Review of Educational Research
Fall 1989, Vol. 59, No. 3, pp. 297–313

Teacher-Based Judgments of Academic Achievement: A Review of Literature

Robert D. Hoge
Carleton University

and

Theodore Coladarci
University of Maine

The focus of this paper is on data reflecting the match between teacher-based assessments of students' achievement levels and an objective measure of student learning. These data are treated as relevant to the validity or accuracy of the judgmental measures. The paper begins with a discussion of two contexts in which such judgments are relevant: the teacher decision-making and assessment contexts. The second section presents a review of studies in which data are presented on the match between judgments and test scores. Two types of studies are reviewed. The first represents an indirect test of validity in the sense that there is a discrepancy between the judgmental measure (usually a rating of achievement) and the criterion measure (a score on a standardized achievement test). The second provides a more direct test of validity in that teachers are directly asked to estimate the achievement test performance of their students. On the whole, the results revealed high levels of validity for the teacher-judgment measures. The studies revealed, however, some variability across teachers in accuracy levels and suggested the operation of certain other moderator variables. The paper concludes with a set of recommendations for future research on the judgments and a set of recommendations for improvements in the teacher-assessment process.

In this paper we examine the empirical literature on the match between teacher-based assessments of student achievement levels and objective measures of student learning. Our specific concern is with the examination of concurrent relationships: the extent to which a teacher's a priori judgment of a student's achievement corresponds to the student's actual achievement on a measure administered at approximately the same time. These data are treated as reflecting on the validity or accuracy of the teacher-judgment measures. We begin with a discussion of the contexts in which teacher judgment emerges as an important question.

The Context of Teacher Judgments

Teacher Cognition

Models of teacher cognition suggest that teachers base their instructional decisions, in part, on judgments they make of student comprehension (e.g., Borko,

A version of this paper was presented at the annual meeting of the American Educational Research Association, San Francisco, March 1989. Thanks are due to D. A. Andrews and Lynda Robertson for their comments on an earlier draft.

Cone, Russo, & Shavelson, 1979; Clark & Peterson, 1986; Peterson, 1988; Shavelson & Stern, 1981). In the preinstructional, or preactive, phase of teaching, for example, teachers form judgments about their students' relative reading abilities before making decisions about instructional groupings (Shavelson & Borko, 1979).

There also is evidence that these judgments influence decisions in the interactive phase of teaching. McNair (1978–1979), through stimulated recall interviews, found that teachers' main consideration in making decisions during reading instruction was student achievement. Her 10 teachers "based many of their decisions on *what they surmised* was happening" with each student (p. 32; italics added). Research on "steering groups" further illustrates the role teacher judgments can play in the classroom. Specifically, Dahllof and Lundgren (1970, cited in Clark & Peterson, 1986, p. 256) found that teachers paced whole-class instruction according to whether a reference group of students (the steering group) "seemed to understand what was being presented." If teachers judged sufficient comprehension on the part of the steering group, a new topic was introduced; if not, the pace was slowed.

Through a series of stimulated recall interviews with six teachers, Colker (1984) found that 41% of the teachers' interactive thoughts pertained to student cognition. And 61% of these thoughts were categorized by Colker as "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')" (Colker, 1984, Table 3). Indeed, in their review of this literature, Clark and Peterson (1986) reported that the largest proportion of teachers' interactive thoughts pertained to the "learner" (p. 269).

Thus, it is apparent that the decision-making process of teachers, particularly in the interactive context, is influenced by the judgments they make about their students' cognitions. In turn, it seems reasonable to suggest that the decision-making process proceeds differently when based on accurate teacher judgments than when based on inaccurate teacher judgments (cf. Clark & Peterson, 1986; Peterson, 1988). This, then, is the primary context in which the accuracy of teacher judgments surfaces as an important question.

The Assessment Issue

The accuracy of achievement judgments may also be viewed as relevant in an assessment context. We clearly depend on teacher-based assessments of academic achievement in making educational decisions regarding students and for providing feedback to children, parents, and school psychologists (Elliott, Gresham, Freeman, & McCloskey, 1988; Gerber & Semmel, 1984; Hoge, 1983). These judgments probably constitute the primary source of information in such contexts. Similarly, there is a heavy dependence on these achievement judgments in research settings; teacher ratings of performance levels frequently appear as measures in research and evaluation studies (cf. Gresham, 1981; Hoge, 1983).

Often, these teacher-based measures are treated in a very casual way. For example, teachers are often asked to designate the students in their classroom possessing high "gifted potential" without being provided any real guidance in defining the construct (Hoge & Cudmore, 1986). There is, however, an increasing recognition that the judgments and assessments of teachers are being used as psychological measures and that the same psychometric criteria that apply to other measures, such as tests

or observation schedules, should apply here as well (Edelbrock, 1983; Gerber & Semmel, 1984; Gresham, 1981; Hoge, 1983, 1984; Hoge & Cudmore, 1986).

There is another sense in which the accuracy issue is important within this assessment context. There seems to be a widespread assumption, particularly among school psychologists, educational researchers, and other professionals, that teachers are generally poor judges of the attributes of their students—that their perceptions are often subject to bias and error. This assumption is rarely given explicit acknowledgment, but it does exist, and it has been discussed in connection with the decision-making literature by Egan and Archer (1985), the expectancy literature by Brophy (1983) and Hoge (1984), and the assessment literature by Hoge (1983) and Hoge and Cudmore (1986). One form of the criticism has been expressed as follows:

> Directly or indirectly, the accuracy of teachers' assessments of student ability is often an issue in educational research. It is commonly argued that commercial tests provide teachers with valuable information about the abilities and deficiencies of their students, from which it follows that teachers who rate their students without such information will often be in error. (Egan & Archer, 1985, p. 25)

Such an assumption represents a rather serious criticism of teachers, and a careful examination of the evidence bearing on it is therefore in order.

Review of the Research

Terms of the Review

The studies reviewed here are ones in which data are presented regarding the relationships between teacher judgments of student achievement and the student's actual performance on an independent criterion of achievement. The studies were located through a search of *Psychological Abstracts* and ERIC databases and a manual search of key journals. With three exceptions, the studies focused on students within regular classrooms. The exceptions are Gresham, Reschly, and Carey, (1987), who included both learning disabled (LD) and non-LD students in their sample; Leinhardt (1983), whose sample solely comprised LD students; and Silverstein, Brownlee, Legutki, and MacMillan (1983), who used only educable mentally retarded (EMR) students.

Three constraints were imposed for selecting studies. First, only studies employing naturalistic data were included; thus, analogue and simulation studies are not represented in the review. Second, the review focuses on cases where judgmental and test data were collected concurrently; thus, expectancy-type studies where teachers were asked to make a prediction of future performance are not included. Third, the review includes only published studies. This last criterion was introduced to ensure that some minimal level of methodological standards was met and that readers have access to original sources. (See Achenbach, McConaughy, & Howell, 1987, for an elaboration of this and related points.)

Methodological Considerations

The 16 studies included in the review have a common focus: the relationship between teachers' judgments of their students' academic performance and the students' actual performance on an achievement criterion. These studies are methodologically similar in some general respects: Each contains a variable representing

a teacher's judgment of a student's academic performance and each examined the correspondence between the teacher-judgment measure and student performance on a standardized achievement test.

There are also a number of methodological differences among these 16 studies that affect interpretation of results. These methodological characteristics are summarized in Table 1. A synthesis of the research results follows a discussion of these characteristics.

Direct versus indirect evaluations of teacher judgments. Nine of the studies summarized in Table 1 entailed relating teacher ratings or rankings of achievement levels to standardized achievement test scores. For example, Airasian, Kellaghan, Madaus, and Pedulla (1977) had teachers rate, on a 5-point scale, the performance of their students in English and mathematics and then related those ratings to standardized achievement test scores. These ratings are viewed as indirect evaluations of teacher judgments insofar as teachers were not asked specifically to estimate achievement test performance.

The direct judgments, in contrast, asked teachers specifically to estimate their students' performance on a concurrently administered achievement test. Helmke and Schrader (1987), for example, had teachers estimate the number of problems on an achievement test that each student would solve correctly. Teacher judgments of this kind represent direct judgments in that there is a stronger logical link between judgment and criterion. Seven of the studies summarized in Table 1 employed direct assessments; Wright and Wiese (1988) included both types.

Judgment specificity. The direct/indirect distinction also has implications for the specificity of the judgment, although the link is not entirely consistent. By definition, indirect measures of teacher judgments are less specific than direct measures in that the former are not explicitly tied to any one criterion in the judgmental process. Nonetheless, the degree of specificity varies among studies employing only indirect measures of teacher judgements. Luce and Hoge (1978), for example, asked teachers to rank order their students on various academic abilities. This kind of judgment, albeit indirect, requires teachers to make finer discriminations among students than those required by a 5-point rating scale.

Similarly, degree of specificity varies among studies involving direct measures of teacher judgments, although these direct measures are, in general, more specific than the indirect. For example, Hoge and Butcher (1984) asked teachers to estimate, in grade-equivalence scores, the likely performance of each of their students on an achievement test administered concurrently. Although direct, this summary index is less specific than the format employed by Coladarci (1986) and Leinhardt (1983), where teachers were asked to make judgments on an item-by-item basis.

Five types of judgment measures were employed in the studies reviewed, and these can be ordered roughly by the level of specificity the judgment entailed: (a) *ratings* (low specificity), where teachers rated each student's academic ability (e.g., "lowest fifth of class" to "highest fifth of class"); (b) *rankings,* where teachers were asked to rank order their students according to academic ability; (c) *grade equivalence,* where teachers estimated, in the grade-equivalent metric, each student's likely performance on a concurrently administered achievement test; (d) *number correct,* where teachers were asked to estimate, for each student, the number of correct responses on an achievement test, administered concurrently; and (e) *item responses* (high specificity), where teachers indicated, for each item on an achievement test

TABLE 1
Methodological characteristics of the studies reviewed

Author	Direct vs. indirect	Judgment measure	Reference group	Accuracy assessment	Unit of analysis
Airasian, Kellaghan, Madaus, & Pedulla (1977)	I	Ratings	NR	C	Pooled
Coladarci (1986)	D	IR	PI	C & PA	WC
Doherty & Conolly (1985)	D	GE	NR	C	Pooled
Farr & Roelke (1971)	D	Ratings	NR	C[a]	WC
Gresham, Reschly, & Carey (1987)	I	Ratings	NR	C	Pooled
Helmke & Schrader (1987)	D	NC	PI	C	WC
Hoge & Butcher (1984)	D	GE	NR	C & MR	Pooled[b]
Hopkins, Dobson, & Oldridge (1962)	I	Rankings	NR	C	Pooled
Hopkins, George, & Williams (1985)	I	Ratings	NR	C	WC
Leinhardt (1983)	D	IR	PI	C & PA	Pooled
Luce & Hoge (1978)	I	Rankings	NR	C	Pooled
Oliver & Arnold (1978)	I	GE	NR	C	Pooled
Pedulla, Airasian, & Madaus (1980)	I	Ratings	NR	C	Pooled
Sharpley & Edgar (1986)	I	Ratings	NR	C	Pooled
Silverstein, Brownlee, Legutki, & MacMillan (1983)	I	Ratings	NR	C[a]	Pooled
Wright & Wiese (1988)	I & D	Ratings, GE	NR	C	Pooled

Note. I = indirect, D = direct, IR = item response estimates, GE = grade equivalence or percentile estimates, NC = number correct estimates, NR = norm-referenced estimates, PI = peer-independent estimates, C = correlational analysis, PA = percent agreement, WC = within class.

[a] Complete multitrait-multimethod analysis performed.

[b] Analyses were based on both pooled and within-class data.

administered concurrently to the students, whether they thought the student would respond correctly to the item or had sufficient instruction to respond correctly.

Norm-referenced versus peer-independent judgments. Some teacher-judgment measures had a decidedly norm-referenced flavor, whereas others did not. Regarding the former, for example, 1 and 5 on the 5-point teacher judgment scale in the Airasian et al. (1977) study signified a student in the *lowest fifth* and *highest fifth* of the class, respectively. Rankings, as well as estimates of grade equivalents and instructional levels, also reflect a norm-referenced judgment. In contrast, a peer-independent judgment is called for where, for example, the teacher is asked to estimate the number of test problems a student will solve correctly. This judgment does not require the teacher to compare one student with another.

Assessing the accuracy of teacher judgments. Where teacher judgments were expressed as ratings, rankings, grade equivalents, or total-score estimates, the accuracy of the judgments was assessed by examining the correlation between judgment and criterion. Thus, accuracy is operationally defined as the correspondence between the relative standing of two sets of values: (a) the teachers' judgments of their students and (b) the students' actual performance on a relevant standardized test. Fourteen of the 16 studies reported correlations (or regression coefficients) as the sole index of accuracy.

Offering an alternative operational definition of accuracy, Coladarci (1986) examined the percentage of items for which (a) the teacher reported the student would answer the item correctly and (b) the student, in fact, answered the item correctly. Leinhardt (1983) also obtained item-level judgments: She determined the percentage of items—what she called the "hit rate"—for which (a) the teacher indicated sufficient instruction had been provided for the student to answer the test item correctly and (b) the student, in fact, answered the item correctly. In addition to examining accuracy in this way, both Coladarci and Leinhardt reported correlations between summary measures of teacher judgment and student achievement. That is, a "total" teacher judgment was derived by summing the teacher's item-level judgments, which, in turn, were correlated with the students' total scores on the achievement criterion. (A parallel procedure was followed to construct subscale teacher judgments.)

Unit of analysis. Researchers took one of two general approaches in calculating correlations between judgment and criterion. Some investigators combined K teachers and N students into a single, undifferentiated group. That is, class membership was ignored. Such a procedure can either overestimate or underestimate the judgment/criterion relationship. For example, where teacher judgments are in the form of ratings, judgment/criterion correlations based on a single, undifferentiated group will be attenuated by individual differences among teachers in how each calibrates the rating scale (Hopkins, George, & Williams, 1985).

Irrespective of calibration error, these correlations also will be underestimated where there is a positive correlation between judgment and criterion when computed for each of the K classes separately, but the scatterplot with all classes combined is considerably less elliptical. This could occur, for example, where there is little variability among class means on either the criterion measure or the judgment measure.

A similar phenomenon can *over*estimate the judgment/criterion relationship. That is, one might obtain a significant correlation when based on a single,

undifferentiated group of N students; when computed for each of the K classes separately, however, the correlation is zero. (Imagine a series of circles, sloping upward at a 45° angle.) Thus, within any one class, a teacher's judgments about student knowledge could be quite inaccurate. By determining the relationship across a wide range of classes, however, the investigator artificially inflates the judgment/criterion correlation.

To address these concerns, some investigators have incorporated class membership into their statistical analyses. In three studies, for example, the investigators computed judgment/criterion correlations separately for each of K classes and then, using the r to z transformation, determined the mean within-class correlation (Coladarci, 1986; Farr & Roelke, 1971; Hopkins et al., 1985). Choosing an alternative to this procedure, Helmke and Schrader (1987) simply reported the median of K correlations. Finally, Hoge and Butcher (1984) presented K within-class regression equations, where the dependent variable was a teacher judgment measure and one of the predictors was the student's performance on an achievement test. (Hoge and Butcher also presented the regression equation on the basis of a single, undifferentiated group of N students.)

The Correspondence Between Teacher Judgments and Student Achievement

Table 2 contains a summary of the principal findings of the studies, divided according to whether they called for direct or indirect teacher judgments of student achievement. Taken as a whole, these studies yielded judgment/criterion correlations ranging from 0.28 to 0.92. The median correlation, 0.66, suggests a moderate to strong correspondence between teacher judgments and student achievement. Instead of reporting a judgment/criterion correlation, Hoge and Butcher (1984) presented the results of a multiple regression analysis in which achievement test, IQ, and gender served as the predictors of teacher judgments. The standardized partial regression coefficient associated with achievement test was 0.71, which, like the correlations above, suggests a strong correspondence between teacher judgments and student achievement.

The percentage-agreement statistics reported by Coladarci (1986) similarly point to the validity of teacher judgments. Teachers, on the average, correctly judged their students' responses to at least 70% of the items on reading and mathematics subtests. Somewhat analogous to this statistic, Leinhardt (1983) found a "hit rate " of 64% on a reading comphrehension test. That is, for roughly two thirds of test items, teachers were correct in determining whether sufficient instruction had been provided for the student to answer the item correctly.

As noted above, however, these 17 studies vary methodologically in several basic ways. Do these methodological differences affect the results of these studies, particularly those involving judgment/criterion correlations? To address this question, we determined the median judgment/criterion correlation for the following methodological groupings: (a) indirect versus direct teacher judgments, (b) levels of judgment specificity, (c) norm-referenced or peer-independent teacher judgments, and (d) statistical analyses based on a single, undifferentiated group versus those that took class membership into account.

Indirect versus direct teacher judgments. Direct teacher judgments entailed an explicit link between criterion and judgment. In contrast, indirect evaluations did not involve an explicit criterion. Instead, the teacher was asked to provide an

TABLE 2
Summary of Results

Study	Results	
	INDIRECT ASSESSMENTS	
Airasian, Kellaghan, Madaus, & Pedulla (1977)	Reading	$r = .64$
	Math	$r = .62$
Gresham, Reschly, & Carey (1987)	Reading recognition	$r = .62$[a]
		$r = .67$
	Reading comprehension	$r = .64$
		$r = .66$
Hopkins, Dobson, & Oldridge (1962)	Reading	$r = .79$ (Grade 1)
		$r = .74$ (Grade 2)
		$r = .86$ (Grade 3)
		$r = .86$ (Grade 4)
		$r = .85$ (Grade 5)
Hopkins, George, & Williams (1985)	Reading	$r = .73$
	Language arts	$r = .74$
	Math	$r = .72$
	Social studies	$r = .64$
	Science	$r = .60$
Luce & Hoge (1978)	Reading	$r = .41$
	Math problem solving	$r = .28$
	Math concepts	$r = .29$
Oliver & Arnold (1978)	Reading	$r = .74$
Pedulla, Airasian, & Madaus (1980)	Reading	$r = .65$
	Math	$r = .63$
Sharpley & Edgar (1986)	Reading vocabulary	$r = .42$ (boys)
		$r = .44$ (girls)
	Reading comprehension	$r = .50$ (boys)
		$r = .56$ (girls)
	Math	$r = .45$ (boys)
		$r = .38$ (girls)
Silverstein, Brownlee, Legutki, & MacMillan (1983)	Reading	$r = .55$[b]
		$r = .61$
		$r = .48$
	Math	$r = .44$
		$r = .55$
		$r = .37$
Wright & Wiese (1988)	Reading	$r = .71$
	Language arts	$r = .70$
	Math	$r = .71$
	Social studies	$r = .57$

TABLE 2 (Continued)

Study	Results	
	DIRECT ASSESSMENTS	
Coladarci (1986)	Reading vocabulary	$r = .67$ (74%)[c]
	Reading comprehension	$r = .70$ (73%)
	Math concepts	$r = .72$ (70%)
	Math comprehension	$r = .70$ (76%)
Doherty & Conolly (1985)	Math	$r = .67$
	English	$r = .72$
	Reading	$r = .68$
Farr & Roelke (1971)	Reading vocabulary	$r = .92$
	Reading comprehension	$r = .59$
	Reading word analysis	$r = .48$
Helmke & Schrader (1987)	Math	$r = .67$
Hoge & Butcher (1984)	Reading	$\beta = .71$[d]
Leinhardt (1983)	Reading	$r = .67$ (64%)[e]
Wright & Wiese (1988)	Reading	$r = .82$
	Math	$r = .77$
	Language arts	$r = .76$
	Social studies	$r = .67$

[a] Separate ratings were collected for (a) pupils judged relative to classmates and (b) relative to grade-level expectations.
[b] Based on data collected on a group of pupils in each of 3 successive years.
[c] Agreement between teachers' item judgments and students' item responses in parentheses.
[d] Beta based on teacher estimate of performance with pupil IQ the other independent variable; multiple $R = .85$.
[e] Agreement between teachers' item judgments regarding sufficiency of instruction and students' item responses in parentheses.

achievement judgment, but with little guidance as to the nature of the construct. Yet, in both cases, teacher judgments were related to a single criterion: a score on a standardized achievement test.

Insofar as an ambiguous link between judgment and criterion should attenuate resulting correlations, one might expect *indirect* teacher judgments to correlate less with actual achievement than do *direct* teacher judgments. Interestingly, although this was the case, the differences were not dramatic. Among the studies calling for indirect teacher judgments, the judgment/criterion correlations ranged from 0.28 to 0.86; the median correlation was 0.62. In contrast, the studies involving direct assessments of teacher judgments yielded a range of judgment/criterion correlations of 0.48 to 0.92, with a median correlation of 0.69.

Judgment specificity. As indicated above, the operational definitions of teacher judgments in these studies differed in their specificity. For example, ratings required the least specificity: Teachers merely were asked to place each student on a scale ranging from, say, 1 to 5. All other forms of teacher judgments, on the other hand, called for considerably greater specificity. In ranking students, for example, the teacher must consider each student relative to his or her classmates; in predicting a student's actual achievement score, the teacher selects from a full continuum of possible values.

Among studies employing ratings—the predominant form of teacher judgment—the median judgment/criterion correlation was 0.61, with a range from 0.37 to 0.92. In fact, these correlations were generally lower than those associated with ranks (median $r = 0.76$; range: 0.28 to 0.86), grade equivalents (median $r = 0.70$; range: 0.67 to 0.74), number correct (single study $r = 0.67$), and item judgments (median $r = 0.70$; range: 0.67 to 0.72). The lower correlations associated with ratings probably reflect teachers' disinclination to use the full range of rating categories, which reduces the variability among teacher judgments and, consequently, the judgment/criterion correlations (Hopkins et al., 1985). The relative value of these correlations notwithstanding, there is strong correspondence between teacher judgment and student achievement, irrespective of how the former is operationally defined.

Norm-referenced versus peer independent. Again, some teacher judgments were measured in a norm-referenced fashion (e.g., rankings, grade equivalents); others called for peer-independent judgments (e.g., number correct, item judgments). This distinction, however, did not appreciably affect the judgment/criterion correlations. Among studies employing peer-independent ratings, the median judgment/criterion correlation was 0.68, with a range from 0.67 to 0.72; for norm-referenced judgments, the median correlation was 0.64, with a range from 0.28 to 0.92.

Unit of analysis. Most researchers based their correlational analyses on a single, undifferentiated group. Some, however, took class membership into account by determining the mean (or median) within-class correlation. Interestingly, both kinds of analyses produced similar judgment/criterion correlations. Among studies where the analyses involved a single, undifferentiated group, the median judgment/criterion correlation was 0.64, with a range from 0.28 to 0.86; for within-class analyses, the median correlation was 0.70, with a range from 0.48 to 0.92.

Moderator Variables

Some researchers explored the possible effects of additional variables on the accuracy of teacher judgments.

Differences among teachers. Research on teacher decision making has pointed to the hazards of pooling data across teachers in reporting summary statistics such as correlations or regression coefficients. Specifically, such a practice fails to recognize individual differences among teachers in their cognitions and instructional strategies (e.g., Borko & Cadwell, 1982; Clark & Peterson, 1986; Shavelson, Webb, & Burstein, 1986).

Although most of the studies in our review pooled data across teachers, there were some exceptions. For studies solely reporting judgment/criterion correlations, these exceptions entailed the separate calculation of K correlations, where K corresponds to the number of teachers. Variability among the K correlations, of course, speaks to the question of individual differences among teachers in the accuracy of their judgments.

In short, these data suggest that teachers do, in fact, differ in how accurately they judge their students' achievement. For example, Hopkins et al. (1985) obtained a range of within-class correlations of 0.44 to 0.88 across their 42 teachers. Even greater variability among teachers was found by Helmke and Schrader (1987), who reported within-class correlations ranging from .03 to .90 ($K = 31$). Finally, Hoge

and Butcher (1984) also uncovered individual differences among teachers in judgment accuracy. As shown above, these researchers estimated separate within-class multiple regression equations, where the dependent variable was teacher judgment and the predictor variables were IQ, achievement test, and gender. Hoge and Butcher reported standardized partial regression coefficients for achievement test ranging from 0.40 to 0.87 ($K = 12$).

Finally, Coladarci (1986) investigated teacher effects on his percentage-agreement index by treating "teacher" as the independent variable in an analysis of variance; the dependent variable, percentage agreement, was the mean percentage of items for which the teacher correctly judged the student's item-level responses. A significant teacher effect was found for one of the four achievement areas: mathematics concepts.

Although the 16 studies generally point to the validity of teacher-based achievement judgments, the results of the four studies just discussed are important insofar as they demonstrate that not all teachers are equally adept at making these judgments. "Teacher judgment accuracy," then, appears to be an individual-difference variable that is worthy of further examination in research on teaching.

Student gender. In all three studies examining it, student gender failed to show a significant effect on the judgment/criterion relation (Doherty & Conolly, 1985; Hoge & Butcher, 1984; Sharpley & Edgar, 1986). These essentially negative results are consistent with the general findings within the teacher-judgment literature. Thus, although Dusek and Joseph (1983) concluded that teachers hold differential social-behavioral expectations for boys and girls, their meta-analysis yielded no significant gender differences in expectations for academic performance (also see Brophy & Good, 1974.)

Subject matter differences. Although a number of researchers reported analyses separately by subject matter, in only two cases was this variable systematically analyzed. Hopkins et al. (1985) found that judgment/criterion correlations for achievement in social studies and science were significantly lower than for achievement in language arts, reading, and math. However, the magnitude of correlations in all five content areas was appreciable.

Coladarci (1986) calculated his percentage-agreement index separately for four subtests: reading vocabulary, reading comprehension, mathematics concepts, and mathematics computation. An analysis of variance resulted in a significant effect of subject matter: Teachers were considerably more accurate in judging performance in mathematics computation than in mathematics concepts. He saw this difference as due, in part, to the greater amount of direct instruction involved in teaching computation skills compared with mathematics concepts.

Student ability. Student ability, broadly conceived, was explored as a potential moderator variable in two studies. Although Leinhardt (1983) did not report detailed data on the issue, she did indicate that her "hit rate" index correlated $r = 0.39$ with actual achievement. That is, teachers of the learning disabled were somewhat more accurate in judging the sufficiency of instruction for higher-achieving than for lower-achieving students.

Coladarci (1986) found substantial correlations between his percentage-agreement index and student ability: rs ranged from 0.78 to 0.89, depending on the subject matter. Across all items on the four subtests, the mean percentage agreement was roughly 60% for students in the lowest quartile and 88% for students in the

highest quartile. Clearly, these eight teachers were less able to judge the performance of their lower-achieving students.

Student IQ was explored as a moderator of the judgment/criterion relation in one study. Because they included student IQ as a predictor variable in their study, Hoge and Butcher (1984) were able to separate the effect of student "intelligence" from student "achievement" in predicting teacher judgments of the latter. With 12 teachers pooled, IQ made a significant contribution to the prediction of the achievement judgments ($\beta = 0.18$), although the magnitude of the contribution was far less than that of the achievement test scores ($\beta = 0.71$). It is interesting to observe, however, that the extent of the independent contribution of IQ scores to the achievement judgments varied widely across the 12 teachers. This is evident both from the separate within-class multiple regression analyses and from an analysis of residual scores in which it was revealed that 4 of the 12 teachers displayed a tendency to overestimate the performance of high-IQ students.

There is, then, a relatively strong suggestion from these four studies that students' academic ability may influence the accuracy with which teachers judge student achievement. Coladarci (1986) suggested that the higher levels of accuracy observed for high-performance pupils may arise from a response set that operates with the achievement judgments; however, as he admitted, the whole issue requires further exploration.

Discussion

The 16 studies reviewed in the previous section yielded data indicating generally high levels of agreement between the judgmental measures and the standardized achievement test scores. The range of correlations for the indirect comparisons was 0.28 to 0.86, with a median correlation of 0.62, whereas the direct tests yielded a range of correlations from 0.48 to 0.92, with a median of 0.69. In our view, these data support the validity of the teacher judgments of academic achievement. The correlations certainly exceed the convergent and concurrent validity coefficients normally reported for psychological tests, and it is encouraging that the correlations remain strong irrespective of methodological distinctions among the studies. Finally, it is worth noting that the levels of association between teacher judgment measures and the criterion measures uncovered in this review were similar to those reported in teacher expectation studies (cf. Brophy, 1983).

This overall positive conclusion regarding the accuracy of these judgments must, as we have shown, be interpreted in light of some methodological considerations and the operation of moderator variables. (These considerations are dealt with in the following sections detailing research and practical implications.) Still, the conclusion that these achievement judgments are generally veridical has important implications for the teacher decision-making and assessment contexts discussed at the beginning of our review.

Teacher cognitions about student attributes and performance levels constitute only one element in the teacher decision-making process. It can be argued, however, that they are critical elements in the process: Other things being equal, decisions based on accurate assessments of student attributes will be more functional than those based on inaccurate assessments. Our conclusion that the achievement judgments are generally veridical is an encouraging one in this context, but it also

highlights the importance of considering this aspect of the decision-making process in future analyses.

Our conclusion that the performance judgments are, by and large, valid also has important implications for the practical use of teacher-based assessments. In particular, it speaks to members of the public and to educational professionals (e.g., university-based researchers, school psychologists) who express doubts regarding the quality of teacher-based assessments of students. Although the studies in our review by no means provide a final evaluation of the accuracy of achievement judgments or any evidence that the judgments are without error, this literature does not support the total rejection of teacher judgments that one sometimes encounters.

Implications for Research

The first recommendation is that research be guided by more explicit statements of the achievement construct. Much of the research has entailed global achievement judgments being assessed against scores from standardized achievement tests that are sometimes of questionable construct validity (Linn, 1986; Sattler, 1988; Snow, 1980). It is, therefore, not always clear in this research just what aspect of student performance is being assessed.

A related point is that, in developing this achievement construct, efforts should be made to ensure that the construct is one relevant to the teaching process. There are several aspects to this issue. First, questions can be raised about the extent to which the definition of achievement represented in standardized achievement tests corresponds to the learning objectives of a particular classroom. Second, questions can be raised about the meaningfulness of the global achievement judgments collected in many of the studies reviewed. As Coladarci (1986) noted in his study, "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" (p. 142). Future research should probably employ specific rather than global judgment indices.

A second recommendation for future research in this area is that closer attention be paid to the match between judgment and criterion. Although our analysis did not reveal major differences in the outcomes of direct and indirect tests of validity, the use of parallel judgment/criterion dimensions facilitates a less equivocal interpretation of findings. A related suggestion is that the measurement scales underlying judgment and criterion should correspond (Egan & Archer, 1985).

Third, we recommend that both convergent and discriminant validity of teacher judgments be examined. We have seen that most of the validity evaluations focused on convergent validity. The two cases using complete multitrait-multimethod matrices, Farr and Roelke (1971) and Silverstein et al. (1983), found support for convergent validity but reported less impressive levels of discriminant validity. Evaluations focusing on both convergent and discriminant validity actually have two advantages. First, they provide us with more complete validity evaluations of the judgmental measure and, second, they encourage a more thorough exploration of the achievement construct.

A fourth recommendation is that further attention be paid to the operation of moderator variables in the judgment/criterion relation. For example, rather strong

suggestions were obtained in this literature to the effect that student ability and achievement levels might be functioning as moderators and that teachers might be more accurate at assessing achievement in high- than low-performing students (Coladarci, 1986; Hoge & Butcher, 1984; Leinhardt, 1983). Unfortunately, in no case were variables that might be associated with this effect investigated, and it certainly merits further exploration. It should be noted, however, that there are nagging methodological problems with analyses of this kind that make these findings difficult to interpret (cf. Gage & Cronbach, 1955; Kenny & Albright, 1987).

The most convincing evidence for the operation of a moderator was in connection with the teacher variable; Coladarci (1986), Helmke and Schrader (1987), Hoge and Butcher (1984), and Hopkins et al. (1985) all presented evidence of individual differences among teachers in judgmental accuracy. Unfortunately, the research included in the review provides few clues as to whether these differences arise from characteristics of teachers (e.g., experience, training, teaching philosophy, measurement policy), the composition of the class, or some other variable. It is worth noting, however, that useful information regarding individual differences in teacher judgments can be found in the teacher expectancy (e.g., Babad, Inbar, & Rosenthal, 1982; Tom, Cooper, & McGraw, 1984) and teacher cognition (e.g., Carpenter, Fennema, Peterson, & Carey, 1988; Carpenter, Fennema, Peterson, Chiang, & Loef, in press; Peterson, Carpenter, & Fennema, in press) literatures. The latter studies are especially interesting in that they are able to link individual differences among teachers in cognitions about pupils to differences in teacher effectiveness.

Implications for Teaching

The achievement judgments revealed themselves to be generally accurate. Still, there was clearly some degree of error operating, and, further, levels of accuracy varied across teachers. There is, therefore, room for improvement.

There are several directions these efforts might take. First, greater efforts should be made to sensitize teachers to the extent and importance of the assessment role in the teaching process (Hoge, 1983; Hoge & Cudmore, 1986). Second, more intensive experience with the basic principles of measurement and assessment should be provided. Third, teachers should be familiarized with the interpretation of different types of assessment devices, including norm-referenced tests, observational procedures, and jugmental measures. Fourth, improved judgmental tools should be developed and made available to teachers. Finally, recently developed programs of the sort described by Carpenter et al. (in press) and Peterson et al. (in press) for enhancing teachers' abilities at diagnosing cognitions and knowledge states in children should be expanded and encouraged.

In summary, parents, researchers, school psychologists, and others depend very heavily on the assessments of achievement provided by teachers. Further, these assessments constitute important elements in the teaching process. It is time that we began giving these measures the same attention accorded other types of measuring instruments.

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational spec-

ificity. *Psychological Bulletin, 101,* 213–232.
Airasian, P. W., Kellaghan, T., Madaus, G. F., & Pedulla, J. J. (1977). Proportion and direction of teacher rating changes of pupils' progress attributable to standardized test information. *Journal of Educational Psychology, 69,* 668–678.
Babad, E., Inbar, J., & Rosenthal, R. (1982). Pygmalion, Galatea, and the Golem: Investigations of biased and unbiased teachers. *Journal of Educational Psychology, 74,* 459–474.
Borko, H., & Cadwell, J. (1982). Individual differences in teachers' decision strategies: An investigation of classroom organization and management decisions. *Journal of Educational Psychology, 74,* 598–610.
Borko, H., Cone, R., Russo, N. A., & Shavelson, R. J. (1979). Teachers' decision making. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 136–160). Berkeley, CA: McCutchan.
Brophy, J. E. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75,* 631–661.
Brophy, J. E., & Good, T. L. (1974). *Teacher-student relationships: Causes and consequences.* New York: Holt, Rinehart and Winston.
Carpenter, T. P., Fennema, E., Peterson, P. L., & Carey, D. A. (1988). Teachers' pedagogical content knowledge of students' problem solving in elementary arithmetic. *Journal for Research in Mathematics Education, 19,* 385–401.
Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C., & Loef, M. (in press). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal.*
Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.), *Third handbook of research on teaching* (pp. 255–296). New York: Macmillan.
Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78,* 141–146.
Colker, L. (1984, April). *Teachers' interactive thoughts about pupil cognition.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.
Doherty, J., & Conolly, M. (1985). How accurately can primary school teachers predict the scores of their pupils in standardised tests of attainment? A study of some non-cognitive factors that influence specific judgments. *Educational Studies, 11,* 41–60.
Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75,* 327–346.
Edelbrock, C. (1983). Problems and issues in using rating scales to assess child personality and psychopathology. *School Psychology Review, 12,* 293–299.
Egan, O., & Archer, P. (1985). The accuracy of teachers' ratings of ability: A regression model. *American Educational Research Journal, 22,* 25–34.
Elliott, S. N., Gresham, F. M., Freeman, T., & McCloskey, G. (1988). Teacher and observer ratings of children's social skills: Validation of the Social Skills Rating Scales. *Journal of Psychoeducational Assessment, 6,* 152–161.
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.
Gage, N. L., & Cronbach, L. J. (1955). Conceptual and methodological problems in interpersonal perception. *Psychological Review, 62,* 411–422.
Gerber, M. M., & Semmel, M. I. (1984). Teacher as imperfect test: Reconceptualizing the referral process. *Educational Psychologist, 19,* 137–148.
Gresham, F. M. (1981). Social skills training with handicapped children: A review. *Review of Educational Research, 51,* 139–176.
Gresham, F. M., Reschly, D. J., & Carey, M. P. (1987). Teachers as "tests": Classification accuracy and concurrent validation in the identification of learning disabled children. *School Psychology Review, 16,* 543–553.
Helmke, A., & Schrader, F-W. (1987). Interactional effects of instructional quality and teacher

judgment accuracy on achievement. *Teaching and Teacher Education, 3,* 91–98.
Hoge, R. D. (1983). Psychometric properties of teacher-judgment measures of pupil aptitudes, classroom behaviors, and achievement levels. *Journal of Special Education, 17,* 401–429.
Hoge, R. D. (1984). The definition and measurement of teacher expectations: Problems and prospects. *Canadian Journal of Education, 9,* 213–228.
Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology, 76,* 777–781.
Hoge, R. D., & Cudmore, L. (1986). The use of teacher-judgment measures in the identification of gifted pupils. *Teaching and Teacher Education, 2,* 181–195.
Hopkins, K. D., Dobson, J. C., & Oldridge, O. A. (1962). The concurrent and congruent validities of the Wide Range Achievement Test. *Educational and Psychological Measurement, 22,* 791–793.
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.
Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social-relations analysis. *Psychological Bulletin, 102,* 390–402.
Leinhardt, G. (1983). Novice and expert knowledge of individual student's achievement. *Educational Psychologist, 18,* 165–179.
Linn, R. L. (1986). Educational testing and assessment: Research needs and policy issues. *American Psychologist, 41,* 1153–1160.
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.
McNair, K. (1978–1979). Capturing inflight decisions: Thoughts while teaching. *Educational Research Quarterly, 3,* 26–42.
Oliver, J. E., & Arnold, R. D. (1978). Comparing a standardized test, an informal inventory and teacher judgment on third grade reading. *Reading Improvement, 15,* 56–59.
Pedulla, J. J., Airasian, P. W., & Madaus, G. F. (1980). Do teacher ratings and standardized test results of students yield the same information? *American Educational Research Journal, 17,* 303–307.
Peterson, P. L. (1988). Teachers' and students' cognitional knowledge for classroom teaching and learning. *Educational Researcher, 17*(5), 5–14.
Peterson, P. L., Carpenter, T., & Fennema, E. (in press). Teachers' knowledge of students' knowledge in mathematics problem solving: Correlation and case analysis. *Journal of Educational Psychology.*
Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Sattler Publishing Co.
Sharpley, C. F., & Edgar, E. (1986). Teachers' ratings vs standardized tests: An empirical investigation of agreement between two indices of achievement. *Psychology in the Schools, 23,* 106–111.
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.
Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. C. Wittrock (Ed.), *Third handbook of research on teaching* (pp. 50–91). New York: Macmillan.
Silverstein, A. B., Brownlee, L., Legutki, G., & MacMillan, D. L. (1983). Convergent and discriminant validation of two methods of assessing three academic traits. *Journal of Special Education, 17,* 63–68.
Snow, R. E. (1980). Aptitude and achievement. In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade* (pp. 39–59). San Francisco: Jossey-Bass.
Tom, D. Y. H., Cooper, H., & McGraw, M. (1984). Influences of student background and

teacher authoritarianism on teacher expectations. *Journal of Educational Psychology, 76,* 259–265.

Wright, D., & Wiese, M. J. (1988). Teacher judgment in student evaluation: A comparison of grading methods. *Journal of Educational Research, 82,* 10–14.

Authors

ROBERT D. HOGE is Professor, Department of Psychology, Carleton University, Ottawa, Ontario, K1S 5B6, Canada. He specializes in educational psychology and psychological assessment.

THEODORE COLADARCI is Associate Professor, College of Education, University of Maine, Orono, ME 04469. He specializes in educational psychology.