Chapter 14

Using Teacher Rating Scales in the Identification of Students for Gifted Services

Karen L. Westberg, University of St. Thomas

Toua, a young Hmong boy, was identified for gifted services just six months after being relocated from a refugee camp in Thailand to Minnesota and enrolling in school. Did he score above the 95th percentile on an aptitude or achievement test? No, but his teacher observed his dramatic progress in learning English and his amazing ability in mathematics, spatial learning tasks, and problem solving tasks. When completing a teacher rating form for screening students for gifted education services, she rated him highly on specific traits and behaviors she observed in the classroom and recommended him for services. This true story, along with less dramatic examples, indicates that obtaining teacher input is valuable when considering students for gifted education services.

Experts in the field of gifted education have long recommended using teacher judgment measures among the multiple sources of information for screening and identifying students for gifted education services. According to the most recent *State of the States in Gifted Education* Report (NAGC, 2009), teacher judgment information and test score information are the two most commonly used sources of information when identifying students for gifted education services. There appears to be universal agreement by experts about the need to include teacher judgment in the identification process. Shore, Cornell, Robinson, and Ward (1991) published a seminal book on 101 recommended practices in gifted education. Among these practices were the need to base identification on multiple criteria (p. 48), and the importance of including teacher nominations in the identification process (p. 65). After reviewing the evidenced-based support for these recommended practices, they concluded, “Nominations forms and questionnaires should address specific characteristics or subject matter, and especially abilities not addressed by formal tests” (p. 65). Lohman and Lakin (2007) also argue for the inclusion of teacher judgment measures when identifying students for gifted services, stating, “Combining evidence of current achievement, reasoning abilities, and teacher ratings can help increase the diversity of gifted programs while also identifying the students in all ethnic groups most likely to benefit from special instruction” (p. 22). The recent *2010*

Pre-K–Grade 12 Gifted Programming Standards (NAGC, 2010) underscore this by stating that comprehensive practices and multiple assessments from different sources should be used in the identification process.

Historical Perspectives

Although widely used today, teacher judgment instruments for identification have not always been a recommended practice because of concerns about the validity and reliability of teachers' input. This view could be traced to Terman's (1925) research published in the *Genetic Studies of Genius*. When gathering data for this study, Terman asked teachers to refer the brightest child, the second-brightest child, the third-brightest child, and the youngest child in their classrooms for assessment on the *Stanford-Binet Intelligence Scale*, the instrument he developed. Because he found that the youngest children, more so than the other children, met his criterion of having IQs of 140 and above on the Stanford-Binet, he concluded that teachers were not particularly skilled in predicting which children would score highest on his intelligence scale. This raises the issue of the *criterion* problem, namely, what are we trying to predict with teacher ratings and what should be used as the criterion when validating teacher judgment measures? In Terman's situation, teachers were asked to predict who would score highest on a particular intelligence test (the criterion), which he equated with giftedness.

Pegnato and Birch's (1959) study on the effectiveness and efficiency of using teacher ratings in the identification process, unfortunately, has had a long-lasting impact on views about using teachers' input when identifying students for services. When conducting a study to identify junior high students, they concluded that teacher ratings lacked validity and reliability and, therefore, should not be used. This single, brief study has been cited over the years as a rationale for excluding or providing little weight to teachers' ratings. After years of mistrust about the value of including teacher judgment information in the identification process, a few researchers examined the Pegnato and Birch study more closely. Borland (1978) challenged their findings by stating that if the teachers in the Pegnato and Birch study had rated students on specific behaviors rather than on general ability, the results would have been different. Gagné (1994) conducted a re-analysis of the Pegnato and Birch data, which revealed major methodological flaws in their study. Gagné illustrates how effectiveness (absence of false negatives) and efficiency (absence of false positives) cannot be independent of each other and, therefore, should not have been measured as such. Gagné concluded his investigation by stating, "Educators in the field should stop citing Pegnato and Birch's (1959) study as proof of poor teacher judgment in identifying gifted and talented children; their data do not support such a sweeping judgment" (p. 126). And, finally, Birch (1984) himself, 25 years later, questioned whether there was any value in formal identification at all.

Rationale for Using Teacher Judgment Measures

Why use teacher judgment measures when identifying students for gifted education services?

The most common rationale is that they provide additional and different information about the characteristics and behaviors we associate with giftedness, and we should not rely on just one source of information when selecting students for gifted services. Most psychologists and educators no longer believe that a high IQ on an intelligence test, as was Terman's assertion, is equated with giftedness (e.g., see Sternberg and Davidson, 2005). The problem, of course, is that there is limited consensus on what constitutes giftedness. Nonetheless, by using teacher judgment measures, it is anticipated that teachers' observations of traits and behaviors not tapped by traditional ability or achievement tests, such as perseverance, intellectual playfulness, and focused interests, will be illuminated, and students who exhibit capabilities in different ways will be identified for gifted education services.

A variety of teacher judgment measures for screening and identifying gifted learners have been developed over the years. Unfortunately, many have limited or no empirical support. Much too often, we find that consultants or school districts have created their own teacher rating forms or checklists, which have absolutely no support for their reliability and validity. In many cases, these forms have been created in an earnest attempt to find students who demonstrate strengths not addressed on aptitude or achievement measures, but school personnel need to realize that, when using teacher judgment instruments with no empirical support, they are using a highly crude measurement tool, much like using one's arm span to measure the length of a football field. Only published teacher judgment measures with empirical support will be discussed in this chapter, and only instruments with empirical support should be used in a formal screening and identification process. Other, non-researched instruments (e.g., *Kingore Observation Inventory*, the *Kranz Talent Identification Instrument*) may be helpful for other purposes (e.g., for discussions in professional development sessions, for developing curricular experiences aligned with certain traits), but non-researched instruments should not be used when identifying students for formal gifted education services.

Scales for Rating the Behavioral Characteristics of Superior Students

In 1976 Renzulli, Smith, White, Callahan, and Hartman published the *Scales for Rating the Behavioral Characteristics of Superior Students (SR-BCSS),* a series of 10 separate teacher judgment scales designed to obtain information about the manifestations of students' characteristics, which were learning, motivation, creativity, leadership, artistic, musical, dramatics, communication-precision, communication-expressiveness, and planning. The first three or four scales—learning, motivation, creativity, and leadership—are most commonly used. The other scales are used when appropriate for programs that focus on those traits. Readers familiar with Renzulli's (1978) three-ring definition of giftedness will recognize that his conception of giftedness underlies the theory behind these scales (see Chapter 2 of this volume). Two items on the 1976 learning scales include: "Possesses a large storehouse of information about a variety of topics (beyond the usual interests of youngsters his age)," and "Displays a great deal of curiosity about many things; is constantly asking questions about anything and everything." Each characteristic listed on a scale in 1976 was selected because of the empirical support for it; for example, the aforementioned characteristic about curiosity is referenced to work done by several researchers, including Torrance (1962). To respond to the items on the scale, teachers were instructed to rate the frequency with which they observe each characteristic manifested in a student on a 4-point scale (1 = never, 2 = rarely, 3 = occasionally, and 4 = always).

These scales have been arguably the most widely used teacher judgment rating scales for gifted programming in the US and have been translated and researched for use in several countries (e.g., Kalatan,1991; Nazir, 1988; Subhi, 1997; Srour, 1989). The research conducted with the original scales is described in the technical and administration manual for the scales (Renzulli, Smith, White, Callahan, and Hartman, 1976). A few years after *SRBCSS* was originally published, Renzulli and Reis (1985) published teacher-training exercises to accompany the learning, motivation, creativity, and leadership scales. Unfortunately, many users of the scales have not been aware of them nor have they used the teacher-training materials for the first four scales (the most widely used of the 10 scales). The teacher-training exercises were designed to increase teachers' understanding about the key concepts underlying the items and to increase the reliability of teachers' ratings.

The *Scales for Rating the Behavioral Characteristics of Superior Students* were revised and published in 2002 (Renzulli, Smith, White, Callahan, Hartman, & Westberg). When conducting the literature review for the *SRBCSS* revision (examining articles published between 1976 and 2001), studies were organized into two categories: those examining the construct validity of teacher judgment measures and those in which a teacher judgment measure was used in criterion-related validity studies. Construct validity refers to the extent to which the operationalization of a construct on a test or scale actually supports the construct—that is, does a measure of critical thinking really measure what we mean by critical thinking, and does a scale on motivation really measure motivation (see also Chapter 7 of this volume for a discussion of validity)? Summaries of the limited studies exploring construct validity of all teacher judgment measures are summarized in the *SRBCSS Technical and Administration Manual* (Renzulli, Smith, Callahan, White, Hartman, & Westberg, 2002). Conclusions from these construct validity studies were taken into account when revising the *SRBCSS* scales.

Many of the studies on teacher judgment conducted between 1976 and 2001 involved using the *SRBCSS* scales or other scales as predictors in a criterion-related validity study. Criterion-related validity refers to the degree to which a measure is correlated with another measure presumed to be related to the first measure. Quite often the criterion in investigations of teacher judgment measures has been an intelligence test. Many researchers (e.g., Borland, 2008; Renzulli & Delcourt, 1986) believe that the selection of an intelligence test as a criterion for a teacher judgment measure simply does not support logical inferences. If teachers' ratings are used to predict performance on intelligence tests, what is the rationale for even using the teachers' ratings? In other words, why second guess intelligence tests? This is referred to as the *criterion* problem. Despite this, many studies involving the use of teacher judgment measures have used intelligence tests or achievement tests as the criterion, which the authors of *SRBCSS* believe is inappropriate.

When preparing the revised scales for field tests, a few new items (characteristics) with empirical support were added; scales were modified to include a 6-point response scale (Never, Very Rarely, Rarely, Occasionally, Frequently, Always), as opposed to the original 4-point response scale, which was criticized in the literature as being not on an interval scale; compound items were separated into separate items; and item stems were worded into gender-neutral language (Renzulli, Smith, Callahan, White, Hartman, & Westberg, 2002). Details about the sampling and data-gathering procedures for the field tests of the revised scales with Grade 3–12 teachers are described in the *SRBCSS Technical and Administration Manual* (Renzulli, Smith, Callahan, White, Hartman, & Westberg, 2002). The manual also contains details about the judgmental and empirical procedures used to provide evidence for the content validity (ratings by 60 experts in the field of gifted education), construct validity (principal components analysis), and criterion-related validity of the scales. The procedure for investigating the criterion-related validity warrants some discussion here because it was designed to address the *criterion* problem mentioned earlier. Instead of using an intelligence or achievement test as the criterion, another instrument was developed for this purpose: *Rating Student Performance in a Gifted Program (RSP/GP)* (Renzulli & Westberg, 1991). The *RSP/GP* contains 10 items on a 5-point response scale, such as "This year, [the student] created quality projects." Classroom teachers completed the *SRBCSS* scales (learning, motivation, creativity, and leadership) in the fall, and a sub-sample of gifted education specialists completed the *RSP/GP* in the spring of that same year on the students who had been receiving gifted education services, resulting in a moderate correlation.

Details about the procedures used to support the alpha and inter-rater reliability of the revised *SRBCSS* are also described in the *SRBCSS Technical and Administration Manual* (Renzulli, Smith, Callahan, White, Hartman, & Westberg, 2002). Strong alpha reliability coefficients (ranging from r = .84 to r = .97) and moderate inter-rater reliability coefficients were obtained (r = .50 to r = .65) on the revised scales. Hence, the above analyses provide technical support for the revised *SRBCSS.*

Four New *SRBCSS*

Four new *SRBCSS* teacher-rating scales were developed recently for obtaining teacher ratings on Grade 3–8 students in four content areas—reading, mathematics, science, and technology (Renzulli, Siegle, Reis, Gavin, & Sytsma Reed, 2009). These areas were selected for the new scales for two major reasons. The authors realize that variations exist among learners; namely, some students demonstrate strengths in one domain and not another, and the authors wanted to support teachers' attempts to differentiate instruction in specific content areas. To support the content validity of the new scales, experts' ratings (25 experts for each scale) were obtained, and the new scales were field tested in several schools throughout the country. A total of 187 teachers completed ratings on 726 Grade 4–6 students. Confirmatory factor analysis was conducted to examine the construct-related validity support of the new scales. Initially, separate confirmatory factor analyses were conducted for each of the four domains, and the number of items was reduced in each scale to establish the model of best fit. Then, a confirmatory factor analysis was conducted of a model that included all four scales. The fit index of the combined model, $X^2(371) = 1541.22$, was significant ($p<.001$), providing support for the construct validity of the scales, and all alpha reliabilities of the scales exceeded $r = .97$. Additional support for the validity of the scales was established by correlating the ratings on the scales with students' grades in academic subjects, resulting in moderate to strong correlations (e.g., $r = .453$ for technology and $r = .731$ for mathematics.) Additional details about the research procedures and findings can be obtained in the third edition of *Scales for Rating the Behavioral Characteristics of Superior Students Technical and Administration Manual* (Renzulli et al., 2010).

Authors' Recommendations for Using *SRBCSS*

The third edition of the *SRBCSS* manual (Renzulli et al., 2010) includes an explanation of the procedures used to develop the 2002 revised scales, procedures for developing the four content scales in 2009, and recommendations for using the scales. The manual also includes teacher-training exercises for all 14 scales, which were designed to improve teachers' understanding of the behaviors and traits on the scales as well as improve the reliability of their ratings. Before teachers complete the scales, the authors highly recommend that the teacher-training exercises be used (on different days, not all in one sitting, to address teacher fatigue). Three general guidelines for using the scales are: (1) consider the type of program for which students are being identified when selecting the scales to use (e.g., use the creativity scale if the goals of the program include the development of creativity); (2) examine each scale separately—do not add the scores from the scales together to form a total score (the dimensions on the scales represent relatively different sets of behavioral characteristics, and a composite or total score would overlook unique student strengths); and (3) do not modify or abbreviate the scales by reducing the number of items on each scale (doing so will definitely lower the reliability estimates on the scales).

National norms are not provided in the manual for *SRBCSS* because Renzulli et al. (2010) believe that this information is not meaningful or useful. Instead, the authors believe local norms should be established because *SRBCSS* is purposefully designed to assess students' characteristics within a local reference group. Lohman (2009a) advocates developing local norms when selecting students for gifted education services, stating, "There is a tradeoff between getting a *more precise but less valid* estimate of the student's talent by using an inappropriate national norm group and getting a *less but more valid* estimate by using a more appropriate local or subgroup norm" (p. 238; see also Chapter 10 of this text). The *SRBCSS Technical and Administration Manua*l includes information on how to establish local percentile ranks. In order to establish local norms, the teacher ratings need to be completed on a variety of students, including students who do not demonstrate the characteristics to a high degree. Therefore, to establish local norms initially, it is recommended that a subset of teachers in a district complete the scales on all of their students because a large and varied sample is necessary for calculating norms. (It should be noted that the scales are now available online through Creative Learning Press, and when teachers complete the scales online, the system calculates and provides local norms.)

The final recommendation when using *SRBCSS* is this: "As with other test score information, a *SRBCSS* rating should not be used as the single criterion for selecting students for special programs. The information should be used in conjunction with other information" (Renzulli et al., 2010, p. 25). Once again, we are reminded that we should be using multiple sources of information when identifying students for gifted services.

Scales for Identifying Gifted Students

The *Scales for Identifying Gifted Students (SIGS)* is a series of scales "designed to assist school districts in the identification of students as gifted" (Ryser & McConnell, 2004, p.1). The *SIGS* contains items on seven separate scales (general intellectual ability, language arts, mathematics, science, social studies, creativity, and leadership) to which teachers respond on a 5-point scale (0 = never, 1 = rarely, 2 = some, 3 = somewhat more, 4 = much more). Teachers are asked to respond to items by keeping in mind how each child compares to his or her peers on the characteristic being rated. The authors developed these seven scales because they "recognize these as being seven areas of giftedness," and they developed two versions of the scales, the *School Rating Scales (SRS)* form and the *Home Rating Scales (HRS)* form. The items on the scales are identical on both forms. For example, one of the general intellectual ability items states, "Demonstrates a healthy skepticism and curiosity," and one of the language arts items states, "Is able to discuss literature or other issues at an interpretive (explanatory) level." The *SIGS* are designed for ages 5–18 and contain 12 items on each scale. Based on the authors' review of the literature in each of the seven areas, the authors selected characteristics for the scales that indicated strengths within each area. The citations for the literature support are provided in the technical manual accompanying the scales.

When developing the *SIGS*, (Ryser & McConnell (2004) piloted the scales with two groups to establish national norms for "general" and "gifted" students. To obtain the pilot groups, the authors solicited participants who had purchased tests previously from the publisher. Once selected for participation, teachers were asked to complete the scales on students who were already participating in a gifted program and on the general population of their students. The technical manual contains tables for converting raw scores into standard scores and percentile ranks on each scale for the various age groups.

Technical Support for *SIGS*

The *SIGS* technical manual (Ryser & McConnell, 2004) includes summary information on the procedures used to support the validity of the scales. Using sub-samples from the pilot group, scales were correlated with students' scores on the *WISC-III, Test of Cognitive Skills, Otis-Lennon School Ability Test, Cognitive Ability Test-2,* and *Torrance Tests of Creative Thinking-Figural* scores to support criterion-related validity. These various analyses resulted in moderate to high correlations on the *School Rating Scale,* with the highest correlations obtained between the seven *SIGS* and the *Test of Cognitive Skills-2.*

The *SIGS* technical manual (Ryser & McConnell, 2004) also includes information on the procedures used to support the reliability of the scales. Internal consistency, test-retest, and inter-rater reliability procedures resulted in moderate to high reliability coefficients. For example, the alpha reliabilities ranged from r = .93 to .96 on the scales from the *School Rating Scale*-gifted subsample. Using a two-week interval on the test-retest procedures, reliabilities ranged from r = .58 to .93 on the scales from the *School Rating Scale*-gifted sample. Inter-rater reliability of the school and home versions was examined, resulting in correlations between the teacher and parent ratings of r = .43 to .53 on the gifted sample.

Authors' Recommendations for Using *SIGS*

Ryser and McConnell (2004) do not suggest summing the scores on the scales. Norms are provided for the seven scales only and not for the composite score. The authors explain that all scale ratings do not necessarily need to be completed on students. For example, if a school has a program for students gifted in mathematics and science, perhaps only the mathematics and science scales should be used.

Ryser and McConnell included a Summary Form along with the scales and technical manual in the kit (2004). They recommend that a screening/identification committee use this form when selecting the students who will be identified for services. The Summary Form includes an area for recording the *School Rating Scale* and *Home Rating Scale* results as well as areas for recording additional information about a child being considered.

Gifted Rating Scales

The *Gifted Rating Scales* (*GRS*) were developed to help teachers to "assess observable student behaviors indicating giftedness" (Pfeiffer & Jarosewich, 2003, p. 1). The *GRS-School Form* contains six scales based on areas mentioned in the 1972 and 1978 federal definition of giftedness: intellectual, academic, creativity, artistic, leadership, and motivation. The authors' rationale for using these areas is based on the assumption that most states or districts use the 1978 federal definition or parts of it. In addition to developing a *GRS-School Form (GRS-S)*, the authors developed a *Preschool/Kindergarten Form (GRS-P).* The two versions are similar in format, but only 29% of the items overlap, and the leadership scale is not included on the *GRS-P*. Sample items on the *GRS-S* are "Thinks insightfully, intuitively understands problems" (intellectual ability scale); "Completes academic work correctly" (academic ability scale); and "Displays an active imagination, thinks or acts imaginatively" (creative scale). The *GRS-S* is designed for children in Grades 1–8, ages 6.0–13.11. The authors state that the *GRS-P* "identifies giftedness in children between the ages of 4.0–6.11." When rating 6-year-olds, teachers should use the *GRS-P* if the children are in kindergarten and use the *SRS-S* if the children are in Grade 1. The *GRS-P* contains items such as, "Learns difficult concepts easily" (intellectual ability scale), "Completes activities correctly" (academic ability scale), and "Engages in elaborate imaginative play" (creativity scale).

Both the *GRS-S* and *GRS-P* contain 12 items per scale and instruct teachers to rate characteristics along a range of 9 points (Pfeiffer & Jarosewich, 2003). When doing the ratings, teachers are directed to first consider whether the students' characteristics are below average, average, or above average, and then select one of the three points within that category. Ratings of 1, 2, and 3 are in the below average category; ratings of 4, 5, and 6 are categorized as being average; and ratings of 7, 8, and 9 are categorized as being above average. Both Korean and Chinese versions of the GRS have been developed and researched (Lee & Pfeiffer, 2006; Li, Pfeiffer, Petscher, Kumtepe, & Mo, 2008).

Technical Support for the *GRS*

Pfeiffer and Jarosewich (2003) used various procedures to support the validity inferences on the *GRS,* beginning with expert ratings on the items (content validity evidence). Convergent and discriminant validity were examined by correlating responses on all *GRS* scale scores (intellectual ability, academic ability, creativity, artistic talent, motivation, and leadership scales) with measures of intelligence (Wechsler tests), achievement (Wechsler tests), creativity (*Torrance Tests of Creative Thinking*), artistic talent (*SRBCSS* Artistic and Creativity scales, *Expert Art Panel* ratings), motivation (*Academic Competence Evaluation Scales* and *SRBCSS* Motivation scale), and leadership (*SRBCSS* Leadership scale and number of students' leadership activities). These analyses were conducted with subsets of the standardization sample and resulted in a plethora of correlations presented in 11 tables in the technical manual (Pfeiffer & Jarosewich, 2003). The results of the analyses of the various GRS scales with measures of intelligence generally demonstrated low to moderate correlations. The five *GRS-P* scales were correlated with the *Wechsler Preschool Primary Intelligence Scale-III (WPPSI-III)* subtest and composite scores, resulting in correlations generally in the moderate range (r = .40s). The six *GRS-S* scale scores were correlated with the *Wechsler Intelligence Scale for Children-IV (WISC-IV)* subtest scores, index scores, and full scale score, resulting in correlations in the low to moderate range (r = .30s and .40s).

In addition to looking at the relationship with measures of intelligence, the *GRS* scales were correlated with an achievement measure, the *Wechsler Individual Achievement Test-II* (*WIAT-II*) subtests and composite scores. The *GRS-P* academic ability and motivation scales correlated most strongly with the *WIAT-II* subtests, with correlations in the low to moderate range (r = .30s and .40s). The *GRS-S* scales correlated more strongly than the *GRS-P* scales with the *WIAT-II* subtests, resulting in correlations in the moderate range (r = .50s), with the strongest correlations between the *GRS-S* intellectual and academic scales and the *WIAT-II* subtests and composite scores.

To examine the predictive validity of *GRS* with creativity, the authors examined the correlations between *GRS* scales with both the *Torrance Test of Creative Thinking (TTCT), Figural Form B* and the *SRBCSS* creativity scale. Interestingly, all five *GRS-P* scales correlated most highly with the *SRBCSS* creativity scale, with r = .76–.88. The same was found for the *GRS-S,* with all six scales correlating more highly with the *SRBCSS* creativity scale, r = .67 on the GRS-S artistic scale and r = .86 on both the *GRS-S* academic and creativity scales. Correlations between the *GRS* with the *TTCT-Figural* were all very low, r = .10s.

To examine the relationship between the *GRS* with measures of artistic talent, correlations were performed between all *GRS* and ratings of students' art samples as well as the *SRBCSS* artistic scale. The results indicated the highest correlations between the five *GRS-P* scales and the *SRBCSS* artistic scale scores, r = .77–.91. Correlations on the six *GRS-S* scales with the *SRBCSS* artistic scale ranged from r = .39 (GRS-S academic scale) to r = .86 (*GRS-S* artistic scale).

The authors also examined the relationship between the *GRS* with measures of motivation, namely, the *Academic Competence Evaluation Scale (ACES)* motivation scale and the *SRBCSS* motivation scale. Similar results were obtained for both the *GRS-P* and *GRS-S* with high correlations (r = .70s and .80s) found on both measures of motivation. The strongest correlations were between the *GRS* motivation scale and the *SRBCSS* motivation scale (r =.90 on both).

The relationship between the *GRS-S* scales and measures of leadership was examined by correlating *GRS* scales with the number of students' leadership activities and teachers' ratings on the *SRBCSS* leadership scale. As with the correlations on creativity and motivation, the strongest correlations were found between the *GRS-S* scales and the *SRBCSS* leadership scale, r = .62–.90.

Pfeiffer and Jarosewich (2003) concluded that these correlation analyses demonstrated convergent and divergent validity evidence for the *GRS* scale scores, illustrating *convergent* validity when, for example, the *GRS-S* creativity scale correlated highly with the *SRBCSS* creativity scale (r = .86) and illustrating *divergent* validity when the *GRS-S* artistic scale correlated somewhat lower with the *SRBCSS* creativity scale (r =.67). This concept would have been better supported if the correlations between the other *GRS* scales and the *SRBCSS* creativity scale had been much lower. The correlations of the five or six *GRS* scales with external measures of intelligence, achievement, motivation, and leadership demonstrated overall evidence for convergent validity and, in some case, for divergent validity, most notably between the *GRS* leadership scale and the intelligence and achievement scores.

In addition to providing support for the validity of the *GRS* scales, Pfeiffer and Jarosewich (2003) conducted procedures to provide evidence for the reliability of the scales. The alpha reliability coefficients on the *GRS-P* scales for the standardization sample were all r = .98 or .99. As with the *GRS-P*, the alpha reliability coefficients on the *GRS-S* scales were also very high, r = .97–.99. Test-retest reliability was also conducted on the *GRS-P and GRS-S* using a subsample of 124 students and 154 students, respectively. Using an average retesting interval of 18 days on the *GRS-P* scales, the test-retest reliability estimates ranged from r = .91 to r = .95 for the entire *GRS-P* subsample. Using a median retesting interval of 7 days on the *GRS-S* scales, the reliability estimates ranged from r = .83 to r = .90 for the entire subsample. Thus, the test-retest reliability estimates were high.

Inter-rater reliability on the *GRS-P* and *GRS-S* scale ratings was also examined by having two teachers/raters complete the *GRS-P* ratings on 56 students and *GRS-S* ratings on 147 students. The intraclass correlation coefficients on the *GRS-P* ranged from r = .62 on the artistic scale to r = .80 on the intellectual ability scale, and on the GRS-S, they ranged from r = .68 on the artistic scale and r = .77 on the academic ability scale. Therefore, these coefficients indicate adequate consistency across different teachers' ratings of the same students.

Pfeiffer and Jarosewich (2003) established national norms using data from the standardization samples. Specific details as to how the standardization samples were recruited and selected are not described in the technical manual, but the authors report that both student samples were stratified to match the US census by ethnicity (White, African American, Hispanic, Asian, and Other) and by parent education level. A total of 90 teachers participated in the *GRS-P* standardization, and a total of 382 teachers participated in the *GRS-S* standardization. The *GRS-S* student sample was stratified within eight 12-month age bands from 6.0 to 13.11 years.

To obtain national norms on the *GRS*, scale raw score totals are converted into a *T* score (which has a mean of 50 and standard deviation of 10) and into cumulative percentages for the *T* scores. The technical manual (Pfeiffer & Jarosewich, 2003) contains conversion tables for determining the *T* scores and cumulative percentages for each age level on the appropriate *GRS* scale. Complete details used to establish the standard scores (*T* scores) are not provided in the technical manual, but the authors state that norms were based on the performance of the students in the standardization samples (n = 375 on the *GRS-P* sample and n = 600 on the *GRS-S* sample.) The authors classify *T* scores of 70 and above as having a "very high probability" of gifted classification, scores of 60–69 as a "high probability" of gifted classification, scores of 55–59 as a "moderate probability of gifted classification, and below 55 as a "low probability" of gifted classification.

Authors' Recommendations for Using GRS

Pfeiffer and Jarosewich (2003) provide a few guidelines for using the *GRS* in screening students for gifted programs. They recommend that the teacher/rater complete the entire instrument in a single session to ensure consistency when completing the ratings. The authors believe ratings on the 60 items on *GRS-P* can be completed in 10 minutes or less, and ratings on the 72 items on the *GRS-S* can be completed in 15 minutes or less. When asking teachers to complete the ratings, the raters should be instructed to complete their ratings by comparing the child being rated with "typical" students of the same age in a regular classroom setting. When collecting the completed ratings from teachers, the authors suggest the scales be returned to teachers if more than one item is missing from a scale. If a scale is missing two or more ratings, the *T* score and cumulative percentage should not be calculated. If one item is missing, the average of all items on that scale should be inserted for the missing item before totaling the scores on a scale. The authors also note in the technical manual that consumers might want to develop local norms rather than use the national norms provided. They acknowledge that "local norms take into account the unique characteristics of the school district and its community" (p. 20). And, finally, Pfeiffer and Jarosewich want consumers to realize the *GRS* is designed to be an initial screening instrument, and decisions about placement of students in gifted programs should be based on a comprehensive selection process.

Conclusions About Using Teacher Judgment Measures

As described above, the three instruments—*Scales for Rating the Behavioral Characteristics of Superior Students*, *Scales for Identifying Gifted Students*, and *Gifted Rating Scales*—all have empirical support for their use. In addition to reviewing the technical support for instruments, how do school personnel make a decision for which instrument to use? The best advice is to consider, first of all, the needs of their gifted learners and the definition of giftedness being used to develop program services, and then to develop screening and identification procedures and instruments aligned with the definition. If a district is providing advanced classes in language arts and mathematics to its gifted learners, then certain types of teacher rating instruments will be better suited for identifying talent in those areas. In other words, we don't identify students until we know what services we are identifying students for.

When decisions have been made as to the sources and types of information to be considered in the screening procedure, school personnel should be reminded that modifying teacher judgment instruments is not permissible. Removing or adding some items to a teacher rating scale changes the technical support for the instrument. It is analogous to saying that when buying new tires for a car, "Oh, the tires are so expensive, I will just buy three new tires and get along with just three new ones." The vehicle (or program) may suffer greatly because of the change in the support.

Something else that consumers might consider when using teacher judgment measures is the use of local norms. Many scholars and researchers now recommend that contextual assessment and local norms be used when making interpretations from instruments to assist when identifying students for gifted services (e.g., Lohman, 2009b); Lohman & Renzulli, 2007; Peters & Gentry, 2011; Sternberg, 1998). In fact, the National Association for Gifted Children (2010) includes a statement about using local norms in the program standards. Within the standards we find, "Evidenced-based Practice 2.3.1: Educators select and use non-biased and equitable approaches for identifying students with gifts and talents, which may include using locally developed norms or assessment tools in the child's native language or in nonverbal formats." Lohman argues convincingly that "the need for special services depends not so much on a student's standing relative to age or grade mates nationally, but on the student's standing relative to the other students in the class" (2009b, p. 49; see also Chapter 12 of this text). It is the students at the top, regardless of the reference group, whose needs are most likely not to be met in a regular classroom. Lohman and Lankin (2007) explain, "Local score distributions generally provide a better way to determine which students are most likely to be mismatched with the instruction they are receiving than will national norms" (p. 16). Lohman (2009a) also proposes that using local norms is the best way of being more inclusive when selecting students who have had fewer opportunities to learn. It remains to be seen if more developers of teacher judgment measures begin to advocate for greater use of local norms.

In addition to using teacher judgment instruments with a clear purpose, technical support, and local norms, developers of teacher judgment measures all recommend that consumers do not sum scores across scales. The individual scales were developed to assess different traits, characteristics, and domains, and summing the scores across scales in not advised because information about a student's unique strengths would be lost.

Some research suggests that teacher training is very important before asking teachers to complete teacher-rating forms. Hunsaker, Finley, and Frank (1997), in an investigation of teacher nominations and student performance in gifted programs, concluded from their investigation that helping teachers focus on particular manifestations of traits in specific cultural or socioeconomic settings would improve the predictive validity of the ratings. Gear (1978) found that trained teachers, versus untrained teachers, nominate more students. Johnson (2004) recommends that professional development training on the characteristics of gifted and talented students be employed whenever teachers are involved in the nomination process.

Just as using a single test score is not recommended when identifying students for gifted services, using just a teacher rating scale is not advisable either. Toua, the child described at the beginning of this article, scored at the 82nd percentile using local norms on a standardized test in his school district. Because his score wasn't at the highest levels, the district screening and identification committee spent more time examining other sources of information about him. When examining these other data, the committee members noted the *SRBCSS* ratings provided by Toua's classroom teacher. She rated him very highly on the creativity scale and motivation scale and submitted examples of his classroom work for consideration. After a comprehensive look at several sources of information, including the fact that Toua was just learning English, the committee determined that Toua should be selected for gifted services. This illustrates how important it is to have teachers' input when making decisions about the selection of students for gifted services.

Resources

Three Teacher Rating Instruments Discussed in the Chapter

Pfeiffer, S. I., & Jarosewich, T. (2003). *GRS: Gifted Rating Scales* [published instrument]. San Antonio, TX: Pearson. Available from http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8130-502&Mode=summary

Renzulli, J. S., Smith, L. H., White, A. J., Callahan, C. M. Hartman, R. K., & Westberg, K. W., Gavin, M. K., Reis, S. M., Siegle, D., & Systma Reed, R. E. (2010). *Scales for Rating the Behavioral Characteristics of Superior Students* [published instrument]. Mansfield Center, CT: Creative Learning Press, Inc. Available from http://www.creativelearningpress.com/scalesforratingthebehavioralcharacteristicsofsuperiorstudents--50scales.aspx

Ryser, G. R., & McConnell, K. (2004). *SIGS-complete kit: Scales for Identifying Gifted Students* [published instrument]. Waco, TX: Prufrock Press. Available from http://www.prufrock.com/productdetails.cfm?PC=212

References

Birch, J. W. (1984). Is any identification procedure necessary? *Gifted Child Quarterly, 28*, 157–161.

Borland, J. H. (1978). Teacher identification of the gifted: A new look. *Journal for the Education of the Gifted, 2,* 22–32.

Borland, J. H. (2008). Identification. In J. A. Plucker & C. M. Callahan (Eds.), *Critical issues and practices in gifted education* (pp. 261–280*)*. Waco, TX: Prufrock Press.

Gagné, F. (1994). Are teachers really poor talent detectors? Comments on Pegnato and Birch's (1959) study of the effectiveness and efficiency of various identification techniques. *Gifted Child Quarterly, 38,* 124-126.

Gear, G. (1978). Effects of training on teachers' accuracy in identifying gifted students. *Gifted Child Quarterly, 22*, 90–97.

Hunsaker, S. L., Finley, V. S., & Frank, E. L. (1997). An analysis of teacher nominations and student performance in gifted programs. *Gifted Child Quarterly, 41*, 19–23.

Johnson, S. K. (Ed.). (2004). *Identifying gifted students. A practical guide*. Waco, TX: Prufrock Press.

Kalatan, A. R. (1991). *The effects of inservice training on Bahrani teachers' perceptions of giftedness.* Unpublished doctoral dissertation. University of Connecticut.

Lee, D., & Pfeiffer, S. I. (2006). The reliability and validity of a Korean-translated version of the *Gifted Rating Scales*. *Journal of Psychoeducational Assessment, 24*, 210–224.

Li, H., Pfeiffer, S. I., Petscher, Y., Kumtepe, A. T., & Mo, G. (2008). Validation of the *Gifted Rating Scales—School Form* in China. *Gifted Child Quarterly, 52*, 160–169.

Lohman, D. F. (2009a).The contextual assessment of talent. In MacFarlane, B. & Stambaugh, T. (Eds.). *Leading Change in Gifted Education: The Festschrift of Dr. Joyce VanTas-*

sel-Baska (pp. 229–242). Waco, TX: Prufrock Press.

Lohman, D. F. (2009b). Identifying academically talented students: Some general principles, two specific procedures. In L. Shavinina (Ed.), *Handbook of giftedness* (pp. 971–998). Amsterdam: Elsevier.

Lohman, D. L., & Lakin, J. (2007). Nonverbal test scores as one component of an identification system: Integrating ability, achievement, and teacher ratings. In J. Van Tassel Baska (Ed.), *Alternative assessments for identifying gifted and talented students* (pp. 41–66). Waco, TX: Prufrock Press.

Lohman, D. F. & Renzulli, J. (2007). *A simple procedure for combining ability test scores, achievement test scores, and teacher ratings to identify academically talented children.* Unpublished paper. Retrieved from http://faculty.education.uiowa.edu/dlohman/

National Association for Gifted Children. (2009). *States of the states in gifted education report: National policy and practice data* [CD Rom]. Washington, DC: Author.

National Association for Gifted Children. (2010). *2010 pre-k–grade 12 gifted programming standards.* Washington, DC: Author. Retrieved August 8, 2011, from http://www.nagc.org/index.aspx?id=6500

Nazar, F. A. (1988). Teachers' and parents' perceptions of the behavioral characteristics of third-grade gifted students in Kuwait. Unpublished doctoral dissertation, University of Miami.

Pegnato, C. W., & Birch, J. W. (1959). Locating gifted children in junior high schools–A comparison of methods. *Exceptional Children, 25*, 300-304.

Peters, S. J., & Gentry, M. (2011, March). *Group-specific norms and teacher rating scales: Implications for underrepresentation.* Paper presented at the American Education Research Association Annual Conference, New Orleans, LA.

Pfeiffer, S. I., & Jarosewich, T. (2003). *GRS: Gifted Rating Scales* manual. San Antonio, TX: Pearson.

Renzulli, J. S. (1978). What makes giftedness. Reexamining a definition. Kappan, *60*(3), 180–184.

Renzulli, J. S., & Delcourt, M. A. B. (1986). The legacy and logic of research on the identification of gifted persons. *Gifted Child Quarterly, 30*, 20–23.

Renzulli, J. S., & Reis, S. M. (1985). *The schoolwide enrichment model: A comprehensive plan for educational excellence*. Mansfield Center, CT: Creative Learning Press.

Renzulli, J. S., Siegle, D., Reis, S. M., Gavin, K. M., & Systma Reed, R. E., 2009). An investigation of the reliability and factor structure of four new *Scales for Rating the Behavioral Characteristics of Superior Students*. *Journal for Advanced Academics, 21*, 84-108.

Renzulli, J. S., Smith, L. H., White, A. J., Callahan, C. M., & Hartman, R. K. (1976). *Scales for Rating the Behavioral Characteristics of Superior Students*. Mansfield Center, CT: Creative Learning Press.

Renzulli, J. S., Smith, L. H., White, A. J., Callahan, C. M. Hartman, R. K., & Westberg, K. W. (2002). *Scales for Rating the Behavioral*

Characteristics of Superior Students. Revised edition. Mansfield Center, CT: Creative Learning Press, Inc.

Renzulli, J. S., Smith, L. H., White, A. J., Callahan, C. M. Hartman, R. K., & Westberg, K. W., Gavin, M. K., Reis, S. M., Siegle, D., & Systma Reed, R. E. (2010). *Scales for Rating the Behavioral Characteristics of Superior Student: Technical and administration manual* (3rd ed.). Mansfield Center, CT: Creative Learning Press, Inc.

Renzulli, J. S., & Westberg, K. L. (1991). *Rating Student Performance in a Gifted Program.* Unpublished instrument. Storrs, CT: The National Research Center on the Gifted and Talented.

Ryser, G. R., & McConnell, K. (2004). *SIGS-complete kit: Scales for Identifying Gifted Students.* Waco, TX: Prufrock Press.

Shore, B. M., Cornell, D. G., Robinson, A., & Ward, V. S. (1991). *Recommended practices in education: A critical analysis.* NY: Teachers College Press.

Subhi, T. (1997). Who is gifted? A computerized identification procedure. *High Ability Students, 8*(2), 189–211.

Srour, N. H. (1989). *An analysis of teacher judgment in the identification of gifted Jordanian students.* Unpublished doctoral dissertation. University of Connecticut.

Sternberg, R. J. (1998). Applying the triarchic theory of human intelligence in the classroom. In R. J. Sternberg & W. M. Williams (Eds.), *Intelligence, instruction and assessment: Theory into practice*. Mahwah, NJ: Erlbaum.

Sternberg, R. J., & Davidson, J. E. *(2005). Conceptions* of giftedness (2nd ed.). NY: Cambridge University Press.

Terman, L. M. (1925). *Mental and physical traits of a thousand gifted children: Genetic studies of genius* (Vol. 1). Stanford, CA: Stanford University Press.

Torrance, E. P. (1962). *Guiding creative behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Chapter 14 Study Guide

Prompt 1 ***Knowledge***

Prepare a chart on which you summarize the strengths and weaknesses of the three teacher rating scales reviewed in this chapter.

Prompt 2 ***Opinion***

The *criterion problem* suggests that it is not appropriate to evaluate the validity of teacher judgments about student giftedness against an IQ score. What, in your opinion, would be an appropriate criterion?

Prompt 3 ***Affect***

Describe the pressures you feel or would feel if asked to complete a teacher rating scale on students in your class. What could be done to alleviate those pressures?

Prompt 4 ***Experience***

Describe any experience you or a colleague has had in creating a teacher rating scale for gifted identification or in using a locally created scale. Why, according to the author, is this a problem? Were these problems apparent with your local instrument? What should a local educational agency do to verify the validity and reliability of any locally produced scale?

Prompt 5 ***Preconception/Misconception***

Some critics feel that introducing teacher judgment into gifted identification injects additional biases into the system; others believe that teacher judgment is one solution to overcoming the bias inherent in testing. Where do you stand on this issue and why?