DOCUMENT RESUME

ED 233 817 PS 013 818

AUTHOR MacPhee, David
TITLE What Do Ratings of Infant Temperament Really Measure?
PUB DATE Apr 83
NOTE 24p.; Paper presented at the Biennial Meeting of the Society for Research in Child Development (Detroit, MI, April 21-24, 1983).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Error of Measurement; Experimenter Characteristics; *Infants; Knowledge Level; Mothers; *Personality Measures; Racial Differences; Rating Scales; Social Differences; *Test Reliability; *Test Validity
IDENTIFIERS *Infant Temperament Questionnaire

ABSTRACT

As data on the reliability and validity of ratings of infant temperament have accumulated, researchers have begun to ask what caregiver ratings really measure. An argument has been made that ratings of social behavior are less a reflection of enduring individual differences than a measure of rater characteristics and error variance. This study investigated the possibility that extraneous sources of variance would be found in Carey's Infant Temperament Questionnaire (ITQ). A large, diverse sample of mothers of 6-month-old infants completed ITQ's. Item analyses first identified a cluster of items left blank by a select, middle-class group of mothers. A large number of items (40 percent) were found to exhibit marked response biases, and parental characteristics were shown to affect response biases and styles, implying that systematic biases are built into the scale. Temperament ratings also varied with social class, ethnicity, and knowledge of infant development. Infants judged as being temperamentally easier tended to have mothers who were white, middle-class, and more knowledgeable than others. Results suggested that caregiver ratings on the ITQ reflect more than mere stylistic differences in infant behavior. (Author/RH)

What do Ratings of Infant Temperament Really Measure?

David MacPhee

Department of Psychology

University of North Carolina

Chapel Hill, North Carolina 27514

Paper presented at the biennial meeting of the Society for Research in Child Development, Detroit, MI, April, 1983.

Abstract

As data on the reliability and validity of ratings of infant temperament have accumulated, researchers have begun to question what caregiver ratings really measure. Cairns and Green (1979) have argued that ratings of social behavior are less a reflection of enduring individual differences than a measure of rater characteristics and error variance. This study investigated these extraneous sources of variance in Carey's Infant Temperament Questionnaire (ITQ).

A large, diverse sample of mothers completed ITQs on their 6-month-olds. Item analyses first identified a cluster of items that were left blank by a select, middle class group of mothers. Next, a large number of items (40%) were found to exhibit marked response biases. Finally, parental characteristics were shown to affect response biases and styles, implying that systematic biases are built into the scale. Temperament ratings also varied with social class and ethnicity as well as knowledge of infant development. Infants judged as easier tended to have mothers who were white, middle class, and more knowledgeable. These results suggest that caregiver ratings on the ITQ reflect more than mere stylistic differences in infant behavior.

A major issue in the study of infant temperament concerns the extent to which caregiver ratings accurately portray infant behavior. Confidence in the validity of such ratings has been eroded by recent findings from two areas of research. First, studies of mother-father and parent-observer agreement typically find correlations on the order of .20 to .50 (Bates, Freeland & Lounsbury, 1979; Wilson & Matheny, 1983). Given that less than 25% of the variance is being explained, this suggests that something other than stable individual differences is being assessed. Second, attempts to document relationships between parental attributes (e.g., SES; race; attitudes; personality traits) and ratings of infant temperament usually are quite successful (Bates, et al., 1979; Crockenberg & Credolo, 1983; Simonds & Simonds, 1982; for reviews, see Bates, 1980, and Hubert, Wachs, Peters-Martin & Gandour, 1982). In fact, parental attributes have been found to be more predictive of temperament ratings than infant characteristics (Sameroff, Seifer & Elias, 1982), raising the possibility that ratings may only be projections of the parent onto the rating scale. The lack of validity and interrater reliability has been variously attributed to flaws in the instruments; to differences in familiarity with the target child; to behavior that is person- or situation-specific; and to the influence of subjective perceptions by the parent.

Taking a broader view of ratings of social behavior clarifies the sources of variance in temperament ratings. The analysis by Cairns and Green (1979) of ratings versus observations of social behavior (see Figure 1) may not change some researchers' antipathy toward ratings, but at least the reasons for any pessimism are apparent. These authors argue that while the goal of ratings is to capture enduring individual traits, most of the variance actually is due to characteristics of the rater or to error variance. Rater characteristics include idiosyncratic interpretations of the construct, knowledge of the individual child and the reference population, and scaling of the individual onto the group distribution.

Parental expectations and beliefs also may mediate temperament ratings, and would be most evident when looking at the effects of parity (Goodnow, in press) or subcultural differences. In contrast, observational techniques primarily reflect the influences of the setting and the dyad. Thus, strong correlations between parent ratings of temperament and observational measures would not be expected because the two techniques serve different purposes. Furthermore, two raters might differ in their familiarity with the target infant and infants in general; in their interpretation of the scale items and metric; and in their expectations and beliefs. All of these will conspire to depress interrater reliability. Such an analysis of the task confronting raters calls into question the validity of infant temperament ratings as pure measures of infant behavior.

This study is particularly concerned with two extraneous sources of variance mentioned by Cairns and Green. First, if "instrument failure" is the primary culprit for the poor reliability and validity of temperament scales, as Hubert, et al. (1982) concluded, then this area of inquiry could benefit from an analysis of the error variance that reflects poor methodology. That is, a consideration of response biases and styles, problems of scaling, and item selection might illuminate techniques for reducing error variance, increasing our confidence that we indeed are measuring infant temperament. Among some of the effects contributing to error variance are response styles and biases. A response style is an individual difference "which is an artifactual product of the measurement methods and is at least partially independent of the trait which the measurement methods are intended to measure" (Nunnally, 1967, p. 593). On a temperament scale with 6 options, for instance, some parents may select only the extreme choices while others may consistently choose the middle two. A response bias, in turn, is a group measurement artifact, such as choosing the "Almost always" option.

Finally, biases may be introduced through the selection of items that do not apply to a select group of subjects, or items that are highly correlated with variables unrelated to temperament. All of these extraneous factors become a property of the rating scale during its construction and inflate the error variance to a greater or lesser extent.

The second source of variance this study investigates is rater characteristics, especially those that might affect the "anchor". Anchoring effects, which pertain to the "origin and unit of the subjective continuum" (Torgerson, 1958, p. 79) the parent uses in rating the infant, will arise from the individual's experiences that shape expectations, or perceptions of what is typical and atypical. These expectations then may be used as a reference point for deciding where the target infant should be placed on the item. Parents who are well-versed in infant behavior and development might have a different anchor--and so rate their infants differently--than parents who have been exposed to only a few infants other than their own. Similarly, cultural differences in expectations may affect ratings.

In sum, Cairns and Green's (1979) analysis of sources of variance in ratings of social behavior is applied to Carey and McDevitt's (1978) Infant Temperament Questionnaire (ITQ). The first issue to be addressed will be properties of the ITQ that might contribute to error variance, such as response styles, response biases, and biases due to item selection. The second area of investigation concerns the effects of parental expectations on temperament ratings. These expectancies for infant behavior--hypothesized to vary with social class, ethnicity and knowledge of infant development--are assumed to mediate temperament ratings.

Method

Subjects

Mothers of 6-month-old infants were selected from the birth records of a regional referral hospital in Chapel Hill, subject to the following criteria:

(1) no life-threatening complications during pregnancy and delivery; (2) birthweight greater than 2700 grams; and (3) 5-minute Apgar greater than 5. Of the 434 potential subjects, 26% could not be contacted because they did not have a telephone or they had moved from the area. Four percent of those contacted declined to participate in the study, and an additional 16% failed to return the questionnaires, leaving a sample of 256 subjects. The sample represents a diverse group of mothers in terms of education ($\bar{X}$ = 13.5 yrs.; range = 7-22); age ($\bar{X}$ = 26.2 yrs.; range = 16-43); race (60% white); parity ($\bar{X}$ = 1.66 children; range = 1-6); and social class (Duncan SEI $\bar{X}$ = 37.6; range = 0-96).

Procedure

All mothers who agreed to participate were mailed the Knowledge of Infant Development Inventory (MacPhee, Note 1), a 75-item questionnaire that assesses familiarity with infant norms and milestones; principles of growth and development; parenting strategies and skills; and health care and safety issues. In addition, 196 of the subjects completed the ITQ. The ITQ contains 95 items in 9 categories. Subjects rate the infant's behavior on a 6-point scale ranging from "Almost never" to "Almost always". An additional series of questions solicits the rater's general impressions of the infant's behavior on each of the 9 temperament dimensions. The completed forms were returned in envelopes supplied to each mother.

Results

Response Biases and Styles

Biases may be built into a rating scale by including items with skewed frequency distributions (due to anchoring effects), or by choosing items that are not relevant to a select group of parents or infants. Items that are often left blank on the ITQ provide an illustration of this latter type of bias. Item analyses identified a cluster of 11 items that were not answered by at least 10% of the sample; these items accounted for 55% of the missing data on the ITQ (see

Appendix A). Most of them (6) pertained to adaptability and approach to new people, procedures, or settings. Five items concerned bottle or solid feedings, experiences that may not apply to most breastfed 6-month-olds--as mothers' comments on the items implied. Finally, the correlation between social class and the number of items left blank ($r = .35$, $p < .0001$) suggests that much of the missing data on the ITQ can be attributed to events not encountered by a select, usually middle class group of infants.

Biases due to anchoring effects will be observed on items where one response option is chosen with disproportionate frequency. On the ITQ, for example, each of the 6 options would be chosen 16.67% of the time under a random model. In fact, however, 38 of the 95 items had an option that was chosen at greater than twice the expected frequency. In general, mothers had a predeliction for the extremes, especially the "Almost always" option. As Figure 2 demonstrates, reversing the items merely transplants the response set from one extreme to the other, with only a modest reduction in bias. Finally, 12 items exhibiting markedly skewed distributions were examined (see Appendix B). It becomes apparent upon reading the items that they describe some general properties of infants rather than behaviors that capture individual differences in temperament. Evidently, most mothers believe that infants are active and distractible, which places the "anchor" for these items at the extreme.

The particular strategy or response style the rater uses also will introduce rater variance that is unrelated to the characteristics of the infant being rated. Mothers, like many raters, appear to choose either the extreme or middle options. As Figure 3 illustrates, mothers who chose "Almost never" (N1) were also likely to select "Almost always" (N6), but not options 3 or 4 which, in turn, were correlated with each other. Middle class mothers tended to be more conservative in that they were more likely to use option 3 ($r = .34$, $p < .0001$ for SES and N3)

while lower class mothers showed a greater preference for "Almost always" (r = -.29, $p < .0001$ for SES and N6). Again, attributes and strategies of the rater are found to influence ratings of infant temperament.

If response bias on the ITQ is jointly determined by the inclusion of items with skewed distributions and by some raters who rigidly select only one or two options, perhaps this source of variance can be reduced by eliminating certain items and subjects from the analyses. Accordingly, the dozen items in Appendix B and 14 subjects who chose one option at least 50% of the time were dropped. Now when the frequency of choosing each option is examined, the largest deviation from the expected rate is 3.5%, as compared to 8.9% in the original raw data. Ratings of infant characteristics on the ITQ thus might be more valid if error and rater variance were reduced.

Maternal Characteristics and Temperament Ratings

Because several of the ITQ category scores are intercorrelated, the first step in the next stage of analyses was to factor analyze the data. Four of the five ITQ categories on the first factor (see Table 1) are used in calculating the infant's Easy/Difficult classification. While this lends some validity to the procedure, it also suggests a more parsimonious scheme for summarizing the data--one using a continuous rather than discrete variable. Factor 2 is consistent with Buss and Plomin's (1975) discussion of activity level in terms of vigor and tempo. Finally, an infant with a high score on Factor 3 could be described as apathetic or phlegmatic: low in persistence and relatively unresponsive to stimuli.

The factor scores then were used as the dependent variables in a 2 (SES) X 2 (Race) MANOVA. A marginal main effect for social class and a strong effect due to race were observed on the Easy/Difficult factor (see Figure 4), replicating Sameroff, et al.'s (1982) results with an earlier version of the ITQ. No differences

emerged on Factor 2 but, on Factor 3, lower SES infants regardless of race were rated as more phlegmatic. As in the Sameroff, et al. (1982) study, the latter effect seems to be carried more by scores on Threshold (r = -.21, p < .005 with SES), rather than Persistence (r = .10, p = .15 with SES). Thus, lower SES mothers rated their infants as being somewhat more apathetic and difficult. Differences became particularly evident when examining the relationship between race and the Easy/Difficult factor. These effects cannot be explained by a confounding of race and social class because there were enough middle SES Blacks in this sample (N=16) to explore the independent contributions of the two variables to temperament ratings. Whether these racial differences are attributable to constitutional differences in infant temperament or to cultural influences on parental interpretations of infant behavior is not clear but, in light of contemporary interactionist positions (Freedman, 1974; Bates, 1980), the issue may be moot.

Cairns and Green (1979) have argued that one task confronting a rater is to interpret the target infant's attributes within the framework of a reference population. That is, the rater must be familiar with the group distribution--or norms--in order to scale the individual onto it. Parents' knowledge of infant development, therefore, should be related to their ratings of infant temperament. Correlations between scores on the knowledge scale and ratings on the ITQ partially confirm this hypothesis (see Table 2). Mothers having more accurate knowledge rated their infants as being easier in temperament. Neither method covariance--a mother's skill on paper-and-pencil tests--nor the process described by Cairns and Green, though, can explain the lack of significant results on Factors 2 and 3. It is plausible, therefore, that more knowledgeable mothers are more skilled at maintaining their infants in a easy, contented state; or these mothers may be more confident and relaxed with their babies and so rate them as being easier.

The latter interpretation is supported by Kronstadt, Oberklaid, Ferb, and Swartz (1979) and by Sameroff, et al. (1982), who found correlations of similar magnitude between maternal anxiety and ratings of the infant as difficult on four of the same ITQ categories (Approach; Adaptability; Mood; and Rhythmicity).

An indication of how strongly maternal characteristics contribute to ratings of infant temperament can be gleaned from 3 multiple regressions, using SES, race, and knowledge scores to predict the ITQ factor scores. For the Easy/Difficult factor, $R^2 = .24$ [$F(3,190) = 9.96$, $p < .0001$] with all three predictors making significant contributions, especially knowledge. The equation for Factor 2 was non-significant but for Factor 3, $R^2 = .08$ [$F(3,190) = 2.70$, $p < .02$] with SES making the only significant contribution. Rater characteristics thus do influence ratings of infant temperament to a modest but significant degree.

Discussion

In conclusion, what do ratings of infant temperament really measure? Several different answers to this rhetorical question were suggested by Cairns and Green (1979), each with different implications for how future research and instrument development should proceed. First, parental ratings may faithfully record the infant's temperament, measured across time and settings. Previous research (see Bates, 1980) and the results of this study, though, suggest that such ratings are an amalgamation of parent and infant characteristics; an assessment of parental interpretations of behavior. Thus, a second source of variance will be the skills, biases and expectations of the rater. Problems with interrater reliability then should lead to questions about the processes involved in rating infants that might cause different raters to reach different conclusions. In this study, expectations about infants in general--the reference point or "anchor"--were hypothesized to influence the placement of the individual child on a subjective continuum. The correlations between mothers' knowledge of infant development and ratings of infant temperament partially support this thesis.

The contribution of rater variance to temperament ratings highlights an issue that has been the cause of heated debate in some circles: how to capitalize on the parent's vast experience with the infant while ensuring some measure of accuracy in reporting the infant's behavior (see Carey, 1981). The present study hints at several techniques that might minimize the effects of rater variance. First, training parents as raters would result in some uniformity if a way could be found to provide similar anchors and units of scaling. This might be accomplished by providing each parent with vignettes of the prototypic "difficult" and "easy" infant. Or, more complete descriptions of each response option could be written. Second, raters, whether mothers and fathers or parents and observers, differ in their familiarity with the target infant and infants in general. Covarying scores on some assessment of expectations or perceptions of the average infant thus may enhance interrater reliability. In either case, the contribution of parental characteristics to temperament ratings will remain an obstacle until standardization of raters is achieved.

Another reply to what temperament ratings really measure involves flaws in the instrument, an area of inquiry that has been relatively neglected. "Instrument failure", to use Hubert, et al.'s (1982) term, fundamentally is a problem of item wording and selection, and the choice of response options. The clustering of items into categories or scales--guided either by theory and/or empirical relationships--also would be considered under this rubric. On the ITQ, for example, faulty selection led to the inclusion of some very skewed items that are descriptors of general infant characteristics. The pattern of missing data--again, a product of item selection--also introduces extraneous variance into the scale. Factor analyses of the ITQ category scores in this study and in previous ones (e.g., Sameroff, et al., 1982) suggest a more parsimonious and possibly more valid method for reducing the data. Finally, the response bias toward the

"Almost always" option, and the two response styles of choosing either the extreme or middle options, again raises the problematic general issue of whether such ratings are an accurate measure of a "difficult" or "easy" infant, or a flaw in the instrument. It is impossible to decide post hoc which explanation is correct without some independent assessment of the infant. Therefore, the validity of the ITQ as a measure of infant behavior remains to be verified.

There are several potential techniques for minimizing error variance in temperament ratings. First, disposing of extreme responses by deleting subjects who select a given alternative more than 50% of the time, as in this study, may decrease the skewed response profile in Figure 2 but at the expense of throwing out the baby with the bathwater. A more defensible approach would be to eliminate items with skewed frequency distributions or a high probability of being left blank. Or, a validity scale similar to the L scale of the MMPI could be developed to identify raters who fail to respond to the range of options. One also could use interrater correlations as a criterion in selecting items, keeping those that have high parent-observer or mother-father agreement. This would provide some assurance that temperament ratings are more a product of individual differences in behavior across settings and raters, rather than error or rater variance. Whatever tactic is adopted, though, it is clear that revisions of the instrument are in order.

The results of this study reinforce the conclusion of Hubert, et al. (1982) that the issue of the validity of temperament ratings ultimately is entangled with psychometric concerns. That is, the presence of biases due to item selection, response styles, and a consistent pattern of relationships with rater characteristics imply that ratings of infant temperament also are measuring error and rater variance. We will have valid measures of infant temperament only when these extraneous influences are minimized.

Reference Notes

1. MacPhee, D. The Knowledge of Infant Development Inventory. Unpublished questionnaire and manual, 1981.

2. Fullard, W. The assessment of infant temperament: can the psychologist trust the mother? Colloquium presentation at the University of North Carolina, September, 1981.

References

Bates, J. E. The concept of difficult temperament. Merrill-Palmer Quarterly, 1980, 26, 299-319.

Bates, J. E., Freeland, C. A. B., & Lounsbury, M. L. Measurement of infant difficultness. Child Development, 1979, 50, 794-803.

Buss, A. H., & Plomin, R. A temperament theory of personality development. New York: Wiley, 1975.

Cairns, R. B., & Green, J. A. How to assess personality and social patterns: Observations or ratings? In R. B. Cairns (Ed.), The analysis of social interactions: methods, issues, and illustrations. Hillsdale, N. J.: Lawrence Erlbaum, 1979.

Carey, W. B. The importance of temperament-environment interactions for child health and development. In M. Lewis and L. A. Rosenblum (Eds.), The uncommon child. New York: Plenum, 1981.

Carey, W. B., & McDevitt, S. C. Revision of the Infant Temperament Questionnaire. Peditarics, 1978, 61, 735-739.

Crockenberg, S., & Acredolo, C. Infant temperament ratings: a function of infants, of mothers, or both? Infant Behavior and Development, 1983, 6, 61-72.

Freedman, D. G. Human infancy: an evolutionary perspective. Hillsdale, N. J.: Lawrence Erlbaum, 1974.

Goodnow, J. J. Change and variation in beliefs about childhood and parenting. In I. Sigel (Ed.), Parents' beliefs about childhood and parenting. New York: Wiley, in press.

Hubert, N. C., Wachs, T. D., Peters-Martin, P., & Gandour, M. J. The study of early temperament: measurement and conceptual issues. Child Development, 1982, 53, 571-600.

Kronstadt, D., Oberklaid, F., Ferb, T., & Swartz, J. Infant behavior and maternal adaptations in the first six months of life. American Journal of Orthopsychiatry, 1979, 49, 454-464.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Sameroff, A. J., Seifer, R., & Elias, P. K. Sociocultural variability in ratings of infant temperament. Child Development, 1982, 53, 164-173.

Simonds, J. F., & Simonds, M. P. Nursery school children's temperament related to sex, birth position and socioeconomic status. Journal of Pediatric Psychology, 1982, 17, 49-60.

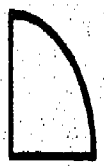
Torgerson, W. S. Theories and methods of scaling. New York: Wiley, 1958.

Wilson, R. S., & Matheny, A. P., Jr. Assessment of temperament in infant twins. Developmental Psychology, 1983, 19, 172-183.

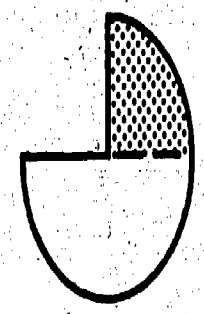
SOURCES OF VARIANCE IN RATINGS AND OBSERVATIONS

RATINGS

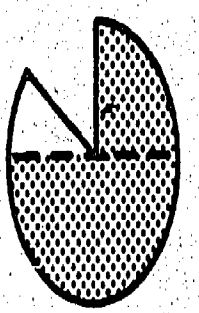
Stable Characteristics of The Child



Characteristics of The Rater + Meas. Error



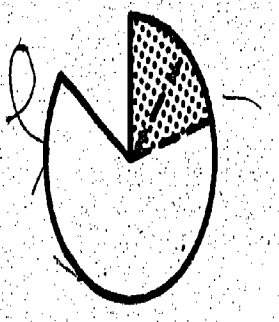
Situational and Interactional Influences



OBSERVATIONS

Stable Characteristics of The Child

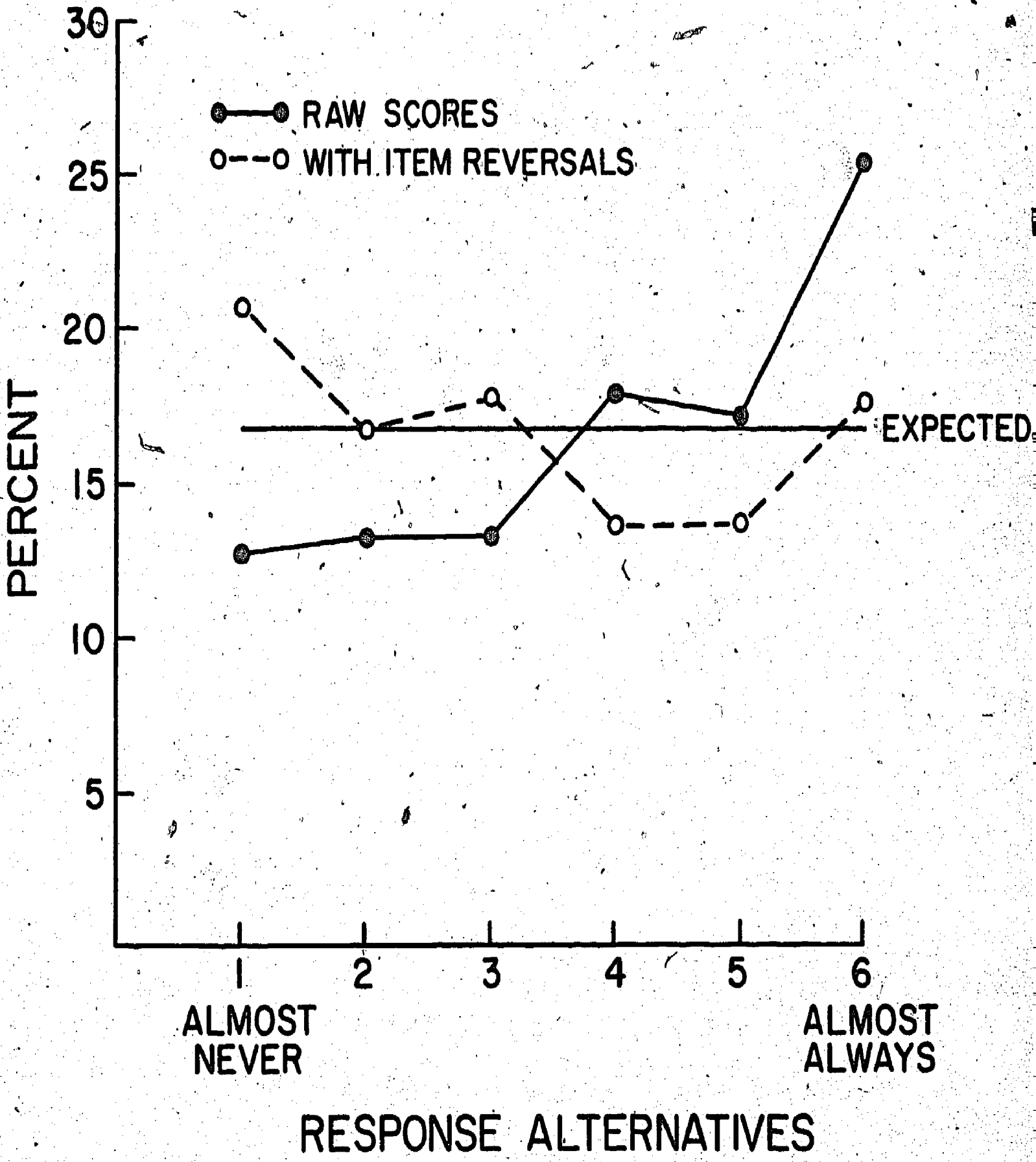
Characteristics of The Observer + Meas. Error



Situational and Interactional Influences

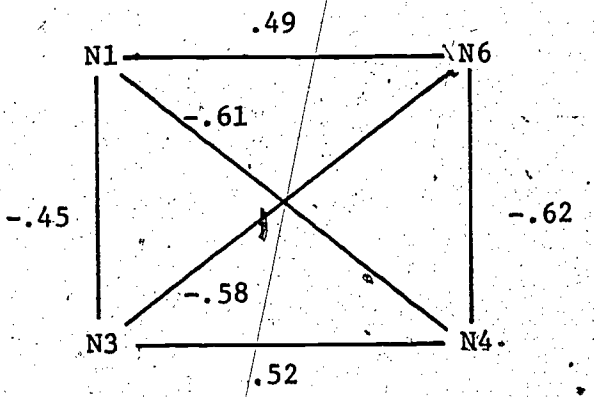
From Cairns & Green (1979)

RESPONSE BIAS IN RATING OF INFANT TEMPERAMENT

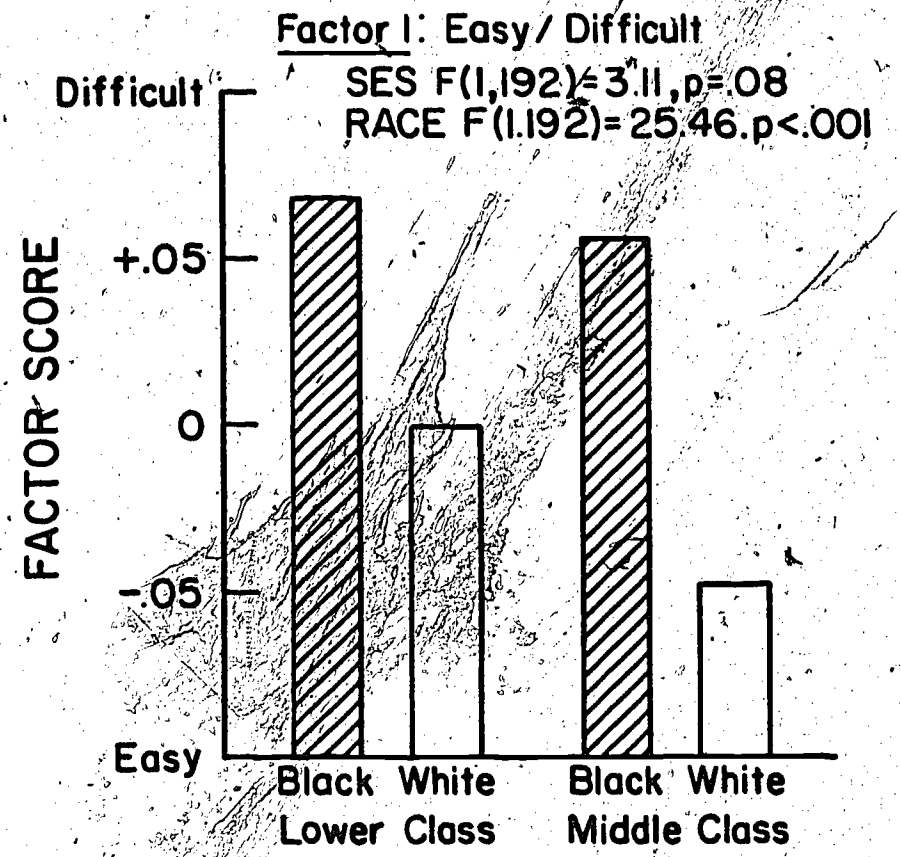


Response Styles on the ITQ:

Correlations Among the Extreme and Middle Options.



SOCIAL CLASS AND ETHNIC DIFFERENCES IN RATINGS OF INFANT TEMPERAMENT



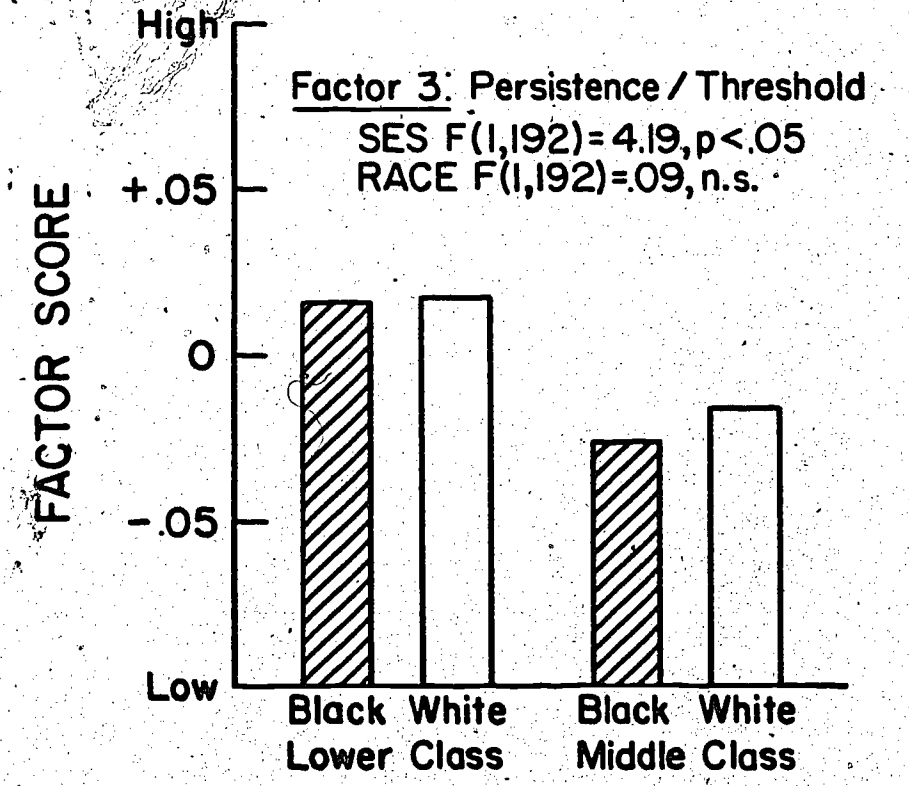


Table 1

Factor Analysis of the ITQ Categories

	ITQ Factor		
	I	II	III
Easy/Difficult			
Approach*	.86		
Adaptability*	.84		
Mood*	.75		
Distractibility	.61		
Rhythmicity*	.57		
Activity/Reactivity			
Activity Level		.80	
Intensity*		.77	
Phlegmatic			
Persistence			.74
Threshold			-.63

* Category is used in calculating the infant's Easy/Difficult classification.

Table 2

Knowledge of Infant Development and

Ratings of Infant Temperament

Carey ITQ	Accuracy on the Knowledge Scale
Easy/Difficult	
Approach	-.31*
Adaptability	-.40*
Mood	-.29*
Rhythmicity	-.30*
Distractibility	-.25*
Activity/Reactivity	
Activity Level	.11
Intensity	-.06
Phlegmatic	
Persistence	-.03
Threshold	-.12

* $p < .0002$

Lower temperament scores are achieved by infants who are rated as easier, less active and intense, and less phlegmatic.

APPENDIX A

Items Frequently Left Blank on the ITQ

Item No.	Category	Item
19	Adaptability	Resists changes in feeding schedule.
31	Approach	Initially rejects a new babysitter.
34	Adaptability	Objects to being bathed by a new person; in a new place.
35	Rhythmicity	Amount of milk taken is unpredictable.
56	Adaptability	Adjusts easily to changes in place or time for sleep.
59	Distractibility	Can be calmed if fussing about a soiled diaper.
62	Approach	Accepts changes in place or time for bath.
68	Distractibility	Rejects disliked food or medicine.
76	Mood	Fussy or moody when has a cold or virus.
83	Threshold	Rejects disliked food even if mixed with a preferred one.
94	Adaptability	Accepts changes in solid food feedings.

APPENDIX B.

Items Exhibiting Marked Response Bias on the ITQ

Item No.	Option chosen* and %	Category	Item
9	6 - 71%	Adaptability	Accepts bath at any time of day without resisting.
17	6 - 56%	Activity	Moves about much during diapering and dressing.
18	6 - 50%	Intensity	Resists additional food or milk when full.
21	6 - 57%	Distractibility	Stops play and watches when someone walks by.
28	6 - 46%	Rhythmicity	Gets sleepy at about the same time each evening.
33	6 - 61%	Activity	Moves much when lying awake in crib.
34	1 - 47%	Adaptability	Objects to being bathed by a new person or in a new place.
43	6 - 72%	Activity	Plays actively with parents.
44	6 - 51%	Distractibility	Watches another toy even though already holding one.
71	6 - 81%	Activity	Actively grasps or touches objects within reach.
87	6 - 54%	Distractibility	Stops sucking and looks when hears an unusual noise.
95	6 - 58%	Activity	Moves much when playing by self.

* The parent is given 6 options for each item, ranging from 1 (Almost never) to 6 (Almost always).