THE BILINGUAL RESEARCH JOURNAL
Summer/Fall 1996, Vol. 20, Nos. 3 & 4. pp. 433-463

ATTITUDES AND BEHAVIORS TOWARD TESTING-THE-LIMITS WHEN ASSESSING LEP STUDENTS: RESULTS OF A NABE-SPONSORED NATIONAL SURVEY

Virginia Gonzalez
University of Arizona

Jaime A. Castellano
Palm Beach County School District, West Palm Beach, Florida

Patricia Bauerle & Ricardo Duran
University of Arizona

Abstract

The purpose of this NABE-sponsored national survey was to describe the relationship between practitioners' and administrators' attitudes and their practices when assessing LEP students. The survey constructed included 9 multiple-choice demographic questions, and 29 affirmative/negative and Likert-Scale questions measuring attitudinal and behavioral components. Exploratory first-and-second-order factor analyses, and one-way ANOVAS with the demographic questions and the four second-order factors were used. The four factors found included attitudes toward: (1) psychometric test properties, (2) adaptation of administrative practices, (3) accommodating for cultural and linguistic differences, and (4) translations and dialectal variations. Evidence for the survey's construct validity was demonstrated because only a unidimensional first-order factor was found underlying the four second-order factors identified. Results have important practical implications for improving current practices when assessing LEP students.

Introduction

The purpose of this NABE-sponsored national survey study is to describe practitioners' and administrators' attitudes toward the technique called "testing-the-limits" and whether they use it or not when assessing limited English proficient (LEP) students. To accomplish this purpose, we stated the following two research questions: (1) What attitudes do examiners hold in relation to psychometric properties of standardized tests such as appropriateness of translations, given the existence of different dialects, and norming samples used? (2) Do examiners use the testing-the-limits technique for adapting standardized administration practices? Thus we attempt to explore the relationship between examiners' attitudes and the current practices that examiners use when assessing LEP students.

Testing-the-limits as an assessment technique for providing additional clinical information useful for placement and instructional decisions received some interest during the 1960s and 1970s. For instance, during the late 1960s, Eysenck (1969; Eysenck & Eysenck, 1969) stated that individual differences in intelligence may be significantly related to changes in test performance when a testing-the-limits approach was used. He proposed that some personality traits of examinees, such as introversion-extroversion and neuroticism, may be better measured with the testing-the-limits technique. Thus, we can expect individual differences in how children will respond and react when using testing-the-limits. For instance, the rapport built between the examiner and the child, the strategies and learning process used by the child, and in general the level of responsiveness and verbalizations made by the child will most likely be influenced by intrinsic personality factors in the child such as introversion-extroversion tendencies. In addition, Kagan (1964) proposed that another personality factor, impulsivity-reflectivity can also influence performance in both intelligence testing and in school in general.

However, somehow this interest was buried during the 1980s and 1990s as the lack of literature that we have found in our exhaustive search attests. After searching the major databases in educational and psychological research, we were surprised to find only a few articles, and most much older than 5 years, that dealt with researching the use of

testing-the-limits as a viable technique for improving the assessment of LEP students. The lack of literature on this topic attests to the fact that this movement was overshadowed by the trend of adhering tightly to standardized procedures followed by the majority of practitioners and researchers. Only few people have reinitiated the interest in testing-the-limits during the 1980s and 1990s with new possible applications for LEP and special education students. In this literature review, we will critically examine the definition of the testing-the-limits technique and we will discuss its advantages and disadvantages when administering standardized tests. Next, the contextual variables surrounding test administration, such as the quality and kind of feedback provided to the examinee, will be reviewed. Finally testing-the-limits as a form of using "scaffolding" and "dynamic assessment" will be discussed.

The Testing-the-Limits Technique

According to Sattler (1982) the standardized procedures that accompany a test should be followed as specified in the manual or test instructions, and only those modifications permitted in the test manual for the reasons stated by the authors should be made. However, as stated by Sattler (1982), "to gain additional information about a child's abilities, procedures known as testing-the-limits, may also be used following the complete standardized administration of the test as specified in the manual" (p. 169), because these procedures "can facilitate interpretation of the test results" (p. 150).

Testing-the-limits consists of re-administering portions of the test while modifying one or more of the standardized testing procedures. Although results of testing-the-limits may not be incorporated into the child's test scores, the results may be used to help determine what conditions, that were not part of the standardized test procedures, may facilitate the child's performance. According to Sattler (1982), modifications that can be useful during testing-the-limits can include any of the following five procedures: (1) providing additional clues; (2) re-administering failed items; (3) changing the modality of stimuli; (4) eliminating time limits; and (5) reconstructing the children's errors and asking them to detect and then correct their errors by providing

additional information or asking probing questions. In addition, Holtzman and Wilkinson (1991) include the following three strategies used when using the testing-the-limits technique with LEP students: (1) to substitute words or phrases in the instructions or questions to facilitate students' comprehension of what type of answers are being requested by examiners; (2) to administer additional items beyond the ceiling or cut- off point to see if students can answer correctly any of the more difficult items; and (3) to teach students how to answer certain items after the standardized testing has been completed. We consider that this third strategy can be very useful for exploring LEP students' learning potential.

We consider that testing-the-limits is a very useful and effective technique for linking assessment with instruction because valuable qualitative information can be obtained for developing individualized educational programs for LEP students. However, as pointed out by Holtzman and Wilkinson (1991), even though testing-the-limits can help to solve the methodological problems of standardized testing procedures, it is not widely used by examiners with LEP students. Thus, it is our objective in this survey study to explore and document the reasons underlying current attitudes and behaviors of administrators and practitioners toward the use of the testing-the-limits technique with LEP students.

Advantages and Disadvantages of Using Testing-the Limits

Some advantages and disadvantages have been reported by several researchers when using testing-the-limits. Sattler (1982) recommended that examiners be cautious regarding a possible disadvantage when using testing-the-limits, because providing additional clues may invalidate a future administration of the same test to the child (within six months). For instance, Sattler (1969) found that providing clues on the Picture Arrangement and Block Design of the Wechsler Intelligence Scale for Children can significantly affect retest performance of eighth and ninth graders. In relation to advantages obtained when using testing-the-limits, and as stated by Sattler (1982), providing additional clues may help an evaluator "determine how much help is necessary for the child to solve a problem" and "the more clues that are needed before success is achieved, the greater the possible degree of learning disorder or cognitive deficit" (p. 139). However, no mention of dialectal or

linguistic variables was made by Sattler (1982), even when additional clues may also help a second language learner. Thus, perhaps this practice of not including linguistic and cultural differences as possible adaptations that the examiner can make when using testing-the-limits needs to be changed. In fact, information provided below refers to research studies using the testing-the-limits technique only with regular and special education children, all from a majority background. Thus, we believe that it is necessary for evaluators of LEP children to examine how much help is necessary for them to solve a problem for improving their educational services.

Carlson and Wiedl (1979) used testing-the-limits procedures with the Ravens Matrices and found significant increases in scores of second and fourth graders with specific problem-solving strategies, including: (1) "verbalizations" made by the child of the pattern used and the reasons for choosing the strategy during and after the problem is being solved; (2) "elaborated feedback" provided by the examiner for explaining to the child which answer was correct and why; (3) "verbalization plus elaborated feedback" during and after problem solution. They also found that verbalization seemed to help second and fourth graders with higher verbal abilities, but that feedback helped better second and fourth graders with higher non-verbal abilities. In contrast, for fourth graders, they reported that the combination of verbalization and feedback helped the ones with higher verbal abilities the most. Thus Carlson and Wiedl (1979) demonstrated that "test performance can be affected by various administration techniques," and also by "the interactions between certain personality variables and the testing procedures employed" (p. 343). Within the personality variables, they included introversion-extroversion, aspects of intelligence such as verbal and nonverbal skills, and cognitive style such as impulsivity and reflectivity. Carlson and Wiedl (1979) concluded by stating that "when testing-the-limits procedures are used, the testor should be aware of the reasons or specific goals for their employment and how they interact with sources of individual variation to affect performance" (p. 343).

Bethge, Carlson, and Wiedl (1982) tested the effects of dynamic assessment procedures, such as verbalization and elaborated feedback, on the performance of third-graders on the Raven Matrices. They found that dynamic assessment modified visual search behaviors, reduced test

anxiety and negative orientation to the testing situation, and produced higher test performance. They concluded that dynamic assessment increases the examinee's motivation to succeed and more positive attitudes toward test performance. Then, modifying the situational variables when using dynamic assessment affects the examinee' 5 evaluation of the problem-solving tasks and the general atmosphere of the testing situation (called "orientation" by these authors).

In another study, Dash and Rath (1986) found significant increases in scores for the same three strategies (i.e., verbalization, elaborated feedback, and the combination of the two) when children ages 8 and 9 were tested in India with the Raven Matrices. These children were matched for initial scores prior to being randomly assigned to one of six experimental conditions. These authors explained the significant effect of verbalization as the influence of: (1) the examiner directing the attention of the child to analyze the information analytically; (2) the presence of more organization in solving the tasks as a result of self-regulation; and (3) the increase of insight and flexibility of thinking in the child, ability that could be transferred to other tasks. They explained the positive effect of elaborate feedback on the performance of the children as the result of inducing in children higher expectations of their testing performance and improving their level of knowledge. Dash and Rath (1986) stated that "(t)he application of certain approaches leads to more accurate, thus fairer assessment of intellectual functioning" (p. 87). Their objective was accomplished by the results obtained, indicating that test scores were indeed affected by variations in testing procedures.

Moreover, verbalizations used as specific problem-solving strategies within the testing-the-limits technique have also been used in second language research by Færch and Kasper (1987) and have been called "thinking aloud protocols." They describe thinking aloud protocols as an introspective method or procedure for data collection that use retrospection, self-report, self-observation, and self-revealment. This introspective data collection method is used within an exploratory-interpretative paradigm and leads to qualitative research that has as a major purpose theory construction (Færch & Kasper, 1987). When using thinking aloud protocols, participants are not called "subjects" but "informants" because their "subjective theories are of central importance for the process of theory construction" (Færch & Kasper, 1987, p. 21).

Thus, when using testing-the-limits and particularly "verbalizations," examiners can also be considered, from a research perspective, to be collecting qualitative data within an exploratory-interpretative paradigm in the form of thinking aloud protocols.

Testing-the-Limits as a Form of "Scaffolding" and "Dynamic Assessment"

The modifications used when using testing-the-limits can also be interpreted as a process called "scaffolding" by Vygotsky (1978) that occurs within the "Zone of Proximal Development" (ZPD). According to Vygostky (1978), the ZPD is "the difference between the level of problem difficulty that the child could engage in independently and the level that could be accomplished with adult help." As stated by Newman, Griffin, and Cole (1989), "Another kind of assessment called `dynamic assessment' derives from a particular interpretation of Vygostky's ZPD" (p. 77). Thus, one way of linking assessment with instruction is by using the testing-the-limits technique as a form of "dynamic assessment" in which assessment occurs while teaching the examinee on a one-to-one tutorial situation. According to Newman et al. (1989), dynamic assessment includes two aspects related to the ZPD, including the assessment of: (1) "the child's current state in relation to the zone available for acquiring the concept" and (2) "the child's 'modifiability' or readiness to learn" (p. 79).

Pascual-Leone and Ijaz (1991) described "capacity testing" as a form of dynamic assessment because the examiner adjusts the child's knowledge and abilities before or during assessment. This adaptation made by the examiner uses task modeling and interpretation of the child's performance within a developmental theory framework. As explained by Pascual-Leone and Ijaz (1991) there are different methods of capacity testing, including for instance: (1) the "train-test-train" procedure that consists of using learning as a control for testing the capacity of conceptual problem-solving of well-structured developmental tasks; (2) the "dynamic assessment via human mediation" procedure that consists of testing a child twice, the first time to determine innate capacity with no external help, and the second time to give tutorial guidance to assess learning potential; and (3) the "qualitative stage assessment" that evaluates intelligence using problem- solving mental capacity and developmental stages that are found across

knowledge domains, and that is based on a theory-guided method for conducting task analysis.

The testing-the-limits technique can also be conceptualized as a dynamic assessment approach, defined as a test-teach-retest approach. Peña, Quinn, and Iglesias (1992) found that dynamic assessment can help determine whether lack of experience with the contextual variables surrounding the assessment process (e. g., familiarity or unfamiliarity with the examiner, day and time of assessment, cultural and linguistic content of the test, etc.) influence the test results obtained by language disordered and non-disordered children. Peña et al. (1992) also reported that the test-teach-retest approach has been found to be effective in discriminating between those who are language-disordered and those who were not among the children enrolled in a Head Start program from a Puerto Rican and African-American background.

Dynamic assessment or the test-teach-retest approach has also been viewed as a way to distinguish between a child's actual development and his or her potential, with the difference being termed by Vygotsky as the ZPD (Samuda, King, Cummins, Lewis, & Pascual-Leone, 1991). Determining the ZPD in a child consists of figuring out the number of prompts necessary to teach a child a skill or concept during readministration of the items to which the child did not respond correctly during the initial standardized administration (Samuda et al., 1991). Thus, when doing dynamic assessments, the clinical observation of the learning process will help diagnose the child's ability to transfer learning.

In summary, in the critical literature review presented above, it is clear that more studies on the effect of using the testing-the-limits technique with LEP children are needed. By discussing available findings of studies exploring the use of testing-the-limits with majority children in regular or special education, we have learned about the effect of contextual variables surrounding the testing situation on children's test performance. Findings of these studies revealed that when testing- the-limits is used as a "dynamic assessment" method, a natural link between assessment and instruction is established. Thus, testing-the- limits can be a valuable tool for examiners of LEP students for gathering qualitative information from non-standardized administrations of tests.

Methodology

Subjects

A total of 125 surveys were completed by volunteers; of these, 101 respondents were NABE members and 24 respondents were not.

Instruments

In the survey we included two sections: (1) demographic questions, and (2) pairs of questions with an attitudinal and behavioral component. The demographic questions section included 9 multiple-choice questions previously stated above which provided information about the respondents on five areas, including: (1) position, (2) description of the school/school district (e.g. number of years assessing LEP students, region of the country, location - urban, suburban, or rural), (3) ethnicity/cultural background, (4) familiarity with other cultures, and (5) languages other than English spoken (See Appendix A). Twenty-nine pairs of questions with an attitudinal and behavioral components were included, of which 18 were 5-point Likert Scale type items (including from "strongly agree" to "strongly disagree" points), and 11 were closed-ended affirmative-negative (yes/no) questions. The survey mapped behaviors and attitudes toward the use of the testing-the-limits technique tapping the two research questions stated above.

Procedure

The construction of the survey included several steps for assuring internal validity. Following Anastasi's (1988) recommendations, construct validity of the survey was assured by the participation of two subject-matter experts, one a researcher and one a practitioner (first and second authors of this paper), who classified items based on different aspects of the operational definitions of the constructs to be measured. For constructing the survey we designed pairs of items that included attitudinal and behavioral components. For example, item 9 reads "Testing the limits should include rewording instructions when necessary," and item 10 reads "I reword instructions when assessing LEP students." In addition, special attention was given to the use of language when constructing items for avoiding biases related to unclear statements and using words with positive or negative connotations.

In order to avoid respondents' misunderstandings of the meaning of the definition of the testing-the-limits technique, we considered it

important to include an explanation of this term in the survey. We adapted a definition of testing-the-limits provided by Holtzman and Wilkinson (1991) who described it as an informal clinical procedure in which the examiner purposely changes the standardized administration of the test in some way in order to explore the student's abilities further. Thus, following Holtzman and Wilkinson (1991), we described testing-the-limits in the survey as "(a)n assessment technique in which the examiner changes standardized assessment conditions in some way" (Castellano & Gonzalez, 1994, p. 21). In the survey we also considered it important to describe five procedures used when applying the testing-the-limits technique, including: (1) to provide additional clues or to omit items for matching the children's cultural and linguistic backgrounds, and developmental levels; (2) to change modality (e.g. from written to oral language, from English to the child's first language, from verbal to non-verbal forms, from more difficult to easier words for giving instructions) involved in tasks administered; (3) to study methods and processes that children used for approaching and trying to complete tasks (i.e., strategies and styles for learning); (4) to eliminate time limits so that examiners can obtain much needed information about the children's abilities to accomplish specific tasks, and (5) to ask the children probing questions after the standardized testing had been completed to give examiners the opportunity to explore further the children's responses.

The survey was pilot-tested with a group of 30 part-time graduate students enrolled in a Midwestern university. These 30 graduate students were engaged in coursework leading toward the teaching endorsement in bilingual education or English as a second language. Twenty-five of these students held bilingual (Spanish) teaching positions, two were preschool teachers of at-risk children, one was a school librarian, and the remaining two were in the business world working toward a teaching certification. Of the 25 students working as teachers, years of experience ranged from a first-year teacher with no experience to those with more than 10 years of experience.

The survey was modified according to the results of the pilot test for improving its format and clarity of items. The final version of the survey for data collection was published in NABE News (see Castellano & Gonzalez, 1994) and was sent by mail (a total of 200 surveys were

mailed, with 125 being returned) mainly to practitioners and administrators in the Chicago and Tucson areas who were affiliated with NABE.

Data Analysis Design

This descriptive study used exploratory first-and-second-order factor analyses with varimax and Harris-Kaiser rotations in order to determine the structure of the 29-item survey and produce independent or orthogonal factors. However, since 53% of the responses for item 2 were found to be missing, it was omitted from the factor analysis. Item 2 inquired about the respondents' testing in Spanish. The first-order factor analysis produced four independent factors which all loaded on a single second-order factor. Thus the construct validity of the survey was demonstrated by the existence of a unidimensional factor underlying the four first-order factors. In addition, one-way ANOVAS were conducted entering the nine demographic items as independent variables and the 29 survey items (clustered in the four first-order factors) as dependent variables.

Results

The exploratory first-and-second-order factor analyses conducted showed that the four different orthogonal factors corresponded to one unidimensional construct measured by the survey. The first-order factor analysis conducted revealed four independent factors including: (1) attitudes toward psychometric properties of assessments (i.e., the use of English and the native language with criterion-referenced and normed- referenced tests, and the appropriateness of using the testing-the-limits technique); (2) attitudes and behaviors toward the adaptation of administrative practices (i.e., additional testing time, testing over several sessions, and repeating stimuli), and its implications for the performance of LEP students in standardized testing; (3) attitudes toward accommodating for cultural and linguistic differences (i.e., the use of standardized and alternative assessments that account for students' culturally appropriate responses); and (4) attitudes and behaviors toward translations and dialect variations (i.e., differentiating dialects and acknowledging heterogeneity within ethnic and cultural groups). The survey items related to the attitudinal and behavioral components

measured included: (1) items 1, 3, 6, 7, 12, 13, 14, and 29 for factor 1; (2) items 15, 16, 19, 20, and 28 for factor 2; (3) items 4, 9, and 23 for factor 3; and (4) items 21, and 25 for factor 4.

One-way ANOVAS revealed that only 1 out of 9 demographic factors made a significant difference ($p < .05$) in responses to the survey in relation to the first-order factor 1 (e.g., attitudes toward psychometric properties of assessments). The demographic factor that was found to make a significant difference was whether or not respondents were NABE members ($F = 4.49$, see Table 1). That is, NABE members significantly differed from the non-NABE members in relation to their attitudes toward psychometric properties of assessments used with LEP students. We explain the existence of a difference in responses given by NABE and non-NABE members as the presence of commonalties or differences in ideologies and attitudes toward the psychometric properties of assessments used with LEP students. NABE members tend to endorse multiculturalism and bilingualism, philosophies that recognize linguistic and cultural diversity in students as an enrichment that needs to be portrayed in assessment practices. Moreover, it was interesting to find that NABE membership and not ethnicity or familiarity with other cultures was the demographic independent variable that made a difference in responses to the first-order factor found.

Table 1
Summary of F value, Mean of Squares (MS), Least Square Mean (LS), d (degrees of freedom), and p (level of Significance) for the Demographic Variable of NABE Membership/or not NABE Membership Using the First Factor as a Dependent Variable

Demographic Variable	MS	LS	d	F	p
NABE Membership (n=11)	235.19	35.12	2	4.49	.157
Not NABE Membership (n=25)					

Discussion

We can conclude that the survey demonstrated construct validity because only one unidimensional construct or trait was found underlying the four first-order factors identified (Anastasi, 1988). Thus the construct of attitudes and behaviors toward the testing-the-limits technique can be used to explain the four first-order factors found. Furthermore, these four factors found have important practical implications for improving current practices when assessing LEP students. These four factors highlight the value consequences of testing, including: (1) attitudes toward psychometric properties of tests such as validity, (2) adaptation of standardized administration procedures, (3) accommodating for linguistic and cultural differences, and (4) translations and dialectal variations.

First Factor: Attitudes Toward Psychometric Properties of Tests

One of the most important psychometric properties of tests is construct validity in relation to which Messick (1989) considered that test interpretations and uses have implications and consequences of a social, educational, ethical, and moral nature. Thus, the examiners behaviors involved in test use and interpretation, and their attitudes and values held, are all subsumed within the concept of construct validity. Moreover, Messick (1989) pointed out his concern for the potential harm derived by the misuse of standardized tests which lack construct validity for minorities. In addition, Moss (1992) considered that construct validity is a central property of tests which needs to include consequences of test use such as the justification of interpretations of behaviors and the evaluation of social values. That is, criteria or standards on which criterion or norm-referenced tests are constructed reflect ideologies, beliefs, and sociopolitical structures of socially constructed diagnostic categories; the criteria used in traditional standardized tests are the product of the particular ideologies and values reflected in the "medical model" that emphasizes the quantitative measure of abilities and skills considered to be innate with fixed values (for a more extended discussion of philosophical assumptions of assessment models see Gonzalez, 1994; Gonzalez & Yawkey, 1993).

Moreover, in relation to this first factor found, Oiler and Damico (1991) pointed out that examiners make practical decisions on how to assess and interpret students' behaviors using particular theoretical conceptualizations of constructs measured, hypotheses and expectations, beliefs, attitudes, and in general their mental ability. They relate that in the case of LEP students, a solid theoretical understanding of how bilingual children develop linguistically and cognitively is central for assuring that assessments have construct validity. They point out, as do the standards for psychological testing published by the American Psychological Association, American Educational Research Association, and National Council of Measurement and Evaluation (APA, AERA, & NCME, 1985), that when assessing a student who is a speaker of an English dialect or is LEP with standardized English tests of intelligence or any other ability, the examiner may be measuring instead English language proficiency.

In relation to psychometric properties of tests such as their appropriateness for assuring "objective" measurements, Roth (1988) stated that "(o)bservation has a dual nature and cannot exist outside that duality" (p. 128) and that "(t)o examine a child is to examine a child being examined" (p. 125). Roth (1988) considered that when assessing a child, the situation created by the evaluation process will in turn influence the child's behavior and that the evaluator interprets the child's behaviors examined in relation to his own behaviors, value systems, and attitudes.

Second Factor: Attitudes and Behaviors Toward the Adaptation of Administrative Practices

This second factor refers to whether or not evaluators can act as empathic advocates for LEP students given that the technique of testing-the-limits can be used with the purpose of compensating for lack of learning opportunity by measuring learning potential. According to Lewis (1991), testing-the-limits is a clinical assessment procedure to help the evaluator diagnose the child's ability to transfer learning. Lewis (1991) related clinical assessment and testing-the-limits to Vygotsky's conceptualization of intelligence, as the Zone of Proximal Development could be measured as a dynamic process that changes with development and learning. Thus, the decision made by evaluators to use or not use the technique of testing-the-limits is related to their attitudes

and epistemological conceptualizations of constructs measured, such as intelligence. For instance, an evaluator who has a positive attitude tow&d the use of testing-the-limits with LEP students has a dynamic, multidimensional, qualitative view of intelligence as an idiosyncratic, developmental process. Relatedly, Marland (1987) stated that "(f)urther study is needed into...the relationship between attitudes and unconsciously held theories and pedagogical strategies to intervene" (p. 127).

Third Factor: Attitudes Toward Accommodating for Cultural and Linguistic Differences

The third factor relates to content, construct, and external validity. Moss (1992) defined content validity as the "(d)egree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content" (p. 240). Thus, LEP students need to be assessed with culturally and linguistically appropriate items so that they are familiar with the content represented in assessment instruments. In relation to the need for assessment instruments and administration procedures to accommodate the students' linguistic and cultural differences, APA, AERA, and NCME (1985) highlighted the need to recognize "(t)he limits of interpretations drawn from tests developed without due consideration for the influence of the linguistic characteristics of some test takers" (p. 73).

Moreover, when accommodating cultural and linguistic differences in relation to construct validity, Laosa (1991) pointed out that population generalizability becomes an ethical issue for practitioners. According to him, population generalization refers to the analysis of the significance of the effect of treatment variables on different populations that can be predicted by theoretical constructs. Particularly, he pointed out that the decision to use a particular assessment procedure that has been standardized with a different sociocultural population than the one to which the examinee belongs becomes an issue of professional ethics. It is ethically inappropriate for an evaluator to use a standardized assessment procedure when there is no evidence of construct validity to its practical application for making diagnostic and placement decisions. Relatedly, Washington and McLoyd (1982) stated, "Constructs represent the basic building blocks of psychological theory and as such have to be validated across populations

and ecologies" (p. 337). As they explain, ecological validity refers to how the internal characteristics of an individual interact with the particular situations experienced within the natural and social environment throughout the life span. What Washington and McLoyd (1982) called "population validity" is comparable to Laosa's (1991) term "population generalization." As explained above, in order to be able to generalize the effect of some variables across populations comparative research needs to be done leading to new norms and standardization procedures.

Furthermore, in reference to accommodating cultural and linguistic differences, Washington and McLoyd (1982) proposed a new conceptualization of external validity that encompasses cultural and interpretative validity including intentionality and meaning of the cultural contexts, viewpoints, and experiences of minority populations. They define cultural validity as "(t)he procedures necessary to identify the rules which regulate conduct as well as those rules which define various practices and institutions" (p. 325). They considered that the rules forming part of construct validity are in fact norms or social standards that regulate social expectations and behaviors performed by members of a particular sociocultural group. Thus, according to them, a bilingual and bicultural individual can adapt to the acquisition and use of rules in an appropriate manner for two distinctive sociocultural contexts. Moreover, these authors also acknowledged the multidimensional factors affecting the learning process of bilingual and bicultural individuals for adapting to different sociocultural groups. These factors are not only of a cognitive nature, but they are also of an emotional and social nature. Thus, by including affective processes influencing learning, an interface between cognitive and social domains can be created that leads to the possibility of linking rules and social norms with attitudes, value systems, and personal and social norms.

Relatedly, Banks and McGee Banks (1989) considered that standards and criteria for determining whether an individual belongs to one of the psychological diagnostic categories created by the "medical model" are socially constructed; that is, criteria for determining handicapping conditions, disabilities, giftedness, and normal development among minority and majority populations are subjective and culturally loaded. Moreover, they assert that values are preferences

about how to adapt to the environment, ideals, ethical and aesthetic standards, and knowledge developed by the social group to modify the environment and to create products.

In addition, six different value orientations of individuals within cultures have been proposed by Banks and McGee Banks (1989) for studying cultures, including: (1) supernatural beliefs such as religion, (2) respect for nature and the ecosystem, (3) use of the human-made habitat, (4) relational systems created such as the family, (5) activity level of its members such as work ethics and the value given to effort, and (6) the conceptualization of time. Applying cultural value orientations to assessment issues, we can observe that rules and social norms embedded within the construct validity of assessments are created by specific value orientations of cultural and ethnic groups. Rules and norms are socially constructed categories that can be unfair and biased for diagnosing and placing minority students when they represent cultural value orientations of different ethnic and social groups.

Furthermore, according to Washington and McLoyd (1982), interpretive validity is related to the intentionality of human actions and goal setting and achievement. These authors explained racism and stereotyping as the result of biased interpretations of majority people of the experience of being a minority individual. They pointed out that myths and distorted images that misrepresent the experience of being a minority within the mainstream American society are the result of partial interpretations. Explanations of minority issues and problems tend to take into consideration only internal factors focusing on the results of victimization, while ignoring external factors that caused the process of oppression and the meaning of being a minority. Thus, when interpreting and making inferences about test scores and performances of language-minority students, it is important to accommodate diversity and differences between the mainstream and the minority cultural and linguistic realities.

Fourth Factor: Attitudes and Behaviors Toward Translations and Dialectal Variations

In relation to the fourth factor, APA, AERA, and NCME (1985) stated, "One cannot assume that translation produces a version of the test that is equivalent in content, difficulty level, reliability and validity" (p. 73). When translating a test, words selected for items may have

differences across languages such as frequency rates, difficulty levels, acoustic properties, and length. For instance, Valencia and Rankin (1985) reported that the McCarthy Scales of Children's Abilities translated to the Spanish language showed biases against Mexican-American Spanish speaking children in the verbal and numerical memory subtests due to the effect of word length and acoustic similarity on information-processing load. These content biases can be explained as the effect of a predictable phonetic structure that uses a consonant-vowel syllable consistent pattern, and words translated into Spanish tend to be longer than words in English. Valencia and Rankin (1985) concluded that the problem of item inequivalence in the McCarthy Scales translated into Spanish, and not genuine limitations in Spanish- speaking children's cognitive abilities, generated content biases. Thus, in order to have validity and reliability, translated standardized tests need to be normed again with a sample that has the same idiosyncratic characteristics of language-minority students (APA, AERA, NCME, 1985).

Implications of Comments Made by Survey Respondents

Almost all respondents were motivated to include some comments voluntarily. We have categorized these comments into the following six topics: (1) the lack of knowledge or familiarity with testing-the-limits, (2) misconceptions about using testing-the-limits when having different educational purposes (e.g., special education, gifted classrooms), (3) misconceptions about the lack of need for using testing-the-limits for second language learners and LEP students, (4) advocacy for alternative assessment instead of using testing-the-limits with traditional standardized tests for LEP students, (5) the need to include other assessment strategies to test LEP students more appropriately, and (6) the need for alternative assessment for monolingual English children, as well.

The first nominal category, lack of knowledge or familiarity with testing-the-limits, can be exemplified by the following responses: (1) "You cannot get valid results if you change tests, and how reliable are they if we do change assessment procedures?"; and (2) "I feel that practices like testing-the-limits invalidates the standardized concept. How can we compare ourselves or our students to the norm if we change 'the norm'?"

Examples of the second nominal category regarding misconceptions about using testing-the-limits when having different educational purposes (e.g., special education, gifted classrooms) include: (1) "Testing-the-limits yields very helpful information when used for educational programming. But, testing-the-limits is not useful when trying to determine whether a student is eligible for special education services;" and (2) "The degree to which I test-the-limits depends on the way I plan to interpret the results. On standardized achievement tests, we allow only the directions to be re-read, translated, or explained. On individual assessments we modify to a greater degree."

Examples of the third nominal category, misconceptions about the lack of need for using testing-the-limits for second language learners and LEP students, is illustrated by the following comments: (1) "I evaluate many clients whose first language is not English and many who have limited English, but I do not assess the degree of language proficiency. I assess other aspects;" and (2) "I make allowances for my students on criterion-referenced tests, but not on norm-referenced tests or English as a second language evaluation tests."

The fourth nominal category, advocacy for alternative assessment instead of using testing-the-limits with traditional standardized tests for LEP students, can be best exemplified by the following quotes: (1) "I believe we tend to over-rely on standardized instruments for all students, and LEP students are disadvantaged because so few educators understand first/second language acquisition"; and (2) "I feel that alternative, native language tests should be available and should be reliable and valid..."

The fifth nominal category refers to the inclusion of other assessment strategies more appropriate for testing LEP students; it is illustrated by the following comments: (1) "It is hoped that the result of this study will clarify what is currently happening in the field, increase the research database in the filed of assessment, and benefit the language-minority students in the United States as we attempt to determine the most appropriate testing strategies for this subpopulation;" and (2) "I would like to learn about real and effective assessment that would help my bilingual students..."

The sixth nominal category, the need for alternative assessments also for monolingual English children, is best illustrated by the

following comments: (1) "We should look more at authentic assessment of children who speak a language other than English and also of English speakers." (2) "I do not assess students in bilingual programs; however, I do work with monolingual students who have language deficits. I find these students are at a disadvantage when using standardized tests. I question how much leeway should be given and still have the test results valid according to the norming criteria. It is also apparent that there is cultural bias on standardized tests even for the majority population."

These comments illustrate in a qualitative manner the statistical results explained above related to the attitudes of practitioners and administrators regarding the testing-the-limits technique. In these comments we can observe that educators have the need for learning about the use of clinical assessment techniques, such as testing-the-limits, so that their misconceptions can be dispelled. We can also observe that some educators are aware of the importance of using alternative assessments for all students, including minority and majority, so that fairness in the form of valid and reliable results can be achieved.

Conclusions

The four factors found and the nominal categories presented above highlight the need for evaluators to be knowledgeable about psychometric properties of assessment instruments. For instance, evaluators need to be knowledgeable about the validity of standardized administration procedures used with LEP students, and the extent to which clinical assessment techniques, such as testing-the-limits, can be useful and appropriate for accurately diagnosing LEP students. In the responses of administrators and practitioners completing this attitude survey regarding testing-the-limits, we have documented a level of uncertainty about the appropriateness of this clinical assessment technique (as shown in the factors as well as in their comments). In view of this uncertainty, we want to point out that the technique of testing-the-limits can be extremely useful for linking assessment with instruction. This technique gives evaluators the opportunity to discover the individual strategies for thinking and problem-solving used by LEP students, which can be identified as strengths that have major instructional implications for placement recommendations and

educational program development. Thus, testing-the-limits can be used also as a "dynamic" or "clinical assessment" technique that uses the "teach-test-teach" approach to the identification of potential for learning in culturally and linguistically diverse students.

Currently, there is need to educate administrators and practitioners about the use of clinical assessment techniques, such as testing-the-limits, that can be extremely useful for the accurate diagnosis of LEP students. At the same time, we also need to dispel the myths surrounding the psychometric appropriateness or inappropriateness of clinical assessment techniques. Thus we believe that more positive attitudes toward clinical assessment can be nurtured in evaluators and administrators by increasing their knowledge level about psychometric properties of assessments, and by raising their awareness about the possible positive impact on the appropriate education of LEP students that the use of these techniques can have. We need to acknowledge that the surrounding contextual variables when administering tests, such as the personality traits and the knowledge level of the examiner, do influence the examinee' 5 test performance. Moreover, we also need to remember that testing-the-limits as a clinical assessment technique can be used for both standardized and alternative assessments if the examiner's purpose is to conduct a more accurate assessment and to link assessment with instruction.

Limitations of the Study and Future Research Directions

We believe that the survey would have produced more significant findings in the ANOVA tests if it had been conducted on a larger and more representative sample. A future study should gather data on two groups of NABE members and non-members of at least 200 subjects each. This larger sample can help us understand better what kind of attitudes and behaviors practitioners and administrators who endorse and do not endorse NABE philosophies and values show when assessing LEP students. We would also recommend that this future study include face-to-face interviews with a smaller portion of the respondents to the surveys to gather more qualitative information about the administrators' and practitioners' values and beliefs about how to assess LEP students. These interviews should focus on understanding the reasons administrators and practitioners make assessment, diagnostic, and placement decisions for LEP students.

Authors' Note

In a collaborative effort, Dr. Castellano and Dr. Gonzalez developed this survey and used the columns they edited in the NABE News for the survey's national dissemination among bilingual educators.

Acknowledgments

The authors wish to thank Dr. Patricia Jones, Research Specialist at the Principal User Support Center for Computer and Information Technology, at The University of Arizona, for her time and valuable guidance in the statistical analysis of data in this paper. The authors also want to express their appreciation to the volunteer anonymous practitioners and administrators working with LEP students throughout the nation who filled out the surveys for this study.

References

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan Publishing Co.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: APA.

Banks, J. A., & McGee Banks, C. A. (1989). Multicultural education: Issues and perspectives. Boston, MA: Allyn & Bacon. Bethge, H. J., Carlson, J. S., & Wiedl, K. H. (1982). The effects of dynamic assessment procedures on Raven Matrices performance, visual search behavior, test anxiety, and test orientation. *Intelligence 6*, 89-97.

Castellano, J. A., & Gonzalez, V. (1994). Testing the limits in the assessment of LEP students: Research on the perceptions and applications of practitioners, *NABE NEWS*, 17(5), 21-22 & 27-28.

Carlson, J. S., & Wiedl, K. H. (1979). Toward a differential testing approach: Testing-the-limits employing the Raven Matrices. *Intelligence 3*, 323-344.

Dash, A. S., & Rath, R. (1986). Testing-the-limits of Raven's Progressive Matrices: An experiment. *Psychological Studies, 31*, 82-89.

Eysenck, H. J. (1969). Intelligence assessment: A theoretical and experimental approach. *British Journal of Educational Psychology, 37*, 81-98.

Eysenck, H. J., & Eysenck, 5. (1969). *Personality structure and measurement*. London: Routledge & Keegan Paul.

Færch, C., & Kasper, G. (1987). From product to process: Introspective methods in second language research. In C. Færch, & G. Kasper (Eds.), *Introspection in second language research* (pp. 5-23). Clevedon, England: Multilingual Matters.

Gonzalez, V., & Yawkey, T. D. (1993). The assessment of culturally and linguistically diverse students: Celebrating change. *Educational Horizons*, 72 (1), 41-49.

Gonzalez, V. (1994). The assessment of language-minority students: A critical discussion of models and problems. *NABE News, 17* (8), 5-6,&38.

Holtzman, W., Jr., & Wilkinson, C. Y. (1991). Assessment of cognitive ability. In E. V. Hamayan, & J. S. Damico, *Limiting bias in the assessment of bilingual students* (pp. 247-180). Austin, TX: Pro-Ed.

Kagan. J. (1964). Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monograph, 78*, 578.

Laosa, L. M. (1991). The cultural context of construct validity and the ethics of generalizability. *Early Childhood Research Quarterly, 6*, 313-321.

Lewis, J. (1991). Innovative approaches in assessment. In R. J. Samuda, S. L. Kong, J. Cummins, J. Pascual-Leone, & J. Lewis (Eds.), *Assessment and placement of minority students* (pp. 123-142). Toronto, Canada: C. J. Hogrefe.

Marland, M. (1987). The education of and for a multiracial and multilingual society: Research needs post-Swan. *Educational Research, 29* (21), 116-129.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: Macmillan Publishing Co.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62 (3), 229-258.

Newman, D., Griffin, P., & Cole, M. (1989). *The construction zone: Working for cognitive change in school*. New York: Cambridge University Press.

Oller, J. W., Jr., & Damico, J. 5. (1991). Theoretical considerations in the assessment of LEP students. In E. V. Hamayan, & J. S. Damico (Eds.), *Limiting bias in the assessment of bilingual students* (pp. 77-110). Austin, TX: Pro-Ed.

Pascual-Leone, J., & Ijaz, H. (1991). Mental capacity testing as a form of intellectual-developmental assessment. In R. J. Samuda, & S. L. Kong *Assessment and placement of minority students* (pp. 143-171). Toronto, Canada: Hogrefe and ISSP.

Peña, E., Quinn, R., & Iglesias, A. (1992). The application of dynamic methods to language assessment: A non-biased procedure. *The Journal of Special Education, 26*, 269-280.

Roth, J. (1988). Some assumptions about the evaluation of children. *Reading, Writing and Learning Disabilities*, 4, 125-131.

Samuda, R. J., King, S. L., Cummins, J., Lewis, J., & Pascual-Leone, J.(1991). *Assessment and placement of minority children*. Lewinston, NY: Intercultural Social Sciences Publication.

Sattler, J. M. (1969). Effects of clues and examiner influence on two Wechsler subtests. *Journal of Consulting and Clinical Psychology, 33*, 716-721.

Sattler, J. M. (1982). *Assessment of children's intelligence and special abilities* (2nd ed.). Boston, MA: Allyn & Bacon.

Valencia, R. R., & Rankin, R. J. (1985). Evidence of content bias on the McCarthy Scales with Mexican-American children: Implications for test translation and non-biased assessment. *Journal of Educational Psychology, 77*(2), 197-207.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Cambridge, MA: Harvard University Press.

Washington, E. D., & McLoyd, V. C. (1982). The external validity of research involving American minorities. *Human Development, 25*, 324-339.

Appendix A
Survey on Attitudes and Behaviors of Practitioners and Administrators on the Testing-the-Limits Technique Used with LEP Students

Demographic Information

Instructions: Please, check the most appropriate response for each item and return this section with your completed survey.

1) My current position is
   - TBE teacher
   - Regular Ed. Teacher
   - TPI Teacher
   - Principal
   - Coordinator
   - Central Office Administrator
   - Director
   - Other (Specify)

2) My school/district can be best classified as:
   - Urban
   - Suburban
   - Rural

3) My school/district is a:
   - Unified School District
   - Elementary District
   - High School District
   - Other (Specify)

4) Number of years assessing LEP students:
   - No experience
   - 1-5 years
   - 6-10 years
   - 11-15 years
   - 16 years plus

5) Number of undergraduate or graduate courses taken on the assessment of LEP students:
None
One
Two
Three
Four or more

6) In what part of the United States are you located?
Southeast
South
East Coast
Midwest
Mideast
West Coast
Southwest
Northwest

7) Ethnic/cultural background:
Native-American
African-American
Hispanic
Asian-American
Anglo
Other (Specify)

8) List the language, other than English, with which you have the most experience and knowledge:
Spanish
Italian
Korean
French
Japanese
Chinese
Other (Specify)

9) Personal familiarity with other cultures:
Native-American
Hispanic
Asian-American
African-American
Other (Specify)

SURVEY QUESTIONNAIRE

Definition: Testing the limits is defined as an assessment technique in which the examiner purposefully changes standardized testing conditions in some way.

For example, some procedures used when applying the testing-the-limits technique include:

(1) to provide additional clues or to omit items for matching the children's cultural and linguistic backgrounds, and developmental levels.

(2) to change modality (i.e., from written to oral language, from English to the child's first language, from verbal to non-verbal forms, from more difficult to easier words for giving instructions) involved in tasks administered.

(3) to study methods and processes that children used for approaching and trying to complete tasks (i.e., strategies and styles for learning).

(4) to eliminate time limits so that examiners can obtain much needed information about children's abilities to accomplish specific tasks.

(5) to ask children probing questions after the standardized testing has been completed to give examiners the opportunity to explore further children' 5 responses.

Instructions: Please, respond to the following items regarding testing-the-limits in the assessment of LEP students. Circle your responses.

Key: SD-Strongly Disagree, D-Disagree, N-Neutral, A-Agree, SA-Strongly Agree.

1) Testing-the-limits in the assessment of LEP students should be an acceptable practice in bilingual education.

SD D N A SA

2) When assessing LEP students in Spanish, I test the-limits.

Yes No

3) Testing-the-limits with LEP students invalidates their results.

SD D N A SA

4) Vendors account for the needs of LEP students when developing tests for them.

SD D N A SA

5) LEP students are a norming population in the tests used by my school or district.

Yes No

6) Testing-the-limits should occur in native language assessments.

SD D N A SA

7) Testing-the-limits should occur in the English language assessments of LEP students.

SD D N A SA

8) When assessing LEP students in English, I test the limits.

Yes No

9) Testing-the-limits should include rewording instructions when necessary.

SD D N A SA

10) I reword instructions when assessing LEP students.

Yes No

11) Testing-the-limits should occur when using teacher-made tests.

SD D N A SA

12) Testing-the-limits should occur when using criterion-referenced tests.

SD D N A SA

13) Testing-the-limits should occur when using norm-referenced standardized tests.

SD D N A SA

14) Testing-the-limits with LEP students should not occur in bilingual education.

SD D N A SA

15) Additional time, beyond that specified in the test manual, should be provided.

SD D N A SA

16) I provide additional time for students to respond.

Yes No

17) The students' answers in dialect should be compared to first language or second language learning features.

SD D N A SA

18) I compare the students' dialect to first language or second language learning features.

Yes No

19) Testing of LEP students should occur over several sessions.

SD D N A SA

20) I test LEP students over several sessions.

Yes No

21) You should omit items you expect the child to miss because of age, language, or culture.

SD D N A SA

22) I have omitted test items in the past.

Yes No

23) You should accept culturally-appropriate responses as correct.

SD N A SA

24) I have accepted culturally appropriate responses as correct.

Yes No

25) Changing the language/vocabulary of test items is appropriate when testing-the-limits.

SD D N A SA

26) I change the language/vocabulary of test items when testing-the-limits.

Yes No

27) Repeating the stimuli more than specified in the test manual is appropriate when testing-the-limits.

SD D N A SA

28) I repeat the stimuli more than specified in the test manual when testing-the-limits.

Yes No

29) Testing-the-limits should be an expected practice in standardized assessments.

SD D N A SA