Journal of Criminal Law and Criminology

Volume 54
Issue 1 *March*

Article 2

Spring 1963

An Analysis of Rater Reliability on the Glueck Scale for the Prediction of Juvenile Delinquency

Charles S. Prigmore

AN ANALYSIS OF RATER RELIABILITY ON THE GLUECK SCALE FOR THE PREDICTION OF JUVENILE DELINQUENCY*

CHARLES S. PRIGMORE

The author is Associate Professor in the School of Social Welfare of Louisiana State University, with responsibility for the programs in corrections and delinquency. He is also co-director of the Louisiana Juvenile Officers Training Institute. From 1956 to 1959, Professor Prigmore served as superintendent of the state correctional school for delinquent boys in Tennessee, and from 1951 to 1956 he was Supervisor of Training for the Wisconsin Bureau of Probation and Parole, as well as Program Coordinator, Juvenile Institutions. Earlier, he was a social worker at the Wisconsin School for Boys.

In the following article, Professor Prigmore reports on a study concerning the Social Prediction Table developed by Sheldon and Eleanor T. Glueck for the prediction of delinquency. What is the reliability of raters in the use of the Glueck scale as a predictive instrument? What does rater reliability or lack of it tell us about the validity, accuracy, and efficiency of the Glueck scale? In an effort to answer these questions, the author made comparisons among the ratings of eight male social workers, matched for education and experience. Two of the raters were Northern-educated Negroes, two were Northern-educated whites, two were Southern-educated Negroes, and two were Southern-educated whites. The author here presents his findings and discusses their implications with respect to possible refinements of the Glueck approach.—EDITOR.

The prediction of social phenomena is increasingly felt to be the ultimate goal of social science.[1] As has been true in other areas of human behavior, efforts have been made to predict juvenile delinquency. In the past, these efforts—by psychologists, sociologists, and others—have not succeeded in developing instruments that have predicted efficiently and accurately.[2] One of the more recent attempts, however, has resulted in the development of a scale for which considerable validity has been claimed.

Two gauges of the efficiency and accuracy of any measuring or predicting instrument are its reliability and its validity. These two characteristics are interrelated, and an instrument lacking in reliability cannot have validity.[3] A predictive instrument is reliable if repeated measurements provide similar results.[4] An instrument is valid if it actually measures what it claims to measure and if correct predictions can be based on it.[5] Although psychological research has focused on reliability, sociological work in prediction has been confined primarily to validation studies. This has been true even though there is every reason to believe that unreliable instruments are wasteful of research time and money.[6] Only recently has an interest in independent rating or assessment, in order to avoid bias and improve reliability, appeared in the sociological literature.[7]

The present study is focused on the reliability of a current prediction instrument of some importance.

* This is also the title of the author's unpublished Ph.D. dissertation, Department of Sociology, University of Wisconsin, August, 1961, upon which this paper is based. The writer is indebted to Associate Justice Joe W. Sanders, Louisiana Supreme Court, formerly Judge of the Family Court of East Baton Rouge Parish, and to his successor, Judge C. Lenton Sartain, for their encouragement and support.

[1] See, *e.g.*, BECKER, THROUGH VALUES TO SOCIAL INTERPRETATION 101 (1950).

[2] See KVARACEUS, THE COMMUNITY AND THE DELINQUENT ch. V (1954); Volkman, *A Matched-Group Personality Comparison of Delinquent and Nondelinquent Juveniles*, 6 SOCIAL PROBLEMS 238–45; BALOGH & RUMAGE, JUVENILE DELINQUENCY PRONENESS: A STUDY OF THE KVARACEUS SCALE (1956); Weeks, *Predicting Juvenile Delinquency*, 8 AM. SOC. REV. 40–46 (1943); Morlock, Predicting Delinquency in a Homogeneous Group of Pre-Adolescent Boys, May 1947 (unpublished Ph.D. dissertation in Ohio State University Library); POWERS & WITMER, AN EXPERIMENT IN THE PREVENTION OF DELINQUENCY: THE CAMBRIDGE-SOMERVILLE YOUTH STUDY 291–92 (1951).

[3] JAHODA, DEUTSCH & COOK, RESEARCH METHODS IN SOCIAL RELATIONS 94 (1951).

[4] *Id.* at 93.

[5] *Ibid.*

[6] SELLTIZ, JAHODA, DEUTSCH & COOK, RESEARCH METHODS IN SOCIAL RELATIONS 167 (Rev. ed. 1959).

[7] See, *e.g.*, GORDON, SOCIAL CLASS IN AMERICAN SOCIOLOGY 111 (1958). In criticizing W. Lloyd Warner's Index of Social Characteristics as a rating scale to determine social class membership, Gordon states that studies of rater reliability are needed.

The Glueck Prediction Scale for Delinquency

The Gluecks turned to a study of delinquency causation in 1940, after some fifteen years of research into the effectiveness of different methods of correctional treatment. In 1950 they published *Unraveling Juvenile Delinquency*, a ten-year study of the genesis of delinquent behavior.[8] Their approach was an eclectic one, without any articulated theory or hypotheses except the general impression that delinquency is a function of sociocultural, somatic, intellectual, and emotional-temperamental influences. They have defended this eclectic approach, claiming that our present state of knowledge requires meaningful integration of data from various levels of inquiry.[9] Generally, it has been felt by other researchers that hypotheses are desirable in most instances, but not always necessary to social research.

The Gluecks matched 500 institutionalized delinquent boys with 500 Boston public-school boys with respect to age, general intelligence, national (ethnico-racial) origin, and residence in underprivileged neighborhoods of Boston. The average age of the boys was fourteen. They gathered very complete data on all 1000 boys, through interviews with parents and boys, physical and psychiatric examinations, and the study of case materials. After developing some 216 tables comparing delinquent boys and non-delinquent boys in the study, they constructed three predictive instruments utilizing factors from the tables. One instrument involved social factors, another Rorschach results, and the third some of the psychiatric findings.

The social factors scale has been the one most earnestly proposed by the Gluecks and others as a means for predicting delinquency when a child is six years of age.[10] The particular social factors selected were chosen by the Gluecks for their presence at age six, for their relative independence from each other, for the ease in gathering information about them, and for their degree of discrimination between delinquents and non-delinquents. The specific social factors chosen were: (1) Discipline of Boy by Father, (2) Supervision of Boy by Mother, (3) Affection of Father for Boy, (4) Affection of Mother for Boy, and (5) Cohesiveness of Family.

The Glueck study has been severely criticized for the lack of a clear-cut research design, for the inadequate control of neighborhood influences, for the use of only institutionalized delinquents, for the possible lack of representativeness of the non-delinquents, and for the neglect of normative and control factors in favor of a bias for early parent-child relations.[11] The prediction scale has been criticized as being useful only in areas of very high delinquency, as presuming intra-family relations to be causes rather than effects, as using data about fourteen-year-olds to predict for six-year-olds, as omitting such significant factors as companions and gang affiliation, as representing an effort to focus remedial attention on some children instead of all children, and as being harmful in view of the tendency of prediction scales to overpredict.[12]

The United States Children's Bureau has taken the position that any attempt to predict delinquency in children at age six will result in the false labeling of many children.[13] The position of the National Institute of Mental Health, on the other hand, is that prediction provides a promising direction for delinquency control analogous to the public health control methods used for diabetes.[14]

The Gluecks have encouraged validation studies, i.e., applications of the Glueck predictive

[8] S. & E. T. Glueck, Unraveling Juvenile Delinquency (1950).

[9] *Id.* at 7.

[10] The list of articles and books written by the Gluecks emphasizing the value of the scale is a long one. The most recent book is S. & E. T. Glueck, Predicting Delinquency and Crime (1959). A very recent article is E. T. Glueck, *Efforts To Identify Delinquents*, 24 Fed. Prob. 49–56 (1960).

[11] Only a few of the many critiques are listed here, due to space limitations. Clinard, *Review of Delinquents in the Making: Paths to Prevention, by Sheldon Glueck and Eleanor T. Glueck*, 17 Fed. Prob. 50–51 (1953); Reiss, *Unraveling Juvenile Delinquency: An Appraisal of the Research Methods*, 57 Am. J. Soc. 115–120 (1951); Rubin, *Unraveling Juvenile Delinquency: Illusions in a Research Project Using Matched Pairs*, 57 Am. J. Soc. 107–14 (1951); Taft, *Implication of the Glueck Methodology for Criminological Research*, 42 J. Crim. L., C. & P.S. 300–16 (1951); Thrasher, *Review of Unraveling Juvenile Delinquency by Sheldon Glueck and Eleanor T. Glueck*, 16 Am. Soc. Rev. 264–65 (1951).

[12] Burgess, *Can Potential Delinquents Be Identified Scientifically*, Proceedings of Twenty-Fourth Annual Governor's Conference on Youth and Community Service (Illinois) 38 (1955); Kahn, *Analysis of Methodology of Unraveling Juvenile Delinquency*, in An Approach to Measuring Results in Social Work 170 (French ed. 1952); Moore, Juvenile Delinquency: Research, Theory and Comment 23 (1958); as well as articles listed in previous footnote.

[13] Herzog, Identifying Potential Delinquents 5 (1960).

[14] U.S. Dept. of H.E.W., Report to the Congress on Juvenile Delinquency 16 (1960).

scale on new samples. Approximately 15 are known to have been made.[15] Thirteen of these, however, are retrospective validations, that is, the Glueck scale has been applied to cases of children who have already become delinquent, to see if the scale would have predicted them correctly. Horst and others have stressed the serious drawbacks in retrospective validations.[16] For example, there is a loss of knowledge about the children who did not become delinquent. Could the scale have predicted them? Further, there is always the question whether even correct predictions really were based on the scale, or whether they might have been based on some other intuitive or factual assessment not formulated from the factors of the scale. Such a process can conceivably occur without conscious knowledge of the rater or user of the scale. The rater, for example, who knows he is applying the Glueck scale to a group of delinquents may well be influenced by that fact in applying the scale.

Only two retrospective studies have used control groups. Eleanor Glueck has served as rater or adviser to the rater in many of the retrospective studies, a rather questionable procedure. All but three of the retrospective studies have been carried out in the states of Massachusetts, New York, and New Jersey; the remaining three were Japanese and French studies.

Of the two truly predictive validation studies attempted, only one, conducted by the New York City Youth Board, has published interim findings.[17] Considerable disagreement has arisen regarding the interpretation of these interim findings, the Youth Board claiming an over-all success rate of 89 per cent, but the United States Children's Bureau claiming the success rate to be only 37 per cent for delinquency prediction to date.[18] The disagreement hinges on the definition of delinquency, the Children's Bureau stressing that the Youth Board's 89 per cent success rate is predicated on a much looser definition of delinquency than used in the original *Unraveling Juvenile Delinquency* study.

[15] For references to these, see E. T. Glueck, *supra* note 10, or S. & E. T. GLUECK, PREDICTING DELINQUENCY AND CRIME (1959).

[16] HORST, THE PREDICTION OF PERSONAL ADJUSTMENT 43 (1941).

[17] N. Y. CITY YOUTH BD., AN EXPERIMENT IN THE VALIDATION OF THE GLUECK PREDICTION SCALE, PROGRESS REPORT (1957).

[18] HERZOG, *op. cit. supra* note 13, at 2.

PROBLEMS AND HYPOTHESES

Since validation studies of the Glueck scale have been so difficult of execution and so subject to error, it seems logical to consider a test of the reliability of raters[19] in the use of the Glueck scale as a way of assessing the accuracy and efficiency of the scale. Since reliability is necessary to validity, establishing a lack of reliability would cast serious doubt on validity. The problems attacked in this study, then, were:

1. What is the reliability of raters in the use of the Glueck scale as a predictive instrument?

2. What does rater reliability or lack of it tell us about the validity, accuracy, and efficiency of the Glueck scale?

The methodological studies of rating scales and reliability of ratings have been carried out primarily by social psychologists.[20] Findings have emphasized the general lack of reliability in rating scales, since judges or raters become measuring instruments themselves in the rating process. Findings have largely favored the use of objective materials rather than ratings.[21] When ratings are used, reliability can be improved through use of clearly defined categories, trained raters, forced-choice rating procedures, pooling of independent ratings, and common frames of reference.[22]

Implicit in any rating is a frame of reference on the part of the rater.[23] The rating takes its meaning

[19] A distinction can be made between the reliability of the *rating* of a particular behavior and the reliability of the *behavior* itself. See Heyns & Zander, *Observation of Group Behavior*, in RESEARCH METHODS IN THE BEHAVIORAL SCIENCES 381 (Festinger & Katz eds. 1953).

[20] However, a number of sociologists have employed judges or raters in their studies, the most common use of them perhaps being with college classes. Smith, *Empirical Scale of Prestige Status of Occupations*, 8 AM. SOC. REV. 185–92 (1943), is an example. At least one study has used a sample of the American population as raters. North & Hatt, *Jobs and Occupations: A Popular Evaluation*, in SOCIOLOGICAL ANALYSIS 464–74 (Wilson & Kolb eds. 1949).

[21] Kelly & Fiske, *The Prediction of Success in the V A Training Program in Clinical Psychology*, 5 AM. PSYCHOLOGIST 395–406 (1950).

[22] See, *e.g.*, THORNDIKE & HAGEN, MEASUREMENT AND EVALUATION IN PSYCHOLOGY AND EDUCATION 367 (1955); Sisson, *Forced Choice—The New Army Rating*, 1 PERSONNEL PSYCHOLOGY 365–81 (1948); Stockford & Bissell, *Factors Involved in Establishing a Merit-Rating Scale*, 26 PERSONNEL 94–116 (1949).

[23] Social psychologists have only recently given recognition to the effect of cultural values and norms on ratings. See Sherif, *Introduction* in SOCIAL PSYCHOLOGY AT THE CROSSROADS 18 (Rohrer & Sherif eds., 1951). Lazarsfeld has commented on the ambiguity in the social psychologists' use of the term "frames of reference." Lazarsfeld, *Problems in Methodology*, in

from the rater's reference groups. Actually, in a broad sense, only two kinds of influences could affect a rater's judgment: social psychological influences and cultural influences. The social psychological influences include motivations, perceptions, personality traits, and the reactions to interpersonal experiences. The cultural influences include inculcation of norms and values from reference groups such as social institutions, kinship groups, and subcultural groups (social class, region, rural-urban, race).[24] If the social psychological influences were controlled in the study, and only certain variables in the cultural component left uncontrolled, then variation in ratings would be due to these uncontrolled variables.

Three underlying hypotheses were then established in the study: (1) Cultural background differences account for different attitudes and judgments regarding parental factors. (2) We will find systematic differences in ratings, if the raters vary in regional background and in race with its sociocultural implications.[25] These differences will emerge more clearly if the raters are matched for social class, level of professional education, years of experience, urban residence, sex, and absence of disrupting social psychological factors.[26] (3) Knowledge of reference subgroup norms will enable us to hypothesize specific systematic differences in ratings.[27]

The hypotheses to be tested were then formulated as follows:

H_1 With respect to Negro ratees, the expectations of Negro raters within region are greater than are those of white raters.[28]

H_2 With respect to Negro ratees, the expectations of Northern white raters are greater than are those of Southern Negro raters.

H_3 With respect to white ratees, the expecta-

SOCIOLOGY TODAY: PROBLEMS AND PROSPECTS 41 (Merton *et al.*, eds., 1959).

[24] Race is used here as a cultural rather than a biological concept. See CLINARD, SOCIOLOGY OF DEVIANT BEHAVIOR 438 (1957).

[25] The question might be raised: why region and race as the uncontrolled variables? Why not social class or the particular school of social work or rural-urban background? Actually, there is no implication here that region or race are any more important than class or any other reference group. The researcher happened to be teaching in a southern state and could easily use region and race as the experimental variables.

[26] It can with some merit be argued that social class cannot be *fully* controlled simply by selecting Negro and white social workers all of whom term themselves upper-middle-class in orientation and all of whom share an upper-middle-class occupation. The Negro class structure is considered a half to a full class level lower than the white. See MARDEN, MINORITIES IN AMERICAN SOCIETY 34–36 (1952).

[27] These hypotheses were developed on the basis of a survey of the literature, particularly that relating to values and norms held by reference groups, significant to the raters, which were not controlled in the study. These reference groups included: (1) the white social work professional group in the South, which has become conciliatory and compromising in the matter of parental standards for Negroes in the effort to reconcile professional norms and the "Southern way of life." The result has been a lowering of standards for both white and Negro parents, but particularly for Negro parents. (2) The Negro professional group, particularly Negro social workers, who have become increasingly motivated to help lower-class Negro parents to raise their standards of child care. The leadership role of Negro social workers in the community reinforces the professional norms to raise expectations. (3) The Negro professional group working in the South, which has had to adapt to Southern white middle-class norms and thus lower expectations as compared with Negro social workers in the North. It should be recognized that Southern Negro hostility toward whites coupled with the deference to whites imbedded in the Southern Negro subculture, can be expected to result in somewhat higher expectations for white parents as compared with Negro parents. (4) The Southern schools of social work, as a subgroup of American schools, which have had to compromise between the over-all professional norm of high expectations for Negro and white parents and the Southern norm of lowered expectations for Negro (and as a consequence, white) parents. Southern schools have tried to avoid controversy by stressing class rather than race, so that students tend to see much Negro behavior (correctly) as lower-class behavior. (5) The Negro middle-class, which has tended to have a very critical attitude toward the Negro lower-class, although there has been ambivalence due to the desire to hold on to petty advantages resulting from Negro lower-class subordination. The Negro middle-class in the South has had to adapt itself to Southern white middle-class norms and generally has expected lower standards of parental care for Negroes than has the Negro middle-class in the North. (6) The white middle-class in the South has generally been strongly anti-Negro and paternalistic, although at least theoretically it has favored improvement in both Negro and white parental standards. The attempt to reconcile these pressures in an atmosphere of strong family relationships and extended kin groups has resulted in the dragging down of lower-class white standards along with lower-class Negro parental standards.

[28] The term "expectations" is used often in the present paper. Although some confusion arose from the use of the term in the formulation of the hypotheses, the term can be conceived as referring to the perception or judgment of a rater as to the level of parents' affection, discipline, supervision, or cohesiveness. That rater who chooses the subcategory "hostile" for the factor Affection of Father for Boy in a given case is holding higher standards for, or expecting more of, the parents than would another rater who chooses the subcategory "warm" for the same data in the same case. The term "expectations" has been used in preference to "standards" since the reference is to the judgment of the *rater* rather than any other gauge of parental adequacy. The use of the term "standards" is apt to evoke a picture of level of parental adequacy, however determined, rather than rater judgment.

TABLE I
PLACE OF EDUCATION AND AGE OF RATERS

	Northern-Educated White		Northern-Educated Negro		Southern-Educated White		Southern-Educated Negro	
Rater Number.	1	2	3	4	5	6	7	8
Year of Birth..	1929	1931	1926	1916	1926	1930	1913	1926
State MSW Obtained....	Michigan	New York	Michigan	Nebraska	Louisiana	Tennessee	Georgia	Georgia

tions of Negro raters within region are greater than are those of white raters.

H_4 With respect to white ratees, the expectations of Northern white raters are greater than are those of Southern Negro raters.

METHOD OF COLLECTING AND PROCESSING DATA

Factual information on the five Glueck factors was obtained in September and October, 1960, for 60 delinquent boys. The five male probation officers in the Family Court Probation Department of East Baton Rouge Parish, Louisiana, furnished this information on boys currently supervised. Both white and Negro cases were used, and the race was designated on the forms subsequently used by the raters. A systematic stratified random sample was taken of each officer's caseload, so that the 60 cases were representative of the 180 cases under supervision.

Information was obtained on all five factors in each case, the officers being asked to provide information about former father-figures, for example, if no father was presently in the home. All cases met the Gluecks' definition of delinquency: "... repeated acts of a kind which when committed by persons beyond the statutory juvenile court age of sixteen are punishable as crimes ... except for a few instances of persistent stubbornness, truancy, running away, associating with immoral persons, and the like."[29]

Eight raters were selected, all male, middle-class social workers employed in urban areas of the South.[30] All had professional education at the level of a Master's degree in social work, and all had between two and eight years' experience subsequent to receiving the degree. The eight raters comprised four cells: two raters were Northern-educated Negroes, two were Northern-educated whites, two were Southern-educated Negroes, and two were Southern-educated whites. The raters were selected on a random basis from the 1960 Directory of Professional Social Workers. Table I provides information about age of each rater and the state in which each rater obtained the Master's degree in social work.

Each rater completed a questionnaire, designed to see that the variables of sex, social class, level of education, length of experience following the M.S.W. degree, and potentially biasing social psychological influences were controlled. For example, the questionnaire included questions to get at unusual childhood experiences, such as early death of parents or foster care, in addition to questions regarding social class, experience, and education.[31] The questionnaire also was designed to see that information was obtained regarding race and regional background of the rater, so that the raters could reflect the particular combinations of these variables demanded by the research design. Data collection and rating procedures were pretested and appropriate modifications were made in forms and procedures before final data collection.[32]

[29] S. & E. T. GLUECK, UNRAVELING JUVENILE DELINQUENCY 13 (1950).

[30] The Gluecks state that "certainly trained case workers could readily gather and interpret the materials." *Id.* at 269. Another question might be raised: Why only eight raters? Why not 80 if reliability of raters is to be evaluated? Actually, most rater reliability studies have used fewer than eight, although Goode & Hatt, and Symonds, recommend eight as the average number of independent ratings required for reliability. GOODE & HATT, METHODS IN SOCIAL RESEARCH 260 (1952); SYMONDS, DIAGNOSING PERSONALITY AND CONDUCT 96 (1931).

[31] No prospective rater with a background, attitude, or characteristic that would prevent the testing of the hypotheses was used as a rater. It was planned that another social worker of the appropriate racial and regional background would be used in his stead if a rater, for example, had been adopted as a child and might have thus conceivably been influenced by this fact in his judgments regarding parental factors. Actually, all the raters responded to the questionnaire in such a way that they were considered suitable for the rater panel, except that the researcher had a question as to whether Rater #2 was really Northern-educated.

[32] In the final collection of case information, each probation officer was given fifteen dollars for the time and effort invested, and in the final rating of cases each

Very full instructions were provided to raters for the final rating. For example, the explanation was given that in some cases various father-surrogates may have been in the home and the rating should reflect the total situation over the boy's lifetime. That is, paternal discipline and paternal affection should be viewed from the standpoint of the total effect of all the father-surrogates. (A very strong effort had been made to provide ample information about the discipline and affection of each of the father-surrogates.) Attention was called to the occasional presence of information relevant to "discipline" under "affection" and vice versa, so that it was suggested that the entire Information Form be read before a subcategory was checked under any one factor on the Rating Form. Explicit directions were given regarding procedure for rating and for completing the forms, the letter of instructions being a three-page letter. Raters were encouraged to contact the researcher for clarification of any areas of question.[33] The final rating involved a forced choice of seven subcategories under each of the Glueck factors, instead of the three the Gluecks proposed.[34] This change was made in order to make the instrument sensitive enough to detect the differences. Thus, under Discipline of Boy by Father, the raters were asked to choose between "overstrict or erratic," "fairly overstrict or erratic," "a little overstrict or erratic," "firm but kindly," "a little lax," "fairly lax," or "lax." The original Glueck scale contains only the subcategories "overstrict or erratic," "firm but kindly," and "lax." A Rating Form was checked on each of the 60 cases by each rater.

The raters were also asked to complete a Rater Judgment Questionnaire on each of the 60 cases, clarifying the reason for their judgments: race, social class, other factors. After indicating the reason for their judgment, the raters were also asked to indicate whether they expected more or less as a result of race, class, or whatever reason affected their judgment.

The method of processing the data involved the use of the sign test, which is a nonparametric statistical test. The test can be used without any statistical inference as to the representativeness of either the raters or the delinquent boys, although the latter group was representative of the population of delinquent boys currently under super-

rater was given forty dollars, using a grant from the Aquinas Fund.

[33] About half the raters did so. On the basis of the detailed instructions and the individual assistance by telephone, letter, and personal interview, the researcher feels that the raters had the benefit of sufficient training to be able to rate adequately. It should also be recognized that the making of judgments about parental affection, discipline, supervision, and family cohesiveness is an almost daily responsibility of social workers, so this kind of rating was not a new activity for the raters.

[34] Research has indicated the value of a forced-choice technique. See Sisson, *supra* note 22. A question may also be raised as to the use of seven subcategories instead of the three the Gluecks employed. Offhand, one would postulate a decrease in reliability from the increase in subcategories, particularly since the new subcategories involve the use of expressions such as "a little lax" and "fairly lax," which could be claimed to have an uncertain meaning to raters. The literature, however, is clear that too few categories in a rating scale produce a coarse scale in which we lose much of the discriminative power of which raters are capable. Early studies indicated seven to be an optimum number, further steps beyond that number *increasing* the reliability too slightly to justify the extra effort. Symonds, *On the Loss of Reliability in Ratings Due to Coarseness of the Scale*, 7 J. EXP. PSYCHOL. 456–61 (1924). Later studies have indicated that the optimal number of categories in a scale may be much greater than seven, often as high as 22. Champney & Marshall, *Optimal Refinement of the Rating Scale*, 23 J. APPL. PSYCHOL. 323–31 (1939). Guilford sums up by stating that the optimal number is a matter for empirical determination, fortunately there being "a wide range of variation in refinement around the optimal point in which reliability changes very little. It can be said, however, that the number 7 recommended by Symonds is usually lower than optimal and it may pay in some favorable situations to use up to 25 scale divisions." GUILFORD, PSYCHOMETRIC METHODS 291 (1954). These remarks pertain primarily to relatively well-defined traits. If the trait is rather obscure or vague (as is our case in the present instance), Symonds suggests that fewer steps may be used since reliability in that case is too low for additional steps to be of value. SYMONDS, *op. cit. supra* note 30, at 79.

In order to overcome the loss of reliability from the obscurity of the traits, and in order to compensate for any difficulty that could arise from the wording of the new subcategories, very full information was supplied on all factors in each case, and very explicit directions were given to raters. Also, the gradations "a little lax" and "fairly lax," for example, were shown on the forms as equal steps, respectively, between "firm but kindly" and "lax." In all instances the new gradations were shown as ordered categories on an ordinal scale. No rater expressed difficulty in the use of the new gradations.

It seemed necessary in the present research to have an instrument sensitive enough to bring out the cultural differences between rater judgments in order to test the hypotheses. If some loss of reliability did occur as a result of the new gradations—a questionable occurrence in the light of previous research and the interest, preparation, and prior training of the raters—any such hypothetical loss seems justified in the light of the clear emergence of the cultural differences. At the worst, any substantial loss of reliability might preclude a conclusive appraisal of the rater reliability of a *three-category* Glueck scale, but it would not seem to preclude an appraisal of the rater reliability of the social factors used in the Glueck scale.

vision of the Family Court, and actually no reason existed to cast doubt on the representativeness of either the rater or the ratee groups as samples of larger populations.

The sign test measures the ranking of the two members of any pair with respect to any variable and actually compares the difference between them with what would be expected in a binomial distribution. It is applicable to our case because we want to test the significance of the difference between pairs, when only race and region of the raters differ in the pairs of ratings being compared. In our situation, each rater is compared with each of the other raters for each factor in the 60 cases. The 28 possible combinations of the eight raters are thus compared for each of the five factors, giving 140 sign tests. When separate comparisons are also made for Negro and white ratees, a total of 420 sign tests results.[35]

[35] If we take a specific example of a Northern white rater, A, and a Southern Negro rater, B, in order to test our hypotheses two and four, and compare them on their 60 ratings of one of the five factors, we might have a distribution like this:

	Negro	*White*	*Total*
A > B	.3	.3	.6
A = B	.05	.05	.10
A < B	.15	.15	.3
Total	.5	.5	1.0

In the table, 60% of all ratings were those in which A showed greater expectation than B, 10% were tied ratings, and 30% were those in which B had greater expectation than A. For the sign test, all tied cases are dropped from the analysis, and the sample is correspondingly reduced in size, so the sample becomes the number of matched pairs whose ratings were different.

The sign test involves the subtraction of the expected proportion of disagreement (.5 in all cases) from the observed value (the proportion of all disagreements in which A had greater expectation than B). This difference of observed from expected value is divided by the standard deviation of the theoretical proportion which will consist of the square root of ¼ n times the proportion of disagreements. The resulting statistic, z, is approximately normally distributed with a mean of zero and a standard deviation of one. The formula could be written as:

$$z = \frac{\frac{a}{b} - e}{\sqrt{\frac{1}{4(n)(b)}}}$$

where

a = proportion of A > B
b = total of A > B and A > B expressed as proportion
e = expected proportion under null hypothesis
n = number of comparisons.

FINDINGS

Of the 420 sign tests, computed on IBM equipment, a total of 90 yielded differences significant at the .01 level, at which chance could account for for differences only once in 100 comparisons.[36] Thirty-eight of these were at a very high (.005) level, at which chance could account for differences only once in 200 times, and 34 at an extremely high (.0005) level, at which chance could account for differences only once in 2000 comparisons.[37] In addition to the 90 highly significant differences, there were another 42 moderate differences at the .025 level. Including the moderate differences, about one-third of the comparisons yielded significant differences.

It is considered probable that an even greater number of statistically significant differences would have occurred had it not been for the researcher's desire to provide the raters with ample information. As the design worked out in practice, the Glueck scale was given an unusually fair opportunity to yield similar ratings, since in a great many cases the probation officers were asked to provide additional information.

Six of the 42 significant differences at a high level (.01 or higher) in the comparisons without reference to race of ratees were within cells, indicating the possible presence of uncontrolled

The use of the sign test in this way means that we are actually using the sample of 60 ratees, rather than the sample of eight raters, as the sample about which the sign test is employed. We know more about the population of ratees and have a larger sample, so that our confidence in the sign test for the ratee sample can be greater.

For references on sign test, see SIEGEL, NONPARAMETRIC STATISTICS FOR THE BEHAVIORAL SCIENCES 68–72 (1956); DIXON & MASSEY, INTRODUCTION TO STATISTICAL ANALYSIS 280 (2nd ed. 1957); Moses, *Non-parametric Statistics for Psychological Research*, 49 PSYCHOLOGICAL BULLETIN 22 (1952).

[36] The "Table of Probabilities Associated with Values as Extreme as Observed Values of Z in the Normal Distribution" was used in SIEGEL, *op. cit. supra* note 35, at 247.

[37] Statistical tests for differences between pairs of samples may lead to fallacious conclusions because they can capitalize on chance. See *id.* at 159–60. It can be seen, for example, that 10½ significant differences in the 420 sign tests could occur by chance. Hence our hypotheses are largely tested at an .0005 level of significance, at which chance differences could not be expected to occur at all in our study. The reason the differences at the .025 and .01 level of significance are included in the findings is that the majority of the former and substantially all the latter are unidirectional and further serve to substantiate the findings at the .005 and .0005 levels.

variables.[33] Two of these were at an extremely high level, and three at a very high level. None of these differences within rater cells occurred in the Southern-educated white cell. Three occurred in the Northern-educated white cell, related to Rater # 2's only partial Northern education. He received his second year of graduate training in New York, but his earlier education was in Tennessee, and he rated more similarly to the Southern-educated raters than to the other member of the Northern-educated white cell. The implementation of the design was faulty with respect to this rater.

The other three within-cell differences had to do with the two Negro cells. In each cell the older Negro tended to hold higher expectations than the younger Negro, both for white and for Negro cases. These two raters were Raters # 4 and # 7. It can be tentatively assumed that age was an uncontrolled factor in the Negro cells, and the research design should have held age constant. Age differences did not exist in the study between white raters, and it cannot be stated on the basis of the present research whether intra-cell differences would have emerged.

In relating the findings to the specific hypotheses to be tested, it appears conclusive that Hypotheses 1 and 3 are correct. For both Negro and white boys, the Negro raters expected more of parents in terms of discipline, supervision, affection, and cohesiveness than did white raters, when region was held constant. Table II shows four differences at the .0005 level between Northern-educated Negro raters and Northern-educated white raters. The same table shows two differences at the .0005 level between Southern-educated Negro raters and Southern-educated white raters.[39] No contrary differences occurred in which white raters expected more than Negro raters, region constant, at a high level of significance, except for one instance of Rater # 2 (the poorly selected Northern-educated white rater) holding higher expecations than one Negro rater for one factor.

Hypotheses 2 and 4 were disproved by the findings; the opposite conditions were found to exist. For both Negro and white boys, the Southern Negro raters expected more of parents than did Northern white raters.[40] Table II shows seven statistically significant differences at the .0005 level, and two more at the .005 level. A breakdown by race of ratees indicates the same trend, although the differences were sharper for white boys than for Negro. In no instance does a Northern-educated white rater have higher expectations than a Southern-educated Negro rater.

Other findings included a clear indication that the two Northern-educated Negro raters expected more of Negro fathers in regard to affection for boy than did the two Southern-educated white raters.[41] For white parents particularly, Southern-educated Negro raters expected more in regard to parental affection than did Northern-educated Negro raters.[42] For both white and Negro parents, Southern-educated white raters had higher expectations than Northern-educated white raters. Table II indicates two differences at the .0005 level and two at the .005 level.

An analysis of each of the five factors (in Table III) reveals that there are three levels of variability of ratings based on differences in cultural background of raters. The two factors involving affection show the most variability, i.e., raters are most clearly influenced in their judgment of these factors by their cultural background. The two factors involving supervision of boy by mother and cohesiveness of family show less variability but still a substantial amount. The factor concerned with discipline by father shows the least variability, much of that being in reference to white fathers. (The Negro rater tends to expect more of white fathers in regard to discipline than does the white rater. This is particularly true of the older Negro rater.)

The two Northern-educated Negro raters had a pattern of highest expectation for maternal super-

[33] By "within-cells" is meant that a Southern-educated Negro rater, for example, had a statistically significant difference from the other Southern-educated Negro rater with respect to one of the Glueck factors.

[39] These same trends showed clearly in tables for white ratees only and for Negro ratees only, which are not reproduced in this paper for reasons of space.

[40] Since the original hypotheses were disproved, it can only be said that there is support for a new hypothesis. Further verification would seem to be indicated.

[41] This information is not contained in Table II, but there were three differences significant at the .01, .005 and .0005 levels respectively. This finding, incidentally, appears to relate to the relatively tenuous role of the father in the Negro family in the South, see FRAZIER, THE NEGRO FAMILY IN THE UNITED STATES 219, 362–68 (Abridged ed. 1951), and the Southern-educated white's tolerance toward this weak role.

[42] This information is not contained in Table II, but there were three differences significant at the .005 level. This finding appears to relate to the traditional deference for whites on the part of Southern Negroes as compared with Northern Negroes.

TABLE II

INCIDENCE OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN RATERS FOR ALL RATEES

	Northern-Educated White		Northern-Educated Negro		Southern-Educated White		Southern-Educated Negro	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8
.0005 level		>Rater 1 (4) >Rater 3 (4)	>Rater 1 (5)	>Rater 1 (2) >Rater 1 (3) >Rater 1 (4) >Rater 3 (4) >Rater 6 (3)	>Rater 1 (2)	>Rater 1 (2)	>Rater 1 (1) >Rater 1 (3) >Rater 1 (4) >Rater 1 (5) >Rater 2 (3) >Rater 3 (4) >Rater 6 (3)	>Rater 1 (2) >Rater 1 (3) >Rater 6 (3)
.005 level		>Rater 1 (5)	>Rater 1 (1) >Rater 8 (5)	>Rater 2 (3) >Rater 3 (3) >Rater 5 (3) >Rater 7 (2) >Rater 8 (4)	>Rater 1 (1) >Rater 1 (5) >Rater 3 (4)	>Rater 3 (4)	>Rater 2 (5) >Rater 3 (3) >Rater 5 (3) >Rater 8 (5)	>Rater 2 (3) >Rater 3 (4)
.01 level		>Rater 1 (2)	>Rater 2 (5)	>Rater 6 (4)	>Rater 1 (4)			
.025 level		>Rater 6 (4)	>Rater 1 (2) >Rater 4 (5)	>Rater 1 (1) >Rater 2 (2)		>Rater 1 (5) >Rater 7 (2)	>Rater 2 (1) >Rater 8 (4)	>Rater 1 (4) >Rater 3 (3) >Rater 5 (3)

Note: The number of the Glueck factor is indicated in parentheses after the significant difference. Thus under "Rater 2" the first entry means that Rater 2 held higher expectations than Rater 1 in regard to "Affection of Mother for Boy," which is the Glueck factor #4, at an .0005 level of significance. This table shows only part of the significant differences found. Similar tables for Negro ratees only and for White ratees only are included in the dissertation.

TABLE III

TOTAL NUMBER OF SIGNIFICANT DIFFERENCES

(By Rater, Cell, and Factor)

	Northern-Educated White		Northern-Educated Negro		Southern-Educated White		Southern-Educated Negro		Total
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	
Factor #1 (Discipline)			2	2	1		6		11
Factor #2 (Supervision)		1	2	7	3	5	1	3	22
Factor #3 (Paternal Affection)		1		13*			13*	11	38
Factor #4 (Maternal Affection)		7		10	4	3	8	5	37
Factor #5 (Cohesiveness)		1	9		2	3	9		24
Total		10	13	32	10	11	37	19	132

* From a possible total of 21.

vision, somewhat more for Negro parents than for white parents.[43] The Southern-educated Negro raters had a pattern of highest expectation for paternal affection.[44] More significant differences emerged for this factor than for any other single parental factor.

The Southern-educated Negro raters also held highest expectations of any of the groups for maternal affection, the fourth factor. The two Northern-educated Negro raters had highest expectation for family cohesiveness, the fifth factor.[45]

An analysis of the Rater Judgment Questionnaires revealed that Negro raters did not often relate judgments to race. Generally, the raters attributed judgments more often to class than to race, but to "other factors" more often than to class. All raters except one paid an average of three times as much attention to the father's role than to the mother's, in explaining why they rated as they did.

DISCUSSION

Negro raters in the study clearly had higher parental expectations than white raters, regardless of whether the raters were Southern-educated or Northern-educated. This finding is true for both Negro and white cases, bearing out a tacit hypothesis that raters will endeavor to equalize their ratings as far as racial considerations go.

There appears to be an indication that Southern-educated raters tended to have higher expectations than Northern-educated raters, when race was held constant. In the case of Negro raters, this was particularly true for white parents. Northern-educated Negro raters tended to hold somewhat higher expectations for Negro parents, particularly in regard to maternal supervision and family cohesiveness.

The error made in the set of hypotheses related to region appears to stem from the fact that the literature pointed the way much more clearly to *parental standards* than to *rater expectations*. The fact that Northern standards of family life tend to be higher than Southern standards of family life for the lower-class, which is largely represented on delinquent caseloads, does not necessarily justify the prediction that Northern-educated social workers will expect more from parents than will Southern-educated social workers. Southern-educated social workers may be well aware of the relatively low standards in the South, but as in the case of Negro social workers may be so eager to raise the standards that expectations are raised considerably. From the other direction, the Northern-educated social worker working in the South may well have lowered his expectations considerably in recognition of the culturally-impoverished family conditions in the Southern lower-class.

[43] This finding apparently relates to the Northern-educated Negro's relative rejection of the traditional matriarchal Southern Negro family structure, contrasted with the Southern-educated Negro's tolerance of the strains on the Negro woman in functioning in both bread-winning and child-rearing roles.

[44] Consistently in the research the Southern-educated Negro raters seemed critical of the weak Southern Negro father, and protective of the Southern Negro mother.

[45] Again this finding relates to the differences between Northern Negro raters and Southern Negro raters in regard to the role of the mother in the Southern Negro family.

The three underlying hypotheses were supported by the study. The findings clearly show that cultural background differences account for different attitudes and judgments regarding parental factors. For example, cultural background differences accounted for higher expectations for parental behavior on the part of Negro social workers as compared with white social workers, when both were asked to judge the same information on the Glueck Scale.

We find systematic differences in ratings, if the raters vary in regional background and in race with its sociocultural implications. Within the Negro group, the older Negro expects more from parents than the younger Negro. As cultural differences between Negro and white social workers blur, one would expect differences in culturally-caused variations in ratings to become less sharp. This is precisely the case.

On the basis of the findings, one could predict ratings or judgments as to parental factors on the Glueck Scale with reasonable certainty on the basis of racial (and with less assurance, for regional) background. A rank of expectations might logically be, with highest expectations at top:

1. Southern-educated Negro social workers.
2. Northern-educated Negro social workers.
3. Southern-educated white social workers.
4. Northern-educated white social workers.

(As pointed out earlier, for Negro parents in regard to maternal supervision and family cohesiveness, Northern-educated Negro social workers exceed Southern-educated Negro social workers in expectations; this constitutes an exception to the above generalization.)

Knowledge of reference subgroup norms enabled us to hypothesize specific systematic differences in ratings, although the confusion between parental standards and rater expectations led to one erroneous set of hypotheses. It would seem more empirically valid to sharpen this underlying hypothesis to read "knowledge of reference subgroup norms as to expectations regarding parental factors."

Conclusions[46]

Initially the problems were presented as follows: (1) What is the reliability of raters in the use of the Glueck Scale as a predictive instrument? (2) What does rater reliability or lack of it tell us about the validity, accuracy, and efficiency of the Glueck Scale?

The findings of the present research suggest that the ratings made of the Glueck factors are lacking in reliability. The judges or raters do not agree on the category or rating they assign to specific units of behavior.[47] Actually, each rater may be making fairly consistent judgments as to the Glueck parental factors from one case to the next. But there is disagreement between rater cells (Southern-educated Negro vs. Northern-educated white, for example) because the different raters rate with different reference groups in mind. To insure reliability of ratings, from the standpoint of the variability of reference groups, one might perhaps restrict the use of the Glueck Scale to particular reference groups, or develop a clear-cut system of categories. For example, all Southern-educated Negro social workers employed in urban areas of the South who are male, middle-class, and of the same age might have about the same images, concepts, values, and norms in mind as they use the Glueck Scale to predict delinquency. Their ratings then can be expected to be reliable. But their ratings cannot be pooled with those of Northern-educated white social workers, since the latter group would be seeing different images and norms. The Glueck Scale cannot be used by raters from different cultural backgrounds if reliable ratings are to be obtained.

The reliability of the Glueck factors themselves is open to question as a result of this study. As previously mentioned, the reliability of the rater and the reliability of the behavior being rated are separate problems. But only after rater reliability has been established can one tackle the problem of behavior reliability.[48] The Glueck factors are complex, highly inferential variables for which adequate external criteria are only partially available.[49] It is to be noted from the findings that

[46] Various other implications of the study will be discussed in future articles.

[47] This conclusion is in line with previous studies such as that of Kelly and Fiske, which indicated that the human being is quite fallible as a measuring instrument and that rater reliability tends to be quite low. *Supra* note 21, at 406.

[48] Heyns & Zander, *supra* note 19, at 410.

[49] It is apparent that the Glueck social factors are actually middle-class values, and it is understandable that the lower-class children so often found on delinquency caseloads have not been reared according to them. One could say that these middle-class values, enforced by middle-class judges and police, define a particular kind of delinquency. If we could conceive of *lower-class* parental values as becoming in the future the norms for police and judges, we could then conceive of a boy being committed to a training school

corporal punishment by father or lack of it is such a clear-cut external criterion for parental discipline that that factor showed less variability. But the other factors, particularly the extremely complex factors on affection, lack such clear-cut criteria.

Validity and reliability are very closely related, and it has been stressed that a predictive instrument which is unreliable cannot be valid. Most validity problems arise in connection with rating scales requiring a good deal of inference on the part of raters.[50] In view of the demonstrated lack of reliability in ratings, and the apparent lack of reliability of the behavior being rated in the Glueck Scale, the validity of the Glueck Scale as an instrument for the prediction of delinquency is dubious. In view of the doubt as to its validity and reliability, one cannot accept the Glueck Scale at this time as an accurate and efficient instrument for delinquency prediction. Future instruments refining the Glueck approach must take into account the different cultural backgrounds of raters or a category system with very clear-cut external criteria will need to be developed.[51]

because he is effeminate and overprotected (and happens to break a lower-class valued law). The judge might well feel that he needs institutional care to make a man out of him. Sometimes lower-class police assume this position now.

[50] Heyns & Zander, *supra* note 19, at 409.

[51] This article was prepared in July, 1961, on the basis of research carried out in 1960. Eleanor T. Glueck has subsequently conceded the lack of reliability in ratings of four of the present factors and has proposed a new prediction scale based on the three factors: supervision of boy by mother, discipline of boy by father, and rearing by affectionless parent substitutes. E. T. Glueck, *Toward Improving the Identification of Delinquents*, 53 J. CRIM. L., C. & P.S. 164 (1962). On the basis of the present research, the reliability of ratings of maternal supervision, at least, remains questionable.