Journal of Educational Psychology
1986, Vol. 78, No. 2, 141–146

Accuracy of Teacher Judgments of Student Responses to Standardized Test Items

Theodore Coladarci
College of Education,
University of Maine at Orono

Six students were randomly selected from each of four third-grade and four fifth-grade classes. For each of their six students, teachers were asked to predict whether the student had responded correctly or incorrectly to selected items on a standardized achievement test that recently had been administered in the district. It was found that (a) aggregate measures of teachers' judgments of their students' responses correlated positively and substantially with aggregate measures of students' actual responses; (b) teachers accurately judged their students' responses to individual items for approximately three quarters of the total number of test items; (c) the accuracy of teachers' judgments varied significantly with subtest; (d) there were significant individual differences among teachers in the accuracy of their judgments; and (e) teachers were least accurate in judging low-performing students and most accurate in judging high-performing students. These results are consistent with other research in this area and are discussed within the context of interactive decision making of teachers.

The preinstructional decisions made by teachers are thought to be influenced by several factors. One prevailing model, for example, posits that teachers' decisions are influenced by (a) their beliefs and attitudes about education, (b) the perceived nature of the instructional task, and (c) available information or cues about their students (e.g., Borko, Cone, Russo, & Shavelson, 1979; Shavelson & Borko, 1979; also see Shavelson & Stern, 1981). The influence of these factors, in turn, is mediated through an additional factor: the inferences or estimates that teachers make about their students' cognition. For example, teachers make judgments of their students' reading ability before making decisions about grouping assignments (see Shavelson & Borko, 1979).

There is evidence that the *interactive* decision making of teachers similarly involves a teacher judgment component. (See Brophy, 1984, and Clark & Peterson, 1986, for overviews of research on teacher decision making in the interactive context.) In one study (McNair, 1978–1979), for example, teachers reported in stimulated recall interviews that their main consideration in making decisions during reading instruction was student learning. Specifically, these 10 teachers "based many of their decisions on *what they surmised* was happening with the individual student" (p. 32; emphasis added). Colker (1984) found that 41% of the interactive thoughts reported by six teachers in stimulated recall interviews pertained to student cognition. Of these, the majority (61%) fell into a category Colker (1984, Table 3) labeled "pupil": "The teacher evaluates or questions pupil comprehension, learning, thinking, knowledge, or task performance" (e.g., "I was thinking . . . that they don't understand what they're doing"). Clearly, interactive teaching judgments, such as those observed by McNair and Colker, differ from those made in a preinstructional context in that the former are more immediate estimates of the student's knowledge whereas the latter often are judgments based, in large part, on available documentation (e.g., performance on worksheets, teacher-constructed tests, standardized achievement tests). In a sense, it is more the *teacher* who provides the estimate in the interactive context, whereas in the preinstructional context, it is more a *measure* that provides the estimate from which the teacher forms a judgment.

Given the role of teacher judgments of this kind in the decision making of teachers, it is surprising to find so little empirical research in which the *accuracy* of such judgments is examined. Existing research that provides some import for this question has examined the correspondence between students' performance on standardized tests of academic achievement and the teachers' a priori judgments of that performance. Within the present context, unfortunately, this research carries at least two limitations. First, many of these studies are of only tangential relevance here because the correspondence between teacher judgments and student abilities was examined in a predictive rather than a concurrent context (e.g., Dusek & O'Connell, 1973; Morine-Dershimer, 1978–1979). Morine-Dershimer, for example, asked teachers at the beginning of the school year to group students into five categories reflecting their expectations for the student's end-of-year academic achievement. These categories, in turn, were compared to April test scores. Because of the temporal aspect of a prediction study, however, such studies provide answers to a substantively different question.

A second limitation of existing research is that teacher judgments have been expressed as *rankings* or general *ratings* of student performance (e.g., Farr & Roelke, 1971; Hoge & Butcher, 1984; Hopkins, George, & Williams, 1985; Kellaghan, Madaus, & Airasian, 1982; Luce & Hoge, 1978; Mayfield, 1979; Oliver

The research reported here was funded, in part, by a grant from the Graduate School at the University of Montana. Portions of this research were presented at the 1984 and 1985 meetings of the American Educational Research Association.

I wish to thank the eight teachers who participated in this study: Suzanne Hampton and Janet Spector for their assistance in data collection; Ed Kameenui, Ralph Putnam, Janet Spector, and the three anonymous reviewers for critically reading earlier versions of this manuscript; Nate Gage and Larry Hedges for several enjoyable discussions of methodological issues in research of this kind; and Cheryl Huskins for patiently typing (and retyping) the tables.

Correspondence concerning this article should be addressed to Theodore Coladarci, College of Education, University of Maine, Orono, Maine 04469.

& Arnold, 1978; Stevenson, Parker, Wilkinson, Hegion, & Fish, 1976; Tokar & Holthouse, 1977). For example, Kellaghan et al. (1982) asked teachers to rate their students' abilities as "well above average," "above average," "average," "below average," or "well below average"; Mayfield (1979) asked teachers to rank order their students on academic achievement. Other approaches include asking teachers to estimate the instructional reading level (Oliver & Arnold, 1978) or grade-equivalent score (Hoge & Butcher, 1984) corresponding to a student's performance.

Because of the summary nature of such teacher judgments, little is disclosed about the teacher's specific knowledge of what the student has and has not mastered in some domain. This limitation is particularly relevant to interactive decision making, where the judgments that teachers make about student's knowledge and comprehension doubtless are more specific than those made in the preinstructional context. For example, the preinstructional judgments that Shavelson and Borko (1979) reported teachers made about students' reading ability for grouping purposes are considerably more global than the kind made in the interactive context, the flavor of which is captured by a comment made by one of McNair's (1978–1979) teachers: "I knew she had it so I decided not to discuss it any further" (p. 29).

An additional problem regarding summary judgments is that the *accuracy* of these judgments remains undemonstrated: Despite the fact that teacher judgments of this kind typically have correlated .55 with student achievement (Hoge & Butcher, 1984), it is true nonetheless that a correlation coefficient indicates the degree of correspondence between the *relative standing* of two sets of values, not the degree of similarity between the values themselves. A perfect correlation, consequently, still would not show how accurately a teacher judged a student's performance. And even if judgment and performance were identical for each pair, we still would not know about the accuracy of the judgment: I may judge accurately that my student correctly answered 80% of the items on a test; however, although correctly answering 80% of the items, this student may have missed entirely different items from those I had predicted. (This limitation, of course, applies equally to teachers' estimates of instructional levels, grade equivalents, or any other summary index of student performance.) To this end, a more revealing question might be whether teachers can correctly gauge student responses *item by item* on a valid achievement test that has been administered concurrently to their students.

Leinhardt (1983), in a study involving special-education teachers, did obtain item-level judgments for individual students. In her examination of the "overlap" between what is taught in a classroom and the content of the standardized achievement test used in the classroom, Leinhardt asked teachers to indicate, for each student, the reading test items for which they believed sufficient instruction had been provided for the student to answer the item correctly (although not necessarily getting the item correct). Leinhardt reported that (a) these judgments of overlap were accurate for roughly 64% of the test items, averaged across students; (b) the judged number of overlapping items for a student correlated .67 with the student's actual performance on the test; and (c) the accuracy of the teacher's overlap judgment correlated .39 with the student's actual performance. However, the implications of Leinhardt's results for the question of teacher judgments as defined in the present context are not clear because Leinhardt's study was conducted in a special-education setting. Further, her overlap measure was designed to assess the teacher's perception of the correspondence between instruction and test items rather than, specifically, the teacher's judgment of the student's probable performance on the test item.

The present study was conducted to address the following questions: How accurately do teachers judge their students' probable performance on achievement-test items? Is teacher accuracy related to the particular academic task being judged (e.g., vocabulary vs. reading comprehension)? Are there individual differences among teachers in the accuracy of their judgments? Is teacher accuracy related to the student's general level of academic achievement?

Method

Five third-grade and 5 fifth-grade teachers in a western Montana elementary school district were invited to participate in the present study. Although the third and fifth grades were selected somewhat arbitrarily, these particular teachers were invited because their classes were the most academically heterogeneous of all third- and fifth-grade classes in the district (which was desired for statistical purposes).

Teachers were told that they would be asked questions concerning their impressions of their students' competencies vis-à-vis those measured by the standardized achievement battery recently administered as part of the district's testing program. Of these 10 teachers, 1 declined to participate and 1 was dropped because she could not be scheduled for an interview in the period during which the study was to be conducted. Ultimately, then, 8 teachers participated in this study, distributed equally between grades. Seven of these teachers were female and had been teaching for 12 to 15 years; the male teacher had taught for 8 years. Class size for these teachers ranged from 19 to 27 students (M = 24.0).

The district had just completed its spring assessment of academic achievement, in which the SRA Achievement Series (Science Research Associates, Inc., 1978) — a test of demonstrated content validity with respect to the district's curriculum — was administered to all grades. Each teacher was interviewed after school during the last week of April or the first week in May, 1 to 2 weeks after the SRA test had been administered. Each teacher prepared for the interviewer the names of those students who fell into one of the following groups: those who were generally performing at grade level, those who were generally performing 1 year below grade level, and those who were generally performing 1 year above grade level. Teachers were instructed to make these designations by considering diverse sources of information (e.g., informal observations, performance on quizzes and teacher-constructed exams, standardized test scores from the previous year).

For each teacher, the interviewer randomly selected 6 students (2 from each group) who had taken the SRA and who were reported to have experienced no unusual problems (e.g., due to illness, language deficits, anxiety). For each student, the interviewer asked the teacher to indicate whether he or she thought the student correctly answered specific items on the SRA at the time of testing. This was done for each item on the Reading Vocabulary, Reading Comprehension, Mathematics Concepts, and Mathematics Computation subtests for both the third-grade and fifth-grade tests (Levels D and E, respectively; Form 1).

To reduce the likelihood of a teacher response bias due to characteristics of the individual student, teachers were asked to provide these judgments for each of their 6 students on the same item before moving on to the next item. Because teachers did not score the tests and did not know in advance which students they would be queried about, there was virtually no chance that teachers' judgments were contaminated by prior knowledge of the students' actual responses to these items.

Measures of student performance, teacher judgment, and performance/judgment agreement were obtained to explore the research questions stated

Table 1
Student Performance (N = 48): Means, Standard Deviations, and Intercorrelations

Measure	*M*[a]	*SD*	Range[b]	1	2	3	4	5	6
1. Reading vocabulary	81.53	13.70	33–100						
2. Reading comprehension	81.74	13.16	35–100	.64[c]					
3. Reading total	81.58	12.18	34–98	.91	.90				
4. Mathematics concepts	74.78	14.02	34–97	.61	.50	.62			
5. Mathematics computation	72.62	17.65	20–95	.64	.49	.63	.56		
6. Mathematics total	73.90	14.09	39–94	.70	.55	.70	.84	.92	
7. Total test	77.77	12.26	44–94	.87	.77	.91	.79	.86	.93

[a]Percentage correct. [b]Rounded to nearest whole number. [c]Student-level correlations; all are statistically significant ($\alpha = .05$).

above. *Student performance* was the students' actual performance on the SRA test; it is expressed here as percentage correct. The item-level responses were summed to form reading vocabulary and reading comprehension, which, in turn, were summed to form reading total. For any aggregation of items, the sum was divided by the respective number of items. The same procedure was followed to form mathematics concepts, mathematics computation, and the corresponding mathematics total. A grand total (total test) was obtained by combining reading total and mathematics total. Again, the metric was percentage correct for all item aggregations. *Teacher judgment* was conceptualized in a parallel fashion. For example, a reading vocabulary teacher judgment was formed by summing the teacher judgments regarding a specific student for the items on the Reading Vocabulary subtest and then dividing by the number of subtest items. Thus, as before, the metric was percentage correct — in this case, the percentage of reading vocabulary items that the teacher had judged the student would answer correctly. Finally, *performance/judgment agreement* was established by summing for each subtest the number of items for which a student's response and a teacher's judgment were in agreement. This sum, in turn, was divided by the appropriate number of items. This was done to parallel the student-performance and teacher-judgment measures.

Results

All analyses reported here were conducted with grades pooled. Correlations among the student-performance measures were computed with the student as the unit of analysis. Correlations involving at least one teacher variable (i.e., teacher judgment and performance/judgment agreement) were computed separately for each of the 8 teachers and then averaged using the *r* to *z* transformation; the mean *z* was then transformed back to *r* (as was done in the studies by Farr & Roelke, 1971, and Hopkins et al., 1985). To test for the statistical significance of the final correlation, a 95% confidence interval for the corresponding mean *z* was estimated.

Descriptive statistics and intercorrelations for the student-performance and teacher-judgment measures are presented in Tables 1 and 2, respectively. Results pertaining to teacher-judgment accuracy appear in Table 3.

The relationship between teacher judgment and student performance can be examined by correlating aggregate measures of teacher judgment with aggregate measures of student performance. For the present study, this would mean correlating the sum of item-level teacher judgments with the sum of item-level student responses. Indeed, as was discussed above, variations of this methodology dominate research in this area (e.g., asking teachers to estimate a student's instructional level or grade-equivalent score). The aforementioned problems of interpretation notwithstanding, performance/judgment correlations were computed here for comparative purposes and appear in the diagonal of the correlation matrix in Table 3. Ranging from .67 to .85, these correlations indicate a consistently positive relationship between teachers' aggregate judgments and students' aggregate performance and, in fact, are higher than those typically reported in the earlier research in which rankings or general ratings by teachers were correlated with student performance on an academic achievement test.

The performance/judgment agreement measures, as was argued above, permit greater insights into the accuracy of teachers' judgments than do the corresponding performance/judgment correlations. From the first column in Table 3, we see that, on the average, teachers correctly gauged from roughly 70% to 77% of their students' responses, depending on which measure is examined. Although evaluating the magnitude of these data is like declaring the proverbial glass as being partially filled or partially empty, these teachers were correctly gauging considerably more responses than not. Indeed, across all items on the four subtests, teachers correctly gauged almost three quarters (73.81%) of students' responses.[1]

At the same time, however, there was marked variability across subtests. A repeated measures analysis of variance was conducted in which the four subtests served as the trial factor. Orthogonal contrasts involving reading vocabulary versus reading comprehension, mathematics concepts versus mathematics computation, and the mean of the first two subtests versus the mean of the second two subtests revealed a significant difference for the second contrast: Teachers' judgments of their students' responses to mathematics computation items were significantly more accurate (76.52%) than their judgments of students' responses to mathematics concepts items (70.02%), $F(1, 47) = 9.58, p < .05$.

Considerable variability in performance/judgment agreement

[1] As an illustration of the ambiguous meaning of performance/judgment correlations, compare the two correlations obtained for the mathematics subtests with the corresponding percentage agreements presented in the first column of Table 3. We see that although the Mathematics Concepts and Mathematics Computation subtests had, respectively, the lowest (70.02%) and highest (76.52%) percentage agreement means, the performance/judgment correlations were similar (.72 vs. .70) — in fact, the direction of the difference between the two means is the opposite of that suggested by the difference between the two correlations. Clearly, such performance/judgment correlations — which reflect the prevailing methodology in this area of research — can obscure the very phenomenon they attempt to elucidate.

Table 2
Teacher Judgments of Student Performance (N = 48): Means, Standard Deviations, and Intercorrelations

Measure	*M*[a]	*SD*	Range[b]	1	2	3	4	5	6
1. Reading vocabulary	79.90	17.84	20–100						
2. Reading comprehension	77.82	22.38	7–100	.94[c]					
3. Reading total	78.83	18.94	14–100	.98	.99				
4. Mathematics concepts	77.92	20.82	31–100	.88	.90	.91			
5. Mathematics computation	78.21	19.59	37–100	.77	.86	.85	.91		
6. Mathematics total	78.36	17.95	39–100	.86	.87	.89	.98	.96	
7. Total test	78.22	16.47	30–100	.96	.97	.98	.95	.93	.97

[a] Percentage correct. [b] Rounded to nearest whole number. [c] Mean within-class correlations; all are statistically significant ($\alpha = .05$).

across students is evident from the ranges presented in Table 3. For all subtests, for example, some students were judged correctly for fewer than half of the test items, whereas other students were judged correctly for nearly all of the items (indeed, in one case, for all of the items). It is possible, of course, that this variability in performance/judgment agreement was the result of variability among *teachers* in the accuracy of their judgments. That is, such variability may reflect a teacher effect rather than a student effect. To pursue this possibility, a one-way analysis of variance with teacher as the grouping factor was conducted on each measure in Table 3; a significant teacher effect was obtained for mathematics concepts, $F(7, 40) = 3.23$, $p < .05$.

The student's general level of achievement also was examined as a possible correlate of teacher accuracy in the present context. Correlations were computed between each performance/judgment agreement measure and three indicators of achievement: (a) the same measure on which performance/judgment agreement was established, (b) the student's total score across the four subtests (i.e., total test), and (c) the student's general designation provided by the teacher at the outset of the study ("below," "at," or "above" grade level). The obtained correlations consistently indicated that abler students were judged more accurately than less able students (Table 4). And the correlations were considerable: Across all items, the accuracy of a teacher's judgment correlated .91 with total test and .92 with the designation provided by the teacher. These correlations, furthermore, correspond to sizable differences in judgment accuracy. For example, the performance/judgment agreement means, across all items, were 60.25% and 87.50% for students in the bottom and top quartiles, respectively, on total test, $t(22) = -11.10$, $p < .05$. Similarly, for students whose teachers rated them as being "below" or "above" grade level, these means were 62.37% and 85.44%, respectively, $t(30) = -9.44$, $p < .05$.

Summary and Discussion

It was found in the present study that (a) aggregate measures of teachers' judgments of their students' responses to items on a standardized achievement test correlated positively and substantially with aggregate measures of students' actual responses (*r*s = .67 to .85); (b) teachers accurately judged their students' responses to individual items for approximately three quarters of the total number of test items; (c) the accuracy of teachers' judgments varied significantly as a function of subtest; (d) there were significant individual differences among teachers in the accuracy of their judgments; and (e) teachers were least accurate in judging low-performing students and most accurate in judging high-performing students.

Of the research that has been conducted in this area, the Leinhardt (1983) study was the most similar methodologically to the present study. Despite the aforementioned differences between the two studies concerning sample characteristics and the manner in which the judgment-accuracy measure was conceptualized, the results of the studies converge at several points. First, the correlation of .67 obtained by Leinhardt between overlap judgments and student performance is not markedly different from the performance/judgment correlation of .79 obtained here for reading total (similar to the criterion measure used by Leinhardt). Aside from the fact that Leinhardt was asking a slightly different question from that examined in the present study, the difference between her correlation and the one obtained here probably reflects the manner in which the two correlations were computed: Whereas correlations in the present study were mean within-class correlations, it appears that Leinhardt did not consider class membership when computing her correlations. (Other things being equal, one would expect mean within-class correlations to be larger.)

Second, Leinhardt found that the accuracy of teachers' overlap judgments was positively related to students' actual performance, although her correlation (.39) was considerably smaller than the .88 obtained here for reading total. The magnitude of this difference may reflect, in part, restricted variability on her achievement criterion due to the special-education context. Third, Leinhardt reported that teachers' overlap judgments and students' actual responses were accurate for roughly 64% of the test items. Conceptualization differences notwithstanding, this result is not markedly different from the 75% accuracy obtained here. And, at any rate, if it is true that teachers tend to have difficulty in providing accurate judgments for low-performing students, then one might expect mean accuracy to be somewhat lower for teachers of the learning disabled than for regular-education teachers.

As was reported above, accuracy of teachers' judgments correlated substantially with students' actual performance. One would expect, of course, that a teacher would tend to be accurate in judging a high-achieving student's performance: For any item, (a) the teacher would be inclined to report that the student would select the correct answer, (b) the student probably would get many of the items correct, and (c) the few performance/judgment inconsistencies that did occur could not appreciably lower the overall level of agreement. Thus, with high-achieving students, teachers' judgments probably were quite accurate not because these teachers were superb diagnosticians, but rather because they were operating

Table 3
Performance/Judgment Agreement (N = 48): Means, Standard Deviations, and Intercorrelations

Measure	M^a	*SD*	Range[b]	1	2	3	4	5	6	7
1. Reading vocabulary	74.33	13.22	47–97	.67[c]						
2. Reading comprehension	73.10	15.52	35–100	.78	.70					
3. Reading total	73.79	13.21	41–98	.93	.96	.79				
4. Mathematics concepts	70.02	15.20	37–97	.72	.68	.72	.72			
5. Mathematics computation	76.52	12.34	45–95	.62	.55	.60	.45	.70		
6. Mathematics total	73.40	12.08	50–93	.83	.70	.78	.85	.87	.74	
7. Total test	73.81	11.39	52–92	.92	.93	.98	.83	.77	.92	.85

Note. Performance/judgment correlations appear in the diagonal of the correlation matrix.
[a] Percentage of items for which there is agreement between student performance and teacher judgment. [b] Rounded to nearest whole number. [c] Mean within-class correlations; all are statistically significant ($\alpha = .05$).

with a general response set that, for the high-achieving student, was efficient.

However, no simple response set would work as efficiently for students further down the achievement scale. In making judgments for the moderate- and low-achieving student, teachers doubtless realized that there were many items that the student would not answer correctly. What was difficult for these teachers, apparently, was to decide *where* the errors would occur. And the lower the student's proficiency, the more difficult — and inaccurate — this judgment was. These results point tentatively to the disturbing implication that students who perhaps are in the greatest need of accurate appraisals made by the teacher in the interactive context are precisely those students whose cognition has a greater chance of being misjudged.[2]

Before implications of these results can be drawn for existing models of teacher decision making and cognition (e.g., Clark & Peterson, 1986), further research is needed to explore additional factors that possibly influence the accuracy of teacher judgments. Variability among teachers in judgment accuracy, for example, might be related to differences among teachers in teaching experience, the frequency and nature of both academic and nonacademic interactions between teacher and student, and the nature of teachers' assessment beliefs and practices. Inquiries that address these last two areas, in particular, might serve also to clarify the observed relationship between teacher accuracy and student ability. In order to explore such hypotheses, of course, more data — and on more teachers — are needed than were possible to obtain in the present study.

Teacher accuracy was found to be higher on mathematics computation than on mathematics concepts. The observed relationship between teacher accuracy and task can be explained, in part, by (a) the degree to which teachers provide direct instruction in the task domain and (b) the amount of information teachers have that bears on student proficiency in that domain. In mathematics instruction, for example, typically there is more direct instruction in mathematical computations than in mathematical concepts. Similarly, in mathematical computations, often there are more data available to the teacher that communicate student proficiency (e.g., worksheets, quizzes).

An additional factor to explore in the relationship between teacher accuracy and task is the complexity of the cognitive processes required of the student in selecting the correct answer. For example, recognizing the correct response to factual questions following a reading passage is less complex cognitively than selecting the most defensible response to inferential questions. Similarly, carrying out mathematics computations is less cognitively complex than this same activity embedded in a problem-solving context. One would expect teacher judgments to be less accurate for more complex tasks simply because the teacher, for more complex tasks, must make a number of "subjudgments" before arriving at the ultimate judgment concerning the student's final response. A mathematics problem-solving item, for example, might require the teacher to make subjudgments regarding the student's (a) comprehension of the givens, (b) selection of the computational procedures to employ, and (c) accuracy in carrying out the computations.[3] Clearly, to address the relationship between teacher

Table 4
Correlations Between Performance/Judgment Agreement and Three Indicators of Student Achievement (N = 48)

	Achievement indicator		
Measure	Same[a]	Total test	Teacher rating
Reading vocabulary	.78[b]	.86	.88
Reading comprehension	.80	.78	.84
Reading total	.88	.87	.90
Mathematics concepts	.89	.79	.78
Mathematics computation	.86	.66	.68
Mathematics total	.88	.82	.85
Total test	.91	.91	.92

[a] Achievement indicator is the same measure for which performance/judgment agreement was established. [b] Mean within-class correlations; all are statistically significant ($\alpha = .05$).

[2] Two of the anonymous reviewers asked whether student performance tended to be under- or overestimated. Although not central to this article, analyses of differences between aggregate measures of teacher judgments and student performance have been presented elsewhere (Coladarci, 1984). The results of these analyses indicated that (a) disparities between teacher judgments and student performance for total test were not systematic; (b) students' reading performance (reading total) tended to be underestimated, whereas their performance in mathematics (mathematics total) tended to be overestimated, $F(1, 47) = 10.19, p < .05$; and (c) the degree to which a student's performance was under- or overestimated was largely unrelated to student ability.

[3] In a separate and exploratory analysis of data from this study (Coladarci & Spector, 1985), vocabulary items were grouped as calling for either literal or nonliteral meanings, and mathematics items were grouped as calling for either straight computation or problem solving; other items were ignored. As expected, students' responses were judged more accurately by teachers on literal-meanings than on nonliteral-meanings items, $F(1, 47) = 11.59, p < .05$, and on computations than on problem-solving items, $F(1, 47) = 9.54, p < .05$.

accuracy and the cognitive complexity of test items, a criterion measure is needed in which levels of student cognition are assessed more systematically than is generally the case with standardized tests of academic achievement. Further, because task complexity and the directness of relevant instruction are confounded in classrooms to some degree, these factors need to be manipulated in research in order for their independent effects to be assessed.

Finally, subsequent investigations might examine these research questions by using test items written in an open-ended (rather than a multiple-choice) format. This would reduce considerably the probability of a student's chance successes on test items and, consequently, would clarify our understanding of teacher judgments and their relationships with other constructs.

In summary, results from the present study point to the multifactorial nature of teacher-judgment accuracy. Applied to the context of interactive decision making, these results suggest that the accuracy of a teacher's judgment is influenced by characteristics of the teacher, student, and academic task. It is for subsequent research to clarify the nature of these characteristics, their interrelationships, and their effects on teacher-judgment accuracy.

References

Borko, H., Cone, R., Russo, N., & Shavelson, R. J. (1979). Teachers' decision making. In P. L. Peterson & H. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 136–160). Berkeley, CA: McCutchan.

Brophy, J. (1984). The teacher as thinker: Implementing instruction. In G. Duffy, L. Roehler, & J. Mason (Eds.), *Comprehension instruction: Perspectives and suggestions* (pp. 71–92). New York: Longman.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 255–296). New York: Macmillan.

Coladarci, T. (1984, April). *The accuracy of teachers' judgments of their students' academic-area competencies.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Coladarci, T., & Spector, J. E. (1985, April). *Cognitive requirements of test items and teachers' a priori judgments of students' responses.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Colker, L. (1984, April). *Teachers' interactive thoughts about pupil cognition.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Dusek, J. B., & O'Connell, E. J. (1973). Teacher expectancy effects on the performance of elementary school children. *Journal of Educational Psychology, 65,* 371–377.

Farr, R., & Roelke, P. (1971). Measuring subskills of reading: Intercorrelations between standardized reading tests, teachers' ratings, and reading specialists' ratings. *Journal of Educational Measurement, 8,* 27–32.

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology, 76,* 777–781.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22,* 177–182.

Kellaghan, T., Madaus, G., & Airasian, P. (1982). *The effects of standardized testing.* Boston: Kluwer-Nijhoff Publishing.

Leinhardt, G. (1983). Novice and expert knowledge of individual students' achievement. *Educational Psychologist, 18,* 165–179.

Luce, S. R., & Hoge, R. D. (1978). Relations among teacher rankings, pupil–teacher interactions, and academic achievement: A test of the teacher expectancy hypothesis. *American Educational Research Journal, 15,* 489–500.

Mayfield, B. (1979). Teacher perception of creativity, intelligence, and achievement. *Gifted Child Quarterly, 23,* 812–817.

McNair, K. (1978–1979). Capturing inflight decisions: Thoughts while teaching. *Educational Research Quarterly, 3*(4), 26–42.

Morine-Dershimer, G. (1978–1979). The anatomy of teacher prediction. *Educational Research Quarterly, 3*(4), 59–65.

Oliver, J. E., & Arnold, R. D. (1978). Comparing a standardized test, an informal inventory, and teacher judgment of third grade reading. *Reading Improvement, 15,* 56–59.

Science Research Associates, Inc. (1978). *SRA Achievement Series.* Chicago: Author.

Shavelson, R. J., & Borko, H. (1979). Research on teachers' decisions in planning instruction. *Educational Horizons, 57,* 183–189.

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research, 51,* 455–498.

Stevenson, H., Parker, T., Wilkinson, A., Hegion, A., & Fish, E. (1976). Predictive value of teachers' ratings of young children. *Journal of Educational Psychology, 68,* 507–517.

Tokar, E. B., & Holthouse, N. D. (1977). The validity of the subtests of the 1976 edition of the Metropolitan Readiness Tests. *Educational and Psychological Measurement, 37,* 1099–1101.

Received June 4, 1985
Revision received October 15, 1985 ■