The assumptions underlying the proposed definition of social competence necessitate that data for assessment purposes be obtained by observing children's natural behaviors during social interchanges. The observation system to be used should include referents of (a) general context, (b) specific interaction events, and (c) child behaviors.

The key to the achievement of such an assessment approach rests in the type of observation system that can be constructed and in the treatment of the resultant observation data. The advances of present-day computer technology and capabilities, as well as current knowledge of systems and approaches to observation and ethnography, deem it feasible for an attempt in this direction. The advantage of such an approach is reflected in meaningful measurements relative to psychometric and ethnic differences.

With regard to psychometrics, the repertoire approach facilitates developmental assessment of persons at any level of development. The repertoire score can be meaningfully compared across various age periods to address directly the question of how the repertoire has been enhanced and not just whether or not it has been. This is reflected by the examination of whether the repertoire includes additional modalities and/or more complex/coordinated/integrated modalities of expressions. Once a behavior has been reliably observed, it can clearly be viewed as part of the child's repertoire.

One of the most important advantages is the expectation that qualitatively different repertoires (that is, repertoires composed of quite different subcomponents) at similar levels of complexity would be equally facilitating. The repertoire approach sets the framework for the direct assessment of quite different behavioral mixes according to a cultural norm. Behaviors acquired from minority culture experience are viewed as being as right and as facilitating, given the opportunity to apply them, as are behaviors of comparable levels of complexity and differentiation that are acquired within a mainstream culture. What is important to assess, according to this perspective, is the composition of the repertoire of a given child, not whether any particular norm-based criterion behavior has been acquired by a particular point in time. Any alternative—in the language or nonverbal response variations used by any ethnic or cultural group—is a valid alternative.

In sum, this model of assessment considers all child behavior as potentially useful in constructing indices of competence and thus provides recognition and legitimation of cultural and/or ethnic strengths. This approach also permits the examination and comparison of the behavioral mixes or components that make up a child's repertoire at different developmental periods; hence, the transitions in repertoire development may be traced. These two factors should be especially useful in program planning for intervention programs such as Head Start.

REFERENCES

Mercer, J. R. A policy statement on assessment procedures and the rights of children. *Harvard Educational Review*, 1974, *44*, 125.

Sarason, S. B., & Doris, J. *Educational handicap, public policy, and social history*. New York: Free Press, 1979.

Zigler, E., & Trickett, P. K. IQ, social competence, and evaluation of early childhood intervention programs. *American Psychologist*, 1978, *33*, 789–798.

LEE C. LEE
Department of Human Development and Family Studies
Cornell University

Chance and Interrater Agreement on Manuscripts

Agreement between reviewers over a manuscript's appropriateness for publication has long been of concern to psychology because publications are a reflection of the scientific quality of the profession. Both as a group and individually, psychologists are intimately dependent on professional publications for knowledge of advancements in their field and, often, for personal advancement. It therefore becomes a matter of some personal and professional importance that manuscript reviewers and editors agree on what articles are of sufficient quality to warrant inclusion in scarce journal pages.

Previous investigations of agreement between manuscript reviewers generally sought to quantify agreement by means of correlational statistics but often found coefficients of such low magnitude that concern over the review process was expressed (Hendrick, 1977; McCartney, 1973). In contrast with these pessimistic results, Crandall (June 1978) recently questioned the appropriateness of correlational statistics because of their essentially associational nature; instead, he advocated the percentage-of-agreement statistic as more useful for the measurement of true agreement between raters. The percentage-of-agreement statistic was used by Scarr and Weber (October 1978) to report on *American Psychologist* reviews and by Crandall (1978) to reanalyze the data of Scott (1974) and Hendrick (1977). Based on obtained percentages, Crandall concluded that "the best estimate of agreement on publishability is about 70%" (p. 624), and Scarr and Weber (1978) expressed a new faith in themselves "as casual observers" (p. 935).

Chance as an alternative explanation of high interrater agreement percentages was considered by Crandall but was not incorporated into the calculation of his 70% agreement statistic. In any classification situation a certain amount of agreement among raters would be found by chance alone, and thus any statement of interrater agreement must reflect not only how much agree-

ment is evident but also how much agreement is in evidence beyond what would be expected by chance alone. Such a statement was not provided by Scarr and Weber or by Crandall, and their reported percentages of agreement must be considered with caution because of the unknown influence of chance embedded in each reported percentage.

An index of agreement among observers that takes chance into account was developed by Cohen (1960) and was subsequently extended by Light (1971) and Fleiss (1971). This statistic might prove useful in analyzing agreement between manuscript reviewers. The kappa statistic (κ) indicates the proportion of agreement remaining after chance agreement is removed, ranges from negative values (less than chance agreement) through zero (chance agreement) to +1.00 (perfect agreement), and is distributed as a standard normal variate.

An application of κ to the *Personality and Social Psychology Bulletin* (PSPB) data of Hendrick (1977) was undertaken to determine the percentage of agreement between manuscript reviewers after chance agreement had been excluded. In the case of the PSPB, a 5-point scale was used, where Category 1 was "definitely accept," Category 2 was "probably accept," Category 3 was "reject but recommend revision and resubmission," Category 4 was "reject but a revision may be acceptable eventually," and Category 5 was "definitely reject." Analysis was accomplished with Fleiss's (1971) computational formulas (Watkins & McDermott, 1979). A κ of .15 ($Z = 3.856$, $p < .001$) resulted; this indicates that 15% of the possible greater-than-chance agreement was obtained. To determine if the reviewers agreed on a general accept–reject dimension, kappas were recalculated after collapsing the 5-point scale into a dichotomy by considering all ratings of 1, 2, or 3 as belonging to the "possibly accept" category and all ratings of 4 and 5 as belonging to the "reject" category. A resultant κ of .091 ($Z = 1.18$, *ns*) indicates a lack of agreement beyond chance and, additionally, confirms the large role chance plays in such a dichotomy. It appears that the interrater agreement of the PSPB reviewers is in fact only marginally greater than chance, and Crandall's faith in the review process, as demonstrated by these reviewers, may be premature.

An additional application of κ was undertaken to determine the effects of chance on the percentage-of-agreement statistics reported by Scarr and Weber (1978) for the *American Psychologist* reviewers. In this case, a 5-point scale was used, where Category 1 was "reject," Category 2 was "reject and recommend another journal," Category 3 was "reject and recommend resubmission after revision," Category 4 was "accept with minor revisions," and Category 5 was "accept in present form." A κ of .49 ($Z = 6.75$, $p < .001$) resulted. This reveals that the agreement between reviewers was approximately 50% after chance agreement had been excluded. An accept–reject dichotomy was again constructed by considering all ratings of 1 and 2 as belonging to the "reject" category and all ratings of 3, 4, and 5 as belonging to the "accept" category. A κ of .53 ($Z = 4.67$, $p < .001$) was produced; this indicates that the *American Psychologist* reviewers agreed at a level substantially greater than would have been expected by chance alone, on the general dimension of acceptability for publication. Such results give support to Scarr and Weber's faith in themselves as casual observers.

Strikingly different results emerged from this analysis of percentage-of-agreement statistics among manuscript reviewers, with PSPB reviewers agreeing beyond chance only marginally and *American Psychologist* reviewers agreeing at levels substantially beyond chance. It is apparent that a statistic such as κ should be used in place of simple percentage of agreement so as not to obscure the role played by chance. The present disparate results suggest that true agreement among reviewers cannot be determined until chance agreement is excluded from consideration.

It seems appropriate that journal editors should begin to assess directly the true agreement of their reviewers, rather than relying on the unproven assumption that valid allocation of journal pages is being assured by current procedures. One hopeful sign, suggested by Gottfredson (October 1978), is the construction of scales for judgment of article quality. His provocative results indicated improved reliability because of increased understanding of what constitutes article quality. This line of inquiry could be continued and expanded under the leadership of our professional organization through its journal editors.

REFERENCES

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, *20*, 37–46.

Crandall, R. Interrater agreement on manuscripts is not so bad! *American Psychologist*, 1978, *33*, 623–624. (Comment)

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 1971, *76*, 378–382.

Gottfredson, S. D. Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments. *American Psychologist*, 1978, *33*, 920–934.

Hendrick, C. Editorial comment. *Personality and Social Psychology Bulletin*, 1977, *3*, 1.

Light, R. J. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 1971, *76*, 365–377.

McCartney, J. L. Manuscript reviewing. *Sociological Quarterly*, 1973, *14*, 290; 440–444.

Scarr, S., & Weber, B. The reliability of reviews for the *American Psychologist*. *American Psychologist*, 1978, *33*, 935.

Scott, W. A. Interreferee agreement on some characteristics of manuscripts submitted to the *Journal of*

Personality and Social Psychology. American Psychologist, 1974, *29,* 698–702.

Watkins, M. W., & McDermott, P. A. A computer program for measuring levels of overall and partial congruence among multiple observers on nominal scales. *Educational and Psychological Measurement,* 1979, *39,* 235–239.

MARLEY W. WATKINS
Phoenix, Arizona

Author Review of Reviewers

The system of anonymous manuscript review used by most of our journals has a great weakness: Reviewers of manuscripts are not sufficiently accountable for the quality of their reviews. Reviewers should not, of course, be accountable to particular authors, but they could be made more accountable to the editors and, indirectly, to the author community of which they themselves are members.

I suggest a possible remedy: It is called *author review.* The journal editor would send to the author, along with the letter of decision and the reviews, a postcard questionnaire (one per review) that would request the author to evaluate each review. I suggest that three dimensions of evaluation are necessary: fairness, carefulness, and constructiveness. There should also be a place on the card for comments. The editor would file the returned postcards under the respective reviewers' names, noting the editorial decision and final disposition of the manuscript. At intervals—perhaps once a year—the editor could examine these questionnaires. If a particular reviewer received repeated complaints, he or she could be terminated as a reviewer or could receive admonishment from the editor. Presumably, repeated low ratings would reflect some real shortcomings in a reviewer's habits rather than the resentment of a particular disappointed author.

This procedure would serve several goals: (1) The possibility of evaluation might make reviewers more conscientious than they sometimes are, (2) editors would have some information to use when weeding out or retaining referees, and (3) authors would have a chance to provide feedback to the editors and might therefore feel they play a more useful role in the editorial process. I think most authors would take the process of author review seriously and would be capable of making the requested assessments. It would also sensitize them to dimensions of adequacy that they should consider when reviewing manuscripts.

JUDITH A. HALL
Johns Hopkins University

Graduate Student Success: Sex or Situation?

As a subject in Hirschberg and Itkin's (December 1978) study of cohorts of beginning graduate students from 1965 to 1970 at the University of Illinois at Urbana-Champaign, I wish to add some information that they omit concerning potential causes of differential attrition rates by sex. I am writing as one who is grateful for the education I received there from 1966 to 1970 and who returned to receive the PhD in 1971.

The authors approach the problem of graduate student success from a personnel psychology perspective of the following sort: Females enter with somewhat higher standard qualifications (grade point average, verbal scores on the Graduate Record Examination, overall scores on admissions criteria) than males do, yet only 35% of them obtain the PhD, compared with 68% of the males. How can the selection procedure be altered to deselect (i.e., reject or terminate) greater proportions of potential nonfinishers and thereby boost the proportion of those who complete the program?

An alternative approach is to consider aspects of the departmental environment into which the highly qualified applicants are placed as potential causes of differential attrition rates by sex from 1965 to 1970. These include percentage of full-time female faculty, extent of financial support for part-time students, and extent of university-supported, low-cost child care. Quantification of these representative aspects can be summarized concisely: zero. Other relevant dimensions to be considered include public and private statements of discouragement of female students by a small percentage of faculty who had nothing to do with the female students' achievement.

The loss to psychology of 65% of the highly selected female students (as well as 32% of the highly selected male students) deserves a more thoughtful response than, What was wrong with our selection procedures? It begins with the question, What can our department do to provide a high-quality academic environment that fosters the perseverance of highly academically qualified students of both sexes?

REFERENCE

Hirschberg, N., & Itkin, S. Graduate student success in psychology. *American Psychologist,* 1978, *33,* 1083–1093.

MICHELE ANDRISIN WITTIG
California State University, Northridge

On Selection and "Deselection"

As I read "Graduate Student Success in Psychology" (Hirschberg & Itkin, December 1978), I compared the authors' conclusions with my own memories of the University of Illinois during the time of their research.

The design of their study reflects accurately the tone of the Department of Psychology at that time: Success or failure in graduate school was ascribed to intrapersonal traits such as "conscientiousness." Fac-