ED 255 559 TM 850 202

AUTHOR Fuchs, Douglas; Fuchs, Lynn S.

TITLE The ɔortance of Context in Testing: A Meta ıalysis.

PUB DATE Mar 85

NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).

PUB TYPE Speeches/Conference Papers (J50) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Disabilities; Effect Size; *Examiners; *Experimenter Characteristics; *Individual Testing; Interpersonal Relationship; *Meta Analysis; Research Methodology; Scientific Methodology; Socioeconomic Status; *Test Bias; Testing Problems

IDENTIFIERS *Context Effect; Familiarity; Positivism

ABSTRACT

This article presents a meta-analysis of the effects of examiner familiarity/unfamiliarity on children's performance during individual testing. Data came from 22 controlled studies involving 1489 subjects. In a typical study, the effect of examiner familiarity raised test performance by .35 standard deviations. Differential performance favoring the familiar examiner condition was greater when subjects: (1) were of low socioeconomic status; (2) were tested on comparatively difficult tests; and (3) knew the examiner for a relatively long duration. The relationship of familiarity to examinee's handicapped status was not clear. The effects of examiner familiarity demonstrate the importance of contextual factors in testing and question the positivistic view that the test instrument is the single most important variable determining test performance. (Author/BS)

The Importance of Context in Testing:

A Meta-Analysis

Douglas Fuchs and Lynn S. Fuchs

Peabody College, Vanderbilt University

Requests for reprints should be sent to Douglas Fuchs, Box 328,
Department of Special Education, Peabody College, Vanderbilt University,
Nashville, TN 37203.

Abstract

This article presents a meta-analysis of the effects of examiner familiarity on children's test performance. The data for the meta-analysis came from 22 controlled studies involving 1489 subjects. In the typical study, the effect of examiner familiarity raised test performance by .35 standard deviations. Differential performance favoring the familiar examiner condition was greater when subjects (a) were of low SES status, (b) were tested on comparatively difficult tests, and (c) knew the examiner for a relatively long duration. Implications are discussed for scientism, the popular epistemological basis for understanding testing, and for practice.

The Importance of Context in Testing: A Meta-Analysis

Positivism, or scientism, is the epistomological basis for the mainstream tradition in the social sciences (Adorno, Albert, Dahrendorf, Habermas, Pilot, & Popper, 1976; Bernstein, 1978). The positivistic ideal is the formulation of universal laws, which are free of the restraints of particular contexts, and therefore applicable to all. Hence, limiting, if not eliminating, contextual influence is a key feature of our standard methods of experimental design, measurement, and statistical analysis (Mishler, 1979).

Scientism also appears to govern the manner in which we administer tests, as well as our understanding of what occurs during testing. Evidence for this may be found in the most recent draft of the Joint Technical Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1984), where, on page 1, the test situation is described as a formal experiment. This perspective requires the examiner (i.e., unbiased investigator) to administer the test instrument according to explicit non-varying instructions (i.e., experimental treatment) in a controlled setting (i.e., laboratory). As in all scientific endeavors, these attempts to objectify and standardize the test situation are made, in part, to isolate the variable of interest, the test, from other contextual or situational variables. By promoting the independence and importance of the test instrument, we attempt to demonstrate a cause and effect relationship between test performance and whatever examinee characteristic the test claims to measure.

It is a fundamental presumption of the positivistic perspective that we may conceptualize the test setting in this "decontextualized" manner; that extra-test factors can be controlled, their effects on performance neutralized. Specific related assumptions concerning the behavior of test participants are that (a) the examiner-examinee relationship is static, unidirectional, and predictable, with the examiner controlling the testing by manipulating materials, questions, and feedback, while the examinee passively observes and responds; (b) examiners objectively and reliably administer the instrument and score performance; (c) test developers and test participants share similar interpretations of important elements of testing, such as the purposes of testing and the meaning of test instructions; and (d) the examinee attends to variables in the test setting accorded importance by test constructors and examiners, and ignores those stimuli to which examiners and developers assign scant importance.

It is testimony to positivism's powerful influence on testing that these assumptions infrequently have been explored. Nevertheless, a growing corpus of empirical studies calls these assumptions into question. First, this research suggests that examiners and examinees participate in dynamic, bi-directional, and idiosyncratic relationships, resulting in unpredictable behavior (Fuchs, Zern, & Fuchs, 1983; Mehan, 1978; Roth, 1974). Second, examiners' scoring may be influenced by pretest information on examinees (Bahad, Mann, & Mar-Hayim, 1975; Fiscus, 1975; Hersh, 1971; Schroeder & Kleinsasser, 1972), as well as by examinee characteristics (Fuchs & Fuchs, 1984; Masling, 1957). Third, test performance can be affected: (a) by examinees' interpretation of the purpose of testing (Deyhle, 1983; Goodnow, 1976), comprehension of test instructions (Abramyan, 1977; MacKay, 1974; Mehan, 1978), anxiety (Sarason, 1980), and pretest contact with examiners

(Fuchs, Fuchs, Power, & Dailey, in press); and (b) by examiners' personality (Exner, 1966; Feldman & Sullivan, 1971; Sacks, 1952), reinforcement (Ayllon & Kelly, 1972; Taylor & White, 1981; Tiber & Kennedy, 1964), attitudes about the legitimacy of testing (Horne & Garty, 1981), the order in which they administer tests of varying difficulty (Zigler & Butterfield, 1968), and their choice of test location (Labov, 1973; Seitz, Abelson, Levine, & Zigler, 1975; Stoneman & Gibson, 1978).

Such findings challenge positivism's decontextualized view of testing, and simultaneously corroborate a competing notion that contextual variables, including test participants' unique experiential backgrounds, mediate between the test instrument and performance. Comparative research in cognition (see Cole & Means, 1981) corroborates this idea and suggests further that various groups of examinees may respond differently to contextual variables in assessment. If this were true, then situational factors systematically may enhance the performance of certain groups and/or consistently depress the performance of others. In such cases, situational variables would represent systematic sources of error or bias.

Despite the possibility and importance of such an occurrence, this type of test situation, or test procedure, bias generally has gone unexplored (Flaugher, 1978). One of the few exceptions has been the issue of the effects of examiner unfamiliarity on test performance. Interest in this facet of the test procedure probably has been spurred by one or more of the following. First, examiner unfamiliarity often has been perceived as an important and desirable characteristic of standard testing (cf. Standards for Educational and Psychological Tests, 1974), thereby making it a conspicuous component of the test procedure. Second, and in apparent contradiction, there is a longstanding developmental notion that, because children derive much of their comprehension and feeling about a situation from significant adults in that setting (Freud, 1921/1922; Piaget, 1965), examiner attributes, as well as behaviors, are pivotal to examinee performance. Finally, psychological research into related but substantively different areas, such as the effectiveness of adults' social reinforcement on children's performance (cf. Stevenson, 1965), has demonstrated indirectly the impor. nce of the tester's familiar ty/ unfamiliarity.

Nevertheless, there has been no previous quantitative integration of the effects of examiner unfamiliarity on children's performance. Therefore, the purpose of the present study was to conduct a meta-analysis on this topic, specifically focusing on whether examiner unfamiliarity exerts a bias against select subgroups, such as low-SES and handicapped children.

Methodology

Search Procedure

The search for pertinent studies comprised a five-step procedure. First, employing the Thesaurus of Psychological Index Terms (APA, 1982), multiple descriptors were generated for key topic-related terms. For example, rapport alternately was identified by "examiner-examinee interaction," "interpersonal factors," and "situational factors." Second, in June 1982, the descriptors facilitated a computer search of three on-line data bases: ERIC (Educational Resources Information Center, from 1966); Psych Info (Psychological Abstracts Information Service, from 1967); and Dissertation Abstracts International (from 1927). Following Dusek and Joseph (1983), the descriptors were entered into the computer as isolated words or phrases to promote a comparatively broad search.

Third, employing similar key descriptors, a manual search was conducted of 12 educational, psychological, and speech/language journals for the years 1965-1982, inclusive. (If a journal began publication after 1965, all of its volumes were explored.) These journals were: *American Journal of Mental Deficiency*, *Child Development*, *Developmental Psychology*, *Exceptional Children*, *Journal of Abnormal and Social Psychology*, *Journal of Consulting and Clinical Psychology*, *Journal of Experimental Child Psychology*, *Journal of Genetic Psychology*, *Journal of Speech and Hearing Disorders*, *Language, Speech, and Hearing in the Schools*, *Merrill Palmer Quarterly*, and *Psychology in the Schools*. Fourth, the reference sections were explored for selected textbooks on psychological and educational assessment, such as Sattler's (1974) *Assessment of Children's Abilities*. Finally, titles in the references of investigations discovered by these efforts were pursued.

Criteria for Relevant Studies

A study was considered for inclusion if it compared examiner familiarity to unfamiliarity in terms of effects on examinees' performance during individualized testing. For reasons discussed by Cooper (1982), "familiarity" was defined broadly, including either children's personal acquaintanceship with the examiner or their prior contact with a rather well-defined class of adults, such as white middle-class females, of which the examiner was a member. "Test performance" was defined as examinees' performance on one or more IQ, speech/language, or educational achievement test, or on experimental tasks meant to simulate test items found in such measures. This definition of test performance helps to distinguish the studies in the present review from those that describe determinants of children's responsiveness to adults' social reinforcement (cf. Stevenson, 1965). In similar fashion to some of the investigations under review, the social reinforcement literature explores the effects of negative, positive, and an absence of prior contact with an experimenter on children's performance. However, these studies typically employ persistence and/or rate of performance on relatively simple motoric tasks, such as marble dropping (cf. Stevenson & Kennedy, 1966) or underlining Ss (e.g., Rosenkrantz & Van De Riet, 1974). We believe such tasks are fundamentally different from the more complex and demanding requirements in IQ, speech/language, and educational achievement assessments, and probably contribute to a qualitatively different experience for test participants. The resulting sample included 24 studies of the effects of examiner familiarity/unfamiliarity on children's test performance.

Data Extracted from Each Study

The effects of examiner familiarity and examiner unfamiliarity were noted in each study. Effects for five studies were unclear and, in each case, an attempt was made to obtain additional information from the investigator. One researcher could not be reached and one did not respond, reducing the sample from 24 to 22 studies (see the Appendix). Many of the 22 studies reported more than one effect. In such instances, each effect was coded separately. In all, the 22 studies yielded 38 effects of examiner familiarity/unfamiliarity.

Effects of examiner fariliarity and unfamiliarity were related to one composite procedural variable and nine substantive variables. The composite procedural variable indicates the overall methodological quality of each investigation. It was based on an aggrεgation of nine design-related characteristics. These methodological characteristics, as well as the standards against which they were judged to generate an overall quality index, follow:

1. Assignment of subjects to examiners. It was necessary for subjects to be assigned randomly to examinees.

2. Assignment of subjects to treatments. Investigators were required to assign subjects randomly to experimental conditions, or to use a repeated measures design.

3. Examiner expectancy. Researchers were expected to insure that examiners were blind to the general experimental questions and, specifically, to the familiar/unfamiliar nature of the test conditions.

4. Fidelity of treatment conditions. Investigators employing a personal acquaintanceship definition of familiarity were required to make explicit that unfamiliar examiners were strangers to examinees and that examiner familiarity either represented a long-term acquaintanceship between test participants or was the resultant of an experimentally-induced procedure.

5. Multiple treatment effects. Studies were evaluated as acceptable when effects of the familiar/unfamiliar examiner conditions did not appear to be confounded with other factors, such as the gender of familiar and unfamiliar testers.

6. Number of examiners. It was judged important that there be a minimum of two familiar and two unfamiliar examiners.

7. Order of testing. Studies employing a repeated measures design were required to counterbalance 'esting in familiar and unfamiliar examiner conditions.

8. Scoring. It was necessary that scores be calculated by a blind procedure.

9. Technical adequacy of dependent measure. At a minimum, a study was expected to use measures with indices for internal or test-retest reliability exceeding .69.

Interrater agreement on each of these dimensions, based on two rat rs' scores on six randomly selected studies (26% of the sample), ranged from .67 to 1.00. Average agreement across all nine methodological characteristics was .83.

The substantive variables noted in each study included the following:

1. Duration of treatment. This refers to the amount of time in which either (a) examiners and examinees became personally acquainted or (b) examinees became familiar with a type of examiner. We stratified the duration of the acquaintanceship period into five levels, ranging from less than 16 minutes to more than 20 hours. This stratification does not distinguish between long-term familiarity (such as exists between teacher and student) and experimentally-induced familiarity.

2. Examiners' professional familiarity with subjects. Examiners were classified as "professionally familiar" with subjects if they had previous experience with a type of child of which subjects were exemplars. Examiners were identified as "professionally unfamiliar" if they had no prior experience with a group of children of which subjects were members.

3. Examiners' training. A distinction was made between examiners who were trained formally as professional testers (e.g., school psychologists and speech clinicians) and those who were not (e.g., classroom teachers and mothers).

4. *Familiarity-inducing activity*. This refers to whether the examiner interacted with or simply observed the examinee during the familiarizing phase of the study. Long-term acquaintanceship always was defined as interactional in nature.

5. *Handicapped status*. Subjects were identified as either handicapped or nonhandicapped. No distinction was made with respect to specific categories of exceptionality (e.g., mental retardation vs. learning disabilities) or to degree of handicapping condition (e.g., mild vs. profound).

6. *Subjects' CA*. Subjects' CA, ranging from 2 to 16 years, was converted into months and treated as a continuous variable.

7. *Subjects' SES*. Initially, subjects' SES was classified in terms of either (a) poverty level, (b) mix of poverty level and working class, (c) middle-class, or (d) upper middle-class. For purposes of analysis, *a* and *b* were collapsed, as were *c* and *d*, creating two SES categories: low and high.

8. *Test location*. Location was classified as either familiar or unfamiliar to the examinee.

9. *Type of test*. Dependent variables were classified as IQ tests, speech/language tests, or isolated tasks, which were taken from, or created to closely resemble certain dimensions of IQ, speech/language, or educational achievement tests.

As a reliability check, two raters independently coded the nine substantive characteristics in six randomly selected studies (26% of the sample). Interrat eement for each of the study features ranged from .67 to 1.00. Average ient across all nine substantive variables was .93.

Characteristics of the Sample

Of th. 22 investigations included in this review, 18 were published studies and 4 were unpublished studies. Among the pul.lished articles, 17 appeared in 14 different journals; 1 study was published in a book. Three of the 4 unpublished investigations were doctoral dissertations; 1 study was included in the proceedings of a conference. Nineteen of the 22 studies were dated after 1970; the earliest was dated 1929. Also, 19 of the 22 studies defined examiner familiarity in terms of an examinee's personal acquaintanceship with the examiner; in 3 investigations examinees became familiar with a type of examiner, of which their eventual tester was an exemplar. Among the 19 investigations employing a personal acquaintanceship definition of familiarity, examiners and examinees were long-term acquaintances in 8 studies, familiarity was experimentally induced in 10 investigations, and, in 1 study, the procedure facilitating personal familiarity was unclear. A total of 1489 subjects particioated in these studies. Thirty-two percent of the subjects were male; 30% were female. Researchers did not report the sex of 38% of the subjects.

Results

Overall Effects

Results of the 22 studies were combined to provide three interrelated aggregate descriptions of the effects of examiner familiarity: unbiased effect size, percentage of distribution nonoverlap, and meta-analytic Z.

Unbiased effect size. A mean effect size was derived by determining the standard mean difference between examinees' scores in the familiar and unfamiliar examiner conditions and dividing this difference by the standard deviation of the examinees' scores in the unfamiliar condition (see Glass, McGaw, & Smith, 1981). Before averaging effect sizes, each one was converted to an unbiased effect size (UES) to correct for the inconsistency in estimating true from observed effect sizes (Hedges, 1981). The mean difference between the biased and unbiased effect sizes was small ($\overline{X}$ = .019, SE = .005), as has been demonstrated elsewhere (e.g., Bangert-Drowns, Kulik, & Kulik, 1983). Nevertheless, the UES was employed in all analyses to insure mathematical tractability of the data. For purposes of analysis, an effect was given a positive sign if examinees achieved higher scores in the familiar condition.

For 32 of 38 effect sizes in the sample, examiner familiarity had a positive impact on test performance; 6 effect sizes indicated the effect of examiner familiarity was negative. The average UES was .35 (SD = .47; SE = .076), $\underline{t}(37) = 4.67$, $\underline{p} < .001$.

Percentage of distribution nonoverlap. The percentage of distribution nonoverlap, or U_3 statistic (Cohen, 1977), denotes the percentage of the group with the smaller mean that is exceeded by 50% of the people in the larger-meaned group. The U_3 statistic indicated that the upper 50% of the distribution of scores in the familiar examiner condition exceeded 64% of the distribution of scores in the unfamiliar examiner condition. Given an IQ test with a population mean of 100 and a standard deviation of 15, the use of a familiar examiner would raise the typical score from 100 to 105.25, or from the 50th to approximately the 64th percentile.

Meta-analytic Z. Results from the 22 studies were combined to determine the unweighted Stouffer meta-analytic $\underline{Z}_{ma}$ (Rosenthal, 1978). This statistic permits computation of the probability that the combined effect of children's greater performance in the familiar examiner condition would occur by chance. It was derived by changing the p values of all effects to z scores, summing them, and dividing this sum by the square root of the number of studies included. When calculating a z score for studies in which multiple dependent variables were analyzed, a median p value was calculated for each study and its associated z score was used in the meta-analysis (see Rosenthal & Rubin, 1978). The resulting Z_{ma} was 7.20, $p < .001$.

Credence in a statistically reliable meta-analytic Z may be compromised by the suspicion that researchers do not report nonsignificant results (Greenwald, 1975). Rosenthal (1979) described a method for determining the number of unreported null effects that would be needed to reduce a meta-analytic Z to nonsignificance. The larger this "fail-safe N," the more confidence one can have in the reliability of a meta-analytic result. This investigation's fail-safe N was 418. As a rule of thumb, Rosenthal suggested that a meta-analytic Z be regarded as resistant to the "file drawer problem" of unreported null results if the fail-safe N exceeds 5K + 10, where k is the number of reported effects. In the current study this requisite number was 205. Thus, the fail-safe N of 418 was more than twice as large.

Relation between UES and Study Characteristics

Methodological quality of studies. The methodological quality of each of the 22 studies was quantified employing a four-step procedure. First, every investigation was analyzed in terms of the nine design-related characteristics and criteria described above. These design features were coded acceptable (0 points) unacceptable (1 point), or not applicable. As mentioned, the mean interrater agreement for the codings across the nine methodological characteristics was .83. Second, a weight of 1 or 2 was assigned to each methodological characteristic. "Technical adequacy of dependent measure," "assignment of subjects to treatments," and "assignment of subjects to examiners" received a weight of 2; the remaining six design characteristics received a weight of 1. Third, a composite score was generated for each study by multiplying the coded values (0 or 1) by the assigned weights (1 or 2), summing these products, and then dividing the sum by the number of applicable study characteristics. Finally, a frequency distribution of these composite scores was generated. It indicated that 55% and 45% of [illegible] investigations received composite scores above .7 (low quality) and below [illegible] (high quality), respectively.

Twenty-one effect sizes were assigned the status of low quality, with an average effect size of .51 (SD = .50); 17 effect sizes were assigned the status of high quality, with an average effect size of .17 (SD = .37). The correlation between the studies' quality ratings and UESs was -.38 (p < .05).

Substantive features of studies. Analyses were conducted to determine whether substantive features of the studies mediated the findings of the meta-analysis. Correlations were run to determine which of the substantive variables were related to examiner familiarity outcomes. Table 1 displays the means and standard deviations of the UESs, and correlations of the UESs with the nine substantive features coded in the meta-analysis.

Insert Table 1 about here

Three of the 9 substantive variables correlated significantly and moderately with UES: Duration of Familiarity, SES, and Type of Test (see Table 1).

These correlations indicated that stronger performance with the familiar examiner was related to (a) examiner—examinee familiarity of comparatively long duration, (b) examinees' low SES status, and (c) relatively demanding tests. A substantive feature correlating in weak fashion with UES was Examiners' Professional Training (see Table 1).

Duration of Familiarity, SES, and Type of Test were entered as predictor variables into a forward stepwise multiple regression. Subjects' CA also was employed as a predictor because, among the remaining substantive variables, it demonstrated the highest correlation (_r_ = .21) and claimed 38 effect sizes. These four predictor variables correlated weakly among themselves; correlations ranged from .32 to -.03, with a median correlation coefficient of .12.

Each of the equations, displayed in Table 2, indicate that the predictor variables were statistically significant in explaining the variance in the UES. In the last equation, incorporating all four variables, SES, Duration of Familiarity, CA, and Type of Test explained 22%, 8%, 7%, and 5% of the variance, respectively. However, the regression was calculated on a relatively small number of effect sizes and, as a consequence, findings may be unstable (Kerlinger & Pedhazur, 1973). Thus, in summarizing and decomposing the linear dependency of the UES on the four predictor variables, results from the regression should be viewed as a heuristic addition to the foregoing correlational analysis.

Insert Table 2 about here

Discussion

This meta-analysis indicated that examinees achieve higher scores when tested by familiar than unfamiliar examiners. The magnitude of this differential performance was both statistically and practically significant. However, caution should be exercised in interpreting examinees' stronger performance in the familiar examiner condition because larger effect sizes were associated with studies of relatively weak methodologies. Additionally, it is unclear whether, and if so to what extent, these results are robust. Although examinees' higher scores with familiar examiners appeared unrelated to whether testers were professionally trained or not, the low number of effect sizes associated with trained (N = 3) and untrained (N = 8) testers undermines confidence in this correlation. Similarly, we are unable to determine possible moderating or mediating effects of examiners' professional familiarity/ unfamiliarity with the group of children of which the examinee was a member. This is because only one study reported a controlled contrast of this examiner-related characteristic.

On the other hand, duration of the familiarity-inducing activity was associated in a strong, positive fashion with effect size. This relation suggests examiner familiarity is a legitimate and important construct. In addition to duration of familiarity, the nature of the test instrument seemed to mediate examinees' differential performance: Examinees performed stronger in the familiar condition when tested on a difficult measure (e.g., an IQ test); however, such differential performance lessened when the measure was comparatively simple (e.g., a speech test). This result is consonant with empirical evidence in the social reinforcement literature, which suggests prior contact with an experimenter increases the level of subjects' responding on complex, but not on simple tasks (Crow, 1964; Rosenkrantz & Van De Riet, 1974). Rosenthal (1980) has suggested an explanation for this pattern of findings: Examiner unfamiliarity engenders anxiety in examinees, and whereas this anxiety enhances motivation to do well on simple tasks, it interferes with the higher order thinking required by complex tasks. Thus, examiner familiarity is presumed to vitiate examinees' anxiety and its negative influence on complex task performance.

The most important subject variable to intercede between examiner familiarity and test performance was SES. Correlational analysis indicated that low SES children's differential performance in favor of the familiar examiner was greater than that of high SES children. This result suggests examiner unfamiliarity selectively depresses the scores of low SES children.

Enhancing the importance of this finding is that most examiners in clinical and educational settings are strangers to the children they test. This has been substantiated directly by reports of practicing professionals (Fuchs, 1981). Indirect evidence comes from an analysis (Fuchs, Fuchs, Dailey, & Power, 1983) of the user manuals of 20 well-known intelligence and speech/language measures: Only 2 manuals suggested that examiners establish pretest contact with their examinees. Moreover, the Standards for Educational and Psychological Tests (1974) seem to discourage examiner familiarity, as reflected in a call for "impersonal" procedures (p. 64) and in a recommendation that testers "minimize" (p. 63) any effect they may have on children's performance. Therefore, on normative tests, the suboptimal performance of low SES children may be compared to the maximal performance of other groups, such as high SES examinees. If so, examiner familiarity is a source of systematic error or bias.

Our findings of apparent test procedure bias may explain at least partially why, on average, low SES children obtain lower IQ scores than high SES children, a phenomenon first described by Binet (see Lippmann, 1976) and repeatedly corroborated since then (e.g., Masland, Sarason, & Gladwin, 1978; Tyler, 1965). A frequent estimate of the magnitude of this difference in IQ performance has been one standard deviation (e.g., Christiansen & Livermore, 1970; Jensen, 1970). Low SES children's test performance conventionally has been interpreted as a rather straightforward demonstration of those skills and abilities that the tests claim to measure. Typically, their comparatively poor showing on these tests has been attributed primarily to either poor genes or a disadvantaged environment (see Nichols, 1978).

Nevertheless, current findings question such interpretations that presume a cause and effect relation between children's cognitive processes and their performance on tests that purportedly measure salient cognitive and/or academic abilities. Our results indicate that at least one extra-test factor, examiner unfamiliarity, also affects the performance of select groups of children. For low SES pupils, the effect size associated with examiner familiarity was .53, which is the equivalent of a difference of approximately 8 points on a standardized IQ test with a mean of 100 and standard deviation of 15. Furthermore, as mentioned above, a growing literature suggests there may be additional contextual variables constituting the typical test situation, which influence certain pupils' performance. Thus, one legitimately might wonder how much of the reported difference between low and high SES children's IQ performance may be explained by differential responses to contextual variables. Until we know the answer to such a question, attributing this discrepancy to a difference in the group's ability level seems precipitous.

Although subjects' SES was related strongly to UES, their handicapped/nonhandicapped status was not. However, this finding may be misleading. Among the relatively few studies employing handicapped subjects, speech and/or language-impaired children consistently performed more strongly with the familiar examiner, whereas mentally retarded children either performed stronger with the <u>unfamiliar</u> examiner or did not demonstrate differential performance. Thus, by combining results from the few investigations involving speech and/or language-impaired, mentally retarded, and other handicapped children, this meta-analysis may be masking possible interaction effects between <u>type</u> of handicap and the familiarity/unfamiliarity of the examiner. Future research might experimentally test such a possibility.

In sum, the effects of examiner familiarity demonstrate the importance of contextual factors in testing. Such factors seem to intercede between the test and performance, questioning the positivistic view that the test instrument is the single most important, if not the exclusive, variable to determine test performance. Although this proposition contradicts traditional thinking about the test situation, it is not new. More than a decade ago, Cronbach (1971) stated that the test is only one element in a procedure, and the validity of data obtained in educational and psychological assessment is dependent upon the procedure as a whole. However, adopting this perspective will be difficult. It not only complicates interpretation of test performance, it also presumes the existence of an adequate data base on contextual effects, which has yet to be developed. Nevertheless, accuracy in interpreting test results requires that we acknowledge the importance of context in assessment and continue the challenging task of defining the relation between situational factors and test performance.

References

Abramyan, L.A. (1977). On the role of verbal instructions in the direction of voluntary movements in children. Quarterly Newsletter of the Institute for Comparative Human Development, 1, 1-4.

Adorno, T.W., Albert, H., Dahrendorf, R., Habermas, J., Pilot, H., & Popper, K.R. (1976). The positivist dispute in German sociology. New York: Harper & Row.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). Standards for educational and psychological tests. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1984, February). Draft: Joint technical standards for educational and psychological testing (Available from APA, Office of Scientific Affairs, 1200 17th St., N.W., Washington, DC 20036).

Ayllon, T., & Kelly, K. Effects of reinforcement on standardized test performance. (1972). Journal of Applied Behavior Analysis, 5, 477-482.

American Psychological Association. (1982). Thesaurus of psychological index terms (3rd ed.). Washington, DC: Author.

[illegible], Mann, M., & Bar-Havim, H. (1975). Bias in the scoring of the WISC subtests. Journal of Consulting and Clinical Psychology, 43, 268.

Bangert-Drowns, R.L., Kulik, J.A., & Kulik, C.C. (1983). Effects of coaching programs on achievement test performance. Review of Educational Research, 53, 571-585.

Bernstein, R.J. (1978). The restructuring of social and political theory. Philadelphia: University of Pennsylvania Press.

Christiansen, T., & Livermore, G. (1970). A comparison of Anglo-American and Spanish-American children on the WISC. Journal of Social Psychology, 81, 9-14.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cole, M., & Means, B. (1981). Comparative studies of how people think. Cambridge, MA: Harvard University Press.

Cooper, H.M. (1982). Scientific guidelines for conducting integrative research reviews. Review of Educational Research, 52, 291-302.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), Educational measurement (pp. 443-507). Washington, DC: American Council on Education.

Crow, L. (1964). Public attitudes and expectations as a disturbing variable in experimentation and theory. Unpublished manuscript, Harvard University.

Deyhle, D. (1983). Learning failure: Test-taking and the Navajo student. [Summary]. Proceedings of the Fourth Annual University of Pennsylvania Ethnography in Education Research Forum, 5.

Dusek, J.B., & Joseph, G. (1983). The bases of teacher expectations: A meta-analysis. Journal of Educational Research, 75, 327-346.

Exner, J.E. (1966). Variations in WISC performances as influenced by differences in pre-test rapport. *Journal of General Psychology*, *74*, 299-306.

Feldman, S.E., & Sullivan, D.S. (1971). Factors mediating the effects of enhanced rapport on children's performance. *Journal of Consulting and Clinical Psychology*, *36*, 302.

Fiscus, E.G. (1975). The effects of pre-test information on sch ol psychologists' scoring of the Wechsler Intelligence Scale for Children. *Dissertation Abstracts International*, *36*, 1387A. (University Microfilms No. 75-19-435).

Flaugher, R. (1978). The many definitions of test bias. *American Psychologist*, *33*, 671-679.

Freud, S. (1922). *Group psychology and the analysis of the ego.* London: International Psycho-Analytical Press. (Originally published, 1921).

Fuchs, D. (1981, April). Differential responses of preschool language-handicapped children and familiar and unfamiliar testers as a function of task complexity, length of acquaintanceship, and sex of child. In V. Shinman (Chair), *Client identification and issues of validity: The influence of situational variables on children's cognitive performance.* Symposium presented at the annual meeting of the American Educational Research Association, Los Angeles.

Fuchs, D., Fuchs, L.S., Dailey, A.M., & Power, M.H. (1983). *Effects of pre-test contact with experienced and inexperienced examiners on handicapped children's test performance* (Research Report No. 110). Minneapolis:

University of Minnesota, Institute for Research on Learning Disabilities.

Fuchs, D., Fuchs, L.S., Power, M.H., & Dailey, A.M. (in press). Bias in the assessment of handicapped children. American Educational Research Journal.

Fuchs, D., Zern, D.S., & Fuchs, L.S. (1983). Participants' verbal and nonverbal behavior in familiar and unfamiliar test conditions. Diagnostique, 8, 159-169.

Fuchs, L.S., & Fuchs, D. (1984). Examiner accuracy during protocol completion. Journal of Psychoeducational Assessment, 2, 101-108.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Goodnow, J. (1976). The nature of intelligent behavior: Questions raised by cross-cultural research. In L.B. Resnick (Ed.), The nature of intelligence. Hillsdale, NJ: Erlbaum.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

Hedges, L. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 359-361.

Hersh, J.B. (1971). Effects of referral information on testers. Journal of Consulting and Clinical Psychology, 37, 116-122.

Horne, L.V., & Garty, M.K. (1981, April). What the test score really reflects: Observations of teacher behavior during standardized achievement test administration. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Jensen, A.R. (1970). Learning ability, intelligence, and educability. In V.L. Allen (Ed.), *Psychological factors in poverty* (pp. 106-132). New York: Academic Press.

Kerlinger, F.N., & Pedhazur, E.J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart, & Winston.

Labov, W. (1973). The logic of nonstandard English. In F. Williams (Ed.), *Language and poverty*. Chicago: Markham.

Lippmann, W. (1976). Tests of hereditary intelligence. In N.J. Block & G. Dworkin (Eds.), *The IQ controversy* (pp. 21-29). New York: Pantheon Books. (Reprinted from *Popular Science Monthly*, May, 1915).

MacKay, R. (1974). Standardized tests: Objective and objectivized measures. In A.V. Cicourel et al. (Eds.), *Language use and school performance* (pp. 218-247). New York: Academic Press.

Masland, R.L., Sarason, S.B., & Gladwin, T. (1978). *Mental subnormality*. New York: Basic Books.

Masling, J.M. (1957). The effects of warm and cold interaction on the interpretation of a projective protocol. *Journal of Projective Techniques*, *21*, 377-383.

Mehan, H. (1978). Structuring school structure. *Harvard Educational Review*, *48*, 32-64.

Mishler, E.G. (1979). Meaning in context: Is there any other kind? *Harvard Educational Review*, *49*, 1-19.

Nichols, R.C. (1978). Policy implications of the IQ controversy. *Review of Research in Education*, *6*, 3-46.

Piaget, J. (1965). *The moral judgment of the child*. New York: Free Press.

Rosenkrantz, A.L., & Van De Riet, V. (1974). The influence of prior contact between child subjects and adult experimenters on subsequent child performance. Journal of Genetic Psychology, 124, 79-90.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null effects. Psychological Bulletin, 86, 638-641.

Rosenthal, R. (1980). Experimenter effects in behavioral research (Enlarged paperback edition). New York: Irvington.

Rosenthal, R., & Rubin, D.B. (1978). Interpersonal expectancy effects: The first 345 studies. The Behavioral and Brain Sciences, 3, 377-415.

Roth, D.R. (1974). Intelligence testing as a social activity. In A.V. Cicourel et al. (Eds.), Language use and school performance (pp. 143-217). New York: Academic Press.

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

Sarason, I.G. (Ed.). (1980). Test anxiety: Theory, research, and applications. Hillsdale, NJ: Erlbaum.

Sattler, J.M. (1974). Assessment of children's performance. Philadelphia: Saunders.

Schroeder, H.E., & Kleinsasser, L.D. (1972). Examiner bias: A determinant of children's verbal behavior on the WISC. Journal of Consulting and Clinical Psychology, 39, 451-454.

Seitz, V., Abelson, W.D., Levine, E., & Zigler, E. (1975). Effects of place of testing on the Peabody Picture Vocabulary Test scores of disadvantaged Head Start and non-Head Start children. Child Developmant, 46, 481-486.

Stevenson, H.W. (1965). Social reinforcement of children's behavior. In L.P. Lipsitt & C.C. Spiker (Eds.), Advances in child development, II (pp. 97-126). New York: Academic Press.

Stevenson, H.W., & Hill, K.T. (1966). Use of rate as a measure of response in studies of social reinforcement. Psychological Bulletin, 66, 321-326.

Stoneman, Z., & Gibson, S. (1978). Situational influences on assessment performance. Exceptional Children, 46, 166-169.

Taylor, C., & White, K.R. (1981, April). Effects of reinforcement and training on Title I students' group standardized test performance. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Tiber, N., & Kennedy, W.A. (1964). The effects of incentives on the intelligence test performance of different social groups. Journal of Consulting Psychology, 28, 187.

Tyler, L.E. (1965). The psychology of human differences. New York: Appleton-Century-Crofts.

~~Zigler, E., & Butterfield, E. (1968). Motivational aspects of changes in IQ~~ test performances of culturally deprived nursery school children. Child Development, 39, 1-14.

Appendix

22 Studies of Examiner Familiarity

[1]Back, R.D., & Dana, R.H. (1980). Self-help for male WISC examiners by pretest exposure to children. *Perceptual and Motor Skills*, *51*, 838.

Costello, J. (1970). Effects of pretesting and examiner characteristics on test performance of young disadvantaged children [Summary]. *Proceedings of the 78th Annual Convention of the American Psychological Association*, 309.

Duffy, O.B. (1972). The differential effects of psychologist as examiner and teacher as examiner on word recognition in oral reading of third and fifth grade children. *Dissertation Abstracts International*, *33*, 3375A. (University Microfilms No. 73-00-617.).

DuRant, M.B. (1975). The effect of examiner familiarity on two sub-tests of the Illinois Test of Psycholinquistic Abilities. *Dissertation Abstracts International*, *36*, 3503A-3504A. (University Microfilms No. 75-28-968)

[1]Feldman, S.E., & Sullivan, D.S. (1971). Factors mediating the effects of enhanced rapport on children's performance. *Journal of Consulting and Clinical Psychology*, *36*, 302.

Field, T. (1981). Ecological variables and examiner biases in assessing handicapped preschool children. *Journal of Pediatric Psychology*, *6*, 155-163.

Fuchs, D., Featherstone, N.L., Garwick, D.R., & Fuchs, L.S. (1984). Effects of examiner familiarity and task characteristics on speech- and language-impaired children's test performance. *Measurement and Evaluation in Guidance*, *16*, 198-204.

Fuchs, D., Fuchs, L.S., Garwick, D.R., & Featherstone, N. (1983). Test performance of language-handicapped children with familiar and unfamiliar examiners. Journal of Psychology, 114, 37-46.

Fuchs, D., Fuchs, L.S., Dailey, A.M., & Power, M.H. (in press). The effect of examiners' personal familiarity and professional experience on handicapped children's test performance: A case of who, not what you know? Journal of Educational Research.

Fuchs, D., Fuchs, L.S., Power, M.H., & Dailey, A.M. (in press). Bias in the assessment of handicapped children. American Educational Research Journal.

Irons, D. (1981). The effect of familiarity with the examiner on WISC-R Verbal, Performance, and Full Scale scores. Psychology in the Schools, 18, 496-499.

Jacobson, L.I., Berger, S.E., Bergman, R.I., Millham, J., & Greeson, L.E. (1971). Effects of age, sex, systematic conceptual learning, acquisition of learning sets, and progr mmed social interaction on the intellectual and conceptual development of preschool children from poverty backgrounds. Child Development, 42, 1399-1415.

Kinnie, E.J., & Sternlof, R.F. (1971). The influence of nonintellective factors on the IQ scores of middle- and lower-class children. Child Development, 42, 1989-1995.

Klein, P.S. (1983). Cognitive performance of kindergarten children when tested by parents and strangers. In N. Nir-Janiv, B. Spodek, & D. Steg (Eds.), Early childhood education (pp. 429-440). New York: Plenum.

Marine, E.L. (1929). The effect of familiarity with the examiner upon

Stanford-Binet test performance. Teachers College Contributions to Education, 381, entire issue.

Olswang, L.B. & Carpenter, R.L. (1978). Elicitor effects on the language obtained from young language-impaired children. Journal of Speech and Hearing Disorders, 43, 76-88.

Orost, J.H. (1972). Effects of examiner age and familiarity on test performance of third grade and kindergarten girls. Dissertation Abstracts International, 32, 6011A-6012A. (University Microfilms No. 72-16-092)

Piersel, W.C., Brody, G.H., & Kratochwill, T.R. (1977). A further examination of motivational influences on disadvantaged minority group children's intelligence test performance. Child Development, 48, 1142-1145.

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

Thomas, A., Hertzig, M.E., Dryman, I., & Fernandez, P. (1971). Examiner effect in IQ testing of Puerto Rican working-class children. American Journal of Orthopsychiatry, 41, 809-821.

Tsudzuki, T., Hata, Y., & Kuze, T. (1956). A study on the rapport between the examiner and the subject. Japanese Journal of Psychology, 27, 22-28.

Zigler, E., Abelson, W.D., & Seitz, V. (1973). Motivational factors in the performance of economically disadvantaged children on the Peabody Picture Vocabulary Test. Child Development, 44, 294-303.

Footnotes

[3] Additional information on the Back and Dana, Feldman and Sullivan, and Irons published studies was obtained from the following fugitive sources: Back, R.D., & Dana, R.H. (1980). *The effects of pretest exposure on sex of examiner influence on the Wechsler Intelligence Scale for Children* (Document NAPS-03775). Available from Microfiche Publications, PO Box 3513 Grand Central Station, New York, NY 10017; Feldman, S.E., & Sullivan, D.S. (n.d.). *Factors influencing the effects of enhanced rapport upon children's test performance*. Unpublished manuscript, Northern Illinois University, DeKalb; Irons, D.A. (1980). The effect of familiarity with the examiner on WISC-R Verbal, Performance, and Full Scale scores (Doctoral dissertation, Texas Tech University). *Dissertation Abstracts International*, 41, 1533A.

Table 1

Means, Standard Deviations, and Correlations of UESs by Substantive Features of the Studies

Substantive feature	$\bar{X}$	SD	N	r
Duration of familiarity			36	.47**
Less than 16 minutes	.09	.62	7	
Between 16 and 120 minutes	.13	.13	11	
Between 121 minutes and 10 hours	.62	.41	8	
Between 11 and 20 hours	.75	.46	3	
More than 20 hours	.52	.50	7	
Examiners' professional familiarity with subject type[a]			21	.06
Familiar	.26	.37	20	
Unfamiliar	.17	---	1	
Examiners' Training			11	.20
Professionally trained	.31	.32	3	
Professionally untrained	.06	.52	8	
Familiarity-inducing activity[a]			38	.08
Interaction	.35	.47	37	
Observation	.58	---	1	
Handicapped Status			36	.16
Handicapped	.31	.37	11	
Nonhandicapped	.39	.51	25	
Subjects' CA[b]			38	.21
Subjects' SES			37	-.40**
Low	.53	.50	17	
High	.24	.40	20	
Test location			15	.19
Familiar	.26	.34	13	
Unfamiliar	.43	.17	2	
Type of test			38	-.33*
IQ	.54	.54	18	
Speech/language	.19	.35	18	
Isolated tasks	.24	.19	2	

[a] Given the distribution of effect sizes across values of these variables, the related correlations are likely to be unstable. The same may be true for other variables such as Test Location.

[b] Since subjects' CA was treated as a continuous variable, there are no group means to report.

*$p < .05$.

**$p < .01$.

Table 2

Results of Multiple Regression on Predicting UESs

Source	Multiple R	R^2 Cumulative	R^2 Change	F^a	F^b
SES	.47	.22	.22	10.37**	10.37**
Duration	.55	.30	.08	7.61**	3.99*
CA	.61	.37	.07	6.66**	3.62*
Type of Test	.65	.42	.05	5.91**	2.69*

[a] *F* value is for the regression equation.

[b] *F* value is for the contribution of each variable.

**p* < .05.

***p* < .001.