

Utah State University

DigitalCommons@USU

---

All Graduate Theses and Dissertations

Graduate Studies

---

5-1994

## A Comparative Study of the Effect of Paper-and-Pencil Versus Computer Administration of an Achievement Test

Perry Sailor  
*Utah State University*

Follow this and additional works at: <https://digitalcommons.usu.edu/etd>



Part of the [Psychology Commons](#)

---

### Recommended Citation

Sailor, Perry, "A Comparative Study of the Effect of Paper-and-Pencil Versus Computer Administration of an Achievement Test" (1994). *All Graduate Theses and Dissertations*. 6053.

<https://digitalcommons.usu.edu/etd/6053>

This Thesis is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact [digitalcommons@usu.edu](mailto:digitalcommons@usu.edu).



Copyright © Perry Sailor 1994  
All Rights Reserved

A COMPARATIVE STUDY OF THE EFFECT OF PAPER-AND-PENCIL VERSUS  
COMPUTER ADMINISTRATION OF AN ACHIEVEMENT TEST

by

Perry Sailor

A thesis submitted in partial fulfillment  
of the requirements for the degree

of

MASTER OF SCIENCE

in

Psychology

Approved:

UTAH STATE UNIVERSITY  
Logan, Utah

1994

## ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Blaine R. Worthen, for his suggestion of the topic for this thesis and for his invaluable support and guidance throughout.

I would also like to express my appreciation and gratitude to Drs. Lani Van Dusen and Byron Burnham, members of my committee, for their helpful suggestions for improving this thesis.

My appreciation and gratitude also go to Kevin Hague, principal of Pleasant Green Elementary School in Magna, Utah, to Prent Klag, principal of Edith Bowen Laboratory School at Utah State University in Logan, and to students and faculty of both schools.

Finally, I would like to give special thanks for their love and support to my mother, Wilma Sailor, to my late father, Raymond J. (Al) Sailor, to my wife, Lori Roggman, and to my children, Rae and Mark.

Perry Sailor

## CONTENTS

	Page
ACKNOWLEDGMENTS . . . . .	ii
LIST OF TABLES . . . . .	iv
ABSTRACT . . . . .	v
THE PROBLEM: COMPUTER VS. PAPER-AND-PENCIL TESTING . . . . .	1
REVIEW OF LITERATURE . . . . .	5
Studies Using Elementary-School or Middle-School Students . . . . .	5
Studies Providing Comparable Conditions Between CA and PP Testing . . . . .	9
Studies in Which Administration Conditions Are More Restrictive in the CA Condition . . . . .	13
Comparison of Results of Studies with Comparable PP and CA Conditions to Studies with More Restrictive CA Conditions . . . . .	15
PURPOSE AND OBJECTIVES . . . . .	17
METHOD . . . . .	18
Population and Sample . . . . .	18
Design . . . . .	20
Data and Instrumentation . . . . .	21
Validity and Reliability . . . . .	23
Procedure . . . . .	25
Analysis . . . . .	26
RESULTS . . . . .	28
Time as a Dependent Variable . . . . .	31
Time as an Independent Variable . . . . .	35
SUMMARY AND DISCUSSION . . . . .	39
REFERENCES . . . . .	44
APPENDICES . . . . .	49
Appendix A. Grade 2 Test . . . . .	50
Appendix B. Grade 4 Test . . . . .	77

## LIST OF TABLES

Table		Page
1	Mean Scores, Standard Deviations, and Ns for Each Mode at Each Grade . . . . .	28
2	Summary of Analysis of the Relationship of the Independent Variable Mode to Scores . . . . .	30
3	Mean Scores, Standard Deviations, and Ns for Each Mode at Each Grade--Students with Pretest Scores Only . . . . .	32
4	Summary of Analysis of the Relationship of the Independent Variable Mode to Scores, with Pretest an Additional Control Variable . . . . .	33
5	Mean Time to Completion, Standard Deviations, and Ns for Each Mode at Each Grade . . . . .	34
6	Summary of Analysis of the Relationship of the Independent Variable Mode to Time to Completion . . . . .	35
7	Summary of Analysis of the Relationship of the Independent Variable Mode to Scores, with Time an Additional Control Variable . . . . .	36

## ABSTRACT

A Comparative Study of the Effect of Paper-and-Pencil Versus  
Computer Administration of an Achievement Test

by

Perry Sailor, Master of Science

Utah State University, 1994

Major Professor: Dr. Blaine R. Worthen  
Department: Psychology

The study examined whether, under comparable testing conditions, second- and fourth-grade students who took a computer-administered (CA) achievement test in mathematics achieved the same mean score as comparable students who took the same test by paper and pencil (PP).

For number correct, the CA standardized mean difference effect size was - 0.28, which was larger than the expected effect size of zero, although not statistically significant at .05. It was noted that CA subjects completed the test more quickly, on the average, than PP subjects (CA effect size for time to completion = - 0.79). When time to completion was statistically controlled, the difference in mean scores between CA and PP modes vanished (CA effect size = - 0.02).

Possible explanations for the findings are discussed. It is concluded that, based on these results, one would not be justified in assuming CA and PP scores from elementary school students to be equivalent.

(109 pages)



## THE PROBLEM: COMPUTER VS. PAPER-AND-PENCIL TESTING

One of the many applications for computers in modern society is in the field of testing students' learning. Particularly with the increasing power and availability of microcomputers, the perceived advantages of computer-administered (CA) testing over paper-and-pencil (PP) testing are frequently cited. For example, Mazzeo and Harvey (1988) and Wise and Plake (1989) collectively listed the following advantages for CA testing: (a) increased test security; (b) lowered costs for production, administration, and scoring, which should quickly offset increased development costs; (c) less testing time, particularly for so-called adaptive or tailored tests, in which the computer chooses items of appropriate difficulty based on responses to earlier items, resulting in fewer total items needed for assessment; (d) graphic displays which may realistically depict movement or other important features, in turn leading to better measurement of test takers' understanding in certain fields; (e) more flexible administration schedules; (f) immediate feedback/scoring; and (g) the ability to measure response latency and patterns of skipping and changing answers. To the extent that these advantages are believed to outweigh any perceived disadvantages (such as initial hardware and software costs), the use of computers for testing will continue to proliferate.

While the potential benefits of CA testing are numerous, little is known about the actual effects of the technology itself on student performance. The American Psychological Association, in its Guidelines for Computer-Based Tests and Interpretations (American Psychological Association, 1986), asserted that equivalence of scores from CA and PP administrations of the same test should not be assumed,

but should be established and documented before using PP-derived norms for CA administrations. This is a practical guideline which is appropriate for handling a specific situation, but there are important, broader--and still unresolved--questions as well. For example, given the same content domain, or even the same items, does CA administration produce, on the average, higher scores, lower scores, or the same scores as PP administration? If there are differences, what causes them? Empirical testing is needed to answer these questions.

Possible effects of CA testing could come from two sources: (a) those related to personal characteristics of examinees, and (b) those related to characteristics of the testing situation. Evidence is scanty concerning individual differences. Eaves and Smith (1986) examined the effect of differential familiarity with computers and found it made no statistically significant difference in test performance, a finding corroborated by the results of Wise, Barnes, Harvey, and Plake (1989). Wise et al. (1989) and Ward, Hooper, and Hannafin (1989) also found no effect for another individual difference variable, anxiety, while Llabre et al. (1987), in a correlational study, found that CA examinees had lower scores and more anxiety. Because the present study concerns the testing situation rather than examinee characteristics, the remainder of this review is restricted to the former category.

Wise and Plake (1989) noted that there are three test characteristics that are almost always present on PP tests but often are not characteristic of CA tests: (a) allowing items to be skipped and answered later, (b) allowing the review of items already answered, and (c) allowing examinees to change answers to items. Wise and

Plake reported finding only one study that examined this issue directly, an unpublished dissertation done by Harvey (1987), who compared two versions of the same CA test, one with and one without these three features. Harvey found no statistically significant differences between the two versions, but Wise and Plake noted that college students participated in the study for research credit and may not have been motivated to do well (and hence would be unlikely to review items or change answers anyway).

Although a detailed review of the literature will be presented below, it can be stated here that the present study will contribute to the literature in two ways. First, many previous studies have been characterized by failure either to test or control for the effects of such variables as ability to change answers or review items, either confounding these variables with test mode--usually allowing answer changes and item review in the PP condition only--or not mentioning them at all. Second, only three previous studies have tested elementary school students, as the present study did, and none of these three specified whether or not subjects in the CA condition were permitted to change answers or review past items. The present study controlled subjects' ability to change answers and review past items in both CA and PP conditions, and used an elementary school sample. Therefore, it stands as a relatively pure test of the effects of CA testing on performance, concerning an age group for which the effects of CA testing are little known.

The general purpose of the present study was to see whether mode of test administration is associated with student performance on a test of typical school

subject matter. Specifically, the objective was to determine whether elementary school students obtain different test scores depending on whether the test is administered by computer (CA) or by paper-and-pencil (PP).

## REVIEW OF LITERATURE

In the course of exploring the literature on computer testing, it became apparent to the author that two dimensions were particularly important in making sense of the literature, because they had not been systematically explored: the age of the subjects, and the conditions of testing. These dimensions serve to organize the following review. At the end of the review, the findings of studies cited will be summarized as they relate to the present study.

### Studies Using Elementary-School or Middle-School Students

A review of the literature on possible effects of administering tests by computer reveals that very few studies have used an elementary school sample. Wise and Wise (1987) administered a 32-item multiple-choice arithmetic test to 68 third and fourth graders who were randomly assigned to one of three conditions--paper administered, computer administered with item feedback, and computer administered without item feedback. (The item feedback consisted of informing the subject whether the response was correct or not.) Although the mean score for the computer-no feedback condition was lower than for the paper condition, an overall analysis of variance (ANOVA) revealed that differences in mean number correct for all three conditions did not reach the .05 level of statistical significance. The standard mean difference (SMD) for the paper condition (considered the control) compared to the computer-no feedback condition was -0.22. The SMD is computed by subtracting the control mean score from the treatment mean score, and dividing the difference by the

control standard deviation. In the present review, it will also be referred to as the "effect size" (Glass & Hopkins, 1984). A positive effect size means that subjects in the CA condition achieved the higher mean score; negative means the PP subjects scored higher.

Olsen, Maynes, Slawson, and Ho (1989), in a study that also included adaptive testing, tested nearly 600 third- and sixth-grade students on mathematics application items from the California Assessment Program item bank. (In adaptive testing, the items an examinee receives depend on his or her ability level. There are many different procedural models for this, but in general, a computer is programmed to begin with an item of intermediate difficulty, record whether the response is correct or not, and then select each successive item based on the examinee's total response history up to that point. In this way, an examinee's ability level can be estimated very precisely with many fewer items than in traditional testing.) It is not clear from Olsen et al.'s (1989) report if the items in the paper and computer administrations were identical, but the number of items was identical in each condition. (Because the whole point of computer adaptive testing is to use fewer items, one presumes that in the computer adaptive condition, the number of items was fewer than in the other two conditions.) In Olsen et al.'s (1989) design, each student was randomly assigned to one of four groups. Group 1 took a computer-administered test followed by a computer-adaptive test; Group 2 took a computer-adaptive test followed by a computer-administered test; Group 3 took a paper-administered test followed by a

computer-adaptive test; and Group 4 took a computer-adaptive test followed by a paper-administered test.

For the present review, the key comparisons would be: (a) the computer-administered test taken by Group 1 versus the paper-administered test taken by Group 3--each was the first in the two-test sequence for those groups--and (b) the computer-administered versus paper-administered tests taken by Groups 2 and 4, each taken with a computer-adaptive test preceding it. Unfortunately, Olsen et al. (1989) did not report those comparisons. They did, however, report that in a separate "Test Mode x Order" ANOVA including only the paper- and computer-administered conditions, test mode differences were not statistically significant at either grade. The computer-administered effect size was 0.06 at grade 3, and -0.002 at grade 6. Olsen et al. (1989) reported that their subjects had significant computer experience. However, neither these researchers nor Wise and Wise (1987) reported whether their subjects could change answers and/or review previous items.

Ronau and Battista (1988), as part of a larger study on computer diagnosis of errors in solving ratio and proportion problems, developed computer and paper-and-pencil versions of tests on concepts of ratio and proportion. Two studies were conducted to compare the influence of these two testing modes. Study 1 tested 20 eighth graders in a within-subjects design, with half the subjects taking the computer test first and half taking the paper-and-pencil test first. The interval between tests was not reported. Study 2 used a between-subjects design, with 20 students taking the computer version and a different 20 students taking the paper-and-pencil version.

Students taking the computer version were allowed to use paper and pencil for calculation. In both studies, the mean score on the computer version was lower than on the paper-and-pencil version. Both of these differences were statistically significant at the .01 level; effect sizes were -0.72 for Study 1 and -1.63 for Study 2.

In summary, the three studies using elementary school students (Wise & Wise, 1987 and the two reported in Olsen et al., 1989) reported no statistically significant differences between means on computer- and paper-administered tests, with one reporting a very small positive effect, one negative, and one essentially zero. In contrast, the two middle school studies by Ronau and Battista (1988) found sizable and statistically significant negative effects of computer testing.

At least two possible explanations for the Ronau and Battista (1988) findings can be advanced, based on information in their report. First, students were tested before being taught the concepts, and mean scores on the tests were quite low--below 50%. Beach (1989) has reported that random responding is more likely on a computer-administered test than on a paper-and-pencil test. It seems reasonable that eighth graders being tested on a concept they had not yet been taught may have some tendency to respond randomly in any case; if the computer group did this more than the paper group, as Beach's (1989) findings suggest they might, that alone may have accounted for the computer group's lower mean scores.

A second possible explanation is more general, and therefore potentially more interesting. That explanation is that the difference in conditions of testing between the two modes may have caused the difference in test scores. In Ronau and Battista's



(1988) study, subjects were not permitted to review past items, change previous answers, or return to skipped items in the computer condition, but were permitted to do so in the paper-and-pencil condition. None of the studies using elementary children as subjects reported on these variables, so conditions are not known.

The failure either to equate testing conditions between the computer and paper modes, or even to report whether or not they were controlled, is characteristic of many studies in this area. Of 21 separate studies meeting criteria established for inclusion in the present review (that is, studies including a direct comparison between student performance on a CA and PP aptitude or achievement test of the same length, and including no graphics more complex than simple line drawings), only 8 reported allowing subjects to change answers and review past items in both the computer and paper modes. In other words, only 8 studies provided truly comparable conditions between the CA and PP modes of administration. (Incidentally, none of the 8 used elementary-age students.)

#### Studies Providing Comparable Conditions Between CA and PP Testing

Four of the eight analyses providing comparable testing conditions were reported by Mazzeo, Druesne, Raffeld, Checketts, and Muhlstein (1991). Mazzeo et al. investigated the comparability of scores from paper-and-pencil and computer-administered versions of the College Level Examination Program's (CLEP) General Examinations in Mathematics and English Composition. A within-group design was used, with half the subjects taking the computer version first and half taking the paper

version first. All items were multiple choice. Items on the two versions of each test were not identical, but each test was separately calibrated to the CLEP 200-800 score reporting scale, with a mean of 500 and standard deviation of 100. The Mathematics paper- and computer-administered tests had no items (of 90) in common, while the English Composition tests had 29 of 95 items in common. The average interval between tests is not reported, but the authors report that paper-and-pencil tests were given during a 3-day interval, and that computer testing began 4 days before and ended 4 days after this interval, so the range could have been 0-7 days. In English Composition, all subjects took computer and paper tests the same day.

The results of Mazzeo et al.'s (1991) first study suggested that, despite efforts to design CA versions of the exams that were administratively similar to PP testing (that is, both modes allowing item review and changing answers, and both being comparably timed), statistically significant mode-of-administration effects were found. For the English Composition test, the computer effect size was -0.27, while for the Mathematics test the effect size was -0.13.

For Study 2, Mazzeo et al. (1991) attempted to make the CA and PP tests even more administratively similar. Although the speed factor in the tests was very small, some students were concerned that in Study 1, the clock continued to run during the delay between items on the computer version. In Study 2, the clock did not run between item presentations. Moreover, in Study 2 the computer subjects were given a means to skip items but mark them to return to later, much as students

taking paper-and-pencil tests often do. Finally, practice items were changed so that they more closely matched items on the actual exams.

On the Study 2 English Composition exam, the difference in mean scores by mode of administration was not statistically significant at the .05 level (effect size = -0.005); for the Mathematics test, a slight difference remained in favor of the paper-and-pencil test (effect size = -0.09) but this difference was also not statistically significant.

Harrell, Honaker, Hetu, and Oberwager (1987) administered a CA and PP version of the Verbal scale of the Multidimensional Aptitude Battery (MAB-V) to undergraduates, using a counterbalanced repeated measures design. The two versions used identical items, and the CA version was designed to be highly comparable to the PP version. Administrative conditions were very similar but may have been a bit more restrictive under the CA condition. Subjects taking the test via computer could, after each item response selection, either back up to the previous item, erase the response, or continue to the next item. Presumably subjects in the PP condition could go back to any item, not just the previous one. Also, it is not clear if CA subjects could change the answer to the previous item or merely review it. However, it may be that CA subjects could go back one item at a time, in a successive fashion, thus providing them access to any previous item at any time. This would make conditions of the CA test completely comparable to the PP test. Unfortunately, the report is not written in such a way as to make clear exactly what the administrative conditions were.

Subjects were assigned to one of four groups. Group 1 took the paper-and-pencil version twice. Group 2 took the paper version, then the computer version. Group 3 took the computer version, then the paper-and-pencil version. Group 4 took the computer version twice. Testing sessions were about one week apart.

Mean scale scores for all five subtests of the MAB-V were compared among the four groups using MANOVA; the overall group effect was not statistically significant at .05. An effect size was computed by the present author using the combined Verbal IQ means for the first administration given to Groups 1 and 2 (both PP), compared to the combined Verbal IQ means for the first administration given to Groups 3 and 4 (both CA); the size of the computer effect was 0.27.

Huba (1988) used adults (mean age 34 years) in a study of the comparability of PP and CA versions of the Western Personnel Test, a 24-item test of general ability. The items measure proofreading, cultural knowledge, recognition, computational skills, ability to recognize a numerical sequence, design reorganization, and logical thinking. Subjects were allowed to skip items, jump backward to correct previous items, and review and change all responses. Group 1 took Form A of the test via CA, and Form B via PP, with half receiving Form A first and half Form B first. Group 2 took Form A via PP and Form B via CA, again with one half receiving Form A first, and the other half receiving Form B first. Differences in mean scores between computer and paper modes were not statistically significant for

either form. The effect size for Form A was 0.15; for Form B, the effect size was - 0.38.<sup>1</sup>

Ward et al. (1989) attempted to determine whether a computerized test which "incorporates traditional test taking interfaces" (p. 329) has any effect on students' performance. These traditional interfaces included the ability to skip and review items. The authors do not explicitly state that any answer could be changed at any time, but given that the purpose of their design was to create maximum similarity to a traditional paper-and-pencil test, it seems reasonable to assume this was the case. Ward et al. randomly assigned college students from an advanced-level course in Special Education, in a between-subjects design, to take a 25-item multiple choice class test either by CA or PP. The mean performance difference was not statistically significant at .05; the effect size was -0.27.

#### Studies in Which Administration Conditions Are More Restrictive in the CA Condition

As one might expect, studies which permit answer change and item review in only the PP condition consistently show negative computer effects. In addition to Ronau and Battista's (1988) two studies reviewed above, three other similar studies have been found.

Eaves and Smith (1986) investigated the effects of computer experience as well as mode of administration, using a sample of 96 college students who took a class test

---

<sup>1</sup>Huba (1988) reported means but not standard deviations; effect sizes were computed from the F values using a formula found in Taylor and White (1990).

in an educational media class. Subjects in the PP group could move back and forth on the test, scan the test as a whole, correct errors recognized on later review, etc., while CA subjects could look at only one item at a time, could not change responses once given, and could not scan the test or skip items. Groups of students with no computer experience, 1 to 10 hours experience, and more than 10 hours experience were each randomly assigned to either PP or CA mode. Results of a "Mode x Experience" ANOVA, with number correct as the dependent variable, yielded no statistically significant differences at .05. The overall computer effect size was -0.14, with effect sizes of -0.29, -0.18, and 0.09 for the no experience, 1-10 hours experience, and more than 10 hours experience groups, respectively. This may indicate some negative CA effect for inexperienced computer users, although the lack of statistical significance means that chance cannot be ruled out as the cause of the results.

Lee, Moreno, and Sympson (1984) administered a 30-item test of arithmetic reasoning to 654 male Marine Corps recruits, who were randomly assigned to either the PP or CA mode. They did not allow subjects in the CA group to change answers or refer to previous answers. A statistically significant effect in favor of the PP group was found, both on raw number correct and on number correct adjusted for a covariate.<sup>2</sup> The computer effect size, measured on both the raw means and on the adjusted means, was -0.19.

---

<sup>2</sup>The covariate was number correct on the Arithmetic Reasoning subtest of the Armed Services Vocational Aptitude Battery, which all subjects had taken 2 weeks to 6 months before the experiment.

Lee and Hopkins (1985), as part of a study of the effects of training on computer test performance, administered 30-item tests of arithmetic reasoning to 92 undergraduates in a within-group design. Subjects were randomly assigned to training or no-training groups. Subjects in both groups took the PP version of the test, then an anxiety measure, then, after one week, took the anxiety measure again, then an innocuous "Personal Preference Questionnaire (PPQ)," then the computer version of the test. The "training" consisted of taking the 20-item PPQ either by CA or PP. The CA version of the test, unlike the PP version, did not permit answer changes or review of past items. Results revealed that training did not account for a statistically significant amount of variance on the CA version of the test, so an overall comparison of the PP test mean to the CA test mean seems reasonable. The mean score on the PP test was higher than on the CA test (computer effect size  $-0.29$ ), and the difference was statistically significant at  $.05$ . However, it should be noted that the two versions of the test had no items in common. Items for both versions were drawn from a common pool and "matched judgmentally in terms of apparent difficulty and mathematical principles required" (Lee & Hopkins, 1985, p. 3), and the authors believe the difficulties were "closely equivalent" (Lee & Hopkins, 1985, p. 8), but it is possible that the items on the CA version were simply more difficult.

Comparison of Results of Studies with  
Comparable PP and CA Conditions to Studies with  
More Restrictive CA Conditions

The mean effect size for the five comparisons in which answer changes and review of past items were allowed on only the PP version of a test is  $-0.59$ ; for the

eight studies in which comparable conditions of item review and answer change held between modes, the mean effect size is  $-0.10$ . However, for the three previous studies using elementary school children as subjects, authors did not report whether conditions differed on the two test modes. The present study corrects these deficits by making PP and CA test administration conditions match as closely as possible when administering the test to elementary school students, by precluding answer change and item review in both the PP or CA versions.



## PURPOSE AND OBJECTIVES

The general purpose of the present study was to see whether mode of test administration is associated with student performance on a test of typical school subject matter. Specifically, the objective was to determine whether elementary school students would obtain different test scores depending on whether the test is administered by computer (CA) or by paper-and-pencil (PP). Based on the previous research reviewed on the previous pages, it was expected that if both groups were operating under identical conditions with respect to ability to change answers and review already-completed items, then mean scores on the tests would not differ to a statistically significant degree. This is in accord with results of studies using college students and adults; as mentioned, no studies were found which reported comparable CA and PP conditions and which used elementary students.

The research question to be answered, then, was this: Under comparable testing conditions, do elementary school students who take an achievement test administered by computer achieve the same mean score as comparable students who take the same test by paper-and-pencil? It was predicted that under comparable testing conditions with respect to answer changes and review of previous items, there would be no difference in performance between students taking CA and PP tests--that is, not only would there be no statistically significant difference at the conventional .05 level, but the effect size would be very nearly zero.

## METHOD

In the present study, elementary school students took a 25-item math test, with half of the students randomly assigned to computer and half to paper-and-pencil administration. Test items were visually identical in both formats. Neither group was permitted to change answers or review previous items. In the CA condition, the computer program incorporates this restriction; in the PP condition, the investigator monitored the testing to ensure compliance.

### Population and Sample

All second- and fourth-grade students at Pleasant Green Elementary School were tested. Pleasant Green is located in Magna, Utah, in suburban Salt Lake City, and is part of the Granite School District. This raises the issue of population validity--how comparable are Pleasant Green's students to other students in the Salt Lake area? How confident can one be that findings from Salt Lake City are generalizable to the rest of Utah, or to the rest of the United States?

Bracht and Glass (1968) differentiated between two types of population validity: (1) the extent to which one can generalize from the experimental sample to a defined population, and (2) the extent to which individual differences ("personological" variables) interact with treatment effects. For example, mode of test administration could interact with gender, age, ability, trait anxiety, or various other variables. If so, the differential effects will limit generalizability. Some of these variables--grade level and ability--were measured and their possible effects

tested in the present study. But this reveals nothing about students whose ability or grade level are outside the range of the present study.

Within the first type of population validity--generalizing from sample to population--there are two levels of inference that collectively define generalizability. The first deals with the extent to which the experimental sample is representative of the accessible population of second- and fourth-grade Pleasant Green students, while the second deals with the extent to which the accessible population is representative of a larger target population. The first type should not be an issue. All second- and fourth-grade classes were tested. This does not represent all second and fourth graders who attend Pleasant Green, because the school operates on a year-round schedule, so only about 75% of the students are attending at any one time. However, the "tracks" are formed by an essentially random process, so results should be generalizable to the 25% of students who are "off track." Further, generalizing the results to other Pleasant Green students--grades 1, 3, and 5--is probably safe. None of these grades is more than one grade removed from a tested grade.

Pleasant Green students seem quite representative of the Salt Lake area. The school is located in a middle class, suburban area on the far western fringe of suburban development in the Salt Lake valley, very similar to other suburbs west of the city.

All fifth graders in Utah take the Stanford Achievement Test (SAT) each spring. Pleasant Green's 1992 median percentile of 59 on the SAT's Math Total subtest ranked 31st of the 63 elementary schools in the Granite School District, and

was identical to the percentile of the median student in the district (Granite School District, 1992). The median percentile for the state of Utah was 62 (Granite School District, 1992).

The issue of generalizing to students outside suburban Salt Lake is problematic. Ultimately, the research question pursued in this study should and will be decided by similar, replication experiments performed in a variety of settings with samples differing on such variables as age, socioeconomic status, academic achievement, gender, ethnicity, computer experience, and other relevant variables. Over time, such replication will produce a body of pertinent knowledge. In the meantime, the investigator's judgment is that the results of the present study are applicable to middle class elementary school students who are familiar with computers (Pleasant Green students spend about 45-60 minutes a week in computer lab).

### Design

A posttest-only control group design, with matching on ability (Campbell and Stanley, 1963), was used for this study. There were three classes of second graders and two classes of fourth graders tested. All students at Pleasant Green are taught math by their regular classroom teacher. Scores on the spring, 1992 administration of the Utah Core Assessment Series, Elementary Mathematics, were obtained, and the students were listed in rank order (with tied students listed in random order). Then one of each adjacent pair of students was randomly assigned to take the CA version of the test, while the other took the PP version. Students without scores were randomly assigned.

According to Campbell and Stanley (1963), blocking on a subject variable that is presumed to be related to the dependent variable (an achievement measure, in this case) can provide "an increase in the power of the significance test very similar to that provided by a pretest" (p. 26). Blocking in conjunction with the posttest-only control group design makes an already powerful experimental design even more powerful (Campbell & Stanley, 1963). Matching may be considered a special case of blocking. As Kerlinger (1964) put it:

Instead of splitting the subjects into two, three, or four parts, however, they are split into  $N/2$  parts,  $N$  being the number of subjects used; thus the control of variance is identified and built into the design. Matching is theoretically a more powerful method of achieving this aim, because it uses most of the variance due to the variable. (p. 285)

Campbell and Stanley (1963) agreed: "...matching plus subsequent randomization usually produces an experimental design with greater precision than would randomization alone" (p. 49).

### Data and Instrumentation

There are three variables in the study: (a) the scores on the standardized math test, the Utah Core Assessment Series, which were used for matching; (b) the independent variable, mode of test administration; and (c) the dependent variable, number correct on the 25-item math test. The standardized math test is a criterion-referenced test developed by the Utah State Office of Education and administered each spring to all students in the state. Procedures developed to ensure content validity, described in detail in the technical manual (Utah State Office of Education, 1988), seem very adequate. Information concerning concurrent or predictive validity is not

available. Coefficient alphas for the two parallel forms of the Grade 1 test are .94 and .88; for the Grade 3 test, the alphas are .92 and .93. Parallel form correlation is .74 for Grade 1, .83 for Grade 3 (Utah State Office of Education, 1988).

Mode of test administration is straightforward; one student in each matched pair was randomly assigned to take the test via paper-and-pencil, while the other took the identical test via computer. All test conditions--location, time of day, ability to change answers or review past items--were held constant across groups. The dependent variable, math test score, will now be discussed at some length.

Jostens Learning Corporation has developed an "integrated learning system" which they market to schools across the nation. Becker (1992) summarized characteristics of integrated learning systems as those

...supplied by a single vendor and containing instruction and practice problems covering a multiple-year curriculum sequence. This software is housed on a central server computer linked in an electronic network to fifteen to thirty student computers. Specific lessons are automatically loaded into each student's computer when that student "logs in" based on continuing assessment of that student's previous accomplishments and current learning needs. (p. 2)

The Jostens system consists of instructional lessons and "Unit Tests"--one for each 10 lessons. The present study used two 25-item multiple-choice Unit Tests, one for each grade tested, and paper-and-pencil versions that matched them as closely as possible. Tests were selected by the investigator in consultation with the students' math teachers, to ensure that students had in fact been taught all the skills tested. The second-grade test chosen assesses students' ability in the following five skills: (a) using addition and subtraction facts and tens, (b) adding/subtracting without

regrouping, (c) adding with regrouping, (d) subtracting with regrouping, and (e) applying addition and subtraction. The fourth-grade test assesses students' skills in the following areas: (a) multiplying/dividing without regrouping, (b) multiplying with regrouping, (c) relating multiplication and division, (d) dividing with partial products, and (e) dividing with short form algorithm (Jostens Learning Corporation, 1989).

A paper-and-pencil version of each test was developed using MacDraw II software. To match conditions on the computer test, only one item was placed on each page, and an attempt was made to make each item look as much as possible like the computer version in size, layout, and style (except the paper version does not have color). The paper-and-pencil tests are contained in the appendices.

### Validity and Reliability

Validity. Cronbach (1971) defined validation as the process by which evidence is collected to support the types of inferences to be drawn from test scores. For these tests, the inference intended by the test developer to be drawn is that the scores measure degree of mastery of the five skills listed above for each test. Crocker and Algina (1986) listed a series of steps to be taken in a content validation study, including defining the performance domain of interest, selecting a panel of experts, matching items to the domain in some structured framework, and summarizing the data from the matching process.

For the present study, it is argued that the inference to be drawn is different, and that the requirement for content validity evidence is therefore less stringent. In the present study, the important factor is not whether the test measures any particular

performance domain, but whether the test measures whatever it measures equally for the CA and PP subjects. Validity for the purpose intended by the developer would be sufficient but not necessary to establish validity for the purpose of the proposed study. For the present study, it is submitted that the following procedure is adequate to establish validity: The investigator selected several tests from the curriculum level generally appropriate for second (or fourth) graders, as prescribed by Jostens product documentation manuals. The students' teachers then chose a test for which all items met the following criteria (from Crocker & Algina, 1986):

1. Appropriate subject matter--that is, the skill has been taught
2. Level of cognitive processing required is appropriate to the grade level
3. Appropriate stimulus (question) format
4. Appropriate mode of required response.

In sum, the present validation procedure differs from that of a "classic" content validation study in the following respects. First, rather than writing items and matching them one by one to a performance domain, an entire test was chosen based on the teacher's judgment that it appears to test appropriate material in an appropriate way. Each item was then compared with the test developer's list of skills tested to ensure there is a match with one of these skills. Second, rather than using a "panel of experts," a set of items was chosen by the child's own classroom teacher (in consultation with the investigator). Finally, there was not a summary of item-domain matches. An entire test was chosen (and many other entire tests were rejected).



It is important to remember that the focus of the present study is the mode of test administration. The issue of what particular math domains the test measures is not relevant. The key issue in validity (Crocker & Algina, 1986, Ch. 10; Messick, 1989) concerns the "usefulness of inferences drawn from test scores for a given purpose under a prescribed set of conditions" (Crocker & Algina, 1986, p. 238). It is argued that the procedure for test selection is adequate to ensure that inferences drawn about effects of mode of test administration are valid.

Reliability. KR-21s calculated on pilot data collected at the Edith Bowen Laboratory School at Utah State University were approximately .7 for grade 2 and .8 for grade 4. Considering that the test measures five separate skills, these seem reasonably high. Test-retest reliability has not been assessed, but it is expected that the skills assessed on these tests do not exhibit much random fluctuation over time. There is only one form of each test, so alternate form reliability cannot be assessed.

### Procedure

Pleasant Green was the site of all testing. PP students were tested in their classrooms, with the investigator administering the tests. While PP students were taking the test in their classrooms, their CA classmates were tested in the school's computer lab. The classroom teacher accompanied them to the lab, where they followed their normal "log on" procedure and were presented with the test.

Students were told they were going to take a short math test, that it would have 25 multiple-choice questions, and that it would not affect their grades. They were instructed not to review previous questions and not to change answers once

marked. Students were observed by the investigator during testing to be sure they adhered to these conditions. Only one item appeared on each page, which helped with monitoring and also served to make CA and PP conditions more similar (because the computer showed one item per screen). Students in the CA condition had paper and pencil available for computation.

After testing, students were asked not to talk with students from other classes about the test or anything they did, until all classes had been tested. After all testing was completed, students were debriefed as whole classes (as part of a mini-lesson on the scientific method).

### Analysis

Analyses were done with the General Linear Models procedure (PROC GLM) of the SAS software system (SAS Institute Inc., 1988). The general analytic procedure followed below is a series of comparisons of linear models, as recommended and described by several authors (e.g., Pedhazur, 1982, Ch. 10; Kleinbaum, Kupper, & Muller, 1988, Ch. 20). In general, to test the statistical significance of a particular independent variable, one tests the increment in the proportion of variance in the dependent variable accounted for ( $R^2$ ) when a model containing that variable is compared to one which does not contain it, using the formula (Pedhazur, 1982, p. 62; Jaccard, Turrisi, & Wan, 1990, p. 18; Kleinbaum et al., 1988, p. 156):

$$F = \frac{(\underline{R}_1^2 - \underline{R}_2^2) / (\underline{k}_1 - \underline{k}_2)}{(1 - \underline{R}_1^2) / (\underline{N} - \underline{k}_1 - 1)}$$

where  $\underline{R}_1^2 = \underline{R}^2$  for the model with more predictors (full model)

$\underline{R}_2^2 = \underline{R}^2$  for the model with fewer predictors (restricted model)

$\underline{k}_1$  = number of predictor vectors in the full model (1 for each continuous variable; Number of categories - 1 for each categorical variable)

$\underline{k}_2$  = number of predictor vectors in the restricted model

$\underline{N}$  = total sample size

and  $F$  has  $\underline{k}_1 - \underline{k}_2$  and  $\underline{N} - \underline{k}_1 - 1$  degrees of freedom.

With a simple two-group comparison, this approach is mathematically equivalent to a  $t$  test (or  $F$  test). The advantages of the linear models, or "regression," approach, are that (a) it enables one to test the effect of mode of administration, while controlling for the fact that the design is unbalanced (i.e., the numbers of subjects in each condition at each grade are not equal), and (b) it enables one to easily add other variables (e.g., score on the matching test) to the model, for additional statistical control (Cohen, 1968).

In addition to the tests for statistical significance, effect sizes (with PP considered the control condition) were also calculated for all PP vs. CA comparisons.

## RESULTS

The research question to be answered was: Under comparable testing conditions, do elementary school students who take an achievement test administered by computer achieve the same mean score as comparable students who take the same test by paper-and-pencil?

Means for each mode at each grade are shown in Table 1.

Table 1

Mean Scores, Standard Deviations, and Ns for Each Mode at Each Grade

Mode	Grade						Overall		
	2			4			Mean	SD	N
	Mean	SD	n	Mean	SD	n			
Paper	12.4	5.0	35	16.0	5.3	28	14.0	5.4	63
Computer	11.6	5.9	26	13.6	4.9	25	12.6	5.5	51
Overall	12.1	5.4	61	14.9	5.2	53	13.4	5.3	114

The first model tested had three predictors: mode (computer or paper-and-pencil), grade (2 or 4), and the joint effect, or interaction, between mode and grade. This model was compared to a model which contained only mode and grade as predictors.  $R^2$  for the three-predictor model was .092, compared to .086 for the two-predictor model. The incremental change in  $R^2$  for the third predictor, the interaction between mode and grade, was  $.092 - .086 = .006$ , meaning the interaction accounted for only 0.6% of the variance in scores. This was tested for statistical significance;  $F$

(1 and 110 df) = 0.62, which is not statistically significant at  $p = .05$ . Thus it can be concluded that the effect of mode, if any, is constant across both grades.

Consequently, in subsequent model comparisons the interaction sum of squares was pooled into the error term, as recommended by several authors (e.g., Pedhazur, 1982, p. 377; Kleinbaum et al., 1988, p. 468; Applebaum & Cramer, 1974).

The next test compared a model containing the predictors mode and grade to one containing grade only. This comparison shows the percent of variance accounted for by mode, controlling for grade. (Grade must be controlled for because the design is not balanced; i.e., the cells have unequal ns.)  $R^2$  for the model containing mode and grade was .086, while  $R^2$  for the model containing grade alone was .067; the incremental  $R^2$  was .019, with  $F(1,111) = 2.36$ ,  $p = .13$ . Mode, therefore, accounts for 1.9% of the variance in scores (a measure of effect size), an amount which is not statistically significant at .05. The standard mean difference effect size, computed on the mean for each mode (adjusted for grade), was -0.28.

The reader may recall from the review of literature that the mean effect size from the studies in which the CA test had more restrictive conditions was -0.59, while the mean effect size from studies in which conditions were comparable was -0.10. It was predicted that the effect size in the present study would be nearer the latter value, or, more precisely, "near zero." The obtained effect size of -0.28 was not expected, albeit the difference from zero is not statistically significant; thus chance cannot be ruled out as a cause.

Results of the analysis are summarized in Table 2.

Table 2

Summary of Analysis of the Relationship of the Independent Variable Mode to Scores

<b>Model</b>	<b>Predictor(s)</b>	<b>R<sup>2</sup></b>
1	Mode, Grade, Mode × Grade	.092
2	Mode, Grade	.086
3	Grade	.067

<b>Test</b>	<b>Result</b>
Model 1 vs. Model 2	$F(1,110) = 0.62$ n.s. $R^2$ Change = .006
Model 2 vs. Model 3	$F(1,111) = 2.36$ n.s. $R^2$ Change = .019

<b>Mode</b>	<b>Mean (Adjusted for Grade)</b>	<b>SD</b>	<b>Effect Size</b>
Paper	14.2	5.4	
Computer	12.6	5.5	-0.28

A score on the matching test, the Utah Core Assessment Series, was available for 87 of the 114 students in the study and was used to pair subjects before randomly assigning them to modes of administration. (The students without scores were randomly assigned.) In an attempt to gain additional precision in the analysis, a separate analysis was done for these students, with the test used for matching (hereinafter called pretest) entered in all models as an additional predictor.

The results of such an analysis must be viewed cautiously in this case, because the group of students with a pretest score available are not a random subset of the total sample; one could reasonably suppose, for example, that as a group they are

from less mobile families and have a lower rate of absences, just to name two possibilities. The means and standard deviations for this analysis are shown in Table 3; models tested and their associated  $R^2$  are in Table 4.

As Table 4 shows, the increment in  $R^2$  was very small and not statistically significant for each of the interaction terms (comparisons of Models 1 vs. 2, 1 vs. 3, 2 vs. 4, 3 vs. 4). Consequently, the interaction sums of squares were pooled into the error term, and Model 4 was compared to Model 5 in order to test the effect of mode. The increment in  $R^2$  when mode is added to a model containing grade only was .017, which was not statistically significant at .05 ( $p = .11$ ). The effect size on the group means adjusted for grade and pretest was  $-0.27$ . Overall, the results for the subgroup of students with pretest scores was nearly identical to those obtained on the total sample. Therefore, no further analyses were done on this subgroup; the remainder of the analyses in this report included all students tested.

#### Time as a Dependent Variable

In doing the literature review for this thesis, the author encountered no studies in which time to complete the test was included as a variable. However, in collecting pilot data for the present study, the investigator noticed that the children taking the test via computer took, on the average, less time to complete the test than those using paper and pencil. Consequently, the investigator decided to measure time to completion during testing. Means and standard deviations are shown in Table 5.

Table 3

Mean Scores, Standard Deviations, and Ns for Each Mode at Each Grade--Students with Pretest Scores Only

Pretest

Mode	Grade						Overall		
	2			4			Mean	SD	N
	Mean	SD	n	Mean	SD	n			
Paper	75.1	19.4	28	60.1	21.8	19	69.0	21.5	47
Computer	75.9	17.7	21	60.7	22.4	19	68.7	21.2	40
Overall	75.4	18.5	49	60.4	21.8	38	68.9	21.3	87

Posttest

Mode	Grade						Overall		
	2			4			Mean	SD	N
	Mean	SD	n	Mean	SD	n			
Paper	13.0	4.9	28	15.5	5.3	19	14.0	5.1	47
Computer	12.2	6.1	21	13.7	5.0	19	12.9	5.4	40
Overall	12.7	5.4	49	14.6	4.9	38	13.5	5.2	87



Table 4

Summary of Analysis of the Relationship of the Independent Variable Mode to Scores, with Pretest an Additional Control Variable

<b>Model</b>	<b>Predictor(s)</b>	<b><math>R^2</math></b>
1	Pretest, Mode, Grade, Mode $\times$ Grade, Mode $\times$ Pretest	.470
2	Pretest, Mode, Grade, Mode $\times$ Grade	.470
3	Pretest, Mode, Grade, Mode $\times$ Pretest	.468
4	Pretest, Mode, Grade	.468
5	Pretest, Grade	.451

<b>Test</b>	<b>Result</b>
Model 1 vs. Model 2	$F(1,81) = 0.009$ n.s. $R^2$ Change = .00006
Model 1 vs. Model 3	$F(1,81) = 0.24$ n.s. $R^2$ Change = .002
Model 2 vs. Model 4	$F(1,82) = 0.32$ n.s. $R^2$ Change = .002
Model 3 vs. Model 4	$F(1,82) = 0.08$ n.s. $R^2$ Change = .0005
Model 4 vs. Model 5	$F(1,83) = 2.60$ n.s. $R^2$ Change = .017

<b>Mode</b>	<b>Mean (Adjusted for Grade)</b>	<b><u>SD</u></b>	<b>Effect Size</b>
Paper	14.5	5.1	
Computer	13.1	5.4	-0.27

Table 5

Mean Time to Completion, Standard Deviations, and Ns for Each Mode at Each Grade

Mode	Grade						Overall		
	2			4			Mean	SD	N
	Mean	SD	n	Mean	SD	n			
Paper	12.5	5.2	35	13.2	3.8	28	12.8	4.6	63
Computer	8.3	4.0	26	10.3	5.0	25	9.2	4.6	51
Overall	10.7	5.1	61	11.8	4.6	53	11.2	4.9	114

Models tested and results of the analysis are shown in Table 6.

As Table 6 indicates, the interaction of mode and grade did not add much predictive power ( $R^2$  Change = .004), nor was this addition to  $R^2$  statistically significant ( $F = 0.49$ ). The test for the addition of mode, controlling for grade, revealed a sizeable effect ( $R^2$  Change = .135), which was statistically significant ( $F = 17.51$ ,  $p < .0001$ ). The standard mean difference effect size for mode was  $-0.79$ ; that is, the computer group on the average completed the test  $0.79$  standard deviation faster than the paper-and-pencil group.

This finding may be of import for at least two reasons. First, it is of some interest in its own right. Why should examinees work faster when tested by computer? This question will be explored in the Summary and Discussion section to follow. Second, it raises the obvious question of whether time to completion might be a moderator of the relationship between mode of administration and score. A

Table 6

Summary of Analysis of the Relationship of the Independent Variable Mode to Time to Completion

Model	Predictor(s)	R <sup>2</sup>
1	Mode, Grade, Mode × Grade	.152
2	Mode, Grade	.148
3	Grade	.013

Test	Result
Model 1 vs. Model 2	F (1,110) = 0.49 n.s.      R <sup>2</sup> Change = .004
Model 2 vs. Model 3	F (1,111) = 17.51 p < .0001      R <sup>2</sup> Change = .135

Mode	Mean (Adjusted for Grade)	SD	Effect Size
Paper	12.9	4.6	
Computer	9.2	4.6	-0.79

simple correlation revealed a statistically significant relationship between time and score ( $r = .38, p < .0001$ ). Consequently, further analyses were done.

Time as an Independent Variable

The variable time to completion was examined for possible effects as a predictor of scores, and as a moderator of the mode-score relationship. The question is, what happens to the mode-score relationship when time to completion is controlled (i.e., entered first into the regression equation)? To answer this question, models were created and tested as summarized in Table 7.

Table 7

Summary of Analysis of the Relationship of the Independent Variable Mode to Scores, with Time an Additional Control Variable

Model	Predictor(s)	R <sup>2</sup>
1	Time, Mode, Grade, Mode × Grade, Mode × Time	.200
2	Time, Mode, Grade, Mode × Grade	.200
3	Time, Mode, Grade, Mode × Time	.192
4	Time, Mode, Grade	.191
5	Time, Grade	.191
6	Time	.144

Test	Result
Model 1 vs. Model 2	$F(1,108) = 0.05$ n.s. $R^2$ Change = .0004
Model 1 vs. Model 3	$F(1,108) = 1.08$ n.s. $R^2$ Change = .008
Model 2 vs. Model 4	$F(1,109) = 1.19$ n.s. $R^2$ Change = .009
Model 3 vs. Model 4	$F(1,109) = 0.15$ n.s. $R^2$ Change = .001
Model 4 vs. Model 5	$F(1,110) = 0.01$ n.s. $R^2$ Change = .0001
Model 6	$F(1,112) = 18.83$ $p < .0001$ $R^2$ Change = .144

Mode	Mean (Adjusted for Grade)	SD	Effect Size
Paper	13.5	5.4	
Computer	13.4	5.5	-0.02

As Table 7 shows, neither interaction term (Mode  $\times$  Grade, or Mode  $\times$  Time) added much predictive power, either with the other interaction term already in the model (Model 1 vs. Model 2; Model 1 vs. Model 3), or added to the main effects alone (Model 2 vs. Model 4; Model 3 vs. Model 4). All incremental  $R^2$  values were very small and none were statistically significant. Likewise, the effect of adding mode to the model containing time and grade was small and not statistically significant (Model 4 vs. Model 5).

When one compares the results of the analysis in which time was an additional predictor (Table 7) to that in which time was ignored (Table 2), it appears that the effect of including time is to moderate the effect of mode. When time is ignored (Table 2), there is a computer effect size of -0.28 standard deviation, although this is not statistically significantly different from 0. When time is held constant (i.e., included as a predictor in all models), as summarized in Table 7, the effect of mode almost vanishes (effect size = -0.02). This is completely in accord with the original prediction of no mode effect.

In summary, the difference in mean scores between the CA and PP mode is not statistically significant at the conventional .05 level; however, the CA effect size of -0.28 is larger than expected. When time was statistically controlled, the difference between CA and PP modes disappeared (effect size = -0.02).

It should be noted in passing that time to completion alone accounts for 14% of the variance in scores (Model 6). The simple correlation of time with score is positive,  $r = .38$ ,  $p < .0001$ . Of course, entering time on the "predictor" or

independent variable side of the equation does not make it a causal variable. The logic of the present design does not permit causal attributions for any variables except the manipulated variable mode.

## SUMMARY AND DISCUSSION

The prediction in the present study was that the computer (CA) and paper-and-pencil (PP) groups would not differ in mean scores. It was predicted that not only would the scores not differ to a statistically significant degree, but that the effect size would be very nearly zero. This prediction was based on a review of prior studies in which test-taking conditions (specifically the ability to change answers and review items) were similar in the CA and PP conditions; the mean effect size for these studies, none of which used elementary-age students, was  $-0.10$ .

With respect to the above predictions, results from the present study are somewhat ambiguous. The difference in mean scores between modes is indeed not statistically significant at the conventional  $.05$  level; however, the CA effect size is  $-0.28$ , which is larger than expected.

Subjects in the CA condition completed the test much more quickly, on the average, than subjects in the PP condition. The difference was both sizable (effect size =  $-0.79$ ) and statistically significant at  $p < .0001$ . Because the length of time to complete the test was positively correlated with the score achieved ( $r = .38$ ), it was speculated that time might moderate the relationship between mode and score. Indeed, when time was statistically controlled, the difference in scores between CA and PP modes vanished (effect size =  $-0.02$ ).

These findings, of course, raise more questions than they supply answers. One question has to do with the direction of causality, which is far from obvious in this case. Does taking more time really cause higher scores, in the sense that taking

more time and care leads to fewer mistakes? Or, is the direction of causality the other way around? This could be the case, for example, if students who do not know how to solve a problem just make a random guess, which takes less time than working out a solution. Or is the relationship something more complex? As mentioned previously in the Results section, the logic of the design of the present study does not permit an answer, because time was not manipulated. Further studies, in which time per item is somehow experimentally controlled, are necessary.

A second question raised by the results of the present study is, why should CA students work so much faster? With respect to this question, the literature on computer testing yields no clues, so all one can do is speculate.

One possibility lies in the conditions to which the particular students used in this study are accustomed. The computer lab is a familiar environment for them. According to their principal, they spend about 45-60 minutes a week there, engaged in activities very much like the experimental situation. Indeed, to these students, the test they took in the present study was just another Jostens Unit Test, the kind they take quite often in computer lab.

By contrast, the PP condition might have seemed much more serious and evaluative to the students involved. First, it was more like the usual testing situation. These students understand a "test" to involve sitting at a desk using paper and pencil. Also, the present investigator--an unfamiliar person--administered the PP test, so that answer changing and item review could be monitored. This may have contributed to a sense among the PP students that they were being evaluated (even though they were



told the test would not count toward their grades), and caused them to answer more carefully, which was reflected in longer times and (somewhat) higher scores. One direct way to investigate this possibility would be to interview or survey the subjects after the test; an indirect, but possibly just as valid, method would be to see if the CA group made less use of the scratch paper provided to both groups.

A second possibility may lie in the findings of Beach (1989), who found that undergraduates who fill out an attitude scale on computer are more likely to give random responses than those using paper and pencil; Beach also found that CA subjects, whether they gave any random responses or not, reported being less careful about their responses than PP subjects. (Random responses were defined as nonsense responses, e.g., responding "true" to "I was born on February 30th.") Beach did not measure time to completion in his study, but his findings are consistent with both shorter times and lower scores on a CA achievement test than on a PP version. Beach attributed the increased tendency to respond randomly on a CA test to increased ease of response; that is, pushing a key (or, in the present study, clicking a mouse) is just physically easier than filling in a circle with a pencil. It requires less care, less thought, and (by implication) less commitment to the response. Another, related possibility is that some mouse-click responses were actually made accidentally. Future research in this area should investigate this possibility through postexperiment interviews. It would also be of interest to measure time to respond on questions answered incorrectly, as opposed to time on correct answers. (Unfortunately, the Jostens system does not permit this.) More generally, an obvious way to check if the

mode effect is in fact, more narrowly, a mode of response effect is to include an additional experimental condition in which subjects read items from a computer screen, but make responses with a pencil on paper.

Clearly, it is important to learn more about the role of time (or, to put it another way, speed) in computer testing. First, further CA vs. PP studies should be done with time measured, to see which, if any, of the relationships found in the present study can be replicated: i.e., CA students take less time, time is positively correlated with score, and time moderates the mode-score relationship. Second, experimental studies, with time manipulated, should be undertaken to attempt to establish whether time is causally related to score, and whether the causal relationship, if it exists, is the same for CA and PP administration.

In the meantime, the present study provides evidence that there is no effect of mode of administration, if time to completion is statistically controlled. Of course, in a practical testing situation, where examinees can work as quickly as they like, the mode effect could be quite real, even if mode of administration as such has nothing to do with it. Assuming no control for time, the results of the present study are ambiguous. There appears to be a small negative computer effect of -0.28 standard deviations, relative to PP scores--but the study was not powerful enough for the effect to be statistically significant, so chance cannot be ruled out as the source of the effect. However, even though the results of the present study did not attain statistical significance at the conventional .05 level, if the CA effect size of -0.28 is in fact an accurate population estimate, the implications are rather large, because of the current

prevalence of high-stakes testing throughout American education (e.g., for college admission, high school graduation, and career-ladder eligibility). One would certainly not be justified in blindly treating CA and PP scores as equivalent, based on the present results.

## REFERENCES

- American Psychological Association. (1986). Guidelines for computer-based tests and interpretations. Washington, DC: Author.
- Applebaum, M. I., & Cramer, E. M. (1974). Some problems in the non-orthogonal analysis of variance. Psychological Bulletin, 81, 335-343.
- Beach, D. A. (1989). Identifying the random responder. Journal of Psychology, 123, 101-103.
- Becker, H. J. (1992). Computer-based integrated learning systems in the elementary and middle grades: A critical review and synthesis of evaluation reports. Journal of Educational Computing Research, 8 (1), 1-41.
- Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. American Educational Research Journal, 5, 437-474.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.
- Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.
- Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed. pp. 443-507). Washington, DC: American Council on Education.

- Eaves, R. C., & Smith, E. (1986). The effect of media and amount of microcomputer experience on examination scores. Journal of Experimental Education, 55, 23-26.
- Glass, G. V., & Hopkins, K. D. (1984). Statistical methods in education and psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Granite School District. (1992, December). Report to parents: Special 1992 test score edition. (Available from Granite School District, 340 East 3545 South, Salt Lake City, UT, 84115).
- Harrell, T. H., Honaker, L. M., Hetu, M., & Oberwager, J. (1987). Computerized versus traditional administration of the Multidimensional Aptitude Battery--Verbal Scale: An examination of reliability and validity. Computers in Human Behavior, 3, 129-137.
- Harvey, A. L. (1987). Differences in response behavior for high and low scorers as a function of item presentation on a computer assisted test. Unpublished doctoral dissertation, University of Nebraska, Lincoln.
- Huba, G. T. (1988). Comparability of traditional and computer Western Personnel Test (WPT) versions. Educational and Psychological Measurement, 48, 957-959.
- Jaccard, J., Turrisi, R., & Wan, C. K. (1990). Interaction effects in multiple regression. Newbury Park, CA: Sage.
- Jostens Learning Corporation. (1989). Mathematics curriculum teacher handbook. San Diego: Author.

- Kerlinger, F. N. (1964). Foundations of behavioral research: Educational and psychological inquiry. New York: Holt, Rinehart and Winston.
- Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). Applied regression analysis and other multivariable methods (2nd ed.). Boston: PWS-KENT.
- Lee, J. A., & Hopkins, L. (1985, March). The effects of training on computerized aptitude test performance and anxiety. Paper presented at the Annual Meeting of the Eastern Psychological Association, Boston. (ERIC Document Reproduction Service No. ED 263 889)
- Lee, J. A., Moreno, K. E., & Sympson, J. B. (1984, April). The effects of mode of test administration on test performance. Paper presented at the annual meeting of the Eastern Psychological Association, Baltimore. (ERIC Document Reproduction Service No. ED 246 093)
- Llabre, M. M., Clements, N. E., Fitzhugh, K. B., Lancelotta, G., Mazzagatti, R. D., & Quinones, N. (1987). The effect of computer-administered testing on test anxiety and performance. Journal of Educational Computing Research, 3, 429-433.
- Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Muhlstein, A. (1991). Comparability of computer and paper-and-pencil scores for two CLEP general examinations (College Board Report No. 91-5, ETS RR No. 92-14). New York: College Entrance Examination Board.

- Mazzeo, J., & Harvey, A. L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature (Report No. 88-8). New York: College Entrance Examination Board. (Report No. RR 88-21). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 304 462)
- Messick, S. (1989). Meaning and value in test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5-11.
- Olsen, J. B., Maynes, D. D., Slawson, D., & Ho, K. (1989). Comparisons of paper-administered, computer-administered, and computerized adaptive achievement tests. Journal of Educational Computing Research, 5, 311-326.
- Pedhazur, E. J. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt, Rinehart, and Winston.
- Ronau, R. N., & Battista, M. T. (1988). Microcomputer versus paper-and-pencil testing of student errors in ratio and proportion. Journal of Computers in Mathematics and Science Teaching, 7(3), 33-38.
- SAS Institute Inc. (1988). SAS/STAT user's guide, release 6.03 edition. Cary, NC: Author.
- Taylor, M. J., & White, K. R. (1990, April). An evaluation of alternative methods for computing standardized mean difference effect sizes. Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Utah State Office of Education. (1988). Technical manual: Utah Core Assessment Series. Elementary Mathematics. Salt Lake City: Author.

- Ward, T. J., Hooper, S. R., & Hannafin, K. M. (1989). The effect of computerized tests on the performance and attitudes of college students. Journal of Educational Computing Research, 5, 327-333.
- Wise, S. L., Barnes, L. B., Harvey, A., & Plake, B. S. (1989). The effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. Applied Measurement in Education, 2, 235-241.
- Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computers. Educational Measurement Issues and Practice, 8 (3), 5-10.
- Wise, S. L., & Wise, L. A. (1987). Comparison of computer-administered and paper-administered achievement tests with elementary school children. Computers in Human Behavior, 3, 15-20.



APPENDICES

APPENDIX A  
GRADE 2 TEST

# GRADE 2 MATH

**NAME:** \_\_\_\_\_

Which makes 20?

$10 + 30$

$30 - 20$

$20 + 20$

$60 - 40$



Which makes 16?


$8 + 9$

$7 + 1 + 4$

$9 + 7$

$8 + 2 + 3$

There are 13 . A  eats 6.

How many  are left?

7


5

6

2

There are 17  . A  eats some.

There are 9  left.

How many did the  eat?

7

8

10

9

$$20 + 40 - \blacksquare = 10$$

What goes in the  $\blacksquare$  ?

20

90

40

50



Which makes 34?

$$\begin{array}{r} 34 \\ +10 \\ \hline \end{array}$$

$$\begin{array}{r} 26 \\ +12 \\ \hline \end{array}$$

$$\begin{array}{r} 23 \\ +11 \\ \hline \end{array}$$

$$\begin{array}{r} 34 \\ +20 \\ \hline \end{array}$$

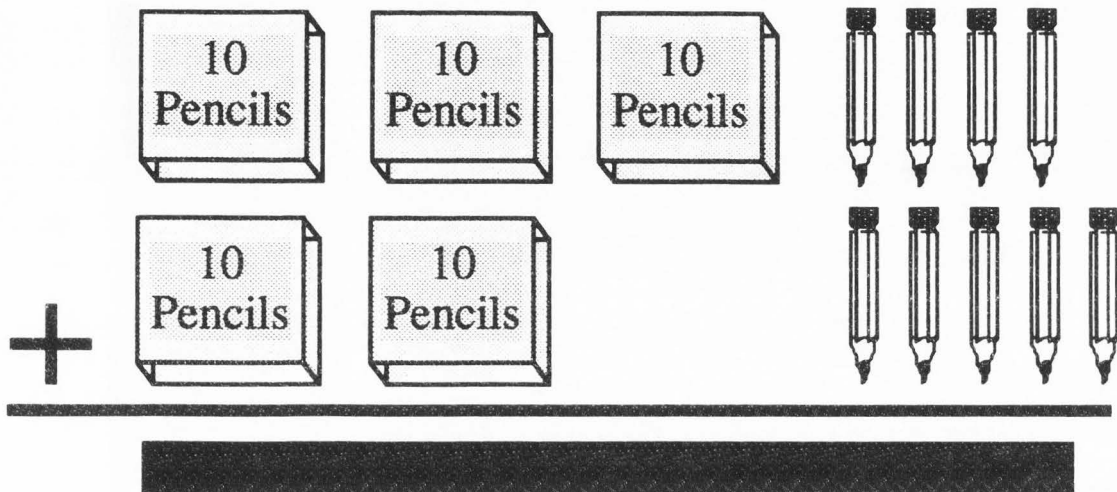
Which makes 22?

$$\begin{array}{r} 60 \\ - 20 \\ \hline \end{array}$$

$$\begin{array}{r} 87 \\ - 65 \\ \hline \end{array}$$

$$\begin{array}{r} 71 \\ - 51 \\ \hline \end{array}$$

$$\begin{array}{r} 42 \\ - 10 \\ \hline \end{array}$$



What goes in the ■ ?

57

39

24

59

$$\begin{array}{r} 86 \\ - 65 \\ \hline \blacksquare \end{array}$$

What goes in the  $\blacksquare$  ?

21

30

11

75

$$\begin{array}{r} 47 \\ + \blacksquare \\ \hline 87 \end{array}$$

What goes in the  $\blacksquare$  ?

47

30

19

40

	tens		ones	
+	3		6	
	1		4	
	4		10	= <span style="display: inline-block; width: 1em; height: 1em; background-color: black; vertical-align: middle;"></span>

What goes in the  ?

- 40
- 50
- 41
- 61

	tens	ones	
	5	6	
+	2	7	
	7	13	= <span style="display: inline-block; width: 1em; height: 1em; background-color: black; vertical-align: middle;"></span>

What goes in the  ?

93

73

101

83

$$38 + \blacksquare = 43$$

What goes in the  $\blacksquare$  ?

81

15

3

5



$$\begin{array}{r} 62 \\ + \blacksquare \\ \hline 70 \end{array}$$

What goes in the  $\blacksquare$  ?

12

8

4

11

Which does **not** need to be regrouped?

$$\begin{array}{r} 26 \\ + 27 \\ \hline \end{array}$$

$$\begin{array}{r} 34 \\ + 8 \\ \hline \end{array}$$

$$\begin{array}{r} 34 \\ + 45 \\ \hline \end{array}$$

$$\begin{array}{r} 26 \\ + 14 \\ \hline \end{array}$$

$$\begin{array}{r} \blacksquare^{12} \\ 72 \\ - 18 \\ \hline \end{array}$$

What goes in the  $\blacksquare$  ?

- 7
- 6
- 8
- 5

Which needs to be regrouped?

$$\begin{array}{r} 25 \\ - 14 \\ \hline \end{array}$$

$$\begin{array}{r} 35 \\ - 29 \\ \hline \end{array}$$

$$\begin{array}{r} 69 \\ - 10 \\ \hline \end{array}$$

$$\begin{array}{r} 25 \\ - 11 \\ \hline \end{array}$$

$$70 - \blacksquare = 61$$

What goes in the  $\blacksquare$  ?

9

11

31


7

$$\begin{array}{r} 61 \\ - 19 \\ \hline \blacksquare \end{array}$$

What goes in the  $\blacksquare$  ?

- 52
- 58
- 48
- 42

There are 24  . 9  run away.

How many  are still here?

15

14

11

9

Which addition matches the subtraction?

$$\begin{array}{r} 73 \\ - 15 \\ \hline 58 \end{array}$$

$$\begin{array}{r} 60 \\ + 13 \\ \hline 73 \end{array}$$

$$\begin{array}{r} 53 \\ + 20 \\ \hline 73 \end{array}$$

$$\begin{array}{r} 48 \\ + 15 \\ \hline 63 \end{array}$$

$$\begin{array}{r} 58 \\ + 15 \\ \hline 73 \end{array}$$



$$\begin{array}{r} 48 \\ - 1\blacksquare \\ \hline 35 \end{array}$$

What goes in the  $\blacksquare$  ?

- 2
- 3
- 6
- 5

$$\begin{array}{r} 3\blacksquare \\ +16 \\ \hline 52 \end{array}$$

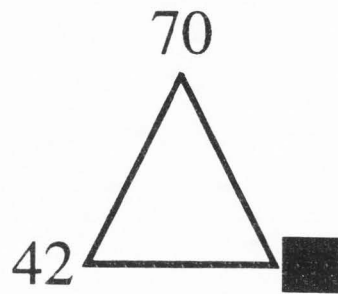
What goes in the  $\blacksquare$  ?


- 8
- 6
- 4
- 2

$$\begin{array}{r} 60 \\ + \blacksquare \\ \hline 78 \end{array}$$

What goes in the  $\blacksquare$  ?

- 8
- 14
- 18
- 10



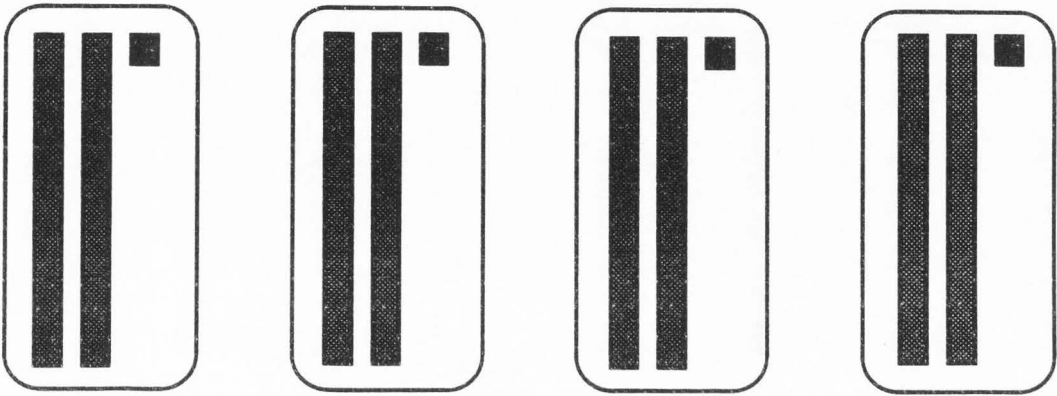
What goes in the  ?

- 32
- 28
- 18
- 30

APPENDIX B  
GRADE 4 TEST

**GRADE 4 MATH****NAME:**

-----



Which division matches the picture?

$8 \overline{) 4 \text{ tens } 4 \text{ ones}}$

$2 \overline{) 4 \text{ tens } 8 \text{ ones}}$

$4 \overline{) 8 \text{ tens } 4 \text{ ones}}$

$2 \overline{) 8 \text{ tens } 6 \text{ ones}}$

$$\begin{array}{r} 4 \text{ hundreds } 3 \text{ tens } 2 \text{ ones} \\ \times \quad \quad \quad \quad \quad \quad 2 \\ \hline \end{array}$$

Which is the same problem in **short form**?

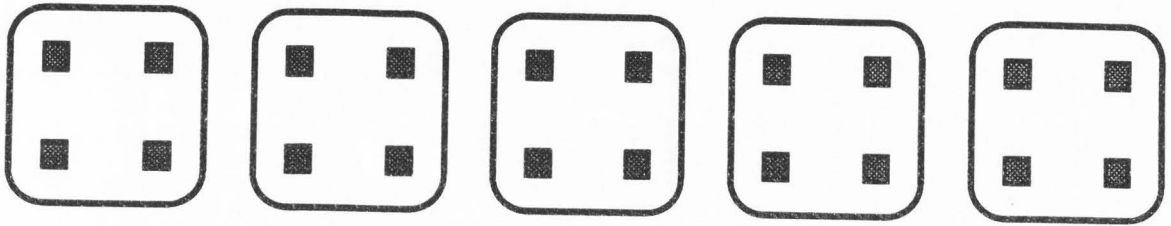
$$\begin{array}{r} 346 \\ \times 2 \\ \hline \end{array}$$

$$\begin{array}{r} 400 + 30 + 2 \\ \times \quad \quad \quad \quad \quad \quad 2 \\ \hline \end{array}$$

$$\begin{array}{r} 434 \\ \times 2 \\ \hline \end{array}$$

$$\begin{array}{r} 432 \\ \times 2 \\ \hline \end{array}$$





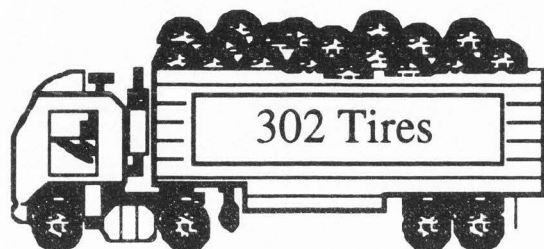
Which division matches the picture?

$4 \overline{)80}$

$6 \overline{)20}$

$5 \overline{)20}$

$2 \overline{)20}$



Which problem shows the number of tires on 4 trucks?

$$\begin{array}{r} 302 \\ \times 4 \\ \hline 1208 \end{array}$$

$$\begin{array}{r} 302 \\ \times 4 \\ \hline 1288 \end{array}$$

$$\begin{array}{r} 302 \\ \times 2 \\ \hline 604 \end{array}$$

$$\begin{array}{r} 302 \\ \times 4 \\ \hline 1248 \end{array}$$



Five children share 250 marbles.  
Each child gets the same amount.  
How many marbles will each child get?

- 5 marbles     15 marbles
- 50 marbles     25 marbles

Which answer must be regrouped?

$$\begin{array}{r} 1 \text{ ten} \quad 2 \text{ ones} \\ \times \quad \quad 4 \\ \hline 4 \text{ tens} \quad 8 \text{ ones} \end{array}$$

$$\begin{array}{r} 8 \text{ tens} \quad 1 \text{ ones} \\ \times \quad \quad 7 \\ \hline 56 \text{ tens} \quad 7 \text{ ones} \end{array}$$

$$\begin{array}{r} 9 \text{ tens} \quad 0 \text{ ones} \\ \times \quad \quad 5 \\ \hline 45 \text{ tens} \quad 0 \text{ ones} \end{array}$$

$$\begin{array}{r} 6 \text{ tens} \quad 3 \text{ ones} \\ \times \quad \quad 6 \\ \hline 36 \text{ tens} \quad 18 \text{ ones} \end{array}$$

Which is the same?

$$\begin{array}{r} {}^2 36 \\ 36 \\ 36 \\ + 36 \\ \hline 144 \end{array}$$

$$\begin{array}{r} 36 \\ \times 4 \\ \hline 124 \end{array}$$

$$\begin{array}{r} {}^2 36 \\ \times 4 \\ \hline 144 \end{array}$$

$$\begin{array}{r} {}^2 24 \\ \times 6 \\ \hline 144 \end{array}$$

$$\begin{array}{r} {}^2 48 \\ \times 3 \\ \hline 144 \end{array}$$

Which has the **wrong** answer?

$$\begin{array}{r} 30\overset{2}{6} \\ \times \quad 4 \\ \hline 1224 \end{array}$$

$$\begin{array}{r} \overset{1}{7}32 \\ \times \quad 4 \\ \hline 2928 \end{array}$$

$$\begin{array}{r} \overset{1}{2}03 \\ \times \quad 4 \\ \hline 812 \end{array}$$

$$\begin{array}{r} \overset{1}{3}30 \\ \times \quad 5 \\ \hline 1550 \end{array}$$

$$\begin{array}{r} \blacksquare 3724 \\ \times \quad 3 \\ \hline 11172 \end{array}$$

What goes in the  $\blacksquare$  ?

- 4
- 2
- 3
- 1

What will I see if I press  
these keys on my calculator?

$$5 \times 400 =$$

- 2000
- 200
- 2500
- 5400



$$6 \times 7 = 42$$

Which division matches this problem?

$42 \overline{)6^7}$         $42 \overline{)7^6}$

$6 \overline{)42^7}$         $3 \overline{)42^{12}}$

Which operation do we use to check division?

- addition
- subtraction
- division
- multiplication

$$4 \times \blacksquare = 28 \qquad \blacksquare \overline{) 28}^4$$

These two problems match.

What goes in the  $\blacksquare$  ?

7

28

6

3

Which division is wrong?

$$\begin{array}{r} 4 \\ 7 \overline{)30} \\ \underline{-28} \\ 2 \end{array}$$

$$\begin{array}{r} 4 \\ 4 \overline{)18} \\ \underline{-16} \\ 2 \end{array}$$

$$\begin{array}{r} 3 \\ 6 \overline{)24} \\ \underline{-18} \\ 6 \end{array}$$

$$\begin{array}{r} 5 \\ 9 \overline{)45} \\ \underline{-45} \\ 0 \end{array}$$

$$\begin{array}{r} 3 \\ 3 \overline{)902} \\ \underline{-\blacksquare} \end{array}$$

What goes in the  $\blacksquare$  ?

- |                           |                         |
|---------------------------|-------------------------|
| <input type="radio"/> 902 | <input type="radio"/> 9 |
| <input type="radio"/> 90  | <input type="radio"/> 3 |

$$\begin{array}{r} 2\blacksquare \\ 4 \overline{)92} \\ \underline{-8} \\ 12 \end{array}$$

What goes in the  $\blacksquare$  ?

- 3
- 4
- 16
- 2

$$\begin{array}{r} 14 \\ 6 \overline{)84} \\ - 6 \phantom{0} \\ \hline 24 \\ - \blacksquare \\ \hline 0 \end{array}$$

What goes in the  $\blacksquare$  ?

- 14
- 24
- 21
- 6



The total of the ages of these animals is 12 years. What is the average age for one of the animals?

- 3 years
- 2 years
- 4 years
- 8 years



$$\begin{array}{r} \blacksquare \\ 3 \overline{) 237} \\ \underline{-21} \phantom{0} \\ 27 \\ \underline{-27} \\ 0 \end{array}$$

What goes in the  $\blacksquare$  ?

- 97
- 73
- 237
- 79

Which equals 90?

$6\sqrt{120}$

$4\sqrt{480}$

$4\sqrt{360}$

$5\sqrt{200}$

Which has the wrong answer?

$4 \overline{)804} \begin{array}{r} 201 \end{array}$

$5 \overline{)2005} \begin{array}{r} 41 \end{array}$

$3 \overline{)126} \begin{array}{r} 42 \end{array}$

$8 \overline{)1608} \begin{array}{r} 201 \end{array}$

$$\begin{array}{r} 6 \blacksquare \\ 4 \overline{)276} \\ \underline{-24} \phantom{0} \\ 36 \\ \underline{-36} \\ 0 \end{array}$$

Which division shows the answer?

$4 \overline{)276} \begin{array}{r} 69 \end{array}$

$4 \overline{)276} \begin{array}{r} 79 \end{array}$

$3 \overline{)276} \begin{array}{r} 92 \end{array}$

$2 \overline{)276} \begin{array}{r} 108 \end{array}$

$$\begin{array}{r} \blacksquare \\ 2 \overline{) 112} \end{array}$$

What goes in the  $\blacksquare$  ?

- 46
- 56
- 61
- 66

Which division is wrong?

$$\begin{array}{r} \phantom{0} \circ 3 \overline{)312} \\ \underline{-3} \phantom{0} \\ \phantom{0} 1 \phantom{0} \\ \underline{\phantom{0} 0} \\ \phantom{0} 12 \\ \underline{\phantom{0} 12} \\ \phantom{0} 0 \end{array}$$

$$\begin{array}{r} \phantom{0} \circ 3 \overline{)195} \\ \underline{-18} \phantom{0} \\ \phantom{0} 15 \\ \underline{-15} \\ \phantom{0} 0 \end{array}$$

$$\begin{array}{r} \phantom{0} \circ 5 \overline{)495} \\ \underline{-45} \phantom{0} \\ \phantom{0} 45 \\ \underline{-45} \\ \phantom{0} 0 \end{array}$$

$$\begin{array}{r} \phantom{0} \circ 4 \overline{)816} \\ \underline{-4} \phantom{0} \\ \phantom{0} 41 \\ \underline{-32} \\ \phantom{0} 96 \\ \underline{-36} \\ \phantom{0} 0 \end{array}$$

$$\begin{array}{r}
 3402 \\
 \hline
 6 \overline{)20412} \\
 \underline{-18} \phantom{00} \\
 24 \phantom{00} \\
 \underline{-24} \phantom{00} \\
 1 \phantom{00} \\
 - \phantom{0} \underline{0} \\
 12 \\
 \phantom{0} \underline{12} \\
 \phantom{00} 0
 \end{array}$$

What is wrong with this division ?

- The answer should have 3 digits.
- The tens digit should be a 2.
- The thousands digit should be a 5.
- Nothing. The answer is correct.