Using the DAS Nonverbal Scales with Spanish-Speaking Children: Translation and Validation


Consultation in New Teacher Groups: School Psychologists Facilitating Collaboration Among New Teachers

Exploring the Association Between Grade Retention and Dropout: A Longitudinal Study Examining Socio-Emotional, Behavioral, and Achievement Characteristics of Retained Students

Best Practices in Assessing Kindergarten Readiness

Using a Formative Program Portfolio to Enhance Graduate School Psychology Programs
Best Practices in Assessing Kindergarten Readiness

Renee Pavelski Pyle
University of California, Santa Barbara

This article addresses the numerous challenges that have developed in response to the first national education goal that states, “All children in America will start school ready to learn.” The definition and measurement of school readiness has been surrounded by controversy for decades. There is little consensus on how it should be defined and even greater uncertainty on how it should be measured. This paper addresses the issues of assessing readiness by reviewing the history, research, and methodological challenges relating to screening instruments for school readiness and then links these data to a best practices model of screening techniques.

Keywords: Preschool, Kindergarten, Assessment, Readiness, Screening Instruments

In the fall of 1989, the President and governors of the 50 states met for the first Education Summit held in nearly 100 years (Meisels, 1998). This meeting established eight “National Education Goals.” First among these goals was the following: “All children in America will start school ready to learn” (National Education Goals Panel, 1991). Clearly, this is a lofty and needed goal. However, beyond broad agreement about the importance of readying children (Bracken & Walker, 1997; Gredler, 1992; Lehr, Ysseldyke, & Thurlow, 1987), the readiness waters are murky. There is little consensus on precisely what constitutes this term (Scarpiti & Silver, 1999) and even greater confusion on how it should be measured.

This paper addresses the definitional and measurement challenges that plague readiness assessments by reviewing the history, research, and methodological issues relating to screening instruments for school readiness. This research is then linked to a best practices model of screening techniques.

HISTORY OF PRESCHOOL ASSESSMENT

The history of early childhood assessment is not lengthy. Kelley and Surbeck (1991), in an extensive review of this history, identified two periods of high productivity. The first period includes the contribution of Arnold Gesell who created a “developmental schedule” that contained approximately 150 items in four areas: motor development, language development, adaptive behavior, and personal-social behavior. His work, which spanned almost 40 years, influenced and is still influencing the construction of tests for preschool children (Gredler, 1992).

The second period, during the 1960s, was a time of high productivity for early childhood assessment (Kelley & Surbeck, 1991). The funding of the 1964 Child Health and Mental Retardation Act and of the Head Start and Follow Through Programs created a demand for preschool tests for diagnosis, monitoring, and program evaluation (Zigler, 1998). Assessments were designed to measure the variety of domains included in Head Start instruction (e.g., affective, intellectual, psychomotor, and subject achievement; Zigler & Styfco, 1994).

Over both periods, it is evident that the domains measured by these early assessments were child-centered and paid little attention to contextual variables. As discussed throughout this paper, the implication of this “within child” model is that the burden is placed on the child to be ready for school while
largely ignoring the influences of family, community, school, and culture on the child’s performance. As outlined in the following section, it is evident that this child-focused model continues to be perpetuated in current assessment practices. In addition, the psychometric properties of many of these instruments are alarmingly inadequate.

CURRENT RESEARCH ON EARLY CHILDHOOD ASSESSMENT INSTRUMENTS

Heterogeneity

The research on current early childhood assessments illustrates further the myriad of problems surrounding these instruments (Bracken, 1987; Lehr, et al., 1987; Thurlow & Ysseldyke, 1979). Two surveys, in New York and Michigan (as reported in Bagnato & Neisworth, 1994), provide insight into the range of instruments used for screening purposes. In the New York study, 177 school districts were polled to identify the instruments currently being used in their pre-kindergarten screening procedures. It was found that 151 separate tests or assessment instruments were reported in use. In the Michigan study, 111 different tests were in use in the state public schools. These studies suggest that the use of locally developed instruments may account for a large number of the measures found in these surveys, and that there is little consensus on how kindergarten readiness should be defined or measured (Bagnato & Neisworth, 1994).

Technical inadequacy

Many researchers report that few individual screening measures have the technical adequacy necessary for early identification (Lehr, et al., 1987; Mercer, Algozzine, & Trifiletti, 1979; Thurlow & Ysseldyke, 1979). In 1979, Thurlow and Ysseldyke evaluated the validity, reliability, and norms of the most frequently used tests in federally funded Child Service Demonstration Centers (CSDC) and found that only seven of the 28 tests reviewed were technically adequate in all psychometric properties. A similar analysis by Lehr, Ysseldyke, and Thurlow (1987) using the Handicapped Children’s Early Education Program demonstration projects revealed that among the 19 most commonly used devices, only three (Vineland, McCarty, and K-ABC) were technically adequate. The criteria used to define “adequate” were compiled from several sources, including the Standards for Educational and Psychological Tests, and Assessment in Special and Remedial Education (Lehr et al., 1987).

Tests of cognitive abilities. In a similar study, Bracken (1987) analyzed the 10 most commonly used early childhood instruments (five used for educational placement decisions, and five used to assess specific skills and/or abilities). The technical adequacy of these instruments was evaluated through various indexes of reliability (median subtest reliability, total test internal consistency, and total test stability coefficients), subtest and total test floors, subtest item gradients, and provision of validity information. For each of these areas, Bracken (1987) outlined minimal standards of technical adequacy. He concluded that many of these tests are severely limited in floor, item gradient, and reliability, especially below the age of 4 years.

Examining many of the same psychometric properties as Bracken (1987), Flanagan and Alfonso (1995) sought to determine whether certain technical limitations of previous instruments were improved with the publication of new or recently revised intelligence tests. These authors reviewed the following tests: Wechsler Preschool and Primary Scale of Intelligence – Revised (WPPSI-R); Differential Ability Scale (DAS); Stanford-Binet Intelligence Scale: Fourth Edition (S-B: IV); Woodcock-Johnson Psycho-Educational Battery: Tests of Cognitive Ability (WJ-R: COG), and the Bayley Scales of Infant Development – Second Edition (BSID-II).
Similar to Bracken (1987), Flanagan and Alfonso (1995) found that most of the tests showed some of the same inadequacies at the lower end of the preschool age range. Problems with test floors and item gradients, in particular, continued to be evaluated as weaknesses for children below the age of 4 years. Although test-retest reliabilities reported in the respective test manuals appeared satisfactory, the authors pointed out a number of methodological concerns about the design of these test-retest studies. These include small sample sizes as well as the use of samples that were either not representative of preschoolers, comprised of too broad an age range, and/or included children beyond preschool age.

Flanagan and Alfonso (1995) found two tests, the BSID-II and the WJ-R:COG, to be technically adequate across most criteria below the age of 4 years. Additionally, the technical qualities of these instruments appeared to be superior to those summarized by Bracken. These authors conclude that the technical qualities of the new and recently revised tests for preschoolers have shown improvement.

Tests of behavior and social-emotional functioning. The evaluative studies of Bracken (1987) and Flanagan and Alfonso (1995) were limited to tests of cognitive ability. Bracken, Keith, and Walker (1998) examined the quality of 13 commonly used or newly developed instruments designed to assess preschool behavior and social-emotional functioning. Using the same criteria as Bracken (1987), Bracken et al. (1998) found that the 13 social-emotional, third-party assessment devices had more psychometric limitations than preschool cognitive ability measures. When comparing more recently published instruments to others with older publication dates, it was found that the newer instruments were generally more technically sound. This latter finding parallels the work of Flanagan and Alfonso (1995). Therefore, despite the substantial limitations among existing early childhood instruments, there may be some optimism for improved quality assessment tools developed in the future. However, it is clear that these cognitive, behavioral, and social-emotional assessments focus on the child’s abilities rather than viewing their performance in context. Little effort has been made to assess environmental factors that prove tremendously influential in a child’s development. In addition, it is evident from the heterogeneity of assessments currently in use that a consistent definition of school readiness remains elusive.

**WHAT PSYCHOMETRIC PROPERTIES ARE MOST IMPORTANT IN SCHOOL READINESS ASSESSMENTS?**

The research reviewed above outlines numerous problems surrounding the technical adequacy of early screening instruments. When choosing a readiness or screening test, standards for professional test development must be considered [see Standards for Educational and Psychological Testing; American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999]. Regarding screening tests, the standards state clearly that no such test should be used to make any decision about an individual other than referral for further evaluation. However, most screening programs today seem to have adopted the model that they must identify those youngsters who, if they do not receive special services, are at risk for school failure (Thurlow & Gilman, 1999). Therefore, screening programs have assumed the burden not just of identifying for further assessment those who may have a disability, but also of predicting which children already have a problem that will continue as the child matures. Given this conflict, the issue of predictive validity becomes paramount in choosing a screening test. This next section will review critical methodological issues that school psychologists should address when choosing a screening instrument for young children. Particular attention will be paid to the issues of predictive validity, sensitivity, and specificity.
Predictive Validity

To establish predictive validity, two things are necessary: the number of children identified as at-risk and not at-risk by the test, and some measure of performance indicating those children who performed adequately in school and those who did not (Gredler, 1992).

Several studies have looked at the positive predictive value (percentage of children originally identified as at-risk who later developed problems) of screening measures. Gredler (1992) analyzed 12 screening measures and found an average positive predictive value of .55 (i.e., 55% of children considered to be at-risk later developed problems). In a similar study, Carran and Scott (1992) found an average positive predictive value of .65 for eight screening measures. These results indicate that screening measures are better at predicting student success than failure. However, this proves problematic because school districts “act mainly on the number of children who are classified as ‘at risk’” (Gredler, 1997, p. 102).

Very few studies have analyzed the predictive validity of preschool screening on placement decisions (Thurlow & Gilman, 1999). In one of the few studies conducted (Kochanek & Hennen, 1988), preschool screening data obtained in the spring prior to kindergarten entrance predicted special education need two years later. However, it is unknown whether the same outcome would be realized if the screening had been conducted earlier (thus allowing time for service provision). It is clear that much less attention has been paid to the efficacy of preschool screening than to the efficacy of early intervention (Lichtenstein & Ireton, 1991; Thurlow & Gilman, 1999).

Sensitivity and Specificity

The desirable outcomes of preschool screening reflect sensitivity (referring to the proportion of children who actually performed poorly who also were originally selected by the screening measure) and specificity (referring to the proportion of children who performed satisfactorily who were originally considered not at-risk by the screening test) (Lichtenstein & Ireton, 1984; Gredler, 1992, 1997). Sensitivity is a major concern for school psychologists conducting readiness assessments because it reflects the extent to which further testing will be given to only those needing it.

Stevenson, Parker, Wilkinson, Hegion, and Fish (1976) analyzed the effectiveness of a prekindergarten battery of cognitive and psychometric tasks and teacher ratings administered in preschool to predict reading and arithmetic achievement in third grade. Results suggest that the index of sensitivity is low. That is, the percentage of poor readers actually found at the end of third grade and who were originally identified by the preschool measures was less than 40%. Similarly, between 56% and 41% of the children predicted to be low achievers performed satisfactorily. When teacher rating scales were used, 85% of the poor readers were not identified (index of sensitivity). Further, only 50% of those children who were earlier identified by the teachers as being at risk subsequently became poor readers. They concluded that more effective prediction could be made from prekindergarten tasks than teacher ratings. After 1st grade, the more effective predictor was derived from scores on prior tests of achievement.

This lack of success is reinforced by the results of the analysis of 33 early identification predictive validity studies by Lichtenstein and Ireton (1984), a comprehensive study by Pianta and McCoy (1997), and in a review of 74 studies by Tramontana, Hooper and Selzer (1988). These authors all conclude that screening measures do a better job at telling us who performed well and was placed correctly than predicting those children who were members of problem groups. These results support the argument that screening measures should not be used to make high-stakes placement decisions such as restricting entry to kindergarten.
Stevenson et al. (1976) further assert that, “batteries of prekindergarten tasks can be used only cautiously in identifying children who need help” (p. 398). These authors conclude that the most reasonable use of predictive measures, such as psychometric screening batteries and teacher ratings, is to identify children who should be considered for further observation and evaluation rather than using the results for assigning children to special groups or classes.

In summary, predictive validity, sensitivity and specificity are key issues in preschool assessments and screenings. Because predictive validity requires longitudinal tracking of children or large retrospective data analyses over time, it has often been ignored. Until these methodological issues are overcome, it remains crucial for school psychologists to avoid the use of these measures for labeling and placement decisions. What follows is a summary of additional methodological issues that should be reviewed when deciding on readiness assessments.

**Reliability, Test Floors, and Item Gradients**

Utilizing the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), it is evident that reliability, test floors, and item gradients are additional key methodological issues that need to be addressed for educational and psychological testing. These issues will be reviewed and addressed as they relate to readiness screening.

Reliability of a test refers to the degree to which a child’s score is consistent (internal consistency) and stable (test-retest reliability) across time (Jacob-Timm & Hartshorne, 1998). Of particular interest to school psychologists is the issue of stability or the degree to which a preschooler’s cognitive test scores are likely to be similar from one measurement to the next. Tests with low reliability produce proportionately large portions of subtest and composite variability that are due to measurement error rather than true differences in the construct (Bracken, 2000). Adequate internal consistency for subtest and total test scores allows the school psychologist to assume that the items that comprise the test are highly related and measure a similar domain of behavior. During the assessment process, this permits a more concise and clear interpretation of test scores (Flanagan & Alfonso, 1995).

Another dimension of technical adequacy involves test floors. The floor of a test is an indication of the extent to which an instrument provides meaningful scores at very low levels of individual functioning. In instances when a poor floor exists, scores may become inappropriately inflated and, consequently, provide misleading information (Bracken & Walker, 1997). This potential shortcoming is particularly salient for preschool and kindergarten children because many assessment cases have the goal of determining developmental delay based on a significant discrepancy between the referred child’s performance and that of same-age peers. Ceilings are not generally as relevant among early childhood tests as are test floors because screening measures are used to predict failure rather than degrees of success. It is easier to develop suitable items for assessing the upper limits of young children’s abilities than it is to create items that discriminate between the lower limits of ability at this age (Bracken, 1987).

Item gradients are an additional technical quality that is crucial in early childhood assessment. An item gradient refers to “how rapidly standard scores increase as a function of a child’s success or failure on a single test item” (Bracken, 1987, p. 322). Ideally, the incremental change in standard scores that results from one raw score unit to another should produce a comparable small standard score increase. Unfortunately, early childhood tests are notorious for having steep item gradients, with correspondingly large standard score changes associated with minor increases or decreases in raw scores (Bracken, 1987). Bracken and Walker (1997) note that an acceptable item gradient requires a sufficient number of nonredundant test items placed throughout the test.
BEST PRACTICES MODEL

The national goal of having all children ready for school, the effectiveness of early intervention and prevention programs, and legal mandates requiring services to preschool children with disabilities have forced school psychologists to examine their assessment practices as they relate to the accuracy of identification and the utility of assessment findings for treatment planning and evaluation. However, research indicates numerous methodological problems surrounding preschool assessments. This paper suggests that such historical problems of child-focused assessments and methodological inadequacies continue to be perpetuated in current practice. In devising a best practices model for early screening projects, four suggestions are offered: (a) defining the intended purpose of the assessment, (b) selecting ecologically focused instruments with multiple raters and follow-up procedures, (c) determining the process for conducting the assessment, and (d) thinking carefully about how to analyze, interpret, and use the results.

Define the Intended Purpose of the Assessment

A recent Child Trends Research Brief, “School Readiness: Helping Communities Get Children Ready for School and Schools Ready for Children,” summarizes recommendations from the National Education Goals Panel (NEGP) related to assessing school readiness. These recommendations stress that assessments should be used only for their intended purposes. For example, assessments designed to track achievement at the school district level need to differ from the tests used to identify learning problems in a particular child. Screening should be viewed as a first step in assessing the needs of a child, not the first step in labeling them for school failure (Pianta & McCoy, 1997; Rafoth, 1997; Shepard, 1997; Tramontana, Hooper, & Selzer, 1988).

Select the Instrument

Recent scholars are advocating the use of an ecological model to guide the assessment of kindergarten children (Vazquez-Nuttall, Nuttall, & Hampel, 1999). This model stems from research asserting that family background factors have been found to be very useful in predicting school achievement. Particularly for lower socioeconomic groups, such factors as parent education for mothers and family income for fathers have been found to influence home environment, cognitive development, and school readiness (O’Brien, 1996).

Ecologically focused. Vazquez-Nuttall and Nuttall (1999) propose an ecologically based assessment approach that has been greatly influenced by the work of Bronfenbrenner (1976; see Figure 1). As evident from this figure, the innermost circle, which addresses the individual child, has been the primary focus of screening procedures. This represents only one-fifth of the model. It is clear from this “ecomap” that current assessment procedures have unfairly weighted these child-focused domains while largely ignoring family, agency, and cultural influences on a child’s performance.

This ecological model offers a comprehensive guide to the assessment and intervention of children. It includes different settings, factors, agencies, and people that need to be considered when designing a reliable, valid, and useful evaluation and service plan (Nuttall, Nuttall-Vazquez, & Hampel, 1999). This type of model recognizes the instability of the very traits most instruments seek to measure as a result of developmental bursts and inconsistencies that defy normative charts. As a result, less emphasis is placed on inappropriate and methodologically unsound assessments and more attention is paid to what parents, teachers and researchers are telling us about these children.
Determine the Process for Conducting the Assessment

The process of conducting the assessment is as important as the assessment itself. Methodological problems that plague these assessments indicate clearly that results may not be reliable. Further, in keeping with the emerging ecological definitions of readiness, assessments should include multiple sources of information over multiple settings. Using single measures restricts the utility of screening programs by failing to recognize the degree to which aspects of school functioning can be interrelated and differentiated (Pianta & McCoy, 1997).

*Multiple raters.* Meisels (1993) has developed a practical assessment utilizing many of these best practices suggestions. Termed “The Work Sampling System,” (Meisels, 1992) this approach hinges on teacher observation, uses checklists to increase the reliability of observations, and gathers samples of

---

*Figure 1.*

The ecomap of child and family functioning. Adapted from Nuttall, Nuttall-Vazquez, and Hampel (1999).
children’s work to gauge performance. This system is a performance assessment system that offers an alternative to standardized tests with young children. Six checklists cover personal/social development, language and literacy, mathematical thinking, scientific thinking, social studies, art and music, and physical development. This system provides information in a contextualized format quite different from the setting of large screenings where children typically move from station to station in a large auditorium. Therefore, results are more valid and predictive of future performance (Meisels, 1995). Parents are also part of the assessment process and contribute developmental information and their perceptions of their child’s progress. Meisels has found the information gathered through the Work Sampling System can be aggregated and analyzed and shows strong reliability and predictive validity (Meisels, 1993).

_Multiple gates._ As part of the best practices model, it is essential that children judged as “at-risk” as a result of such a screening receive appropriate follow-up evaluation. This is especially true when screening has typically involved the administration of only one test that is not part of a more ecologically based assessment. Good screening programs should be structured in a way that specific follow-up evaluations (which may include more in-depth parent interviews, teacher observations, etc.) are mandated at specific points or “gates” in the process (Rafoth, 1997; Walker et al., 1988). As described by Walker et al. (1988), multiple gating is a procedure that contains a series of three progressively more precise assessments, or “gates,” that (a) provide for the sequential assessment and cross-validation of multimethod forms of child assessment, and (b) establish a decision-making structure for the aggregation of information produced by different assessment sources. The procedure relies on teacher judgment of pupil behavior in the first two assessment stages. In stage three, observational data are recorded through direct observation and free play settings by a school professional other than the teacher (school psychologist, counselor, resource teacher, etc.). Play-based assessment procedures are receiving increasing attention as a critical part of this best practices model (for a review, see Athanasiou, 2000). These types of ecologically focused, multi-rated and follow-up procedures tackle some of the capacity challenges that exist with district wide screenings.

**Think Carefully about how to Analyze, Interpret and Use the Results**

Along with considering ecological factors at numerous points, school readiness measures should be used to drive planning, not placement decisions. Due to the bursts and spurts in development, results must be considered flexibly along with multiple sources of information about the child. Such techniques will help to reduce errors that can result from unstandardized instruments or brief encounters with the student.

In summary, the following checklist can be used by school psychologists to develop a best practices approach to school readiness assessment:

- _Technical adequacy_ – review the predictive validity, sensitivity, specificity, reliability, test floors, and item gradients before choosing an assessment.
- _Multiple sources_ – include information on the child, family attributes, strengths, and context.
- _Multiple raters_ – gather information from family members, teachers, and other significant individuals in the child’s life.
- _Multiple gates_ – due to significant variations in child developmental pathways as well as ongoing changes in family status, children should be screened on multiple occasions over time.
- _Families as partners_ – include families not only in the information gathering stage, but also as key members of the decision-making team.
Cultural sensitivity – engender a holistic perspective and attempt to understand disability, resiliency, and vulnerability in the context of broader socioeconomic, religious, and cultural systems.

Coordination – help families link up with additional community-based services and programs. In this sense, readiness assessment should be viewed as the initial step in service provision.

Multivariate decision making – no single condition, risk factor, or protective factor leads irrevocably to a predictable outcome. Therefore, screening models must allow for these multiple sources of evidence to assume different weights in decision making over time.

CONCLUSION

In conclusion, what has emerged from this research isn’t a clear picture of model instruments but rather a model process that begins with the end in mind and incorporates contextual variables at different points. As kindergarten screening has built up a history, it has become increasingly obvious that environmental factors are missing from the picture. By defining the intended purpose of the assessment, selecting ecologically focused instruments with multiple raters and follow-up procedures, determining the process for conducting the assessment, and thinking carefully about how to analyze, interpret, and use the results, screening procedures have the best chance of identifying and appropriately serving children in need.

REFERENCES


