Standardized Language Test Use: A Canadian Survey

Utilisation d'un test de langage normalisé : un sondage canadien

M. Alanna Kerr
Sabina Guildford
Elizabeth Kay-Raining Bird

Abstract

Members of the Canadian Association of Speech-Language Pathologists and Audiologists (CASLPA) with interests in child language were surveyed to examine their current use of norm-referenced tests, current measurement practices, and psychometric knowledge. Primary focus of the study centred on the frequency with which practices characterized as 'misuse' by McCauley and Swisher (1984b) occur. Specifically, these include the use of individual subtest items to establish treatment goals, use of profiles to characterize patterns of deficits, use of repeated standardized test administration to measure treatment progress, and use of age-equivalent scores to summarize test results. Results indicate that clinician awareness of these 'misuses' is variable. Clinicians are aware of the problems associated with the use of individual subtest items to establish treatment goals and with the use of age-equivalent scores, yet continue to engage in these practices nonetheless. A large proportion of clinicians find profiles useful, but few are aware of the cautions which should accompany their use. Most speech-language pathologists use a combination of criterion-referenced procedures and standardized tests to measure treatment progress. As such, the use of the less sensitive standardized tests may be interfering with accurate measurement of the effectiveness of the intervention provided. Consistent with the foregoing results, the survey revealed that few clinicians are fully confident with their psychometric knowledge.

Abrégé

On a sondé des membres de l'Association canadienne des orthophonistes et audiologistes (ACOA) s'intéressant au développement langagier chez l'enfant afin de connaître leur usage actuel de tests normatifs, leurs pratiques de mesure actuelles et leur connaissance en psychométrie. L'étude s'est concentrée principalement sur la fréquence des occurrences de pratiques caractérisées de « mésusages » par McCauley et Swisher (1984b). Plus précisément, ceux-ci comprennent l'utilisation d'éléments individuels de subtests afin de fixer des objectifs de traitement, l'utilisation de profils pour caractériser des modèles de déficiences, l'utilisation d'une administration répétée de tests normalisés pour mesurer le progrès du traitement, et l'utilisation de scores d'équivalence d'âge pour résumer les résultats des tests. Les résultats montrent que la reconnaissance par les cliniciens de ces « mésusages » est variable. Les cliniciens sont conscients des problèmes liés à l'utilisation d'éléments individuels de subtests afin de fixer des objectifs de traitement et à l'utilisation de scores d'équivalence d'âge, mais continuent tout de même d'employer ces pratiques. Une grande proportion de cliniciens considèrent que les profils sont utiles, mais peu sont conscients des précautions qui doivent les accompagner. La plupart des orthophonistes utilisent une combinaison de procédures critérielles et de tests normalisés pour mesurer le progrès du traitement. À ce titre, l'utilisation de tests normalisés moins sensibles peut nuire à l'exactitude de la mesure de l'efficacité de l'intervention. En cohérence avec les résultats susmentionnés, le sondage a révélé que peu de cliniciens ont pleine confiance en leur connaissance psychométrique.

M. Alanna Kerr
Sabina Guildford
Elizabeth Kay-Raining Bird
Dalhousie University
Halifax, Nova Scotia

Key words: language assessment, child standardized tests, psychometrics, measurement practices, norm-referenced tests

Clinical assessment in speech-language pathology forms the basis for decisions regarding whether to intervene and, if so, the nature of that intervention. The assessment process necessitates either implicit or explicit measurement (McCauley, 1989). Standardized tests have become an ever-present component of speech and language assessment protocols. Several decades ago, McCauley and Swisher (1984a, b) identified several ways in which practitioners frequently misuse or misinterpret standardized tests. This study aimed to illuminate the sources of information clinicians use to reach clinical decisions regarding language impairment in preschool and school-aged children through the use of a survey instrument. In particular, we were interested in documenting what tests are used currently, how they are used, and whether the concerns raised by McCauley and Swisher (1984a, b) remain valid, in a Canadian context.

In their 1997 survey of clinicians in Oregon, Huang, Hopkins, and Nippold found that approximately half of the respondents felt neither positively nor negatively about the psychometric characteristics of standardized tests they used. The authors suggested that this neutral stance could result from a lack of knowledge about the clinical implications of low reliability and validity and that clinicians needed to become more aware of the limitations of standardized tests. Huang et al. further investigated clinicians' satisfaction with standardized tests and concluded that approximately three quarters of clinicians feel neutral or dissatisfied with the standardized tests they use. Among other factors, Huang et al. attributed dissatisfaction to the lack of suitable tests and unavailability of multi-cultural material. Canada is also increasingly culturally and linguistically diverse. The present study will also explore the manner in which clinicians are assessing children from linguistically diverse backgrounds. McCauley (1989) also urged clinicians to increase their knowledge of measurement and psychometric principles as a means of ensuring the appropriate use of psychometric tests.

The objectives of clinical decisions in speech-language pathology can be grouped into four general categories: to determine the existence and general areas of language impairment, to describe the language system (to assess specific areas of deficit), to establish goals and strategies for the intervention process, and to measure response to intervention (Lahey, 1988, 1990; McCauley & Swisher, 1984a; Merrill & Plante, 1997). Huang et al. (1997) reported that standardized tests formed the basis for many decisions that the speech-language pathologists (SLPs) made in terms of all the aforementioned intervention objectives. McCauley and Swisher (1984b) raised concerns associated with the misuse of norm-referenced tests. The extent of these 'misuses' in the context of published research has been examined (McCauley and Demetras, 1990), but no studies have examined the extent of these 'misuses' in clinical practice.

McCauley and Swisher (1984b) found fault with the manner in which norm-referenced tests were commonly used by SLPs. They pointed to limitations of norm-referenced tests and urged test users to familiarize themselves with the psychometric properties of the tests they use (1984a). Specifically their concerns pertained to (a) the use of individual test items to establish treatment goals, (b) the use of profiles to characterize a child's overall strengths and weaknesses, (c) the use of repeated standardized-test administration to measure treatment progress, and (d) the use of age-equivalent scores to summarize results. McCauley and Swisher's 1984 publications are considered by many to be benchmark papers in the study of speech-language pathology. These papers may have impacted the profession sufficiently that the concerns raised in them are no longer a problem in current clinical practice, although there is little empirical data to determine whether this is so. Regardless, the concerns raised in these papers receive ongoing discussion in the literature. Each of these is considered in the following sections.

Using individual subtest items to establish therapy objectives. While some have maintained that use of tests and test items to assess a particular structure is a valid practice (Owens, Haney, Giesow, Dooley, & Kelly, 1983) others have adopted a more cautious position and suggested that standardized tests may help SLPs identify general areas of deficit for further probing through criterion-reference measures (Haynes & Pinzola, 1998; Lahey, 1988). McCauley and Swisher (1984b) argue that there are several problems associated with this practice: (a) norm-referenced tests include relatively few items and cannot test all the forms and levels that may be necessary and relevant to establish functional treatment goals; (b) teaching to the test invalidates the test as a tool for reassessment; (c) functional communication skills are not adequately characterized by the restricted context of a norm-referenced test; (d) scoring systems fail to provide descriptive clues useful for establishment of treatment goals because description of responses - correct and incorrect, immediate and delayed - are needed to help to distinguish between different degrees of impairment; and, (e) individual errors and correct responses can have several explanations other than true linguistic competence. They state that "there is probably no circumstance in which norm-referenced test items can profitably be used for this purpose" (1984b, p. 344). Findings of Merrill and Plante (1997) concur. The latter evaluated the suitability of norm-referenced tests to address two separate assessment objectives: determining the existence of a language impairment and describing the specific areas of deficit (the latter being a necessary prerequisite for establishing therapy objectives). They found that norm-referenced tests provide good discriminating ability, but provide inconsistent results at the individual item level. As such, Merrill and Plante (1997) concluded that standardized tests can be appropriate diagnostic tools for determining the existence of and general areas of language impairment, but are not appropriate for addressing specific areas of deficit. Others agree that most standardized tests lack sufficient numbers of test items to provide detailed descriptions of children's abilities and needs (Lahey, 1988; Salvia & Ysseldyke, 1991). Huang et al. (1997) found that a majority of clinicians also felt that standardized tests did not provide the information necessary for establishing intervention goals and strategies, thus demonstrating knowledge of this potential problem of test use.

Use of profiles to characterize a child's overall strengths and weaknesses. McCauley and Swisher (1984b) pointed out that differences between scores within a test profile may result from measurement error rather than from real differences in the behaviours being measured. Valid use of profiles, they suggested, required information on reliability and intercorrelations of subtest scores, which are often lacking. McCauley and Swisher recommended, as a conservative alternative, that profile use be limited to identifying the presence or absence of impairment in different areas rather than their relative degrees. Lahey concurred and succinctly stated: "while test, or subtest, comparisons can be made in a dichotomous manner (i.e., whether or not each test suggested a problem in an area), they cannot be interpreted to indicate degrees of difference within such a dichotomy" (1988, p. 174).

Use of repeated standardized-test administration to measure treatment progress. McCauley and Swisher (1984b) argued that such practice leads to inflated or depressed estimates of treatment progress. This is because norm-referenced tests are designed to measure general, relatively stable behaviours that are representative of major developmental levels. In contrast, measuring treatment progress requires examining specific behaviours that vary within these developmental levels. McCauley and Swisher pointed out that change in these specific behaviours can be best assessed by criterion-referenced tests specifically designed to measure the behaviours targeted by therapy. Thus, McCauley and Swisher concluded that criterion-referenced tests are most appropriately used to measure treatment progress, while infrequent administration of standardized tests can be informative if the purpose is to reveal whether impairment persists rather than record the amount of change. The inadequacy of norm-referenced tests for the aforementioned purpose is echoed by others who assert that standardized tests lack the number and variety of items necessary for monitoring treatment progress (Huang et al., 1997).

Use of age-equivalent scores. With regards to age-equivalent score use, McCauley and Swisher (1984b) noted: (a) if wide variation in skill ability is common at a given age, even large delays indicated by age-equivalent scores are not necessarily indicative of language delay or disorder; (b) the younger age score obtained by an older child does not justify the inference that he/she has the language and world-knowledge that the older child would have; (c) the reliability of age-equivalent scores is poorer for developmentally more advanced test takers because "as age increases, similar differences in age-equivalent scores are the result of smaller and smaller differences in raw scores"; and (d) age-equivalent scores are often estimated by interpolating between ages for which data have been collected which requires assumptions about the continuity of language development which may not be justified (1984b, p. 340). McCauley and Swisher suggested that if age-equivalent scores are used to summarize test results, they should be used only in conjunction with standard scores or percentile ranks and accompanied by a description and explanation of the inferences required to validate their use. Furthermore, they argued that tests for which only age-equivalent scores are available should be avoided. In general, age-equivalent scores as less reliable than standard scores or percentile ranks as they are more open to misinterpretation (Lahey, 1988; Petersen, Kolen, & Hoover, 1989). Anastasi refers to age-equivalent scores as "psychometrically crude" and unable to "lend themselves to precise statistical treatment" (1988, p. 78). Despite these cautions, a review of 72 studies published between 1983 and 1988 found that age-equivalent scores were the most frequent and often the sole scoring system used to identify children with language impairments (McCauley & Demetras, 1990). It is unclear whether such frequent use of age-equivalent scores persists in current clinical practice. Huang et al. (1997) found that 15% of the respondents in their study reported exclusive use of age-equivalent scores.

The purpose of the present study was to examine standardized test use in a sample of SLPs practising in

Canada. In particular, it was of interest to determine whether the concerns around 'misuses' of norm-referenced standardized tests identified by McCauley and Swisher had currency in Canada today. In addition, it was of interest to describe the motivation of clinicians for engaging in these practices should they exist. The foregoing may not only guide further education and training of clinicians, but may also point to deficits in appropriate resources available to clinicians. This may contribute to feedback to test-providers and those who design assessment materials.

Method

Participants

Data were collected by means of a questionnaire survey mailed or faxed to a sample of 507 SLPs in Canada expressing an interest in child language. These were selected from CASLPA's mailing list as of March 1999, which constituted approximately 930 such members. Participants were considered eligible if they identified themselves as currently practising (employed), and indicated proficiency in English. These restrictions were made as English proficiency and knowledge based upon current clinical practice were considered important.

Survey Instrument

The survey instrument was a 25-item survey which covered six general areas.

Demographic information. The first section of the questionnaire requested information on participants such as age, years worked as an SLP with children, ages of clients served, work place, and province of residence.

Ratings of the importance of sources of assessment information. Question 1 of the survey explored the relative importance that respondents gave to six sources of information: (a) information from significant others, (b) case histories, (c) standardized tests, (d) criterion-referenced procedures, (e) observations in context, and (f) language sample analysis for performing each of five clinical assessment tasks. These assessment tasks were as follows: (a) screening, to establish the existence of an impairment; (b) diagnosis, to determine the presence and severity of specific areas of deficit; (c) describing a child's language system; (d) establishing intervention goals; and (e) measuring treatment progress. Specific attention was devoted in the analysis to the relative importance of standardized tests. If they proved to be of little importance, then subsequent questions regarding their 'misuse' would be moot points.

Use of standardized tests in ways identified as problematic. Questions 2 through 5 spoke specifically to the concerns of McCauley and Swisher (1984b) outlined above. Examining responses to each of the questions in this section allowed the researchers to establish the frequency with which practices of "misuse" occurred. Specifically, this section of the survey questioned clinicians regarding their use of individual subtest items (Question 2), profiles (Question 3), and age-equivalent scores (Question 5). Questions 2, 3, and 5 had similar structures: For example the latter question was: (a) When summarizing standardized test results, do you use age-equivalent scores? (b) Do you see benefits to this practice? (c) If so, list two of the most important. (d) Do you feel there are problems with this practice? (e) If so, list two of the most important. Question 4 examined the relative importance of various tools for evaluating treatment progress. Specifically it asked clinicians to rank standardized tests, formal criterion-referenced tests, informal criterion-referenced measures, and "other"(please specify). It also asked clinicians for the frequency with which they assessed progress.

Self-assessments of psychometric knowledge. Question 6 related to the confidence clinicians had with respect to their psychometric knowledge, the sources of that knowledge, and its maintenance. This question, therefore, addressed the ability of clinicians to take advantage of opportunities to increase their psychometric knowledge - a recommendation of McCauley (1989) and Huang et al. (1997).

Use of specific standardized tests. Question 7 pertained to the identification of the tests respondents use most frequently and the purposes for which they use them.

Non-English clients and standardized tests. Question 8 asked participants to report their practices regarding standardized test use with non-English speaking children.

Procedure

The 10-page questionnaire was mailed to 507 SLPs in April 2000. Coded envelopes allowed responses to be tracked while ensuring anonymity of the respondent. A hundred and twenty-five SLPs responded. SLPs who failed to return the questionnaire and for whom fax numbers were available (42) were contacted a second time. Nineteen further surveys were collected as a result. A total of 144 completed surveys were received.

Data Analysis

For descriptive analyses, means, standard deviations, frequencies, and percentages were calculated. Rankings given to the relative importance of tools for clinical decision-making (Question 1) and to the relative importance of various methods of measuring treatment progress (Question 4) were compared using a nonparametric Friedman Rank Sum test (Hollander & Wolfe, 1973). This required the assumption that the rankings represented the relative values of a conceptual underlying continuous variable representing the utility of each of these tools or methods. The errors associated with the conceptual utility are assumed to be independent and identically distributed. When significant differences were found to be present, further two-treatment nonparametric Wilcoxon Signed Ranks tests (Hollander & Wolfe, 1973) were performed to examine whether each of the sources of information (Question 1) or tools for measuring treatment progress (Question 4) differed significantly in relative importance from that assigned to standardized tests. The authors were aware of the implications this would have for increasing the overall error rate. Overall error rate was conservatively estimated as the sum of the Type I errors associated with each set of comparisons performed. In all cases, the overall Type I error rate was below $p = 0.05$; furthermore, the only instances where it exceeded $p = 0.01$ was for the survey Question 1, clinical-task comparison groups: preschool, screening; preschool, diagnosis; and elementary-school, measuring treatment progress.

A Pearson's Chi-Square test was used to examine the independence of problems associated with "use of individual subtest items to establish treatment goals," "use of profiles to establish patterns of impairment," and "use of age-equivalent scores to summarize test results" as identified by the respondents (Questions 2, 3, and 5) from respondents' "reported self-confidence" (Question 6).

Results and Discussion

Sample Characteristics

Numbers of responses and constituent percentages of the sample by province are summarized in Table 1. All participants were speech-language pathologists and members of the Canadian Association of Speech-Language Pathologists and Audiologists (CASLPA). The total number of CASLPA members by province who fit the sample selection criteria is unknown, but the percentage of the response sample constituted by each province is comparable to that reported by Potter and Lagacé (1995) - with the present study having perhaps a slightly greater representation from the maritime provinces.

Table 1
Percentage and Number of Speech-Language Pathologists Receiving and Responding to Questionnaires by Province

Province	# Mailed/ # Returned	Response rate	Percentage of sample
Alberta	118 / 33	28%	23%
British Columbia	98 / 23	23%	16%
Manitoba	18 / 8	44%	6%
New Brunswick	25 / 11	44%	8%
Newfoundland	12 / 4	33%	3%
North West Territories	1 / 1	100%	1%
Nova Scotia	23 / 10	43%	7%
Ontario	157 / 38	24%	27%
Quebec	33 / 7	21%	5%
Saskatchewan	21 / 8	38%	6%
Yukon	1 / 0	0%	0%
PEI	3 / 0	0%	0%

The first section of the survey obtained demographic characteristics of the respondents. All respondents who chose to report their gender (two did not) were female. This is in contrast to the approximately 7% male population reported by Dohan and Schulz (1999) who surveyed Canadian SLPs working with school-based populations. The reason for this difference is unclear. The average time for working with children as an SLP was about 12 years. Thus, clinicians with considerable experience were more likely to complete this demanding survey. This should be kept in mind when interpreting results. Ninety-three percent of respondents held master's degrees, more than in Dohan and Schulz's study (1999) where they constituted approximately 80% of the sample. All respondents in the present study were certified nationally or registered provincially.

Characteristics of the clinicians' work environments and caseloads were also collected. Approximately three-quarters of respondents reportedly worked in urban settings, while the remainder worked in communities of less than 5,000 residents. Many respondents worked in multiple settings (40%). The greatest number (43%) of clinicians reported they worked in schools, 34% indicated that they worked in clinics or hospitals, 33% reported working for a community agency, and 28% stated that they were in private practice. Caseload size varied considerably. A small minority (9%) had caseloads of less than 20 children. Caseloads were distributed fairly evenly from 21 to 80 children, with 18% of clinicians reporting a caseload of 21-40, 20% with 41-60 and 20%

with 61-80 children. One third of respondents reported caseloads of over 80 children. Dohan and Schulz (1999) found an even larger variation in reported caseload size (from 10 to 500), despite surveying only those clinicians who worked in schools. They suggested that the term caseload was open to interpretation. Most clinicians in the current sample served more than one age group: 37% worked with infants, 74% with preschoolers, 70% with elementary school-age clients, and 35% with junior high and high school students. Most clinicians (80%) predominantly saw clients whose first language was English. Thirty percent of clinicians, however, saw mainly children whose first language was French and 35% reported serving children who had first languages other than English or French.

Question 1: Importance of Standardized Tests in Decision Making

The findings for the relative importance of various tools used for the clinical tasks of screening, diagnosing language deficits, describing the language system, establishing treatment goals, and measuring treatment progress are summarized in Table 2.

Screening. Clinicians ranked standardized testing among the least important tools for screening for the existence of a language disorder in the preschool population. "Standardized tests" ranked fifth on average and were not statistically significantly different from "language sample analysis" nor "criterion-referenced procedures." Standardized tests, however, were ranked

Table 2
Average Rank of Decision-Making Tools

Age Group/ Clinical task	Standardised tests	from significant others	Observations in context	Criterion-referenced procedures	Language sample analysis	Case history
Preschool *n* = 112						
Screening	4.26 (5)	2.15*(1)	2.58*(2)	4.37 (6)	4.11 (4)	3.54*(3)
Diagnosis	2.77 (2)	3.50*(3)	2.67 (1)	4.02*(5)	3.76*(4)	4.28*(6)
Describe Lang. System	3.04 (3)	3.90 (4)	2.42 (1)	3.94 (5)	2.76 (2)	4.95*(6)
Est. Tx. Goals	3.09 (3)	3.57 (4)	2.56 (1)	3.63 (5)	3.01 (2)	5.14*(6)
Measure Tx Progress	3.57 (4)	3.63 (5)	2.56*(1)	3.01 (2)	3.09 (3)	5.14*(6)
Elementary School Age *n* = 96						
Screening	2.98 (2)	2.49 (1)	3.12 (3)	4.40*(6)	4.18*(5)	3.82*(4)
Diagnosis	1.92 (1)	3.60*(3)	3.31*(2)	3.99*(4)	4.00*(5)	4.18*(6)
Describing Lang.System	2.32 (1)	3.86*(4)	2.76*(2)	3.90*(5)	3.34*(3)	4.82*(6)
Est. Tx. Goals	2.23 (1)	3.68 (5)	2.90*(2)	3.58*(4)	3.52*(3)	5.09*(6)
Measure Tx Progress	2.93 (2)	3.44 (4)	2.26 (1)	3.20 (3)	3.69*(5)	5.48*(6)
Junior/ Senior Highschool Age *n* =52						
Screening	2.28 (1)	2.89 (2)	3.34* (3)	4.25* (5)	4.69* (6)	3.54* (4)
Diagnosis	1.63 (1)	3.64* (2)	3.70* (3)	3.79* (4)	4.21* (6)	4.04* (5)
Describing Lang.System	1.92 (1)	3.80* (4.5)	3.20* (2)	3.80* (4.5)	3.65* (3)	4.62* (6)
Est. Tx. Goals	1.78 (1)	3.73* (5)	3.08* (2)	3.72* (3.5)	3.72* (3.5)	4.97* (6)
Measure Tx Progress	2.53 (1.5)	3.56* (4)	2.53 (1.5)	3.14 (3)	3.76* (5)	5.48* (6)

Notes. Tools were ranked from 1 to 6 with 1 being most important. Numbers in ( ) denote ranking of importance relative to other tools listed for the same task and age group. Rankings marked with '*' are significantly different in avg rank from standardized tests (*p* = 0.05) for the same clinical task and age group.

Table 3a
Reported Benefits of Using Individual Subtest Items to Establish Treatment Goals

Category (n = 118 of 127 or 92.9% of respondents who said "yes" to benefits answered)	Number of clinicians	Percentage of clinicians	Percentage of responses
Identifies specific, age or developmentally appropriate goals	77	65.3%	46.4%
Easy to reassess	28	23.7%	16.9%
Quick and efficient	19	16.1%	11.4%
Gives a starting point from which to probe further	17	14.4%	10.2%
Beneficial when applied with clinical judgement	11	9.3%	6.6%
Easy to demonstrate to others	9	7.6%	5.4%
Other	5	4.2%	3.0%

higher in importance when screening school-age children. For elementary children, standardized testing ranked second on average - not significantly different from "information from others" and "observations in context." For screening junior high and high school populations standardized testing ranked first, not statistically significantly different from second-ranked "information from others."

Diagnosis. Clinicians reported standardized tests to be an important tool for diagnosing language impairment at all ages. Clinicians working with preschool clients ranked it second, not significantly different in importance from first-ranked "observations in context." Respondents working with all school-aged populations ranked standardized tests first for this assessment task. With the one exception, standardized test use was not only ranked the most important tool, but it ranked significantly above all other decision-making tools.

Describing the language system and establishing treatment goals. When performing these tasks, respondents working with preschoolers considered a range of tools important. Their rankings of tools for "describing the language system" and "establishing treatment goals" were identical. For both, "standardized testing" was no more or less important than any other tool with the exception of "case history" which was significantly less important.

For school-aged populations the rankings were identical: standardized testing was ranked first for both assessment tasks. It was not only the most important tool, but it ranked significantly above all other decision-making tools.

Measuring treatment progress. Respondents regarded a broad range of decision-making tools as important when measuring treatment progress. While clinicians working with preschool populations ranked "standardized tests" fourth, it was only significantly less important than first ranked "observations in context." Clinicians working with school-age populations ranked "standardized testing" as the first (junior and senior high school) and second (elementary) most important tool for this task, but in both cases not significantly different from any other tools with the exception of lower ranked "language sample analysis" and "case history."

Question 2: Use of Individual Subtest Items to Establish Treatment Goals

Respondents were asked to identify the frequency with which they used individual subtest items to establish treatment goals. A majority, 59.5%, reported doing so "sometimes," 28.7% "frequently," and 2.8% "always." Nine percent "never" used individual subtest items to establish intervention goals. Consistent with these numbers, 91% (127/140) felt there were benefits to the aforementioned practice. Respondents were asked to identify two benefits and these benefits are summarized in Table 3a. Most clinicians (89%, 124/140) also felt there were problems associated with using individual subtest items to establish intervention goals. Respondents were asked to identify up to two problems associated with the practice and most clinicians (86%) did. Their answers are summarized in the Table 3b. McCauley and Swisher (1984b) identified five major problems associated with the use of individual subtest items (numbered 1-5 in Table 3b). Seventy-six percent (107/141) of respondents listed at least one of the problems outlined by McCauley and Swisher (1984b) and 30% (42/141) identified two such problems.

Some of the benefits identified by clinicians (Table 3a) are both disconcerting and confusing. A large proportion of respondents (55%, 77/141) felt that using individual subtest items as a basis for establishing therapy objectives was an efficient way to identify age or developmentally appropriate goals. This is confusing

Table 3b
Reported Problems with Using Individual Subtest Items to Establish Treatment Goals

Problem categories **(1.-5. after McCauley and Swisher, 1984b; *n* = 121 of 124 or 97.6% of respondents who said "yes" to problems answered)**	**Number of clinicians**	**Percentage of clinicians**	**Percentage of responses**
1. Norm-referenced speech and language tests include a relatively small number of items and cannot sample all of the specific forms and developmental levels that may be appropriate. Because of the gaps in the skills assessed by available tests, they cannot describe all functional and relevant areas	66	54.5%	37.9%
2. Invalidation of the norm-referenced test as a test of ability, and a tool for reassessment ("Teaching to the test").	56	46.3%	32.2%
3. Norm-referenced tests assess behaviours only within a very restricted range of communicative contexts resulting in an incorrect characterisation of a child's functional communication skills	11	9.1%	6.3%
4. Individual errors and correct responses can have a number of explanations ranging from a momentary lapse of attention to a lucky guess- they may not represent true linguistic competence	9	7.4%	5.2%
5. Scoring systems fail to provide descriptive clues useful in therapy planning. Description of responses, both correct and incorrect, can help to distinguish different kinds and/or degrees of impairment	7	5.8%	4.0%
6. Treatment may not generalize (7)	7	5.8%	4.0%
7. Other: e.g.,- "invalid," "inaccurate," "not reliable," "not following standardized procedure," "overlooks learning concepts," "skills do not relate to a language model, " "reduces clinical practice to rote procedure"	18	14.9%	10.4%

Notes. 53 of the respondents identified two distinct categories of problems as defined in this table. Forty-two respondents identified two of the problems cited by McCauley and Swisher.

given that almost half of these same clinicians (34/77) also recognized that one problem with this practice is tests "include a relatively small number of items and cannot sample all of the specific forms and developmental levels that may be appropriate." Of greater concern is that 24% (28/118) of respondents felt that easy reassessment was one of the most important benefits of using individual subtest items. Again, approximately half of the aforementioned respondents (13/28) also identified "teaching to the test" or "invalidates reassessment" as a problem. More probing may be necessary to further elucidate these issues.

Question 3: The Use of Profiles to Compare Performance across Language Components

Almost half (48.9%) of the respondents reported using profiles to compare language performance across components "sometimes," 28.1% "frequently," 7.2% "always," and 15.8% "never." Many clinicians (89.3% or 117/131) felt there were benefits to the aforementioned practice. They were asked to list two of the most important (summarized in Table 4a). In general, clinicians felt the practice to be advantageous in that profiles quickly capture a holistic picture of language functioning that can be used to facilitate clinical decision-making or to give feedback to others. Respondents were also asked whether they felt there were problems associated with the use of profiles. Of 144 clinicians, 117 answered this question. Of those, 52.1% (61/117) felt there were problems associated with using profiles to compare clients' competency across language components. They were asked to list two of the most important problems. The problems identified by respondents are summarized in Table 4b. Fifty percent (59/117) of all respondents identified problems associated with profile use, although only 61% (36/59) of these problems stated were appropriate and specific to profiles, as opposed to norm-referenced tests in general (Table 4b, categories 1, 2, 3, and 6). Few (13.6%) identified the criticism highlighted by McCauley and Swisher (1984b; Table 4b, category 1) concerning the difficulty establishing that a statistically significant difference exists between components.

Question 4: Importance of Specific Tools for Measuring Treatment Progress

Ninety-four percent (134/143) of respondents reported measuring treatment progress. Some of the few who didn't (6%, 9/143) annotated their response with comments that they were involved only in screening or were at short-term acute-care facilities. Overall clinicians ranked repeated informal testing as being most important, though not significantly more important

Table 4a
Reported Benefits of Using Profiles to Compare Performance Across Language Components

Category (*n* = 109 of 117 or 93% of those surveyed, who said "yes" answered)	**Number of responses**	**Percentage of clinicians**	**Percentage of responses**
The profile allows the clinician to make a quick/easy comparison of strengths and weaknesses to establish appropriate treatment goals	74	67.9%	48.7%
Provides visual feedback to parents, teachers or others that shows strengths as well as weaknesses and it is easy for them to understand. It can help teachers adapt the curriculum and determine the appropriate level for the child	39	35.8%	25.7%
Provides a complete/holistic picture, a summary of comprehension and expressive abilities that gives an impression of what actual communication ability might be like	20	18.3%	13.2%
Allows the documentation/ measuring of treatment progress	5	4.6%	3.3%
Allows clinician to determine differential diagnosis ("Is this is a delay or a disorder?")	5	4.6%	3.3%
Other examples: "to secure funding," "to build upon strengths," "easily compare language to other areas"...	9	8.3%	5.9%

Table 4b
Problems Associated with the Use of Profiles to Compare Across Language Components As Identified by the Respondents

Category **(*n* = 59/61 or 96.7% of respondents who said "yes" answered)**	**Number of respondents**	**Percentage of clinicians**	**Percentage of responses**
1. Two scores within a test profile can seem quite different to one another due to measurement error rather than to real differences in the behaviours being measured or they may not be independent (McCauley& Swisher, 1984b)	8	13.6%	11.1%
2. Not a complete picture - too little room for within sub-test analysis, not detailed enough, need to consider other sources of information	18	30.5 %	25%
3. Profiles lack (construct) validity. "not valid," "can't break language into components that easily," "profiles aren't as comprehensive or balanced as might be suggested," "some subtests test other (sometimes non-language) abilities"	14	23.7%	19.4%
4. Need to investigate language use in context ("not functiona")	10	16.9%	13.9%
5. Too complicated for feedback to others and thus, easily misinterpreted	5	8.5%	6.9%
6. Too much testing required-takes too long	3	5.1%	4.2%
7. Norms aren't accurate or clinician lacks confidence in the norms	3	5.1%	4.2%
8. Other e.g., "can't capture treatment progress," "areas get missed," "sometimes an overall score may be more representative"	10	18.6%	15.3%

than repeated standardized tests ranked second. Both were significantly more important than repeated formal criterion referenced measures ranked third. Forty-six clinicians (39%) identified "other" tools as being of importance in measuring treatment progress. While this category was regarded as least important overall, those respondents who considered other measures as being useful, valued them highly. On the whole, these 46 respondents listed tools (e.g., "information from significant others" (29%), "observations in context" (24%), and "language sample analysis" (21%), and "treatment probes" (21%), which they had already ranked in Question 1 of the survey (Table 2).

It is clear from more informal comments that clinicians tended to measure treatment progress in more than one way. Of the respondents, 128 ranked at least one form of test as being important to measure treatment progress. Five respondents chose tools only from the "other" category and 11 chose not to answer. Approximately 40% of respondents (51/128) chose criterion-referenced tests (either formal and/or informal) as the one and, when a second tool was listed, two most important measures of treatment progress. The foregoing would be consistent with the recommendations of McCauley and Swisher (1984b), who maintain that criterion-referenced measures are designed to be sensitive to subtle changes in ability,

Table 5a
Reported Benefits of Using of Age-Equivalent Scores to Summarise Test Results

Category (*n* = 101 of 108 or 93.5% of respondents who answered "yes" to benefits)	# of clinicians	Percentage of clinicians	Percentage of responses
1. Feedback to parents, teachers and other team members	78	77.2%	61.4%
2. Allows comparison to peers or with competency in non-language areas	17	16.8%	13.4%
3. "To give an indication of severity"; "as a starting point"; " to get an idea of which other tests to administer"; "to determine treatment candidacy"; "as a gross measure of language ability"	9	8.9%	7.1%
4. To secure funding	7	6.9%	5.5%
5. To measure treatment progress	7	6.9%	5.5%
6. Helpful when norms don't apply as in the case of a disordered or severely delayed individual outside the age range of the norms	5	5.0%	3.9%
7. Other	4	4.0%	3.1%

whereas norm-referenced tests, by their very nature, are designed to examine "gross, relatively stable behaviour patterns." (p. 346) Many respondents (52%, 66/128), however, ranked standardized testing in combination with one of the criterion-referenced measures as the two most important test measures and 8% listed only standardized tests to measure treatment progress.

Question 5: Use of Age-Equivalent Scores to Summarize Test Results

Respondents were asked to identify the frequency with which they use age-equivalent scores to summarize test results. About one quarter reported doing so "always" or "frequently" (6% and 21% respectively), almost half (47%) "sometimes," and one quarter (26%) "never." Consistent with these numbers, 76% of respondents (108/142) felt there were benefits to this practice. Respondents were asked to identify two of the most important benefits (Table 5a). The most frequently mentioned advantage was that it was a useful reference when providing feedback to parents, teachers, and other team members. Seventy-four percent of respondents (58/78) who identified "feedback to parents and/or teachers" as a benefit, listed this as the only benefit. One might conclude that these respondents use them primarily for feedback to significant others. Other benefits identified by respondents included "securing funding," "helpful when norms don't apply," and "a gross or initial measure to determine severity and to guide further testing," all of which can be considered appropriate responses. Questionable responses regarding the benefits of age-equivalent score use, including those which could result in conveying misleading information to parents/ teachers or other professionals according to the arguments put forth by McCauley and Swisher (1984b), were listed by 24% (24/ 101) of those who felt that age-equivalent scores are useful. These "inappropriate" benefits included using age-equivalent scores for "comparison to peers" or "measurement of treatment progress." An apparent exception to this warning about the use of age-equivalent scores would be in the case when a child is severely delayed and exceeds two standard deviations from the norm for their age, making standard scores of little use. This latter point was mentioned by 5% of respondents. While this may seem a more appropriate response, one would have to ask the purpose of these scores. They would be superfluous to screening or diagnosis and would presumably be used for profiling or measuring treatment progress.

Respondents were also asked whether they felt there were problems associated with the use of age-equivalent scores. Most, 96.3% (132/137), felt there were. Respondents were asked to list two of the most important problems (Table 5b). Over half (65/131) listed at least one of the points enumerated by McCauley and Swisher (1984b) regarding age-equivalent scores (Table 5b, Categories 1 through 5). Most of the additional problems listed are relevant and valid. Category 7 is appropriate for many norm-referenced tests but is not specific to age-equivalent scores. Many respondents (59/131) provided responses that fell into Category 9 ("other"). Many of these responses (47/59) included enigmatic responses such as "misleading," "misinterpretation," "not valid," "not reliable," and "not accurate" with no· real elaboration. It is very difficult to interpret these answers, but it would appear that a large proportion of respondents consider age-equivalent scores to be problematic.

Table 5b
Reported Problems Associated with the Use of Age-Equivalent Scores to Summarise Test Results

Category (*n* = 131 of 132 or 99.0% of respondents who answered "yes" to problems)	**# of clinicians**	**Percentage of clinicians**	**Percentage of responses**
1. If considerable delay is common for normal children within an age group, even a large age-equivalent delay may not imply language delay or impairment (McCauley & Swisher 1984b)	33	25.2%	19.8%
2. The lower age score obtained by an older child, does not justify the inference that s/he has the language typical of a child of that age. It does not reflect the experience with language and world-knowledge that the older child would have. (McCauley & Swisher 1984b)	16	12.2%	9.6%
3. Summarizing test scores with standard scores or with percentile ranks is preferable (McCauley & Swisher 1984b)*	14	10.7%	8.4%
4. As age increases, similar differences in age-equivalent scores are the result of smaller and smaller differences in raw scores. As a result the reliability of age equivalent scores is poorer for developmentally more advanced test takers (McCauley & Swisher 1984b)	4	2.4%	3.1%
5. Age-equivalent scores are often calculated by interpolation between ages for which data were collected. This may involve assumptions about the continuity of language development that cannot be justified (McCauley & Swisher 1984b)	0	0%	0%
6. Not useful or constructive when used as feedback to others. It may be too threatening or scary to parents. It can evoke negative labelling, pessimism or low expectations. It is not useful in helping others establish goals or understand treatment progress	28	21.4%	16.8%
7. Norms aren't appropriate or not available	7	5.3%	4.2%
8. Not meaningful if it is a disorder rather than a delay	6	4.6%	3.6%
9. Other: Age-equivalent scores are misleading (no elaboration) or not meaningful/ not sensitive/ not accurate/ not valid/ not reliable/ not representative/ not detailed enough/ don't give whole picture/...	59	45.0%	35.3%

Notes. * This isn't as much a "problem" as a "recommendation" made by McCauley and Swisher. Thirty-six respondents listed two distinct problems as defined above. Two respondents listed two of the problems as identified by McCauley and Swisher (1984b).

Question 6: Clinician Confidence and Psychometric Knowledge

Respondents were asked to rate how confident they were that their own psychometric knowledge allows them to evaluate tests adequately. Approximately 17% (24/143) reported feeling "completely confident" in their knowledge, 66% (94/143) reported being "somewhat confident," and 17% (25/143) reported being "not confident." It would appear, therefore, that most clinicians feel they have some knowledge that allows them to evaluate tests, but lack full confidence.

Survey responses were examined to establish whether a relationship existed between clinician's own professed level of confidence in their psychometric knowledge (Question 6) and their ability to identify problems associated with use of individual subtest items to establish treatment goals (Question 2), use of profiles to compare performance across language components (Question 3), or the use of age-equivalent scores to summarize test results (Question 5). The chi-square tests of independence indicated there was no evidence in the present data to suggest the existence of such a relationship.

Ninety-two percent of respondents (132/144) reported that they had received training in the psychometric properties of tests. Respondents were also asked to identify the sources of their psychometric knowledge. Most of the clinicians (86%) received training in the psychometric properties of tests in university courses. For the average respondent, this was 14 years ago. Further sources of knowledge included reading test reviews (62%) as well as books and articles (46%), discussion with peers (57%), and attending workshops (15%).

Respondents were also surveyed as to the frequency with which they engaged in activities that increased their psychometric knowledge. Approximately half of respondents reportedly read journal articles, books, or test reviews at least once per year. Most clinicians reported they refer to the psychometric information pertaining to the tests they use at least once a year or whenever they purchased a new test. Few clinicians reported attending workshops dealing with standardized tests and psychometric concepts. Furthermore, those who did, reported doing so rarely.

Question 7: Use of Specific Standardized Tests

Respondents reported using individual standardized tests to assist in all clinical decisions (summarized in Tables 6 and 7). Frequency of use is highest for diagnosing

Table 6
Number of Respondents Working with Preschoolers who Reported Using Standardised Language Tests for Each Clinical Assessment Task

	Abbreviation	Screening	Differential Diagnosis	Lang. System	Tx Goals	Measuring Tx Progress
Assessment of Children's Language Comprehension-	ACLC	1	1	1		
Assessing Semantic Skills through Everyday Language	Asset		7	6	5	3
Bankson Language Test	BLT	13	3	4	3	4
Boehm Test of Basic Concepts-Preschool Version	BTBC-P	1	22	12	22	13
Bracken Basic Concept Scale	BBCS	3	18	10	21	8
Communication Abilities Diagnostic Test	CADeT					
Carrow Elicited Language Inventory	CELI		4	4	3	2
Clinical Evaluation of Language Fundamentals-Preschool	CELF- Pre	25*	85*	61*	56*	44*
Expressive One Word Picture Vocabulary Test	EOWPVT	25*	68*	40*	27	32*
Expressive Vocabulary Test	EVT	6	17	12	7	8
Houston Test for Language Development	HDLT					
Illinois Test of Psycholinguistic Abilities	ITPA		1	1	1	1
Kindergarten Language Screening Test	KLST	6	1	1		
MacArthur Communication Development Inventories	CDI	19	5	19	14	13
Miller-Yoder Language Comprehension Test	MLCT					
Northwestern Syntax Screening Test	NSST	1	1	2	1	1
Oral Written Language Scales: Listening Comprehension and Oral Expression	OWL		1	2	2	
Preschool Language Assessment Instrument	PLAI	6	23	28	25	12
Preschool Language Scale	PLS	43*	83*	71*	61*	44*
Peabody Picture Vocabulary Test	PPVT	36*	81*	50*	31*	36*
Receptive-Expressive Emergent language	REEL	32*	24	20	17	16
Rhode Island Test of Language Structure	RITL		1	1	1	1
Sequenced Inventory of Communication Developmen -R	SICD-R	7	16	16	9	7
Structured Photographic Expressive Language Test	SPELT	18	41	37	41*	32*
Test for Auditory Comprehension of language	TACL	13	53*	39*	29*	24
Test for Examining Expressive Morphology	TEEM	6	10	7	9	5
Test for Early Language Development	TELD	5	7	2	1	2
Token Test for Children -Revised	Token	7	26	14	9	4
Test of Language Development-Primary	TOLD-P	4	19	14	7	11
Utah Test of Language Development	UTLD		1	1	1	
Vocabulary Comprehension Scales	VCS				1	
Other		16	14	15	15	14

Notes. Test title represents all versions of that particular test. The five tests used most frequently for each task are annotated with an asterix.

Table 7: Part I
Number of Respondents Working with School-age Children who Reported Using Each Standardised Language Test for Each Clinical Assessment Task

Ages 6-12 (Elementary School) ***n* = 106**	**Abbreviation**	**Screening**	**Differential Diagnosis**	**Language System**	**Tx Goals**	**Measuring Tx Progress**
Assessing Semantic Skills through Everyday Language	Asset	1	9	11	9	1
Bankson Language Test	BLT	5	1	3	3	2
Bracken Basic Concept Scale	BBCS	3	17	11	19	10
Communication Abilities Diagnostic Test	CADeT		1	2	1	1
Carrow Elicited Language Inventory	CELI		2	3	1	1
Clinical Evaluation of Language Fundamentals	CELF	15*	72*	62*	58*	43*
Detroit Tests of Learning Aptitude	DTLA		6	3	3	1
Expressive One-Word Picture Vocabulary Test	EOWPVT	23*	49*	30**	23*	22*
Expressive Vocabulary Test	EVT	7	19	16	9	6
Houston Test for Language Development	HTLD			2	1	
Language Processing Test	LPT	3	36	22	27*	12
Illinois Test of Psycholinguistic Abilities	ITPA		1	1	1	1
Miller-Yoder Language Comprehension Test	MLCT					
Northwestern Syntax Screening Test	NSST					
Preschool Language Scale	PLS	13*	30	29	24	17
Peabody Picture Vocabulary Test	PPVT	35*	65*	43*	33*	29*
Structured Photographic Expressive Language Test	SPELT	11	37*	35*	38*	24*
Temporal Analysis of Propositions	TEMPRO	1	1	1	1	1
Test for Auditory Comprehension of Language	TACL	16*	44*	40*	35*	22*
Test for Examining Expressive Morphology	TEEM	2	7	12	10	7
Test of Language Competence-Expanded Edition	TLC-E	1	19	12	12	6
Test of Word Finding	TOWF	2	30	19	17	6
Test of Word Knowledge	TOWK	1	4	3	3	2
Test for Early Language Development	TELD	3	4	3	4	2
The Word Test-R: Elementary	Word	3	34	30**	24	9
Token Test for Children	Token	4	32	21	11	9
Test of Language Development -Intermediate	TOLD-I	2	15	8	6	6
Test of Problem Solving-Elementary	TOPS	1	32	30**	25	12
Utah Test of Language Development	UTLD					
Vocabulary Comprehension Scales	VCS					
Other		6	23	18	18	9

Notes. Test title represents all versions of that particular test.The five tests used most frequently for each task for elementary school-ages and two tests used most frequently for highschool ages are marked in bold with an asterix.

Table 7: Part II
Number of Respondents Working with School-age Children who Reported Using Each Standardised Language Test for Each Clinical Assessment Task

Ages 13-19 (Junior/Senior High) ***n* = 50**	**Abbreviation**	**Screening**	**Differential Diagnosis**	**Language System**	**Tx Goals**	**Measuring Tx Progress**
Adolescent Language Screening Test	ALST	1				
Clinical Evaluation of Language Fundamentals	CELF	8*	35*	32*	26*	14*
Detroit Tests of Learning Aptitude	DTLA	1	2		1	
Expressive One-Word Picture Vocabulary Test-Upper Extension	EOWPVT	5	14	4	4	5
Expressive Vocabulary Test	EVT	4	8	7	3	3
Fullerton Language Test of Adolescent	FLTA		2		1	
Interpersonal Language Skills Assessment	ILSA					
Peabody Picture Vocabulary Test-	PPVT	15*	32*	21*	15*	11*
Screening Test of Adolescent Language	STAL	1				
Temporal Analysis of Propositions	TEMPRO					
Test of Adolescent Language	TOAL		8	7	5	2
Test of Adolescent/Adult Word Finding	TAWF		3	1	1	
Test of Language Competence:	TLC	1	13	11	7	2
Test of Problem Solving-Adolescent	TOPS		10	8	6	1
Test of Word Knowledge	TOWK	1	1	1	2	1
The Word Test-Adolescent	Word Test: A.		9	11	6	2
Other		1	3	5	4	3

Notes. Test title represents all versions of that particular test.The five tests used most frequently for each task for elementary school-ages and two tests used most frequently for highschool ages are marked in bold with an asterix.

the presence and severity of a deficit, describing the child's language system and establishing treatment goals. This was true whether a specific test was designed to describe a language system ,which a limited number are (e.g., MacArthur Communication Development Inventories, Fenson et al., 1993), or not (e.g., Expressive One-Word Picture Vocabulary Test, Gardner, 1990). These results are consistent with the high-ranking position given to standardized tests for these purposes as revealed by Question 1 of the survey. Fewer clinicians reported using standardized tests for screening and measuring treatment progress. This is also in keeping with responses to Question 1 of the survey which indicate a less dominant ranking of standardized tests for these two clinical decision tasks.

These findings are consistent with those of Huang et al. (1997) who reported that while 7% of their survey respondents used tests exclusively for placement (i.e., for determining eligibility for service), a further 74% used standardized tests for other assessment tasks as well. Fifty percent of respondents in their study used standardized tests for screening while 91% used them for establishing treatment goals and measuring treatment progress. Thirty-five percent of respondents in the Huang et al. study used standardized tests for all four traditional assessment tasks: determining the existence and general areas of deficit, describing the language system, establishing intervention goals, and measuring treatment progress.

It is worth noting that a small number of tests dominate for each age group (annotated with an asterix in Tables 6 and 7), and that, in general, the same few tests are used for all tasks. The top five standardized tests for each clinical decision task for preschool and elementary school-aged children accounted for 45% to 65% of all the tests reportedly used. The same was true for the top two tests for each clinical decision task in the junior high/ high school age category. These results are consistent with those of Huang et al. (1997). Indeed, all of the tests that are ranked in the top five (or two) in this study are included in the top ten of Huang et al., and most are ranked in the top five. The most frequently used tests for the preschool and elementary school age categories are also consistent with those found to be most frequently used by California SLPs in the survey of Wilson, Blackmon, Hall, and Elcholtz (1991).

While the ranking of the importance of standardized tests for various clinical decisions is lower for preschool populations than for school-aged populations (Question 1), the proportion of clinicians selecting standardized tests for each of the tasks (Question 7) is at least as large for the preschool population as for the school-age populations. One might conclude, therefore, that standardized test use is comparable for the two populations but that other factors, such as observations, language sample analysis, and information or feedback from parents, are given greater relative importance with regard to preschool populations.

Question 8: Use of English Language Tests with Non-English Speaking Children

Of respondents, 45% (61/137) reported using one or more of the English tests listed in Question 7 of the survey with children whose first language was not English. Of these, 43% (26/61) reported using adapted tests, 64% (39/61) reported using translated tests, 3% (2/61) reported using local norms, and 43% (26/61) reported using original norms.

The motives respondents gave for using English tests varied. A lack of alternative measures was cited by 71%. Twenty-seven percent of respondents who used English tests with non-English speaking clients did so in order to establish appropriate programming. Approximately a quarter (25.4%) of clinicians specifically reported using these tests in a criterion-referenced manner. Furthermore, 54% (33/61) implied using these tests in a criterion-referenced manner as they reported using neither local nor original norms. Nonetheless, 43% (26/61) of these clinicians reported using standardized tests with the original norms for non-English speaking children - an unreliable method of determining the existence of language impairment (Lahey, 1988).

Discussion

Demographically, the respondents appear to be representative of a cross-section of English-speaking Canadian SLPs serving a paediatric population. The survey's length (10 pages which took approximately 40 minutes to complete) could have been responsible for the low return rate of surveys (28%). Response rates for similar surveys range from 49% (Huang et al., 1997) or 53% (Wilson et al., 1991) to 72% (Potter & Lagacé, 1995) or 82% (Dohan & Schulz, 1999). As respondents were essentially self selected, those with less psychometric knowledge may have found the survey even more demanding and thus may have been less likely to complete it. Therefore, as a group, the clinicians who returned this survey may have an above average knowledge of psychometric principles. This is conjecture as, clearly, there is no information available to establish the extent of an individual respondent's knowledge. If true however, the results of this study could overestimate the knowledge of the average English-speaking Canadian SLP. Furthermore, the nature of this survey is such that subsequent survey questions may have cued the answers to previous ones. If so, the result would be to understate the extent to which norm-referenced tests are "misused" by Canadian SLPs. As such, the results of this study lead to conclusions that in some cases give cause for concern.

Standardized tests were ranked among the most important tools used to address clinical decisions by the SLPs sampled (Question 1). This was particularly true for clinicians working with school-age populations, who ranked standardized tests as the most important tool for diagnosing the presence and severity of a deficit, describing a child's language system, and for establishing treatment goals. Of these, only "diagnosing the presence and severity of the deficit" is a task for which most standardized tests are designed. While standardized tests were ranked of lesser importance for preschool populations, the proportion of clinicians selecting standardized tests for each of the assessment tasks (Question 7) was comparable for preschool and school-age children. Thus, standardized tests play a key role in clinical practice especially for school-age populations. It is therefore important to examine the manner in which these tests are being used and whether the caveats raised by McCauley and Swisher (1984b) are being heeded.

A sizable proportion of the respondents are aware of problems associated with the use of individual subtest items to establish treatment goals, yet use them for this purpose nonetheless (Question 2). It may be that a large proportion of clinicians, while aware of the criticisms, fail to believe them. Alternatively, while the vast majority of respondents use individual subtest items as a basis for establishing therapy objectives, it may be that they do not use them as the sole basis for this task. This might also be inferred from the listed benefits of "gives a starting point from which to probe further" and "may be beneficial when applied with clinical judgement such as confirmation for observations in other tests." Both of the aforementioned benefits, as well as responses to Question 1, would indicate that the clinicians are using informal and criterion-referenced measures in conjunction with the standardized tests. A combination of approaches would be consistent with the suggestions of Haynes and Pinzola (1998) and Huang et al. (1997). It may seem that this is contrary to the previously mentioned admonitions of McCauley and Swisher (1984b) who claimed that there were no circumstances under which it would be appropriate to use norm-referenced tests to determine treatment goals, but most certainly they meant exclusive use of norm-referenced tests. When making any clinical assessment decision it behoves clinicians to use absolutely all information available. This will often include norm-referenced tests, criterion-referenced tests, and informal probes of the clinician's own devising. To neglect any one source of information would be illogical. Nonetheless, the greatest concern remains for those 24% of clinicians, who thought that using individual subtest items to establish treatment goals was beneficial because it made reassessment easy. This group may not only be choosing inappropriate intervention goals, but may also be misjudging the progress made on those goals through such practices.

Many of the responding SLPs, approximately 80%, indicated they used profiles (Question 3). When comparing subtest scores of profiles, it is important that clinicians be aware that the minimum difference required to reflect a true disparity is a function of the reliability of each subtest and the correlation of the subtest scores to be compared. The statistical independence of subtest scores becomes particularly important when each of the subtest scores consist of different combinations of the same array of test items. It would appear that few respondents were aware of the need to establish this minimum difference. The authors feel that this may be due to the more abstract, less intuitive nature of profiling problems. It requires a more formal psychometric understanding that extends beyond "common-sense" clinical knowledge. The consequence of such errors would be that a clinician might choose to concentrate efforts on improving an erroneously identified "weaker" area to the neglect of the area "of greater competence." Thus, intervention goals may not be optimally chosen.

Many respondents chose to use standardized tests to measure treatment progress even though they are less sensitive measures (Question 4). Perhaps clinicians establish treatment goals based upon standardized tests and without administration of criterion-referenced measures and then continue this practice to measure progress because the standard for comparison has already been established. This would be consistent with the fact that respondents ranked standardized tests as significantly more important than criterion- referenced tests for the establishment of intervention goals with school-age clients, but criterion-referenced procedures did not differ significantly from standardized tests as tools for measuring treatment progress for any population (Question 1). While virtually all clinicians measure treatment progress, it would seem only 40% do so based solely on criterion-referenced measures. Nonetheless, a large proportion (92%) use criterion-referenced measures to some degree and, for many, these measures are of primary importance. In contrast to the standardized tests, which, as previously mentioned, lack the number and variety of items necessary to monitor treatment progress, criterion-referenced procedures can be designed to probe in detail those specific skills chosen as intervention goals. Few (8%) chose only standardized tests as being of importance (Question 4). The latter would be least advisable.

Respondents were asked to report the frequency with which they measure treatment progress. It was apparent from responses that the question had been unclear. Many respondents explicitly stated so, and annotated their reply with indications to the effect that they used informal criterion-referenced measures very frequently, and more formal measures (not specified whether criterion-referenced or standardized tests) every six months to a year. If periods between successive standardized test administrations are sufficiently large and there is no "teaching to the test," the risk of "learning the test" should be minimized. Furthermore, this may allow time for sufficient progress to be made that might be measured by the less sensitive standardized tests. However, as the time period between successive test administrations increases, the issue of development independent of treatment progress clearly becomes important and adjustments must be made. This mandates the use of age-adjusted standard scores, although interpretation of results may be difficult if the skills tested at the older age level are not related to those addressed in intervention. Thus, ideally, interpretation of these standardized test results should be limited to the continued existence of a language impairment or the absence thereof (McCauley & Swisher, 1984b). In many cases this may well be the motivation for assessing treatment progress - to answer the question of whether to continue therapy or to discharge. This does not negate the fact that the 8% of clinicians who are only using standardized tests to monitor progress are left with no clear indication of the efficacy of their treatment.

Most clinicians felt there were problems with age-equivalent scores, while only three-quarters felt there were benefits (Question 5). Consistent with this, 26% of clinicians reported that they do not use age-equivalent scores. Furthermore, one can infer that 54% of clinicians (58/108) use them mainly for feedback to team members, parents, and teachers. One might conclude that, for these latter clinicians, the use of age-equivalent scores to summarize test results is not adversely influencing their own decisions regarding language impairments. Nonetheless, the information they are providing as feedback may be misleading others by misrepresenting the nature of language impairments, language development, and treatment progress. As such, they run the risk of minimizing the delay in children who are younger, when development is rapid and small differences in age-equivalency relate to large differences in development, and being overly alarmist about delays in older children for whom the converse is true. The remaining 20% of clinicians, despite listing problems associated with age-equivalent score use, are using age-equivalent scores to shape their own clinical decisions. These SLPs may be underidentifying children who are younger and over-identifying older children. Fifteen percent of respondents in the Huang et al. (1997) study used age-equivalent scores more frequently than any other measure. The findings of the present study would appear to be consistent with those of Huang et al.

It would appear that most clinicians lack full confidence in their ability to evaluate the psychometric properties of tests (Question 6). While test manuals and test reviews are important sources of knowledge, books and journal articles, as well as workshops on test use and evaluation, could be sources of more general information. Clearly, there is room to increase the exposure that clinicians have to these more general sources. For those who lack confidence, these could be useful complements to test manuals and reviews that are more test specific. Because of small sample size, some of the conclusions relating to the independence of clinicians' self-reported confidence and ability to identify problems associated with using standardized tests are tentative and require further verification. Nonetheless, it is noteworthy that clinicians' reported self-confidence did not appear to reflect their own ability to identify the problems with the use of individual subtest items to establish treatment goals, profiles to determine patterns of impairment, or age-equivalent scores to summarize test results. This may mean that their own assessment of their abilities is unreliable. The ability of clinicians to identify the extent of their own psychometric knowledge warrants further investigation. If clinicians are unable to identify their own knowledge or lack thereof, they will be unlikely to avail themselves of opportunities to increase that knowledge.

Results indicate that predominant uses of standardized tests are not always those intended by test developers (Question 7). Standardized tests are suited to determining the presence of a language impairment and diagnosing the general nature of that impairment (McCauley & Swisher, 1984a; Merrill & Plante, 1997). Respondents, however, are using a small selection of standardized tests to form a basis for many assessment decisions. The number of clinicians reportedly using each test to diagnose the general areas of impairment - a use for which standardized tests are designed - is usually greater than those who report using it for other clinical decision tasks. This is invariably true for the most frequently used tests. Nonetheless, a large number of clinicians also reported using these tests to describe the language system, establish treatment goals, and measure treatment progress. This situation appears far from ideal. Further investigation may be necessary to reveal the reasons for this. Possibilities include the following: (a) clinicians are unaware of the implications of their actions, (b) clinicians feel they lack the time to administer other types of assessment measures, (c) clinicians are unable to, or lack confidence to, design criterion-referenced assessments tailored to the client and situation, (d) few appropriate formal criterion-referenced materials are available, and/or (e) the quantitative nature of standardized tests give the illusion of greater rigour and credibility. In reference to this last possibility, there is increasing interest in more descriptive methods of language assessment and criterion-referenced measures and in establishing the psychometric validity and reliability of these procedures (Damico, Secord, & Wiig, 1992; McCauley, 1996). It is important to dispel the notion that criterion-referenced procedures are inherently less valid and/or reliable than norm-referenced procedures.

While the composition of non-English speaking populations varies, the problem of assessing such children, in increasingly linguistically and culturally diverse communities, is widespread. The number of respondents in the present study who use English-language standardized tests with children whose first language is not English is small and results should be interpreted with caution. It is worth noting, however, that the percentage of respondents who reported using English standardized tests with non-English children (Question 8), is consistent with the percentage of respondents in the survey of Huang et al. (1997) who were concerned with the lack of materials to assess non-English speaking children. A large proportion of respondents (43.3%) in the present study, who reported using standardized tests to assess children whose first language is not English, use the original norms to do so. This is an unreliable method of determining a language impairment for these children. There is a need for tools designed or carefully adapted for assessing multi-cultural and linguistically diverse populations (Garcia & Derocher, 1997). Further research is needed into the manner in which these tests are being used. It could be that these tests are being used to exclude the possibility of impairment rather than confirm the existence thereof. Assuming that non-English speaking children would typically perform below the norms for their English-speaking peers, then performance within normal range might be a crude technique for imputing the absence of language impairment. As such, this might be considered a valid application. If these tests are being used to establish the existence of an impairment, however, further education in the appropriate use of normative assessments may be warranted.

Summary and Conclusions

This survey examined current use of standardized language tests, measurement practices, and psychometric knowledge among CASLPA members working with preschool and school-age populations. Clinician's awareness of the problems associated with using individual subtest items to establish treatment goals, the difficulties associated with the reliable use of profiles to establish patterns of impairment, and the use of age-equivalent scores to summarize test results is variable. It would appear the concerns raised by McCauley and Swisher (1984b) remain valid. When interpreting the implications of these results, it is nonetheless important to understand that the choice of intervention goals are often the result of a dynamic process. Assessment of appropriate goals is frequently interactive, iterative, and ongoing. These goals are often changed as probes establish the areas in which success is or isn't possible and as interactions in the context of therapy reveal the extent of further deficits. The consequences of choosing inappropriate goals based on the misinformation from profile interpretation or inadequate information from use of individual subtest items may be mitigated by this fact.

A large percentage of clinicians use criterion-referenced tools to some extent to measure treatment progress. As such, it is possible that inappropriate measures of this progress may not have a large impact on the treatment for any one client. Nonetheless, it does compromise the ability of clinicians to evaluate the efficacy of therapy approaches and to accurately document progress. This reduces our ability as practitioners to optimize the therapy we provide and ultimately it undermines our credibility as a profession.

Clinical practice is formed by decisions that are unavoidably based upon measurement. These measurements may range from formal standardized tests to informal treatment probes and subtle clinical observations (McCauley, 1989). Specific assessment tasks must be matched to appropriate measurement tools. The results of this survey would indicate that clinical measurements, and hence clinical decisions, are often not optimally made. Regardless of the reasons for the above listed "misuses" of norm-referenced tests, clinicians could benefit from increasing their knowledge of measurement principles and optimizing their implementation of those principles. Most clinicians are not fully confident in their own ability to evaluate the psychometric properties of tests. The concepts underlying these psychometric principles are not complex. They must be taught convincingly to new graduates of speech-language pathology as well as reiterated and presented in an accessible fashion to those currently practising. There is potential to increase the exposure that clinicians have to journal articles and workshops that deal with psychometric principles. Clinicians, too, must assume responsibility in learning as much as possible about appropriate use of the instruments at their disposal.

Author Notes

Please address all correspondence to Alanna Kerr, School of Human Communication Disorders, Dalhousie University, 5599 Fenwick Street, Halifax, Nova Scotia B3H I R2; email: akerr@navnet.net.

References

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

Damico, J. S., Secord, W. A., & Wiig, E. H. (1992). Descriptive language assessment at school: Characteristics and design. In W. A. Secord (Ed.), *Best practices in school speech language pathology* (p. 1 - 8). San Antonio, TX: The Psychological Corp., Harcourt Brace Jovanovich.

Fenson, L., Dale, P., Reznick, S., Thal, D., Bates, E., Hartung, J., Pethick, S., & Reilly, J. (1993). *The Macarthur Communicative Development Inventories.* San Diego, CA: Singular.

Garcia, L. J., & Derocher, A. (1997). L'evolution des troubles du language et de la parole chez l'adulte fancophone. *Journal of Speech-Language Pathology and Audiology, 21,* 271-293.

Gardner, M. (1990). *Expressive One Word Picture Vocabulary Test-Revised.* Novato, CA: Academic Therapy.

Haynes, W. O., & Pinzola, R. H. (1998). *Diagnosis and evolution in speech pathology.* (5th ed.). Needham Heights, MA: Allyn Bacon.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods.* New York: John Wiley and Sons.

Huang, R., Hopkins, J., & Nippold, M. A. (1997). Satisfaction with standardized language testing: A survey of speech language pathologists. *Language, Speech, and Hearing Services in Schools, 28,* 12-29.

Lahey, M. (1988). *Language disorders and language development.* New York: American Macmillan.

Lahey, M. (1990). Who shall be called language disordered? Some reflections and one perspective. *Journal of Speech and Hearing Disorders,* 33, 612-660.

McCauley, R. J. (1989). Measurement as a dangerous activity. *Journal of Speech-Language Pathology and Audiology,* 13, 29-32.

McCauley, R. J. (1996). Familiar strangers: Criterion-referenced measures in communication disorders. *Language, Speech, and Hearing Services in Schools,* 27, 122-131.

McCauley, R. J., & Demetras, M. J. (1990). The identification of language impairment in the selection of specifically language-impaired subjects. *Journal of Speech and Hearing Disorders, SS,* 468-475.

McCauley, R. J., & Swisher L. (1984a). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49,* 34-42.

McCauley, R. J., & Swisher, L. (1984b). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders, 49*, 338-348.

Merrill, A. W., & Plante, E. (1997). Norm-referenced test interpretation in the diagnostic process. *Language, Speech, and Hearing Services in Schools, 28*, 50-58.

Owens, R., Haney, M., Giesow, V., Dooley, L., & Kelly, R. (1983). Language test content: A comparative study. *Language, Speech, and Hearing Services in Schools, 14*, 7-21

Petersen, N., Kolen, M., & Hoover, H. (1989). Scaling, norming, and equating. In Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221-262). New York: American Counsel On Education and Macmillan.

Potter, R. E., & Lagacé, P. (1995). The incidence of professional burnout among Canadian speech-language pathologists. *Journal of Speech-Language Pathology and Audiology, 19*, 181-156.

Salvia, J., & Yesselldyke, J. E. (1991). *Assessment* (5th ed.). Boston: Houghton Mifflin.

Wilson, K., Blackmon, R., Hall, R., & Elcholtz, G. (1991). Methods of language assessment: A survey of California public school clinicians. *Language, Speech, and Hearing Services in Schools, 22*, 236-241.

Manuscript received: June 15, 2001

Accepted: September 3, 2002

■ ■ ■