EVALUATING BEHAVIORS IN NATURALISTIC SETTINGS: ISSUES OF RELIABILITY, VALIDITY, AND OBSERVER BIAS

Sara S. Sparrow, Yale University
Domenic V. Cicchetti, V.A. Hospital, West Haven, Connecticut
JoAnn Robinson, Cornell University

Introduction

The purpose of this paper is to discuss the issues of reliability, experimenter bias, and validity as they apply to various aspects of the behavior of moderately, severely, and profoundly retarded, institutionalized children. As noted recently by Longabaugh (1977), only a few published studies have even considered these crucial variables. Instead, the observations of single experimenters have usually been assumed to represent a valid assessment of a wide range of behavior occurring in a variety of naturalistic settings. It is our contention that this viewpoint is inconsistent with the reports of broad bodies of literature in both medicine and the behavioral sciences. Specifically, a number of clinical investigators have pointed to the fact that observer variability is an essentially ubiquitous phenomenon spanning multiple and diverse areas of medicine (e.g., Cicchetti, 1977: Cicchetti & Conn, 1976; Etter, Dunn, Kammer, Osmond, & Reese, 1960; and Koran, 1975a and 1975b). In fact, it is rather common for *inter*observer disagreement in medical diagnoses to range between about 25-30%, and for *intra*observer variation to reach proportions of 15-20%. To mention a second major area of clinical investigation, Helzer and associates (1977) report much higher overall agreement in the assessment of neuropsychiatric diagnosis than previous studies. Nonetheless, the extent of interobserver disagreement on *specific* categories of illness showed considerable variability (e.g., 29% disagreement in diagnosing for alcoholism). The same general results occur in the field of mental retardation. For example, Balthazar reports generally acceptable levels of interrater agreement in assessing various behaviors of mentally retarded children. Yet in one reliability study, independent observers agreed only between 42% and 69% in the rating of 16 of 64 areas of behavioral assessment (Balthazar, 1973).

Sources of the Data

This report focuses upon the behavior of mentally retarded children as it occurs and develops in naturalistic institutionalized settings. The data derive more generally from a longitudinal investigation of the effects of a sensorimotor patterning treatment program on the behavior of mentally retarded children in residence at the Seaside Regional Center in New London, Connecticut. Specific sources of data, as displayed in Figure 1, are based upon: (1) Levels of cognitive, psychomotor, social, and self-control behavior, as measured by the Behavior Rating Inventory for the Retarded (BRIR), due to Sparrow and Cicchetti (1977); (2) Results of standardized IQ tests, such as the Catell and the Stanford Binet; (3) Results of performance on unstandardized tests, constructed by the senior author (which included assessment of motor and language development); and (4) Direct behavioral observations (assessing levels of affect, communication, activity, and play). Since the data arising from the first three sources will appear in future publications, this report will be based mainly upon data derived from direct observations of the behavior of mentally retarded children.

Analyses of the Data

In a recent study (Cicchetti, 1977) the issues of reliability, bias, and validity were discussed in the context of medical investigations. This report will focus upon these issues in the field of mental retardation.

Reliability

When we speak of observer reliability, we are concerned with the extent to which independently derived measurements or judgments agree or are interchangeable one with the other. Reliability can be assessed either between two or more independent observations of the same phenomenon (*inter*observer reliability) or within the same observer (*intra*observer reliability). Further, with respect to qualitative data, we can speak in terms of either *overall* agreement or *specific* agreement. Thus, in our research, we assessed interobserver reliability in rating the *communication* level of a group of 49 mentally retarded children, on a six category ordinal scale, as one of the following: (1) none: (2) prespeech sounds only: (3) gestures or sounds; (4) talking to self; (5) noncommunicative speech; or (6) echolalic speech. Using ordinal weighting systems developed by Cicchetti (1976) with the weighted kappa statistic due to Cohen and colleagues (e.g., Cohen, 1968; and Fleiss, Cohen & Everitt, 1969), we assessed both overall observer agreement as well as interobserver *specific* agreement for each of the six categories of the scale. The formulae for the specific agreement indices were recently developed by Cicchetti, Fontana, and Noel Dowds (1977) and are available upon request. The results in Table 1 show that the overall level of agreement is extremely high, even when corrected for the amount of agreement expected by chance alone. Thus, we obtained 95.69% observer agreement (PO); as against 76.56% expected by chance (PC). The level of chance-corrected agreement or kappa (PO-PC)/(1-PC) was .82, with +1 representing perfect chance-corrected agreement. It is interesting to note that the indices of specific rater agreement are also quite respectable, with PO values ranging between 91.67% and 100% agreement, and chance-corrected or specific kappa values ranging between .66 and 1.00. (The value of .66, it should be noted, is consistent with data presented by Koran (1975a; 1975b) for a wide range of clinical judgments, across many diverse fields of medical diagnosis.)

A second variable, level of play, was independently observed in 30 mentally retarded children, and was assessed by a four category ordinal scale as one of the following: (1) does not play at all; (2) plays with non-toy object(s); (3) uses toy(s) inappropriately; or (4) uses toy(s) as intended. These data are given in Table 2 and once again show very high levels of overall agreement, specific agreement and chance-corrected agreement. Thus, PO and overall kappa values are 97.33% and .92, respectively, while specific agreement indices (SO) range between 95.00% and 100%. Chance-corrected specific agreement levels range between .78 and 1.00.

A third variable which independent raters observed was affect, scored as 1 = no or negative affect: and 2 = positive affect. Thirty-nine children were available for assessment on this variable. Results in Table 3 showed that, consistent with the data for communication and play, overall interobserver agreement was high (PO = 94.87%; overall kappa = .72; SO (negative affect) and SO (positive affect) were 97.14% and 75%, with chance-corrected agreement being .72 in each case).

Finally, we assessed levels of interobserver agreement for level of physical activity which could be rated as: 1 = sleeping or no movement; 2 = prone with some movement: 3 = sitting in wheelchair: 4 = sitting or kneeling; 5 = standing; 6 = crawling or creeping; 7 = walking; and 8 = running. As for each of the other variables, overall levels, as given in Table 4, were very high (PO = 99.22%; Overall kappa = .93; SO values ranged between 92.31% and 100%; and chance-corrected specific kappa values ranged between .73 and 1.00).

Observer Bias

The question of observer bias is one of the extent to which one observer evaluates a given phenomenon systematically differently than other observers who have independently assessed the same phenomenon. Thus, to the extent that agreement is very high, and disagreements tend to occur in an essentially random pattern, observer bias does not occur. However, when it does occur it suggests that the observers are not always using the same frames of reference to make the same judgments. As noted by a number of investigators, Longabaugh (1977); Johnson and Bolstad (1973); and Reid (1970), even well-trained observers whose reliability has not been assessed periodically may become biased with respect to their judgments. This phenomenon is referred to as either *observer drift* or *instrument decay* (Campbell & Stanley, 1966). As an example of how pervasive the phenomenon can become, one study reports a drop from 70% to 51% in the extent of observer agreement levels as a function of observer drift or instrument decay (Reid, 1970).

With respect to our longitudinal investigation of mentally retarded children, we made periodic checks upon the reliability of our rater pairs but fortunately found essentially no levels of observer drift which were of clinical concern. At least three plausible reasons why we did not observe a phenomenon which others doing naturalistic observations did indeed experience include the following: (1) The observers were very carefully trained in the use of our rating techniques (both standard and nonstandard): (2) The behaviors we rated were very carefully defined into nonoverlapping categories of classification: and (3) The reported levels of instrument decay cited in the literature were based upon observations of nonretarded samples whose range of expression of behavior tends, in the main, to be more varied, less stereotypic, and hence less clearly delineated than the subjects we refer to in our research.

Validity

The phenomenon of validity is one of answering the question: Does our measuring instrument indeed measure what it purports to measure? There are many different types of validity measures, and these have been discussed by numerous authors, including the following: Balthazar and English (1969); Bechtoldt (1959); Cicchetti (1977): Cronbach (1960 and 1971): French and Michael (1966); Greenwood and Perry (1968); Guion (1974); Nihira (1976); and Nihira, Foster, Shellhaas, and Leland (1974). Some of the more familiar types of validity assessment reported in the literature include: (1) content validity; (2) criterion related validity; (3) construct validity; and (4) factorial validity. The paper by Guion (1974) is an excellent reference for a detailed and comprehensive description of the first three types of validity assessment. Most of these were utilized recently by Sparrow and Cicchetti (1977) in their assessment of the validity of the aforementioned Behavior Rating Inventory for the Retarded (BRIR). As one method of assessment, we used factorial validity, a technique utilized by several investigators in the field of mental retardation (e.g., Balthazar & English, 1969: and Nihira, Foster, Shellhaas, & Leland, 1974). As a result of that experience, we strongly recommend that investigators contemplating using this form of validity assessment heed the following advice (which has *not* as a rule been reported in the mental retardation literature): (1) It is preferable to have *a priori* "factors" against which to compare the empirically derived factors. (2) It is wise to use more than one type of major factor analytic technique (e.g., principal components, principal factors, each with orthogonal and oblique rotations). The purpose of this suggestion is to determine the extent to which different techniques might affect the particular factors obtained (following upon the advice of Frane & Hill, 1975). It was our experience, for example, that a principal components, oblique rotation solution produced fewer BRIR items which overlapped two or more factors than did other types of analyses. (3) It is important to report the percentage of overall variance accounted for by the factor analysis.

In summary, this paper has attempted to discuss the issues of observer reliability, observer bias (or drift), and validity in the context of the behavior of institutionalized retarded

children. Although the high levels of reliability achieved in our sample are somewhat at a variance with the assessment of clinical phenomena based upon nonretarded samples, the central issues discussed here appear to have a broad range of applicability to behavioral science, medicine, and other fields of clinical investigation.

As a final note, computer programs for assessing levels of observer reliability and bias are available upon request.

References

Balthazar, E.E. *The Balthazar scales of adaptive behavior. Section II. The scales of social adaptation (BSAB-II).* Palo Alto, California: Consulting Psychologists Press, 1973.

Balthazar, E.E. & English, G.E. A factorial study of unstructured ward behaviors. *American Journal of Mental Deficiency*, 1969, *74*, 353-360.

Bechtoldt, H.P. Construct validity: A critique. *American Psychologist*, 1959, *14*, 619-629.

Campbell, D.T. & Stanley, J.C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1966.

Cicchetti, D.V. Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry*, 1976, *129*, 452-456.

Cicchetti, D.V. Assessing observer and method variability in medicine. To appear in *Connecticut Medicine*, 1977 (by invitation).

Cicchetti, D.V. & Conn, H.O. A statistical analysis of reviewer agreement and bias in evaluating medical abstracts. *Yale Journal of Biology and Medicine*, 1976, *49*, 373-383.

Cicchetti, D.V., Fontana, A.F., & Dowds, B. Noel. Assessing specific category reliabilities for rating scales in behavioral research. Paper presented at meeting of the American Psychological Association, San Francisco, California, August 1977.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, *70*, 213-220.

Cronbach, L.J. Need for critical evaluation of tests. In L.J. Cronbach (2nd ed.) *Essentials of Psychological Testing.* New York: Harper & Row, 1960, pp. 96-125.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Etter, L.E., Dunn, J.P., Kammer, A.G., Osmond, L.H., & Reese, L.C. Gastroduodenal X-ray diagnosis: A comparison of radiographic technics and interpretations. *Radiology*, 1960, *74*, 766-770.

Fleiss, J.L., Cohen, J., & Everitt, B.S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, *72*, 323-327.

Frane, J.W. & Hill, M.A. Annotated computer output for factor analysis: A supplement to the writeup for computer program BMDP4M. In W.J. Dixon (Ed.) *BMDP Biomedical Computer Programs.* Los Angeles: University of California Press, 1975.

French, J.W. & Michael, W.B. (and the joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education). *Standards for Educational and Psychological Tests and Manuals.* Washington, D.C.: American Psychological Association, 1966, pp. 1-40.

Greenwood, D. & Perry, R. Use of the Adaptive Behavior Checklist as a means of determining unit placement in a facility for the retarded. A paper presented at the meeting of the Rocky Mountain Psychological Association, Denver, Colorado, May 1968.

Guion, R.M. Open a new window: Validities and values in psychological measurement. *American Psychologist*, 1974, *29*, 287-296.

Helzer, J.E., Clayton, P.J., Pambakian, R., Reich, T., Woodruff, R.A., & Reveley, M.A. Reliability of psychiatric diagnosis. II. The test/retest reliability of diagnostic classification. *Archives of General Psychiatry*, 1977, *34*, 136-141.

Johnson, S.M. & Bolstad, O.D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L.A. Hamberlynch, L.C. Handy, & E.J. Mash (Eds.), *International Conference on Behavior Modification, Behavior Change.* Champaign, Illinois: Research Press, 1973.

Koran, L.M. The reliability of clinical methods, data and judgments. *The New England Journal of Medicine*, 1975, *293*, 642-646 (First of Two Parts).

Koran, L.M. The reliability of clinical methods, data and judgments. *The New England Journal of Medicine*, 1975, *293*, 695-701 (Second of Two Parts).

Longabaugh, R. The systematic observation of behavior in naturalistic settings. Manuscript, in preparation, 1977.

Nihira, K. Dimensions of adaptive behavior in institutionalized mentally retarded children and adults: Developmental perspective. *American Journal of Mental Deficiency*, 1976, *81*, 215-226.

Nihira, K., Foster, R., Shellhaas, M., & Leland, H. *AAMD Adaptive Behavior Scale*, 1974 revision. Washington, D.C.: American Association on Mental Deficiency, 1974.

Reid, J.B. Reliability assessment of observation data: A possible methodological problem. *Child Development*, 1970, *41*, 1143-1150.

Sparrow, S.S. & Cicchetti, D.V. The behavior rating inventory for the retarded (BRIR): A scale applicable to moderate, severe, and profound retardation. To appear in *American Journal of Mental Deficiency*, (January) 1978.

FIGURE 1

BEHAVIORAL ASSESSMENT OF MENTALLY RETARDED CHILDREN

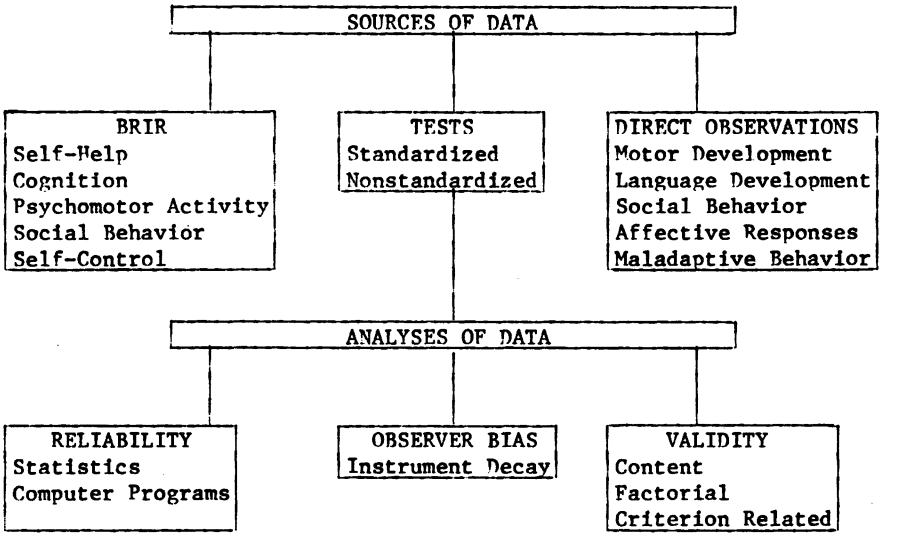

TABLE 1

OBSERVER AGREEMENT IN RATING THE HIGHEST LEVEL OF COMMUNICATION OF SERIOUSLY RETARDED CHILDREN

Category of Communication	Average Rater Frequency of Application	Index of Observer Agreement		
		Obtained	Expected	Chance-Corrected
(1) None	.51	.9733	.8118	.86
(2) Prespeech Only	.31	.9407	.8254	.66
(3) Gestures or Sounds	.06	.9259	.7120	.74
(4) Talking to Self	.08	.9167	.5420	.82
(5) Noncommunicative Speech	.02	1.0000	.3447	1.00
(6) Echolalic Speech	.02	1.0000	.1882	1.00
Entire Scale	1.00	.9569	.7656	.82

TABLE 2

OBSERVER AGREEMENT IN RATING THE HIGHEST LEVEL OF PLAY ACTIVITY OF SERIOUSLY RETARDED CHILDREN

Category of Play	Average Rater Frequency of Application	Index of Observer Agreement		
		Obtained	Expected	Chance-Corrected
(1) No Play	.13	1.0000	.3567	1.00
(2) Play with Non-Toys	.28	.9882	.7035	.96
(3) Inappropriate Play with Toys	.27	.9500	.7725	.78
(4) Appropriate Play with Toys	.32	.9684	.6435	.91
Entire Scale	1.00	.9733	.6567	.92

TABLE 3

OBSERVER AGREEMENT IN RATING AFFECT LEVELS OF SERIOUSLY RETARDED CHILDREN

Category of Affect	Average Rater Frequency of Application	Index of Observer Agreement		
		Obtained	Expected	Chance-Corrected
(1) None or Negative	.10	.7500	.1026	.72
(2) Positive	.90	.9714	.8974	.72
Entire Scale	1.00	.9487	.8159	.72

TABLE 4

OBSERVER AGREEMENT IN RATING HIGHEST LEVEL OF ACTIVITY OF SERIOUSLY RETARDED CHILDREN

Category of Activity	Average Rater Frequency of Application	Index of Observer Agreement		
		Obtained	Expected	Chance-Corrected
(1) Sleeping or No Movement	.02	1.0000	.4301	1.00
(2) Prone with Some Movement	.00	NA[1]	NA	NA
(3) Sitting in Wheelchair	.04	1.0000	.7975	1.00
(4) Sitting or Kneeling	.34	.9930	.9096	.92
(5) Standing	.41	.9923	.9356	.88
(6) Crawling or Creeping	.18	.9915	.8587	.94
(7) Walking	.01	.9231	.7174	.73
(8) Running	.00	NA	NA	NA
Entire Scale	1.00	.9922	.8946	.93

[1]Note. NA denotes not applicable