ABILITY MEASURE EQUIVALENCE OF COMPUTER ADAPTIVE AND PENCIL AND PAPER TESTS: A RESEARCH SYNTHESIS

BETTY BERGSTROM
AMERICAN SOCIETY OF CLINICAL PATHOLOGISTS

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April, 1992.

ABILITY MEASURE EQUIVALENCE OF COMPUTER ADAPTIVE AND PENCIL AND PAPER TESTS: 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 measures 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 (Bunderson, Inouye and Olsen, 1986; Mazzeo and Harvey, 1988; Wise, 1989) only a few studies have been completed which compare the results of actual administrations of pencil and paper tests and comparable computer adaptive tests. This paper reports on existing studies and uses meta-analysis to compare and synthesize the results of twenty studies from eight research reports. All studies compare the equivalency of ability measures from computer adaptive and conventional pencil and paper multiple choice tests.

The substitution of a computer adaptive test for an existing pencil and paper test needs careful consideration. Reasons why ability measures might not be statistically equivalent include: differences in item presentation, differences in cueing due to varying context and location of items on the CAT, and differences in difficulty ordering. Using the research synthesis techniques developed by Hedges and Olkin (1985) it is possible to estimate a "scale-free" index of effect magnitude for each study and thus compare the results obtained in different studies on a common scale.

Method

Data Collection

Studies were collected from a variety of sources including studies reported at meetings

of the American Education Research Association, studies from the published literature on computer adaptive testing, and studies from unpublished research documents. Included in the analysis were studies which reported the mean and standard deviation for person ability measures from a computer adaptive and a pencil and paper test. Studies were categorized according to test content, age of examinees, IRT model used and study design. Several independent effect sizes were obtained from a single research report if data from separate samples were reported, for example, by test content, age of examinees or study design. However, when studies reported the results of several highly correlated tests for the same examinees, only one computer adaptive/pencil and paper difference was used in the research synthesis. Table 1 summarizes the 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 reported as studies 1 and 2 in Table 1.

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 paper test was administered first to all students and counted toward their course grade. The adaptive test was administered anywhere between one day and three weeks after the pencil and 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, and a computer adaptive format in a study by Olsen, Maynes, Slawson and Ho reported in 1986 and 1990. Students in Grades 3 and 6 were randomly assigned to one of four experimental groups. Each group received two different testing formats in a counterbalanced design. The 3 parameter model was used to calibrate items and ability estimates were calculated using marginal maximum likelihood estimation procedures. In Table 1, Grade 3 is reported as study 5 and Grade 6 as study 6.

Comparability of CAT and pencil and paper versions of the math computation section of the College Level Academic Skills Test (CLAST) at the University of Florida were reported by Legg and Buhr in 1987. Two groups of students took the pencil and paper CLAST in June, 1987, followed by a computer adaptive version in July or August, 1987. The 1 parameter logistic model was used to calibrate items and estimate ability measures. One group, reported as study 7 in Table 1, took an adaptive test in which content areas were not specified. A second group, reported as study 8 in Table 1, took an adaptive test in which items were administered within content areas.

The results of computer administered and pencil and paper versions of the Differential Aptitude Test, a battery of eight ability tests, were reported by Henley, Klebe, McBride and Cudeck (1989). Items were calibrated using the Rasch model and estimates of ability were

calculated using a Bayesian updating technique (Owen, 1975). Two groups of examinees (Sample A and Sample B) were administered both the computer and pencil and paper versions of all eight tests. The order of administration was counterbalanced. Seven of the tests were power tests and were administered adaptively in the computerized version; one test, Clerical Speed and Accuracy was administered on the computer, but not adaptively. Of the seven tests administered adaptively, one test, Abstract Reasoning, was randomly chosen to be included in the research synthesis, however, standardized differences for all seven pencil and paper/CAT comparisons for both Sample A and Sample B were statistically equivalent at an alpha level of .05. The test of Abstract Reasoning with examinee Sample A is reported as study 9 and the test of Abstract Reasoning with examinee Sample B is reported as study 10 in Table 1.

Studies 11 through 14 in Table 1 of the research synthesis were reported by Baghi, Gabrys and Ferrara (1991) on research done with the Maryland Functional Testing Program, a statewide competency testing program used as a high school graduation requirement. The study reports on comparisons of pencil and paper versions and computer adaptive versions of math and reading tests. Results are reported separately by order of administration. Thus, in Table 1, study 11 reports on the math test when the CAT was administered first, study 12 reports on the math test when the pencil and paper test was administered first, study 13 reports on the reading test when the CAT was administered first, and study 14 reports on the reading test when the pencil and paper test was administered first. Baghi et.al. note that the administration of the CAT version 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 and Lunz, 1991) in Table 1. 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 done to study the feasibility of computer adaptive testing for RN (nursing) licensure. Items were calibrated and 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.

Limitations

Although attempts were made to include as many studies as possible, only studies which reported means and standard deviations for both versions of the test could be included. Notably missing is data from the Armed Services Vocational Aptitude Battery (ASVAB). Important research on computer adaptive testing, especially in the early stages, was done by the U.S. Armed Services with subtests of the ASVAB (Green, Bock, Humphreys, Linn and Reckase, 1982; Green, Bock, Humphreys, Linn and Reckase, 1984; Green, 1988). Moreno, Wetzel, McBride and Weiss (1984) report on the comparability of pencil and paper/CAT versions of three subtests of the ASVAB but they report the pencil and paper means and standard deviations in raw score units and the CAT means and standard deviations in logit units. Therefore, it was not possible to include this data. However, Moreno et.al. (1984) concluded that the CAT and ASVAB subtests appear to be measuring the same aptitude factors, that CAT subtests correlated as highly with initial ASVAB test scores as did ASVAB retest scores and that CAT was thought to be as valid a measure of abilities tested as the corresponding ASVAB subtests.

Computation of Effect Size

The effect size computed was the standardized mean difference between the ability measure estimated by the computer adaptive test and the ability measure estimated by the pencil and paper test. All statistics were calculated from the formulas presented in Hedges and Olkin (1985). The following statistics were used to calculate the effect (g) for each study:

$$g = (\bar{y}^{CAT} - \bar{y}^{PAP})$$

where $\bar{y}^{CAT}$ is the mean ability measure on the computer adaptive test

and $\bar{y}^{PAP}$ is the mean ability measure 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:

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

with an estimated variance $\hat{\sigma}^2 (d)$:

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

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:

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

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

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

with a variance:

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

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.

Results

Figure 1 and Table 2 show the unbiased point estimator of the effect size (d) for each of the twenty studies. Since the effect size was computed as CAT - PAP, positive values for the effect in Figure 1 indicate that the mean examinee measure was higher on the CAT than on the pencil and paper test while negative values indicate that the mean examinee measure was higher on the pencil and paper test. In 14 of the 20 studies (70%), the 95% confidence interval (See Figure 1) encompasses zero effect, indicating no significant differences in mean ability measures. These fourteen studies vary in test content, age of examinees, IRT model used and the research design of the study (see Table 1).

The effect size of studies 13 and 14 show a large negative effect (-1.170 and -1.093, respectively). This indicates that examinees scored significantly better on the pencil and paper test than on the computer adaptive test in these two studies. Table 1 shows that studies 11 through 14 were reported by Baghi et. al. (1991) on the Maryland Functional Testing Program. Studies 11 and 12, which have small effect sizes are mathematics tests while studies 13 and 14 are reading tests. Baghi et. al. attribute the significantly lower scores on the CAT reading tests to students' lack of practice with the scrollable-text feature of the test.

The effect size of studies 3 (-.543) and 4 (-.492) (Bejar and Weiss, 1978) indicates that, as in studies 13 and 14, mean achievement scores on the pencil and paper tests were higher than mean achievement scores for the computer adaptive tests. Bejar and Weiss hypothesized that since 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 to perform well on the CAT since only the pencil and paper results counted toward their course grades.

Studies 19 (-.350) and 20 (-.241) from the NCSBN research 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 toward licensure while the pencil and paper test did.

Effect of Order of Administration

In ten studies from three research reports, some examinees took the computer adaptive test first while other examinees took the pencil and paper test first. Figure 2 shows that a practice effect is observed in all pairs of studies. When examinees took the CAT first they performed relatively better on the second test, the pencil and paper test. When examinees took the pencil and paper test first, they performed relatively better on the second test, the CAT.

Computation of a weighted Mean Effect Size

Before computing a weighted mean effect size (d_+), the Q statistic was computed to determine whether the studies can reasonably be described as sharing a common effect size (Hedges and Olkin, 1985). The value of the Q statistic for all 20 studies, 281.9 with 19 degrees of freedom, was significant ($p < .001$) and thus the hypothesis of homogeneity of effect size for the 20 studies must be rejected. Five studies must be removed from the analysis, studies 3, 4, 13, 14, and 19 before the value of the Q statistic indicates that the studies can be described as sharing a common effect size (see Table 3). When these five studies are removed from the analysis, the value of the Q statistic is 19.7 with 14 degrees of freedom ($p > .10$). The weighted mean effect size (d_+) for the 15 remaining studies is -0.002.

Discussion

Science builds on replication. The value of this research review lies less in the computation of a common effect size, than in the pattern of comparability of CAT and pencil and paper tests seen in Figure 1. Research into computer adaptive testing is expensive and time consuming. It requires the construction of adequate item banks, the writing of administration software, and the preparation of testing sites in order for the actual administration of the test to

occur. Organizing research studies using research synthesis techniques enables the comparison of the results of different studies on a common scale. Most of the studies included show CAT to be a acceptable alternative to traditional pencil and paper tests.

The Baghi et.al. (1991) reading study, done with the Maryland Functional Testing Program, shows that factors such as content and/or administration conditions, like a scrollable text, may interfere with mode of administration comparability. It may be possible for the National Council State Boards of Nursing (Zara, 1992) to address this question since some items used in their study required the use of a scrollable text while others in the same content area did not require the use of scrolling. Additional research in this area is also needed in regard to the amount of training on the computer required when features such as a scrollable text are included in a CAT administration.

Figure 2 shows that in all studies when some examinees took the CAT first while other examinees took the pencil and paper test first, a practice effect occurred. These studies point to 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 versions. In all cases where mean differences are statistically different, mean measures are higher 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 information on additional studies which can be included in this on-going project.

References

Baghi, H., Gabrys, R. and Ferrara, S. (1991, April). Applications of computer-adaptive testing in Marvland. Paper prepared for the Annual Meeting of the American Educational Research Association, Chicago, IL.

Bejar, I.I. and Weiss, D.J. (1978). A construct validation of adaptive achievement testing (Research Report 78-4). Psychometric Methods Program Department of Psychology University of Minnesota, Minneapolis, MN.

Bergstrom, B.A. and Lunz, M.E. (1991, July). Comparisons of computer adaptive and pencil and paper tests. Unpublished Research. American Society of Clinical Pathologists, Chicago, IL.

Bunderson, V.C., Inouye, D.K. and Olsen, J.B. (1986). The four generations of computerized educational measurement. In R.L. Linn (Ed.), Educational Measurement, New York: MacMillan Publishing.

English, R.A., Reckase, M.D. and Patience, W.M. (1977). Application of tailored testing to achievement measurement. Behavior Research Methods and Instrumentation, 9, 2, 158-161.

Green, B.F. (1988). Critical problems in computer-based psychological measurement. Applied Measurement in Education, 3, 223-231.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L. and Reckase, M.D. (1982). Evaluation plan for the computerized adaptive vocational aptitude battery. Personnel and Training Research Programs Office of Naval Research (Research Report 82-1). Arlington, VA.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L. and Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 4, 347-360.

Hedges, and Okin, K. (1985). Statistical methods for meta-analysis. Academic Press, Inc., San Diego, CA.

Henley, S.J., Klebe,K.J., McBride, J.R. and Cudeck, Robert (1989, December). Adaptive and conventional versions of the DAT: The first complete test battery comparison. Applied Psychological Measurement, 13, 4, 363-371.

Legg, S.M. and Buhr, D. (1987, November). Final report: Feasibility study of a computerized test administration of the clast. (Contract: 5401473-12). Institute for student assessment and evaluation. University of Florida.

Lunz, M.E. and 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 corresponding armed services vocational aptitude battery (ASVAB) and computerized 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 achievement. Paper presented at the American Educational Research Association Meeting, San Francisco, CA.

Olsen, J.B. (1990). Applying computerized adaptive testing in schools. Measurement and Evaluation in Counseling and Development, 23, 31-38.

Owen, R.A. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of American Statistical Association, 70, 351-356.

Reckase, M.D. (1974). An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods and Instrumentation, 6, 208-212.

Wise, S.L. and Plake, B.S. (1989). Research on the effects of administering tests via computers. Educational Measurement Issues & Practice, 8, 3, 5-10.

Zara, Anthony R. (1992, April). A comparison of computerized adaptive and paper- and-pencil versions of the national registered nurse licensure examination. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

TABLE 1
Descriptive Information of CAT Studies

Study	Author	Publication Date	Test	Content	Age of Examinees	IRT Model	Study Design
1	English et. al.	1977	Classroom Achievement	Educational Measurement	College students	1 PL	equivalent groups
2	English et, at,	1977	Classroom Achievement	Educational Measurement	College students	1 PL	equivalent groups
3	Bejar Weiss	1978	Classroom Achievement	Biology Achievement	College Freshmen	3 PL	all students took both tests/PAP FIRST
4	Bejar Weiss	1978	Classroom Achievement	Biology Achievement	College Freshmen	3 PL	all students took both tests/PAP FIRST
5	Olsen et.al.	1986	California Assessment Program	Math Applications	Grade 3	3 PL	equivalent groups
6	Olsen et.al.	1986	California Assessment Program	Math Applications	Grade 6	3 PL	equivalent groups
7	Legg Buhr	1987	College Level Academic Skills	Math Achievement	College students	1 PL	students took both tests/PAP FIRST
8	Legg Buhr		College Level Academic Skills	Math Achievement	College students	1 PL	students took both tests/PAP FIRST
9	Henley et.al.	1989	Detroit Aptitude	Math Ability	Grades 8-12	1 PL	all students took both forms/order of administration was counter balanced
10	Henley et.al.	1989	Detroit Aptitude	Math Ability	Grades 8-12	1 PL	all students took both forms/order of administration was counter balanced
11	Baghi et.al.	1991	Maryland Functional Testing Program	Math Applications	High School	1 PL	students took both tests/CAT FIRST

TABLE 1
Descriptive Information of CAT Studies

Study	Author	Publication Date	Test	Content	Age of Examinees	IRT Model	Study Design
12	Baghi et.al.	1991	Maryland Functional Testing Program	Math Applications	High School	1 PL	students took both tests/PAP FIRST
13	Baghi et.al.	1991	Maryland Functional Testing Program	Reading Achievement	High School	1 PL	students took both tests/CAT FIRST
14	Baghi et.al.	1991	Maryland Functional Testing Program	Reading Achievement	High School	1 PL	students took both tests/PAP FIRST
15	Bergstrom Lunz	1991	ASCP Certification Research	Science Achievement	Medical Tech students	1 PL	students took both tests/CAT FIRST
16	Bergstrom Lunz	1991	ASCP Certification Research	Science Achievement	Medical Tech students	1 PL	students took both tests/PAP FIRST
17	Zara	1992	NCSBN Certification Research	Biology Achievement	Nursing students	1 PL	all students took both tests/CAT FIRST
18	Zara	1992	NCSBN Certification Research	Biology Achievement	Nursing students	1 PL	all students took both tests/PAP FIRST
19	Zara	1992	NCSBN Certification Research	Biology Achievement	Nursing students	1 PL	all students took both tests/CAT FIRST
20	Zara	1992	NCSBN Certifiction Research	Biology Achievement	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

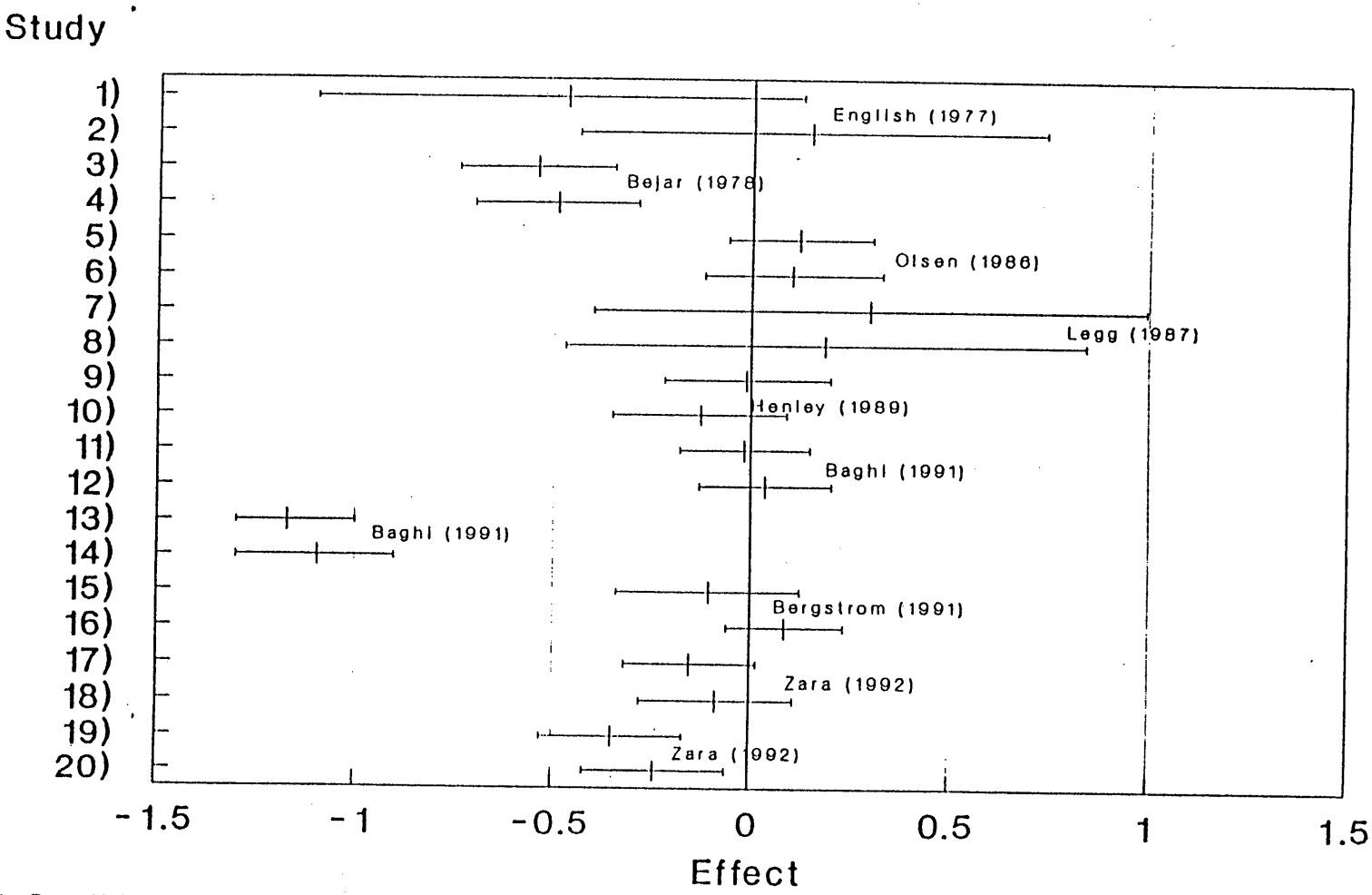
Table 3
Q Statistic Values

No	20 Studies	18 Studies	16 Studies	15 Studies
1	.55	1.38	1.85	2.11
2	1.70	.75	.47	.35
3	9.17	19.01	***	***
4	5.56	13.03	***	***
5	15.71	6.44	3.74	2.65
6	9.13	3.53	1.94	1.32
7	2.32	1.33	.99	.84
8	1.66	.80	.53	.41
9	4.64	.88	.18	.03
10	1.09	.02	.40	.78
11	7.36	1.32	.24	.02
12	10.76	3.05	1.20	.59
13	110.56	***	***	***
14	76.24	***	***	***
15	2.40	.00	.17	.42
16	19.08	6.91	3.58	2.30
17	1.10	.22	1.24	2.07
18	2.49	.07	.09	.33
19	1.38	6.94	10.57	***
20	.00	2.02	4.15	5.51
	Σ281.93	Σ67.71	Σ31.33	Σ19.74

*** removed from analysis

Figure 1

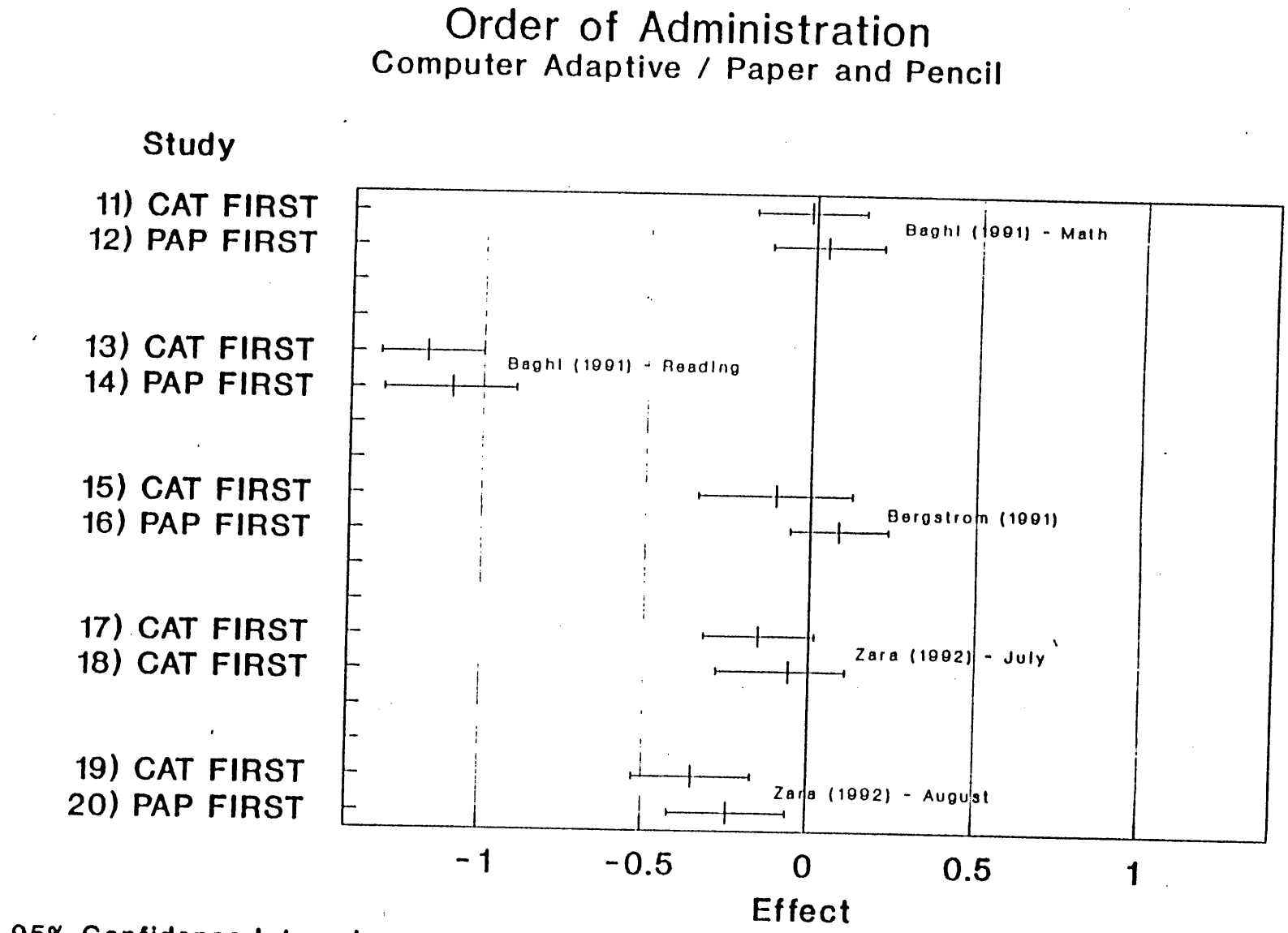
Estimates of Effect Size
Computer Adaptive / Paper and Pencil



95% Confidence Interval

*See Table 1 for study descriptions

Figure 2


Order of Administration
Computer Adaptive / Paper and Pencil
Study
11) CAT FIRST
12) PAP FIRST
13) CAT FIRST
14) PAP FIRST
15) CAT FIRST
16) PAP FIRST
17) CAT FIRST
18) CAT FIRST
19) CAT FIRST
20) PAP FIRST
Baghl (1991) - Math
Baghl (1991) - Reading
Bergstrom (1991)
Zara (1992) - July
Zara (1992) - August
-1
-0.5
0
0.5
1
Effect
95% Confidence Interval