A RESEARCH SYNTHESIS

BETTY BERGSTROM
AMERICAN SOCIETY OF CLINICAL PATHOLOGISTS

ABILITY MEASURE EQUIVALENCE OF COMPUTER ADAPTIVE

A RESEARCH SYNTHESIS

A number of organizations are researching computer adaptive testing (CAT) as an alternative to existing pencil and paper multiple choice tests. If it can be shown that ability measures obtained with computer adaptive tests are statistically equivalent to ability meas

obtained with pencil and paper tests, CAT offers the advantage of shorter, more precise tests. While research has been conducted on the effects of computer administered testing (Bunde studies included in this paper. Studies are presented in publication date order.

In 1977 English, Reckase, and Patience published a study done with undergraduate students enrolled in a course entitled "Introduction to Educational Measurement and Evaluation" at the University of Missouri. Students were randomly assigned to four experimental groups and administered three achievement tests. Group 1 took two traditional pencil and paper tests, Group 2 took a pencil and paper test on the first exam and a computer adaptive test on the second exam, Group 3 took a computer adaptive test on the first exam and a pencil and paper exam on the second test and Group 4 took two computer adaptive tests. All students took a pencil and paper final exam as the third test. Items were calibrated with the Rasch model and the computerized testing algorithm used a logistic tailored testing procedure described in Reckase, 1974. The results of Groups 2 and 3 on tests 1 and 2 were used for the research synthesis. They are

Bejar and Weiss, in 1978, reported on achievement test results for students enrolled in a large introductory biology class at the University of Minnesota during the Fall and Winter quarters of the 1976-1977 school year. The study included comparisons of an adaptive test and a pencil and paper test for both the first mid-quarter and the second mid-quarter for the Fall and Winter groups. The measures from the first mid-quarter and the second mid-quarter tests are highly correlated so only the second mid-quarter tests were used in the research synthesis. In Table 1, the results of the Fall Group are reported as study 3 and the Winter group as study 4. The adaptive tests were administered by the stradaptive strategy (Weiss, 1973) and the 3 parameter logistic model was used to calibrate items and estimate achievement measures. The pencil and [illegible]

[illegible] adaptive test was administered anywhere between one day and three weeks after the pencil and [illegible] paper test and did not count toward course grades.

Mathematics application items from the California Assessment Program item banks were used to create tests in a pencil and paper administered format, a computer administered format, of the math and reading tests differed in that the math items were presented on one screen while

the reading test employed a scrollable text.

A research project done by the American Society of Clinical Pathologists to study the suitability of computer adaptive testing for certification is reported as studies 15 and 16 (Bergstrom

[illegible] Additional results of this research are reported in Lunz and Bergstrom, 1991. Students eligible to take a medical technology certification exam took both a

computer adaptive and a pencil and paper test. Items were calibrated and ability measures estimated with the Rasch model. In Table 1, study 15 reports on students who took the CAT first, study 16 reports on students who took the pencil and paper test first.

The last four studies presented in Table 1 are from research done by the National Council State Boards of Nursing (1991), also reported in Zara, 1992. This research was [illegible]

ability measures estimated with the Rasch model. Some items required scrolling the text to see the entire item. The pencil and paper version counted toward licensure, the CAT version did not. Studies were done in July of 1990 and February of 1991 and reported by order of administration. Study 17 in Table 1 is the July, 1990 report of examinees who took the CAT first, study 18 is the July, 1990 report of examinees who took the pencil and paper test first, study 19 is the February, 1991 report of examinees who took the CAT first and study 20 is the February, 1991 report of examinees who took the pencil and paper test first.

on the pencil and paper test.

and S is the pooled standard deviation calculated as:

$$S = \sqrt{\frac{(n^{CAT} - 1)(s^{CAT})^2 + (n^{PAP} - 1)(s^{PAP})^2}{n^{CAT} + n^{PAP} - 2}}$$

where n^{CAT} is number of examinees who took the CAT

and n^{PAP} is the number of examinees who took the pencil and paper test.

The unbiased (d) effect size (corrected for small sample bias) is calculated as:

$$d = \left(1 - \frac{3}{4N -}\right.$$

$$\hat{\sigma}^2 (\quad = \frac{N}{n^{CAT} \; n^{PAP}} + .$$

where $N = (n^{CAT} + n^{PAP})$

The sample size varies across studies. In order to pool the effects, since estimates from the larger studies are more precise than the estimates from the smaller studies, the larger studies are given more weight with the following formula:

$$W_i = \frac{1}{\sigma^2(d_i)} / \sum_{j=1}^{k}$$

A pooled effect, or weighted mean effect (d_+), can then be calculated as:

$$d_+ = \sum_{i=1}^{k} \frac{di}{\hat{\sigma}^2(d_i)} / \sum_{i=1}^{k} \frac{1}{\hat{\sigma}^2(}$$

with a variance:

$$\hat{\sigma}^2((= \left(\sum_{i=1}^{k} \frac{1}{\hat{\sigma}^2(d_i)} \right.$$

In order to determine whether the studies can reasonably be described as sharing a common effect size the following statistical test for homogeneity of effect size was performed:

$$Q = \sum_{i=1}^{k} \frac{(d_i - d_+)^2}{\hat{\sigma}^2(d_i)}$$

The test statistic Q has an asymptotic Chi-Square distribution with k-1 degrees of freedom.

the adaptive test was taken between one day and three weeks after the pencil and paper test, examinees may have forgotten some of the material. Also, students may have been less motivated

[illegible]

project also indicate higher mean achievement on the pencil and paper version than the CAT. This difference is attributed to the fact that candidates were repeatedly made aware that the CAT examination did not count [illegible]

In ten studies from three [illegible]

test first while other examinees took the pencil and paper test first. Figure 2 shows that [illegible] the importance of accounting for order of administration effects in future research.

Conclusion

Most studies, despite differences in test content, age of examinees, IRT model used and study design, show comparable mean ability measures on the CAT and pencil and paper test

[illegible]

for a pre-existing pencil and paper test. Although these differences are accounted for with reasonable explanations by the authors of the studies, they point to the need for continued research into the comparability of the two modes of administration, especially when the intent of the test developers is to replace an existing pencil and paper test with a computer adaptive test.

The knowledge base on computer adaptive testing will be increased by the addition of more studies to this synthesis. An additional look at CAT is planned by using CAT/pencil and paper ability measure correlations in addition to standardized differences to compute standardized effects. The author welcomes suggestions and/or informati [illegible]

[illegible] Bergstrom, B.A. (1991). Comparability of decision for computer adaptive and written examinations. Journal of Allied Health, 20, 2, 15-23.

Mazzeo, J. and Harvey, A.L. (1988). (College Board Report No. 88-8). New York, NY: College Entrance Examination Board.

Moreno, K.E., Wetzel, C.D., McBride, J.R. and Weiss, D.J. (1984). Relationship between [illegible] adaptive testing (CAT) subtests. Applied Psychological Measurement, 8, 2, 155-163.

National Council State Boards of Nursing, Inc. (1991, July). A psychometric comparison of computerized adaptive and paper-and-pencil versions of the national RN licensure examination. Unpublished report. Chicago, IL: Author.

Olsen, J.B., Maynes, D.D., Slawson, D. and Ho, K. (1986, April). Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of [illegible]

10

11

of CAT Studies

l	Age of Examinees	IRT Model	Study Design
ions	High School	1 PL	students took both tests/PAP FIRST
ent	High School	1 PL	students took both tests/CAT FIRST
ent	High School	1 PL	students took both tests/PAP FIRST
nt	Medical Tech students	1 PL	students took both tests/CAT FIRST
nt	Medical Tech students	1 PL	students took both tests/PAP FIRST
nt	Nursing students	1 PL	all students took both tests/CAT FIRST
nt	Nursing students	1 PL	all students took both tests/PAP FIRST
nt	Nursing students	1 PL	all students took both tests/CAT FIRST
it	Nursing students	1 PL	all students took both tests/PAP FIRST

Table 2
Unbiased Effect Sizes (d)

Study	(d)
1	-.470
2	.148
3	-.543
4	-.492
5	.121
6	.103
7	.297
8	.186
9	-.011
10	-.128
11	-.016
12	.037
13	-1.170
14	-1.093
15	-.105
16	.086
17	-.153
18	-.086
19	-.350
20	-.241

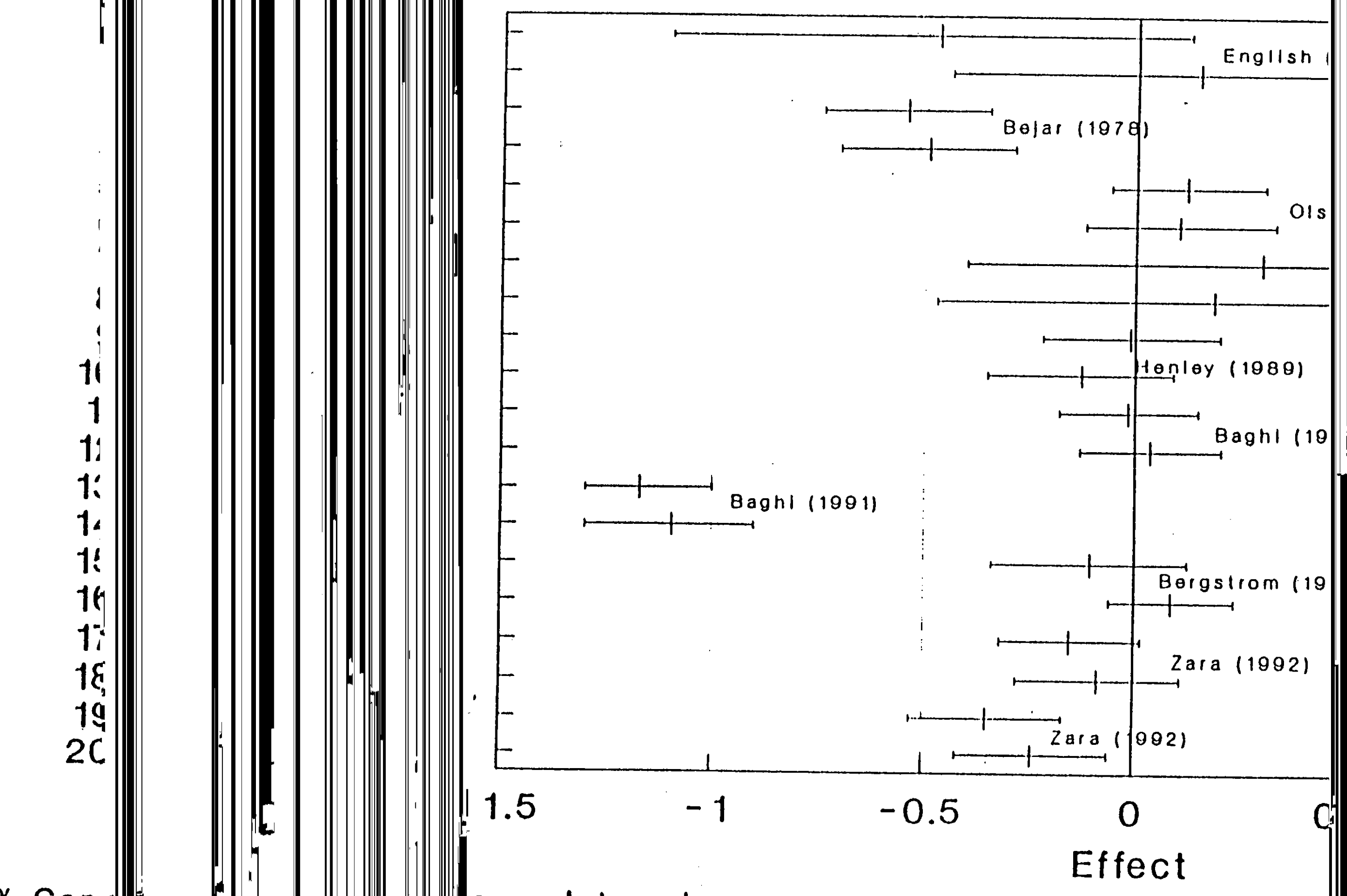
Table 3
Q Statistic Values

No	20 Studies	18 Studies	16 Studies	15 Studies
1	.55	1.38	1.85	2.11
2	1.79	1.53	[illegible]	[illegible]

Figu 1

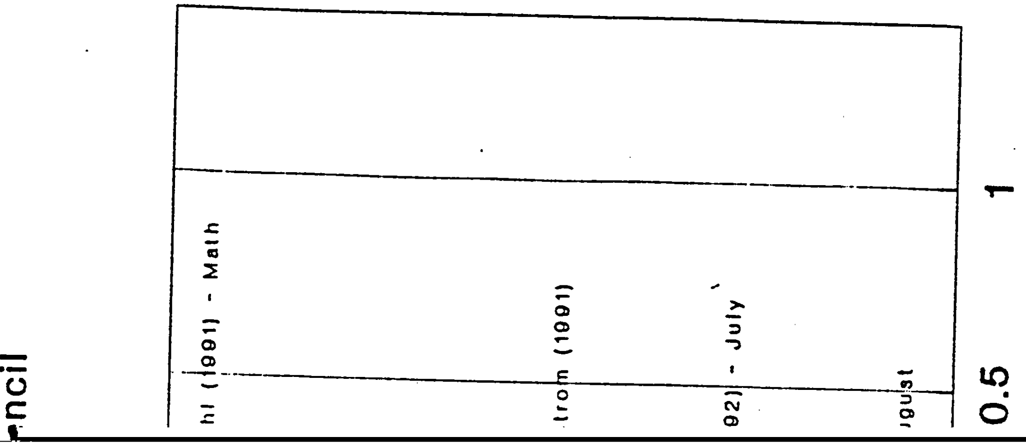
Estimates of Effect Siz
Computer Adaptive / Paper and

Stud



95% Con ence Interval

*Se


ncil
hl (1991) - Math
trom (1991)
92) - July
ugust
0.5
1