAT scores, and those scoring higher on the Masculinity-Femininity scale had higher DAT scores. As the Shipley often is used to estimate Full Scale WAIS-R IQ, it is interesting to note that its overall correlation with WAIS-R IQ was .65. The combination of DAT Language Usage and Space Relations correlated noticeably better (about .80) with Full Scale WAIS-R IQ.

Two factors emerged from a CAT-DAT/WAIS-R factor analysis: the first defined by verbal abilities, math and language skills, recent and remote memory, and freedom from distractibility. The second factor is defined by perceptual abilities, abilities to process non-verbal materials, and psychomotor skills.

Intercorrelations were high within three pairs of CAT-DAT and GATB subtests that share the same names: Verbal, Numeric, and Spatial (Space). A disappointingly low correlation (.53) was found between the two Clerical Speed and Accuracy subtests, which are virtually identical in content and format. This is consistent with clinical impressions we formed in reviewing test results with patients, where performance on the CAT-DAT Clerical test was observed to be highly variable relative to GATB Clerical performance. If a substantially higher score was observed, it was invariably on the GATB. We concluded that the conditions of CAT-DAT administration do not facilitate maximal performance from all subjects because it is the only timed test in the midst of seven pure power tests. Other research has suggested there are substantial differences between printed and computerized versions of the Clerical Speed and Accuracy test.

The CAT-DAT shares little variance with the eye-hand performance GATB subtests (Motor Speed, Finger Dexterity, and Manual Dexterity). These GATB subtests often provide unique variance to the prediction of vocational and training performance. This information has not been tapped by Paper and Pencil tests, although may be tapped by computerized testing currently under development.

Discussion:

CAT's implementation in the VA's more than 170 medical centers where psychological testing via computer is already established should pose no major problems from a computing resources standpoint. The computational demands of the item selection and presentation process is minimal (it is, for practical purposes, unnoticeable on an IBM-PC). From a software development standpoint, products suitable for presentation on personal computers are already available, and preparation of mainframe versions for inclusion in the VA's Mental Health Package could be accomplished without great effort. Staff time required to administer a few CAT tests will be about the same as that required to administer a few Paper & Pencil type tests. For complete multi-aptitude batteries, CAT's staff administrative time is less, and when CAT tests can be substituted for individually administered tests, such as the WAIS-R, the staff time saved can be substantial.

The user interface is a concern. As I noted earlier, the graphic images' quality should be improved. The sheer amount of time subjects spend at the keyboard could tie up computer resources to a significant degree in some settings. We use seven CRT terminals for patient testing, and we would need one more to provide CAT test to our referrals. Most Study 1 patients required 30 minutes or more to complete their two subtests. Testing time could be decreased, perhaps at the risk of decreasing user satisfaction, through greater emphasis on "working as quickly as possible", the incorporation of a time clock as part of the screen display, etc. This might also enhance the validity of the Clerical Speed subtest by promoting maximal performance. Both of these possibilities could be examined in future research.

CAT tests are increasingly commercially available in prepackaged or "build it yourself" systems. They are typically administered on an Apple or IBM-compatible personal computer and require access to a dot matrix printer to print copies of test results. Earmarking a PC for use by patients remains a concern for many sites. A two-floppy drive PC (or a single floppy drive PC with a hard disk) plus monochrome graphics capability is required. Such new equipment may cost less than $500, and used equipment of this type may be available at many sites, being steadily displaced by more advanced equipment. Scored results may be saved on a floppy disk and printed from another PC, or an Epson FX-compatible dot matrix printer with cable may be purchased for under $150 and dedicated to printing test results. As a bonus, such a system would allow the ever-increasing number of non-adaptive psychological tests to be administered and scored on site. The direct cost per CAT-DAT administration is $2.50 to $3.00, depending on the quantity of scorings purchased at one time. This fee is fixed, whether one subtest or all portions of the CAT-DAT are administered. A more sophisticated test-use metering system that would permit administration of partial batteries and reduce administration costs would be desirable.

Under the conditions of this study, CAT produced limited reductions in testing times, although subject satisfaction with the testing process was high. The near-zero rate of incomplete tests also suggests that CAT is a less frustrating method of testing. Further analyses of response patterns are planned to estimate what proportion of subjects exhibited near-random response patterns. We expect this proportion will be small, further bolstering the assertion that CAT is acceptable (and valid) to VA patients. We predict patients would be more accepting of additional testing, allowing test batteries that contain CAT tests to be longer. This would yield more information that is clinically useful than we now obtain.

Our subjects did not exhibit test time savings comparable to those shown by high school students and Dr. Green's military recruits. Much of the reason may be due to the varying contexts of test administration. It appears that students and recruits worked their way quickly through the tests, typically in groups taking the tests simultaneously, which may enhance their performance speed. They were in better physical and psychological health than our subjects. They likely perceived their test outcomes as in some way affecting their futures.

In contrast, our subjects were introduced to the tests without a suggestion that the results would affect decisions made about them, only that they would be helping us to improve methods of testing. They did not begin their tests with others, and in fact most often worked in solitude. Only the Paper & Pencil condition subjects were told they were working under a time limit. A future study that compares CAT testers under time limits to those free from time limits would clarify this uncertainty.

Our unsuccessful post-hoc attempts to account for the higher scores obtained under CAT conditions, combined with our clinical impressions leave us with a concern that our subjects perform somewhat "better" on the CAT-DAT relative to their performance on standard measures. The initial equating process used may be appropriate for high school students, but other equating methods, or additional developmental research may be needed when CAT-based tests are used in clinical settings such as ours with adults tested one at a time.

We concur with Butcher's assertion that those least accepting of computerized assessment are clinicians, not their patients. A few clinicians expressed puzzlement over the meaning and validity of DAT test results throughout our project's course, despite much effort on our part. They appear to be most comfortable with the tests they learned about in graduate school. The majority of our staff, however, were quite receptive and used this new information in their clinical work. This bodes well for acceptance of CAT in other settings. It is clear that patients readily accept computerized adaptive testing