DOCUMENT RESUME

ED 069 689 TM 002 140

AUTHOR	Orost, Jean H.
TITLE	Effects of Age and Familiarity of Examiner on Test Performance: A Draft.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RB-72-29
PUB DATE	Jul 72
NOTE	37p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Age Differences; *Examiners; Grade 3; Grade 6; Individual Tests; Intelligence Tests; Kindergarten Children; *Measurement Techniques; *Peer Teaching; Performance Tests; Predictor Variables; *Testing
IDENTIFIERS	Block Counting (Bussis and Chittenden); Block Sorting (Bussis and Chittenden); Wechsler Intelligence Scale for Children

ABSTRACT

Three third-grade, three sixth-grade, and three adult female examiners tested 108 kindergarten and third-grade girls, half of whom were familiar to them, on three individually administered measures. No differences in performance on any measure as a function of familiarity were found at either grade level. No differences by examiners of different ages were noted on the numerical test, while differences in favor of the third-grade examiners were found on the classification test ($p = .07$, n.s.) and on the Wechsler Intelligence Scale for Children (WISC) vocabulary subtest ($p = .01$). The effects of interpersonal and task-related variables were discussed, along with implications for peer instruction. (Author)

RB-72-29

RESEARCH BULLETIN

EFFECTS OF AGE AND FAMILIARITY OF EXAMINER ON TEST PERFORMANCE

Jean H. Orost

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
July 1972

EFFECTS OF AGE AND FAMILIARITY OF EXAMINER ON TEST PERFORMANCE

Jean H. Orost

Educational Testing Service

Abstract

Three third-grade, three sixth-grade, and three adult female examiners tested 108 kindergarten and third-grade girls, half of whom were familiar to them, on three individually administered measures.

No differences in performance on any measure as a function of familiarity were found at either grade level. No differences by examiners of different ages were noted on the numerical test, while differences in favor of the third-grade examiners were found on the classification test ($p = .07$, n.s.) and on the Wechsler Intelligence Scale for Children (WISC) vocabulary subtest ($p = .01$).

The effects of interpersonal and task-related variables were discussed, along with implications for peer instruction.

EFFECTS OF AGE AND FAMILIARITY OF EXAMINER ON TEST PERFORMANCE

Jean H. Orost[1]

Educational Testing Service

In conducting psychological and educational research, there is often little attention given to the effects which the combination or interaction between subject and examiner produces. In the individual testing situation, there are many factors interacting which could alter the results obtained from the testing. Some of these variables are age, race, sex, and physical appearance of both examiner and subject; situational variables such as room arrangement, rapport establishment, warmth or reinforcement of behaviors; and psychosocial variables such as anxiety, hostility, need for acceptance or achievement (Rosenthal, 1967).

In Anastasi's (1966) review of 25 years of ETS Invitational Conferences on Testing Problems, 1936-1964, there is not one reference to experimenter effects or bias. Since then, however, there has been a growing awareness of the effects on research results of such phenomena as the placebo effect, the Hawthorne effect, and the self-fulfilling prophecy (Coffman & Parry, 1968; Cook, 1967; Rosenthal, 1967; Rosenthal & Jacobson, 1968). The purpose of the present study was to ascertain the effects of child versus adult examiners and familiar versus unfamiliar examiners in individual testing situations.

Much of the literature on examiner effects, especially in the area of bio-social attributes, is in the field of survey research (Benney, Riesman, & Star, 1956; Erlich & Riesman, 1961; Hyman et al., 1954), where large numbers of interviewers and subjects make it more possible to vary such attributes as sex, race,

religion, age, etc., than is usually possible in laboratory experiments. Studies which examine sex, race, or SES effects often show conflicting results or interaction effects (Awkerman, 1970; Pelosi, 1968; Phillips, 1966; Sattler, 1970). Studies which are concerned with the interaction of various attributes, such as sex or social reinforcement, on performance of subjects at different ages (Gewirtz, 1954; Stevenson, 1961) suggest that children at various age levels perform more consistently with female than with male examiners. Sarason (1960) noted the prevalence of test anxiety among elementary school subjects. Although girls tended in general to exhibit more anxiety than boys, the effect of different anxiety levels on performance was not as pronounced as for boys.

Two studies have dealt with the interaction between interviewer's age and age of respondent. In one study by Erlich and Riesman (1961) adolescent girls under 15 gave fewer unacceptable responses to interviewers under 40 than to those over 40, while older girls responded more "unacceptably" toward younger interviewers. Benney, Riesman, and Star (1956) found that, in discussions related to sexual maladjustment, younger subjects (under 40) were more frank with younger interviewers. Rosenthal (1967) notes the dearth of research related to effects of experimenter's age. This may largely be due to the fact that, aside from survey organizations, most research is conducted within university type settings, where the age variation of available examiners is greatly restricted.

Several studies have dealt with the effects of examiner's acquaintance with the subject. Egeland (1969) found that knowledge of a child's past academic performance affected the scoring of the Comprehension and Similarities subtests of the Wechsler Intelligence Scale for Children (WISC) (those subtests most

susceptible to judgmental decisions) but that Vocabulary score was unaffected. Sacks (1952), administering intelligence tests to three-year-olds, found that they performed much better when there had been a period of 10 days (one hour per day) prior contact between examiner and child than when the examiner was a stranger. Rosenthal (1967, p. 88) reported several studies which indicated that the effects of prior contact depended on the nature of the task. With simple, repetitive tasks, an increase in anxiety level, which one would expect to accompany testing by a stranger, produced better performance. On the other hand, with difficult tasks, this same anxiety operated to depress performances.

Procedure

Third-grade, sixth-grade and adult female examiners (n = 9) individually tested 108 third-grade and kindergarten girls, half of whom they had known previously, on a half-hour battery consisting of three measures.

Measures. Three measures were used to assess competence in organizing quantitative concepts, skills, and language:

1. Block Counting task developed for Office of Education evaluation of Follow Through (Bussis & Chittenden, 1970; Chittenden, Amarel, Bussis, Orost, & Tanaka, 1970). Eight buildings constructed from half-inch wooden cubes. The child is asked how many blocks are in each building (range = 3 - 13).
2. Picture Sorting task piloted for OE evaluation of Follow Through, but revised for this study (Bussis & Chittenden, 1970). Sets of three or four cards which "belonged together"

were selected from an array of 30 two-inch cards showing color drawings of simple objects (see Figure 1).

3. Vocabulary subtest of the Wechsler Intelligence Scale for Children (Wechsler, 1949). The child is asked what each of thirty words means.

Insert Figure 1 about here

All tests selected had to meet the criteria of being appropriate for both kindergarten and third-grade subjects, simple enough for third graders to administer, and not over 10 minutes in length. It was anticipated that the counting task, which was the most straightforward and unambiguous in administration and task requirements, and on which perceived failure was rare, would be the least susceptible to interpersonal examiner influences. The classification task consisted of closed sorts in which the child was asked the reason for groupings of picture cards, plus some items on which she was asked to extend the category by selecting an additional picture to go with the others. It was anticipated that the subject's perception of the examiner's expectations for performance level might affect performance on this measure. The one standard measure, the WISC Vocabulary subtest, was chosen not only because of its high correlation to total WISC score and simple administrative format, but also because a verbal performance measure would be anticipated to show interpersonal examiner effects.

Examiners. The three third-grade and three sixth-grade girls used as examiners were selected by their teachers as above average achievers, while the three adults were self-selected from a pool of pupil aides employed in the school system.

Before beginning training, the experimenter administered all three measures to the third-grade and sixth-grade examiners. Training of examiners involved five two-hour training periods, including practice with first-grade children. Examiners were cautioned against discussing the tasks outside the testing period.

Examiners were trained to record responses for the Block Counting and Picture Sorting tasks on the answer sheets provided. However, the entire testing session was tape recorded as a verification precaution as well as to serve a monitoring function promoting task-oriented behavior. The WISC Vocabulary test was transcribed and scored directly from the tape, since it was felt that the young examiners would have difficulty coping with the writing load. Testing took place in large rooms in which one adult, one sixth-grade, and one third-grade examiner were testing simultaneously in different corners of the room. The experimenter was available, but busied herself in seeming inattentive to the testing in progress.

Data collection took place during a period of about one month in the spring. Because of distance and transportation problems, all third-grade subjects were tested before all kindergarten subjects. Examiners typically tested two or three subjects on a given day.

Subjects. All subjects and examiners were from one suburban and rural middle class school environment. A total of 54 kindergarten and 54 third-grade

girls were tested. Each examiner was assigned three third graders she knew. No examiner tested subjects from her own class. Inasmuch as none of the kindergarten subjects were previously known to the examiners, these subjects were randomly selected and then randomly assigned to the familiar or unfamiliar condition. Each examiner was then assigned three subjects in each condition.

Familiarity. On two different days before testing began in the kindergarten, the examiners visited the classrooms and spent the sessions becoming acquainted with the three "familiar" subjects, joining in play, snack, and game activities coupled with personal conversations. Those subjects assigned to the unfamiliar condition were not in the same class as the familiar ones for any given examiner.

Examiners had noted third-grade students they knew from class lists. In order to verify their familiarity or unfamiliarity with the subjects at the time of the actual testing session, examiners completed a Familiarity Questionnaire for all subjects they tested. The questionnaire established the nature, extent, and duration of acquaintanceship, if any.

Other data. After testing was completed for all subjects, nine subjects, one from each of the nine examiners, were interviewed by the experimenter concerning the extent of discussion about the tests which took place outside the testing situation.

Scores from the Iowa Test of Basic Skills, administered the week before this study began, were obtained for all third-grade subjects. No achievement measures were available for kindergarten subjects.

Analysis. A three-way analysis of variance for each of the three tests was employed to discover if there was any relationship between age of subject, age of examiner, or familiarity or their interactions.

Results

In addition to the primary hypotheses tested, several questions were posed. Among these were the intercorrelations among the three measures and their correlations with the Iowa test, which had been administered by the classroom teachers at about the time of the study. In addition, the results of testing the student examiners themselves prior to their training for the project were investigated and compared to subject test results. Contributions of individual examiners to the overall variances in test performance were compared with means for each primary category, i.e., examiner level and familiarity condition. Finally, interviews of a sampling of subjects were inspected for evidence of contamination outside the test situation.

One of the reasons for selecting the three measures used in the study was the disparity in the domains tapped by each. As seen in Table 1, the low correlation among the instruments substantiate this especially for third-grade subjects. Because of the low correlations among measures, analyses of variance were performed to test the various hypotheses separately for each of the instruments. By analyzing the results of such disparate instruments, the differential effects of examiner-induced bias might be seen to be operating more or less strongly with respect to instruments which placed different sorts of demands on the subjects as well as on the examiners.

Insert Table 1 about here

Block Counting. As expected, the third-grade subjects performed very well on the number task, which included a total of eight test items. The mean score for third-grade subjects was 5.74, with a standard deviation of 1.35. Kindergarten subjects achieved a mean score of 4.07 with a standard deviation of 1.55.

As seen in Table 2, the only difference of significance is the difference in performance by third-grade as opposed to kindergarten subjects, which was to be expected. There were no significant effects due to examiner level, familiarity, or any interaction effects.

Insert Table 2 about here

Picture Sorting. The classification measure was scaled to produce a total possible score of 32 points. The ranges for the two grade levels were nearly identical (0-29, 0-30), but the mean for third-grade subjects was 18.85 with a standard deviation of 5.77, while the mean for kindergarten subjects was 11.74 with a standard deviation of 6.40 ($p < .0001$). Subjects tested by third-grade examiners achieved a mean score of 17.14 (SD = 7.31), while sixth-grade examiners produced mean scores of only 14.89 (SD = 7.28). This difference in favor of the third-grade examiners proved to be significant at the .07 level, as can be seen from the analysis of variance results produced in Table 3. The interaction of subject grade level and examiner level proved to be insignificant.

Insert Table 3 about here

Subjects tested by examiners whom they did not know achieved a mean score of 16.07 (SD = 7.07), while subjects tested by examiners they knew achieved a mean score of 14.52 (SD = 6.98). While this difference is in the opposite direction from that which had been anticipated, the analysis of variance showed that this difference was not significant. When comparing the results of the two different types of familiarity conditions at the two grade levels, it was seen that here also there was no difference, as revealed by the lack of interaction effect of subject grade level by familiarity. All other interactions were found to be insignificant also. Therefore, all hypotheses failed to be rejected at the .05 level, although examiner level effects in favor of the third-grade examiners were detected.

Vocabulary. The first 30 vocabulary words from the WISC list of 45 words were used. Using the scoring system in the WISC manual, it was possible to receive one or two points of credit for each item, giving a total possible score of 60 points. Only three of the subjects were able to obtain any points past item number 23 and none past number 27.

The mean score for third graders was 20.50 (SD = 5.12), with a range from 9 to 30. This score is seen to be lower than the standard WISC scores for eight- and nine-year-old subjects with presumably average IQ's, which would be between 25 and 30 points. This discrepancy had been anticipated due to the fact that testers had not been trained to probe adequately. Therefore, reference to scaled scores was not made, and this test was only used in comparison to the others administered under similar conditions.

Kindergarten subjects had a tendency to fail to monitor their own performance. As opposed to third graders who generally admitted when they did not know the meaning for a word, the kindergarten subjects tended to keep responding

long past the point where they no longer knew any meanings. They began to give rhyming or associational responses or just random words in response to the words posed by the examiners. Using the same scoring criteria for these subjects, the mean Vocabulary score was 10.54 (SD = 4.80), with a range from zero to 23 compared to standard WISC norm of 15 to 17 points.

The analysis of variance showed, as seen in Table 4, that there were examiner level effects significant at the .01 level. This difference in performance was shown to be independent of grade level, as there was no interaction between examiner level and grade level.

By using the Scheffé method for detection of significant differences among means it was discovered that third-grade examiners produced significantly higher scores than either sixth-grade or adult examiners at the third-grade subject level. Although adult examiners produced higher scores than sixth-grade examiners at the kindergarten level, the difference was not significant at either subject level. There were no other significant effects, except between grade levels, as expected.

Insert Table 4 about here

A comparison of group means was made between the Picture Sorting and the Vocabulary scores. For both measures, the third-grade examiners produced higher mean scores than either the sixth-grade or adult examiners. On the vocabulary test, the mean scores for sixth-grade and adult examiners was nearly identical for third-grade subjects. On the Picture Sorting task, however, sixth-grade examiners' mean score was 18.61, compared with 20.77 for third grade and only 17.17 for adults. Although only the third-grade versus adult comparison was

significant, this result did represent a different trend than for the Vocabulary test.

When looking at the Picture Sorting means obtained by different examiner levels at the kindergarten level, it was found that the trend for sixth-grade examiners to obtain intermediate scores was maintained. On the Vocabulary test, however, this trend was reversed, and adult examiners' scores exceeded those of sixth-grade examiners. This reversal was investigated by consulting individual examiner means.

In Table 5 the significant relationships between examiner level and vocabulary scores can be further explored. For third-grade subjects, the individual examiner means at each examiner level ranged about four points, with the third-grade examiners having an overlapping but higher range than the sixth-grade or adult examiners, whose mean scores appear to be nearly identical. Third-grade examiner C obtained the highest means at both the third-grade and kindergarten levels, but this difference was not significant in either case.

Insert Table 5 about here

Familiarity. For the Picture Sorting and Vocabulary tasks, the mean scores indicated a difference in favor of examiners of unfamiliar subjects, but this difference did not approach significance ($p = .19$ for Picture Sorting and .45 for Vocabulary). It was also mentioned that there was no difference in scores attributable to the two different conditions of familiarity, i.e., experimentally induced in kindergarten subjects versus long-term acquaintance for third-grade subjects.

The mean scores obtained by the individual examiners for the conditions of familiar versus unfamiliar were examined to see if the results of individual examiners were different. As seen in Table 6, there was a great deal of individual variation among the examiners at all levels and for all tests. The third-grade and adult directions of difference remained the same for the Vocabulary and the Block Counting, while the sixth-grade directions of difference were the opposite for those two measures. All groups changed their patterns for the Picture Sorting task. It can readily be seen from this

---
Insert Table 6 about here
---

table that individual variations among the examiners at each level tended to wipe out any overall effect. Two examiners, third-grade examiner C and adult A were the only ones whose scores consistently and strongly favored the subjects they did not know. None of the other examiners showed consistent patterns across measures, and two of them (third-grade examiner A and sixth-grade examiner B) showed highly inconsistent patterns.

Iowa scores. The week before testing for this study began, the seven third-grade teachers from whose classes the subjects were drawn had administered the Iowa Test of Basic Skills. The total Iowa scores for third-grade subjects were analyzed and compared to data on the measures in this study.

As shown in the correlation matrix in Table 1, the correlations of Iowa scores with the Block Counting and Vocabulary tests were higher than the intercorrelations among those measures used in the study. The results from the Iowa tests were used as an external measure of pupil achievement not contaminated by the examiner bias under this study. The purpose of examining

the Iowa scores was to determine if the discovered difference in scores obtained by examiners at different levels could be explained by the fact that, despite random sampling, there was a difference in achievement level between the three groups.

The results of this inquiry are shown in Table 7. The analysis of variance indicated that there were no significant differences between the familiar and unfamiliar subgroups, and that there was no interaction effect between examiner level and familiarity. Thus, it was established that all subgroups were of approximately equal achievement level.

Insert Table 7 about here

Testing of examiners. All third-grade and sixth-grade examiners were given the three measures by the experimenter before being trained to administer them. It was found that, using the same administrative format and scoring system as that used for the study, the third-grade examiners achieved a mean score significantly higher than that achieved by third-grade subjects for each of the three measures. Whereas the mean Iowa total score for all third-grade subjects was found to be 39.98, the mean of the three third-grade examiners was 52, demonstrating that they were a select group. They had been motivated additionally by being selected by their teachers for this task and had a great deal of personal commitment to this "adult responsibility." Not only did the third-grade examiners excel over their third-grade peers on all three measures, but also surpassed the sixth-grade examiners who were not as select academically on two of the three measures.

In working with the two younger groups of examiners, it was apparent that although they all enjoyed the experience of working as examiners, the third

graders were much more concerned with the quality of their performance than with the amount of time they were being allowed to spend outside of class. There was some tendency, however, for the sixth graders' dedication to the task to wane during the latter part of the study, when the kindergarten subjects were being tested, as partially reflected in the diminishing of performance recorded for kindergarten subjects tested by sixth-grade examiners on the Vocabulary task, which involved the least amount of active examiner involvement.

The adult examiners exhibited a uniform and consistent dedication to the task of testing. They had all had experience in the school system as pupil aides, working with individuals and small groups of children, mostly in the primary grades. While not placed in a position of supervision over the two pupil examiners with whom they worked as a team, the adults nevertheless facilitated the testing by their concern with many administrative details. Their testing performance was generally relaxed and consistently supportive throughout the duration of testing.

Interviews. One third-grade subject tested by each of the nine examiners was randomly selected to be interviewed. Only two girls reported that classmates who had been tested before them said that they had been given different kinds of tests. None of the subjects reported having talked to any of the examiners before being tested, but two reported talking to the examiner about it afterward.

An interview with the teacher revealed that, though there had been much discussion in her classes about the project and its purposes, there had been practically no discussion, to her knowledge, about the specific tasks involved.

Discussion

By using tasks that showed low correlations with each other, which placed different cognitive demands on the subject, and which required different levels of active involvement on the part of both examiner and subject, one would have anticipated getting diversified results.

The most straightforward, uncomplicated task, Block Counting, showed no effects attributable to any of the variables under test. The Picture Sorting task, although it placed greater cognitive and productive demands on the subjects, also showed minimal effects due to examiner age or familiarity. The Vocabulary test alone showed significant effects due to examiner level, but not for any other investigated variable. In view of the planned diversity, then, it is somewhat surprising that, in many respects, the results appear to be quite similar. The effect of being confronted by an examiner in a formal testing situation, in a room apart, complete with answer sheets and tape recorders, must have been sufficiently and uniformly potent that it was able to override considerations of the acquaintance with the examiner outside the testing situation. Hartup (1964) had reported that friendship tended to inhibit task performance on simple, repetitive tasks among preschool peers. The more demanding requirements of the tasks used in the present study coupled with increased test wiseness among older subjects evidently eliminated such effects.

The most important finding of the study was the difference in performance on the Vocabulary test as a function of examiner level. Why did the third-grade examiners produce better results than either sixth-grade or adult examiners, regardless of either familiarity or grade level of subject?

Piagetian theory describes the young egocentric child as one who assumes that other people know what he knows, that everyone sees things from his perspective. Young children may also tend to give minimal responses to those they think are already in possession of the correct responses. There is no need to explain things to those you think already know what you mean. This, coupled with the anticipated tendency for children to feel freer to elaborate with another young child, led to the following investigation.

Since the examiners were not trained to probe restricted responses, those responses obtained on the protocols could be regarded as spontaneous self-regulated answers. Disregarding accuracy, was there any difference in the quantity of verbal output of subjects tested by examiners of different age levels? All but two of the 54 third-grade subjects attempted at least 25 of the 30 items, and 35 of the kindergarten subjects reached at least that point also. Kindergarten subjects who dropped out before that point were evenly divided between third-grade, sixth-grade, and adult examiners. The difference in number of items attempted, then, showed no examiner level effects. The protocols were then examined to discover the number of "don't know" or no answer responses, the number of one-word responses, and the number of responses which exceeded eight words in length. The results can be seen in Table 8.

These results show that third-grade examiners consistently obtained fewer restricted and more elaborated responses than either sixth-grade or adult examiners, who tended to be similar to each other, with differences slightly in favor of the adults. On a vocabulary test such as the WISC, elaboration of responses would in general lead to a greater likelihood of achieving a better score, which indeed turned out to be the case.

Insert Table 8 about here

The question remains, however, if being asked questions by a peer leads to better performance, why did this effect work differently for third and for sixth graders? Age interval cannot be used as an explanation, since the three-year age interval proved to be facilitative for third graders testing kindergartners yet detrimental for sixth graders testing third graders. Identical age facilitative effects would have shown third-grade examiner level effects but not kindergarten effects, which was not the case.

In the testing situations involved in this study, the levels of task orientation and personal-social interaction variables on the part of subjects and examiners showed some interesting effects. The subjects were uniformly task oriented, even at the kindergarten level. Anxiety, which tends to increase in the face of more complex tasks, was probably more potent on the Vocabulary test than for the other tests, since it was not only the hardest test, but the one in which perceived difficulty was most apparent. Third-grade and adult examiners tended to remain more task oriented over time than sixth-grade examiners. Also, it was in the nature of the test administration itself that the Block Counting and Picture Sorting tasks called for much more active examiner involvement than did the Vocabulary test. If an examiner were inclined to be less involved or less enthusiastic on any one test, the Vocabulary test was the easiest one to do this.

The subjects' perceptions of the examiner role were important as well as the examiners' perceptions of their own role. Anxiety levels would be expected to be relatively high when subjects were asked questions by adults who could

be both presumed to know the correct answer and who were probably viewed as teacher-like evaluators of their performance. The sixth graders, perhaps characteristic of girls of 11 or 12, attempted to model this adult teacher-evaluator role. Middle class third graders, on the other hand, are not customarily called on to fill this sort of role. The effect of giving them such responsibility was to increase task orientation and performance level, but not to lose their childlike enjoyment and active involvement with each girl tested as a unique experience. It may be that sixth graders were adopting more of the outward signs of an examiner without many of the ameliorating supportive attitudes of the adult examiners, thus generating more anxiety and consequently poorer test performance.

Some reinforcement to the notion of stereotyped adult role modeling on the part of upper grade children can be found in the Fox and Schwarz (1967) study of peer acceptance in a tutorial program. They found that pupils who had been involved in a tutorial program (in this case, fifth-grade Negro boys) were found to experience lowered peer acceptance and approval when returned to their regular classrooms. Perhaps they, too, had learned to adopt certain "establishment" modes of behavior, somewhat analogous to the teacher's pet syndrome, which were notwell accepted in the traditional classroom milieu.

Certain implications for the current growing trend toward peer instruction can be made. Several studies have shown that when students are used as tutors, the achievement of both the tutors and those they are tutoring shows gains (Allen, undated; Frager & Stern, 1970; Johnson, 1970; National Commission on Resources for Youth, 1969). Harrison et al. (1969), Lippitt and Lippitt (1970), and Niedermeyer (1970) describe the importance of adequate training for the

tutors. Lippitt and Lippitt note, "Protecting the student's self image as an 'expert' is easier if he is at least two grades ahead of the child he is helping [p. 137]." This very factor of "expertise" on the part of tutors may serve to inhibit rather than facilitate performance on the part of the child being tutored. This is not to say that tutors should not be carefully trained to perform their specific tasks, but that an air of superiority must be avoided. The Lippitts go on to say, however, that those students who had themselves experienced some learning difficulties, which was not the case in the present study, were often more sensitive and understanding of the difficulties of those they were helping. It might also be added that an informal classroom or an informal tutorial relationship, where the traditional rigid competitive or evaluative atmosphere is absent, would probably produce a more beneficial learning environment, leading to higher achievement levels. Institutionalizing or formalizing peer instruction, vesting it with an aura of superiority, may contribute to a diminishing of its effectiveness.

Summary and Conclusion

A total of 54 kindergarten and 54 third-grade girls were individually administered three cognitive tests. Nine examiners, three each from the third and sixth grades and three adults, tested these subjects. Results showed no differences in performance levels attributable to familiarity. Results due to the age of the examiner were insignificant except on the Vocabulary subtest of the WISC, where differences indicated that third-grade examiners produce better scores than either of the other ages of examiners for both kindergarten and third-grade subjects. All interactions between variables were insignificant.

These results have generated several questions which need to be investigated more fully. Inasmuch as two of the instruments used were not standardized, and the most pronounced effects were noted on the one standardized measure, several implications for further study could be drawn. Would these results be duplicated if other standardized tests were used, or was there some unique effect due to the subject matter or difficulty level of the test itself? Also, if it can be demonstrated that the research instruments not only tap aspects of cognitive functioning not closely related to the usual achievement or IQ measures, but are simultaneously less sensitive to interpersonal anxiety stresses in the testing situation, then they show great promise as means of assessing performance levels more accurately. It may be, however, that their apparent insensitivity was due, in the case of Block Counting, to a lack of diversity in item difficulty, and in the case of Picture Sorting, to an insensitive scoring system, which, while assessing correctness alone, neglected possible variations in use of certain types of categories and in amount of elaboration. It may later be found that subjects gave more elaborated or more egocentric responses to young examiners on the Picture Sorting task, similar to those findings on the Vocabulary test. Comparisons could be made concerning elaborated versus restricted responses in peer interactions if tasks required even greater levels of productivity or subject-examiner interaction than those used. (See Chittenden et al., 1970, for a description of interpersonal problem-solving tasks.)

Results from this study serve to confirm the theory that there are many variables interacting in the testing situation, some of which tend to operate to counteract the effects of others. That there were few significant effects found to be attributable to the variables investigated can partly be

explained by the very complexity of the interactions between individuals, which, in a study involving a relatively small sample and several variables, could only be reflected but not explained fully.

References

Allen, D. W. Students as teachers. Education cassette series No. 116, Instructional Dynamics Inc., Chicago (undated).

Anastasi, A. (Ed.) Testing problems in perspective. Washington, D. C.: American Council on Education, 1966.

Awkerman, G. L. Testing the effectiveness of auto-instruction in a paired learning arrangement. Paper presented at annual meeting of American Educational Research Association, Minneapolis, 1970.

Benney, M., Riesman, D., & Star, S. A. Age and sex in the interview. American Journal of Sociology, 1956, 62, 143-152.

Bussis, A., & Chittenden, E. A. Analysis of an approach to open education. Project Report-70-13, Princeton, N.J.: Educational Testing Service, 1970.

Chittenden, E. A., Amarel, M., Bussis, A., Orost, J., & Tanaka, M. Specifications of measures for assessing selected cognitive and affective characteristics of children. Interim report, 1970, OE Grant 0-9-526618-4748(100), Educational Testing Service.

Coffman, W. E., & Parry, M. E. Undesirable effects in research results. Unpublished (draft) report, Educational Testing Service, Princeton, N.J., 1968.

Cook, D. L. The impact of the Hawthorne effect in experimental designs in educational research, 1967, Project No. 1757, Contract No. OE-3-10-041, Washington: Office of Education, Bureau of Research.

Egeland, B. Examiner expectancy: Effect on the scoring of the WISC. Psychology in the Schools, 1969, 6(3), 313-315.

Erlich, J. S., & Riesman, D. Age and authority in the interview. Public Opinion Quarterly, 1961, 25, 39-56.

Fox, D. J., & Schwarz, P. M. Effective interaction between older and younger pupils in an elementary school "Peace Corps" project. Final report, 1967, New York, City University of New York School of Education.

Frager, S., & Stern, C. Learning by teaching. Reading Teacher, 1970, 23(5), 403-405, 417.

Gewirtz, J. L. Three determinants of attention-seeking in young children. Monographs of the Society for Research in Child Development, 1954, 19(2, Serial No. 59).

Harrison, G. V. et al. Training students to tutor. Los Angeles: University of California, 1969. (ED 038 329).

Hartup, W. W. Friendship status and the effectiveness of peers as reinforcing agents. Journal of Experimental Child Psychology, 1964, 1, 154-162.

Hyman, H. H. et al. Interviewing in social research. Chicago: University of Chicago Press, 1954.

Johnson, H. Pupils as teachers. Social Policy, Nov.-Dec. 1970, 14, 69-71.

Lippitt, P., & Lippitt, R. The peer culture as a learning environment. Childhood Education, 1970, 47(3), 135-138.

National Commission on Resources for Youth, Inc. Youth tutoring youth. Final report, 1969, No. DOL-43-7-001-34, New York, The Corporation.

Niedermeyer, F. C. Effects of training on the instructional behaviors of student tutors. Journal of Educational Research, 1970, 64(3), 119-123.

Pelosi, J. W. A study of the effects of examiner race, sex, and style on test response of Negro examiners. Doctoral dissertation, Syracuse University, 1968.

Phillips, J. The effects of the examiner and the testing situation upon the performance of culturally deprived children. Phase I--intelligence and language ability test scores as a function of the race of the examiner. Final report, 1966, Nashville, Tenn., George Peabody College for Teachers.

Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1967.

Rosenthal, R., & Jacobson, L. Pygmalion in the classroom. New York: Holt, Rinehart, & Winston, 1968.

Sacks, E. L. Intelligence scores as a function of experimentally established social relationships between child and examiner. Journal of Abnormal and Social Psychology, 1952, 47, 354-358.

Sarason, S. B. Anxiety in elementary school children. New York: Wiley, 1960.

Sattler, J. M. Racial "experimenter effects" in experimentation, testing, interviewing and psychotherapy. Psychological Bulletin, 1970, 73(2), 137-160.

Stevenson, H. W. Social reinforcement with children as a function of CA, sex of E, and sex of S. Journal of Abnormal and Social Psychology, 1961, 63(1), 147-154.

Wechsler, D. Wechsler Intelligence Scale for Children: Manual. New York: Psychological Corporation, 1949.

Footnote

[1]The author gratefully acknowledges the cooperation of Mr. Arthur Mitchell, of the Toms River, N.J., public schools, his staff, teachers, and pupils, without whom this study would not have been possible. Edward Chittenden, William Ward, and Ernest Washington also gave helpful reviews of an earlier version of this manuscript.

Table 1

Correlation between Measures

	Third Grade			Kindergarten	
	Vocab	Picture	Iowa	Vocab	Picture
Block C	0.2897	-0.0705	0.3070	.422	.408
Vocab		0.2792	0.4997		.585
Picture			0.1852		

Table 2

Block Counting

Analysis of Variance Table

Source	Sum of Squares	NDF	Mean Square	F Ratio
Total	2900.0000	108		
Mean	2600.9259	1	2600.9259	1166.7705
Subject Grade Level	75.0000	1	75.0000	33.6449*
Examiner Level	0.5741	2	0.2870	0.1288
SxE	3.1667	2	1.5833	0.7103
Familiarity	0.1481	1	0.1481	0.0665
SxF	0.1481	1	0.1481	0.0665
ExF	4.1296	2	2.0648	0.9263
SxExF	1.9074	2	0.9537	0.4278
Error	214.0000	96	2.2292	

* $p < .0001$

Table 3

Picture Sorting

Analysis of Variance Table

Source	Sum of Squares	NDF	Mean Square	F Ratio
Total	30570.0000	108		
Mean	25269.4815	1	25269.4815	688.4722
Subject Grade Level	1365.3333	1	1365.3333	37.1988 *
Examiner Level	202.3519	2	101.1759	2.7566 **
SxE	3.5000	2	1.7500	0.0477
Familiarity	65.3333	1	65.3333	1.7800
SxF	85.3333	1	85.3333	2.3249
ExF	1.0556	2	0.5278	0.0144
SxExF	54.0556	2	27.0278	0.7364
Error	3523.5556	96	36.7037	

* $p < .0001$.

** $p < .07$.

Table 4

Vocabulary

Analysis of Variance Table

Source	Sum of Squares	NDF	Mean Square	F Ratio
Total	31302.0000	108		
Mean	26009.0370	1	26009.0370	1096.6135
Subject Grade Level	2680.0370	1	2680.0370	112.9978 *
Examiner Level	218.2963	2	109.1481	4.6020 **
SxE	16.0741	2	8.0370	0.3389
Familiarity	13.3704	1	13.3704	0.5637
SxF	2.3704	1	2.3704	0.0999
ExF	25.4074	2	12.7037	0.5356
SxExF	60.5185	2	30.2593	1.2758
Error	2276.8889	96	23.7176	

* $p < .0001$.

** $p < .01$.

Table 5

Vocabulary

Means, Standard Deviations, and Ranges
by Individual Examiners

Subjects	N	Mean	SD(N-1)	Low	High
Third-Grade					
Third-Grade Examiner A	6	22.33	8.36	11.00	30.00
Third-Grade Examiner B	6	20.50	3.15	17.00	26.00
Third-Grade Examiner C	6	25.00	3.22	21.00	29.00
Sixth-Grade Examiner A	6	21.83	4.62	18.00	29.00
Sixth-Grade Examiner B	6	17.83	3.87	12.00	23.00
Sixth-Grade Examiner C	6	18.50	5.96	9.00	24.00
Adult Examiner A	6	18.83	6.40	14.00	28.00
Adult Examiner B	6	21.67	3.20	17.00	25.00
Adult Examiner C	6	18.00	2.90	14.00	22.00
Kindergarten					
Third-Grade Examiner A	6	12.67	5.01	5.00	18.00
Third-Grade Examiner B	6	10.17	5.46	0.00	16.00
Third-Grade Examiner C	6	14.00	7.46	4.00	23.00
Sixth-Grade Examiner A	6	10.00	1.41	8.00	12.00
Sixth-Grade Examiner B	6	10.83	2.40	8.00	13.00
Sixth-Grade Examiner C	6	5.33	4.37	0.00	10.00
Adult Examiner A	6	10.67	2.94	6.00	14.00
Adult Examiner B	6	11.17	3.66	7.00	16.00
Adult Examiner C	6	10.00	5.62	0.00	16.00

Table 6

Comparison of Familiarity Means by Individual Examiners

	Vocabulary			Block Counting			Picture Sorting		
	Not Fam.	Fam.	Diff.	Not Fam.	Fam.	Diff.	Not Fam.	Fam.	Diff.
Third-Grade Examiners									
A	15.5	19.5	+4.0	4.33	5.00	+ .7	16.0	15.83	- .2
B	14.67	16.0	+1.3	4.67	5.67	+1.0	17.33	16.67	- .6
C	21.67	17.33	-4.3	5.17	4.83	- .4	20.17	16.83	-3.4
Sixth-Grade Examiners									
A	16.5	15.33	-1.2	4.83	5.0	+ .2	17.33	16.00	-1.3
B	13.5	15.17	+1.7	4.83	3.67	-1.2	16.83	13.50	-3.3
C	12.83	11.00	-1.8	5.0	5.5	+ .5	12.67	13.00	+ .3
Adult Examiners									
A	16.83	12.67	-4.2	5.0	4.67	- .3	13.83	8.17	-5.6
B	16.17	16.67	+ .6	4.83	4.83	0	15.0	14.83	- .2
C	15.17	12.83	-2.4	5.83	4.67	-1.2	15.5	15.83	+ .3

Table 7

Iowa Test of Basic Skills

Analysis of Variance Table

Source	Sum of Squares	NDF	Mean Square	F Ratio
Total	89165.0000	54		
Mean	86320.0185	1	86320.0185	1670.1113
Examiner Level	68.2593	2	34.1296	0.6603
Familiarity	64.4630	1	64.4630	1.2472
ExF	231.3704	2	115.6852	2.2383
Error	2480.8889	48	51.6852	

Table 8

Number of Restricted Versus Elaborated Responses to Vocabulary Items

Subject Grade Level	Number of "Don't Know" or No Responses			Number of Single Word Responses			Number of Responses Over 8 Words		
	3	6	A	3	6	A	3	6	A
K	132	151	154	64	91	92	45	6	15
3	112	156	136	43	63	49	62	27	33

Figure Caption

Fig. 1. Cards used for the Picture Sorting task.