JSLHR, Volume 41, 834–845, August 1998

SCAN Test-Retest Reliability for First- and Third-Grade Children

Nathan E. Amos
Larry E. Humes
Indiana University
Bloomington

The SCAN is a popular screening test that was developed to provide a rapidly administered, standardized method for determining the potential of central auditory processing disorder (CAPD) in children between the ages of 3 and 11 years. It can be administered in 20 minutes with a portable stereo cassette player and contains three subtests: filtered words (FW), auditory figure ground (AFG), and competing words (CW). Published SCAN test-retest reliability data (Keith, 1986) used a 6-month retest interval and indicated that SCAN scores may be unreliable. No additional reliability data are available, and studies indicate that SCAN has been used by both researchers and clinicians despite reliability concerns. This investigation examined the stability of SCAN outcomes for 25 first-grade and 22 third-grade children (ages 6 to 9 years) using a 6- to 7-week retest interval. Time of day and examiner were held constant, and participants were normal-hearing, were Caucasian, and spoke English as their primary language. ANOVA outcomes indicated that both raw and standard scores improved significantly from Test 1 to Test 2 for two of the three SCAN subtests (FW and CW) and for the composite (COMP) score. Additionally, COMP-percentile-rank and age-equivalent outcomes demonstrated significant improvement from test to retest for both grades. The AFG subtest was the only SCAN measure for which a significant test-retest difference did not emerge. The highest test-retest correlation values (r) were moderately strong ($0.70 \leq r \leq 0.78$) and occurred for the CW and COMP scores. Implications of correlations and factor analyses are discussed. It is suggested that examiners base recommendations for additional testing, follow up, and remediation on the COMP score only. Further, it appears that second administration of the SCAN can provide a better estimate of an individual child's best performance, but lack of second-score norms confounds simple interpretation of such scores.

KEY WORDS: central auditory processing disorder (CAPD), screening, children, SCAN, test-retest reliability

Recently, the Task Force on Central Auditory Processing Consensus Development (1996) defined a central auditory processing disorder (CAPD) as an observed deficiency in one or more of the following behaviors: (a) sound localization and lateralization, (b) auditory discrimination, (c) auditory pattern recognition, (d) temporal aspects of audition, (e) auditory performance with competing acoustic signals, and/or (f) auditory performance with degraded acoustic signals. In addition, they reported that CAPD may result from dysfunction of the processes and mechanisms dedicated to audition and/or from more general dysfunction that affects performance across modalities, such as an attention or neural-timing deficit.

Children who have CAPD have been said to exhibit such characteristics as the following: deficits in the comprehension of speech in competing

 1092-4388/98/4104-0834

background noise, distractibility, reduced auditory attention, inconsistent awareness of auditory stimuli, poor concentration, and academic achievement lower than predicted by intelligence measures (Chermak & Musiek, 1992; Emerson, Crandall, Seikel, & Chermak, 1997). Willeford (1974) first demonstrated the potential value of administering CAPD tests to children. Early intervention for children who are suspected to have CAPD may be critical to academic success. However, there has been considerable debate over whether available tests can reliably and validly measure CAPD. Thus, identification of such children has proven a difficult task (Humes, Amos, & Wynne, in press; Task Force on Central Auditory Processing Consensus Development, 1996).

The SCAN screening test for auditory processing disorders (Keith, 1986) is a CAPD screening test. According to its developer, it may be used to identify potential underlying factors related to poor social skills, language use, and academic performance in children of 3 to 11 years of age (Keith, 1986). As an improvement over other such tests, SCAN was developed to provide a rapidly administered, standardized method for determining potential CAPD in children (Keith, Rudy, Donahue, & Katbamna, 1989). Its specific purposes were reported to be the following: (a) to determine possible central nervous system dysfunction by assessing auditory maturation, (b) to identify children at risk for auditory-processing or receptive language problems who may require additional audiological or language testing, and (c) to identify children who may benefit from specific management strategies (Keith et al., 1989).

SCAN appears to be one of the more thoroughly developed and frequently used tests of its type. It can be administered in 20 minutes using a portable stereo cassette player in any quiet room. SCAN includes three subtests: (1) filtered words (FW), consisting of 40 monosyllabic, low-pass filtered words presented monaurally (20 per ear); (2) auditory figure ground (AFG), consisting of 40 monosyllabic words (20 per ear) presented in a background of multitalker babble (+8 dB signal-to-noise ratio [SNR]); and (3) competing words (CW), in which 100 monosyllabic words are presented as 50 dichotic word pairs. Score outcomes are expressed as number correct and can be examined as raw scores, standard scores, percentile ranks, and age equivalents. Additionally, a composite (COMP) score is determined as the sum of the three subtest raw scores. Again, as outlined in the test manual, it has been suggested that SCAN subtest and COMP score outcomes can profile a child's auditory processing strengths and weaknesses (normal, developmentally delayed, or disordered auditory processing) and suggest potential areas for further assessment and/or remediation (Keith, 1986). For example, poor AFG performance would indicate difficulty hearing in background noise and potential maturational delay of a child's auditory system. Remediation may include management of the listening environment to enhance the teacher's voice and reduce background noise (i.e., improve the SNR).

Given the stated utility and the widespread use of SCAN (Chermak, Styer, & Seikel, 1995; Dietrich, Succop, Berger, & Keith, 1992; Emerson et al., 1997; McCartney, Fried, & Watkinson, 1994), thorough documentation of its reliability (i.e., the stability of test scores across repeated test administrations) is necessary. As with all tests used for research and clinical assessment, high test-retest reliability is desirable to properly interpret outcomes (American Psychological Association, 1985). When examining reliability, test-retest intervals generally range from a few days to a month (Nunnally, 1959) and usually involve a compromise between maximizing practice or memory effects (if too little time passes between tests) and minimizing maturational effects (if too much time passes). Stability of scores is often expressed as a correlation coefficient and/or as a difference in mean scores from test to retest (Humes et al., in press; Nunnally, 1959). The most desirable test-retest outcome would be a high correlation ($r > 0.80$) with no significant difference between means.

Test-retest data have been reported for SCAN by the test's developer (Keith, 1986). The data were obtained by retesting a subsample (37 first-grade and 31 third-grade children) of the original standardization group. Importantly, however, a period of 6 months elapsed between test and retest and no rationale for the interval is provided by the test developer (Keith, 1986). According to Nunnally (1959), this is probably an inappropriately long test-retest interval. With children such an interval may have allowed for a confounding factor, such as maturation, to affect performance. In addition, the average reported Pearson r test-retest correlation across all scores is about $r = 0.40$ (Keith, 1986), which is well below a targeted range of $r > 0.80$ (Nunnally, 1959). Thus, published data in the test manual have indicated that SCAN scores are unstable over time. The true test-retest reliability of SCAN is unknown (Humes et al., in press).

SCAN has been used recently to examine the effects of lead (PbB) exposure (Dietrich et al., 1992), cigarette smoke exposure (McCartney et al., 1994), and chronic otitis media (Emerson et al., 1997) on central auditory processing ability in children. In all cases, the effectiveness of remediation for individual children, or the progress of groups of children in follow-up investigations, would likely be evaluated using retest of SCAN. This approach, however, assumes knowledge of the test-retest reliability of SCAN.

To date, no additional SCAN test-retest data have been reported. It is apparent that SCAN has been used

by both researchers and clinicians and that its reliability concerns may have compromised the integrity of conclusions and recommendations (Dietrich et al., 1992; Emerson et al., 1997; McCartney et al., 1994). This highlights the need for additional reliability data on SCAN. The purpose of this study was to examine the test-retest reliability of SCAN for first- and third-grade children using a 6- to 7-week retest interval. This interval was chosen to minimize potential learning or maturation effects.

Method

Participants

Selection criteria conformed to the SCAN test manual guidelines (Keith, 1986). First, all participants passed an earlier hearing screening bilaterally at 20 dB HL (re ANSI, 1989) for frequencies of 1, 2, and 4 kHz. Second, because it is recommended that test results for minorities and non-native English speakers be interpreted with caution (Keith, 1986), participants were Caucasian, spoke English as their primary language, and were judged to perform at an age-appropriate academic level by their teachers. Additionally, to minimize potential confounding effects, children who exhibited frequent age-inappropriate articulation errors or considerable difficulty performing the task were eliminated from the study (1 child was eliminated). Sex was not constrained because SCAN is reported to yield similar outcomes for boys and girls (Keith, 1986). Finally, as in the published reliability study (Keith, 1986), participants included a group of first graders and a group of third graders, with ages ranging from 6 to 9 years (the primary range of ages appropriate for SCAN).

Children were available at a local elementary school, just outside of Bloomington, Indiana. The school is in a rural, middle-class setting and had a total student body of 460 children (including 75 first-grade and 69 third-grade students enrolled). Participants included 25 first graders (Grade 1) with a mean age of 6.6 years and 22 third graders (Grade 3) with a mean age of 8.6 years for whom the parents had provided informed consent for participation in the study. The children were given a small prize (a pencil) for their participation in the study.

Equipment and Procedures

The SCAN test was administered according to test manual guidelines (Keith, 1986). A high-fidelity Sony Walkman WM-FX221 with Sony MDR-013 stereo headphones was used to present the test items from the cassette tape. Proper headphone placement was ensured by the examiner. To minimize potential variability, the volume setting was fixed at a level of approximately 70 dB SPL for the FW test items, which resulted in a presentation level of 75 to 77 dB SPL for the AFG and CW test items. All SPL presentation levels are referenced to sound levels measured with ER-11 microphones mounted in occluded-ear simulators (Zwislocki couplers) on the Knowles Electronic's Manikin for Acoustic Research (KEMAR). Presentation levels were checked at the initial test (Test 1) and at retest (Test 2). All testing was completed at the school in isolated test rooms where ambient noise measurements taken periodically averaged 64 to 65 dB (C weighting) and less than 50 dB (A weighting).

During testing, each child was seated face-to-face with the examiner, and the scoring form was out of the child's view (cf. Keith, 1986). Both test instructions and practice items were included on the SCAN recording, and clarification was provided only when necessary. The entire test took about 20 minutes per child, and the presentation order of subtests was always FW, AFG, then CW (dichotic). As recommended, breaks were not provided, nor was the tape stopped once a subtest had begun (Keith, 1986). Participants who demonstrated considerable difficulty understanding or performing the task were deemed unreliable and eliminated from the study. Only one such child, a third grader, was eliminated, and this resulted in a total of 47 participants who completed both Test 1 and Test 2.

A period of 6 to 7 weeks elapsed from test to retest for all participants. Also, to minimize potentially confounding effects, both examiner (one CCC-A audiologist and one research assistant, both familiar with SCAN) and time of day (within 1/2 hour, during the regular school day) were held constant across all participants from Test 1 to Test 2.

Data

As stated previously, raw scores are in number correct, and the available appendixes of the SCAN test (Keith, 1986) can be used to convert raw scores to both standard scores and percentile ranks for each subtest. In addition, the COMP score (sum of the three subtest raw scores) can be used to determine age equivalence. Thus, there were numerous data points per individual to compare across Test 1 and Test 2.

Given the data, there was no intention in this study to determine whether individuals "passed" or "failed" the SCAN test. Rather, analyses of variance were conducted to examine potential effects of age or grade (first, third) and test (1, 2) on the FW, AFG, CW, and COMP raw, standard, percentile-rank, and age-equivalent scores. These analyses indicate whether the average first and third graders differed significantly in performance and, more importantly, whether scores changed significantly from Test 1 to Test 2 for either grade. In addition,

Table 1. Grade 1 and Grade 3 means and standard deviations for Test 1 and Test 2 subtest and composite raw scores on SCAN.

		Subtest and composite scores							
		FW		AFG		CW		COMP	
Grade		Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
1	*M*	33.3	35.3	34.6	35.0	77.6	84.2	145.6	154.5
	SD	2.7	1.8	2.3	1.7	8.8	5.7	10.2	6.9
3	*M*	35.3	36.9	35.5	35.6	82.3	86.4	153.0	158.9
	SD	2.1	1.9	2.2	2.0	6.9	5.7	7.9	7.3

Note. FW = filtered words subtest; AFG = auditory figure ground subtest; CW = competing words subtest; COMP = composite.

correlational and regression analyses were performed to examine the consistency of individual data. Low correlations, for example, would indicate that individual scores can fluctuate on retest, which would imply poor reliability and warrant caution by clinicians when making recommendations for specific individuals. Finally, factor analyses were conducted to examine potential redundancy in subtest scores and could indicate, for example, whether just one or two of the subtests are as telling of an individual's performance on SCAN as the entire battery of subtests.

Results

Raw Scores

Table 1 reports Grades 1 and 3 means and standard deviations for Tests 1 and 2 subtest and COMP raw scores. Recall that the highest possible raw score is 40 each for the FW and AFG subtests, 100 for the CW subtest, and 180 for the COMP score. Thus, overall mean scores across all subtests, grades, and test/retest represent performance levels of about 80% correct. Figure 1 illustrates graphically the information reported in Table 1 but includes standard error bars rather than standard deviations.

Analysis of variance (ANOVA) statistics were completed to examine the effects of grade (first, third) and test (1, 2) for the FW, AFG, CW, and COMP raw scores. Table 2 reports a summary of SCAN subtest and COMP raw score ANOVA outcomes for grade (G), test (T), and interaction (G × T) effects. All statistical analyses included significance testing at $p < 0.01$, unless stated otherwise. As indicated in Table 2, FW outcomes revealed significant grade and test effects, with no significant interaction. Thus, third graders had significantly higher FW raw scores than first graders, and scores improved significantly from Test 1 to Test 2. On the AFG subtest, no significant grade, test, or interaction effects occurred. Thus, AFG raw scores did not change significantly from test to retest, and performance by both grades was similar. CW subtest outcomes revealed no significant grade or interaction effects. However, a significant effect of test was observed. Thus, first and third graders performed similarly, and their CW raw scores improved significantly from Test 1 to Test 2. Finally, COMP scores, generated by simply summing the three subtest raw scores, revealed marginally significant grade and significant test effects, with no significant interaction. Thus, third graders performed marginally better than first graders on COMP raw scores, and scores improved significantly from Test 1 to Test 2. Overall, raw score outcomes indicated a significant grade effect on one SCAN subtest (FW) and significantly improved performance from Test

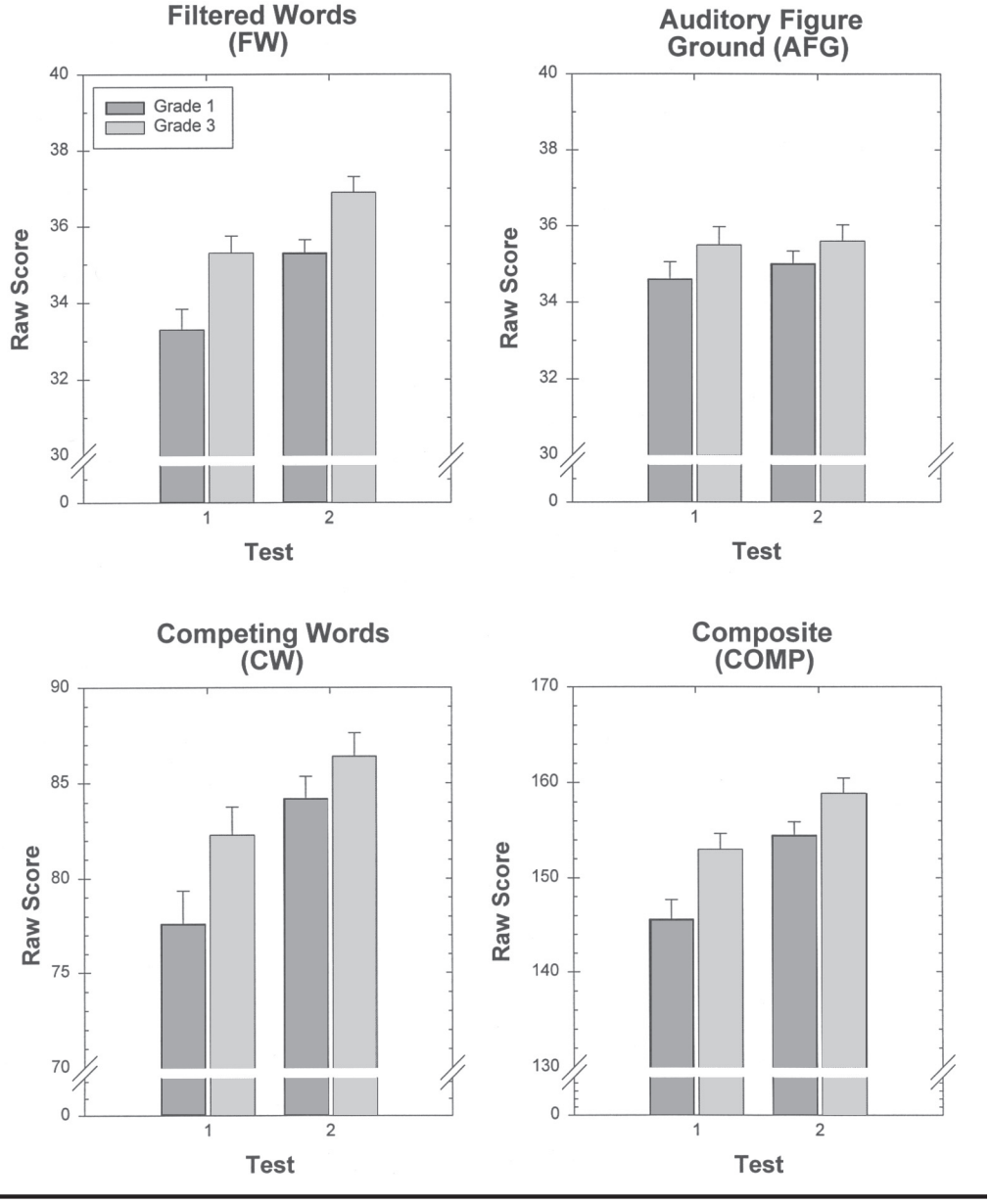
Table 2. Summary of SCAN subtest and composite raw score ANOVA outcomes for grade (G), test (T), and interaction (G x T) effects.

		ANOVA outcomes		
Subtest/effect		*F*	*df*	*p*
FW	G	12.51	1, 45	0.001**
	T	22.24	1, 45	0.000**
	G x T	0.29	1, 45	0.594
AFG	G	2.15	1, 45	0.149
	T	0.46	1, 45	0.501
	G x T	0.09	1, 45	0.761
CW	G	3.35	1, 45	0.074
	T	48.17	1, 45	0.000**
	G x T	2.47	1, 45	0.123
COMP	G	7.16	1, 45	0.010
	T	58.60	1, 45	0.000**
	G x T	2.45	1, 45	0.124

Note. FW = filtered words subtest; AFG = auditory figure ground subtest; CW = competing words subtest; COMP = composite.

**Significant at $p < 0.01$.

Figure 1. Grade 1 and Grade 3 means (with standard error bars) for Test 1 and Test 2 subtest and composite raw scores on SCAN. Error bars correspond to one standard error greater than the mean.



1 to Test 2 on two of the three SCAN subtests (FW and CW) and the COMP score.

Standard Scores

From the available appendixes (Keith, 1986), raw scores can be expressed as age-normalized standard scores, with each subtest having a mean of 10 and a standard deviation of 3 and the COMP having a mean of 100 and a standard deviation of 15 for each of several age groups. The scale was chosen because of its common use in psychoeducational testing, thus allowing for more direct comparison of SCAN scores with composite scores on measures of language, intellectual abilities, and achievement (Keith, 1986). The highest possible standard scores are 17 and 135 for the subtests and COMP, respectively. In the appendixes, a range of raw subtest and COMP scores is sometimes mapped to the same standard score (e.g., a COMP raw score of 108 to 111 results in a standard score of 81). Thus, it seemed prudent to examine whether observed effects of grade and test for the raw scores were lost following conversion to standard scores. In fact, it was expected that the effects of grade would be reduced or eliminated when analyzing the standard scores, because the standard scores are computed using age-based references.

Table 3. Grade 1 and Grade 3 means and standard deviations for Test 1 and Test 2 subtest and composite standard scores on SCAN.

		Subtest and composite scores							
		FW		AFG		CW		COMP	
Grade		Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
1	*M*	11.3	12.9	13.2	13.5	11.8	13.7	111.8	122.3
	SD	1.9	1.6	2.1	1.7	2.2	2.2	11.2	10.3
3	*M*	11.6	13.3	12.5	12.7	11.1	12.7	109.7	119.0
	SD	2.4	2.3	2.4	2.0	2.8	2.4	11.7	11.6

Note. FW = filtered words subtest; AFG = auditory figure ground subtest; CW = competing words subtest; COMP = composite.

Table 3 reports Grade 1 and Grade 3 means and standard deviations for Test 1 and Test 2 subtest and COMP standard scores. Figure 2 illustrates graphically the information reported in Table 3 but includes standard error bars rather than standard deviations.

ANOVA statistics were completed to examine the effects of grade (first, third) and test (1, 2) for the FW, AFG, CW, and COMP standard scores. Table 4 reports a summary of SCAN subtest and COMP standard score ANOVA outcomes for grade (G), test (T), and interaction (G × T) effects. As indicated in Table 4, FW outcomes revealed no significant grade or interaction effects. However, a significant test effect emerged. Thus, both grades performed similarly, and FW standard scores improved significantly from Test 1 to Test 2. An identical pattern of results was observed for the CW subtest and COMP standard scores. On the AFG subtest, however, no significant grade, test, or interaction effects occurred. Thus, AFG standard scores did not change significantly from test to retest, and performance by both grades was similar. Overall, the grade effects (albeit only marginally significant for the COMP) observed previously in the raw scores did not emerge in the standard scores. As noted, this would be expected because the same normalization was completed for each grade. Like the raw scores, however, standard score outcomes indicated significantly improved performance from Test 1 to Test 2 on two of the three SCAN subtests (FW and CW) and for the COMP score. Thus, test effects emerged on the same three outcome measures for both the raw and standard scores.

Percentile Rank and Age-Equivalent Scores

The appendixes of the SCAN manual (Keith, 1986) also provide tables to determine percentile rank and age equivalence on the SCAN test. Percentile ranks can be obtained for the subtest and COMP scores and are based on the standard scores. Subtest standard scores consistently map to the same percentile rank (e.g., an FW, AFG, or CW score of 13 is always the 84th percentile; a score of 10 is always the 50th percentile, etc.). Thus, any observed effects for subtest standard scores also should occur for subtest percentile ranks. However, in some cases for COMP scores, a range of COMP standard scores is mapped to the same percentile rank (e.g., COMP scores of 130 to 132 all correspond to the 98th percentile). Thus, an ANOVA was completed on the COMP percentile rank to examine whether it demonstrated a test effect like the COMP standard score. COMP-percentile-rank outcomes revealed no significant grade [$F(1, 45) = 1.1$, $p = 0.311$] or interaction [$F(1, 45) = 0.0$, $p = 0.958$] effects but did reveal a significant test effect

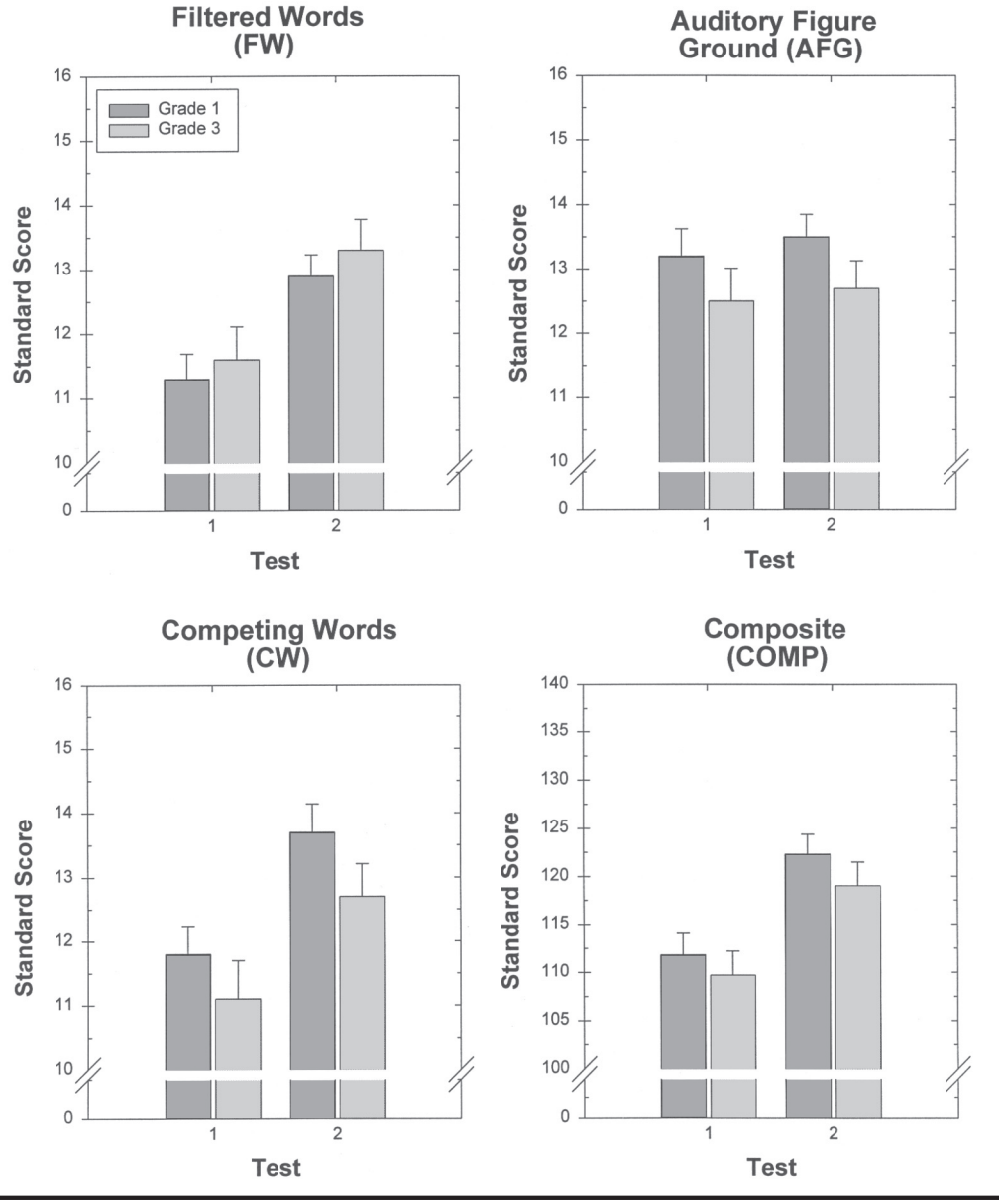
Table 4. Summary of SCAN subtest and composite standard score ANOVA outcomes for grade (G), test (T), and interaction (G x T) effects.

		ANOVA outcomes		
Subtest/effect		*F*	*df*	*p*
FW	G	0.55	1, 45	0.461
	T	19.98	1, 45	0.000**
	G x T	0.00	1, 45	0.996
AFG	G	2.15	1, 45	0.150
	T	0.39	1, 45	0.536
	G x T	0.06	1, 45	0.803
CW	G	1.64	1, 45	0.207
	T	42.46	1, 45	0.000**
	G x T	0.59	1, 45	0.445
COMP	G	0.83	1, 45	0.367
	T	63.92	1, 45	0.000**
	G x T	0.24	1, 45	0.627

Note. FW = filtered words subtest; AFG = auditory figure ground subtest; CW = competing words subtest; COMP = composite.

**Significant at $p < 0.01$.

Figure 2. Grade 1 and Grade 3 means (with standard error bars) for Test 1 and Test 2 subtest and composite standard scores on SCAN. Error bars correspond to one standard error greater than the mean.



[$F(1, 45) = 43.0$, $p < 0.001$]. Thus, like the standard COMP score, both grades had equivalent COMP percentile ranks, and these ranks improved significantly from Test 1 to Test 2.

Age equivalence (AE) is yet another potentially useful SCAN outcome and is based on the overall COMP raw score. In the appendix, however, the COMP raw score range of 0 to 180 is mapped to an AE range of 3 years 4 months to 11 years 0 months (i.e., different raw scores may map to the same AE). Thus, an ANOVA was performed to determine whether the effects observed for the COMP raw score were also evident in AE scores. Outcomes indicated marginally significant and significant effects of grade [$F(1, 45) = 4.8$, $p = 0.033$] and test [$F(1, 45) = 38.6$, $p < 0.001$], respectively, with no significant interaction effect [$F(1, 45) = 3.9$, $p = 0.055$]. Thus, third graders had marginally higher AEs than first graders, and AE scores for both grades improved significantly from Test 1 to Test 2.

Correlations and Scatterplots

In addition to evaluating the stability of SCAN outcomes by analyzing group mean changes in subtest and COMP score performance from test to retest, reliability was evaluated by analyzing test-retest correlations.

Table 5. Grade 1 and Grade 3 Pearson *r* correlations between test and retest for the subtest and composite raw and standard scores on SCAN.

Score/grade		SCAN test score category			
		FW	AFG	CW	COMP
Raw	1	0.33	0.24	0.74**	0.72**
	3	0.25	0.29	0.78**	0.70**
Standard	1	0.27	0.24	0.73**	0.72**
	3	0.25	0.30	0.70**	0.71**

Note. FW = filtered words subtest; AFG = auditory figure ground subtest; CW = competing words subtest; COMP = composite.

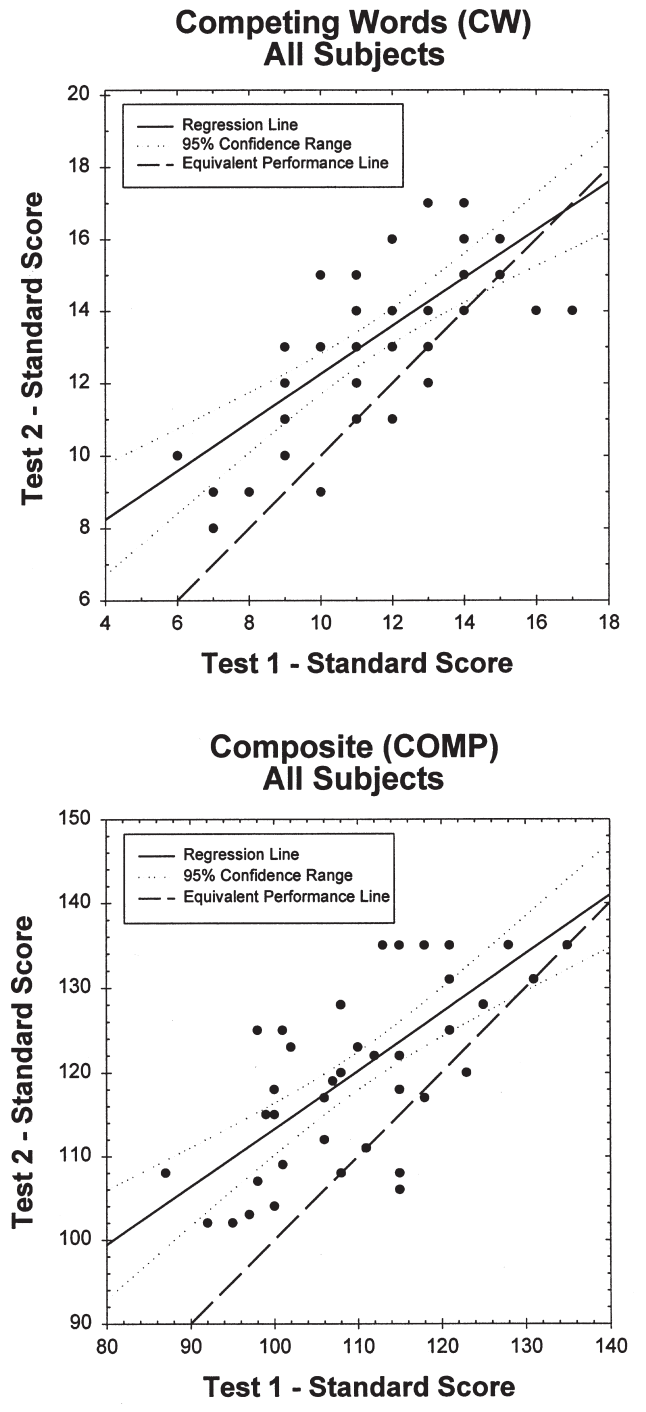
**Correlation is significant at $p < 0.01$ (2-tailed).

Table 5 reports Grade 1 and Grade 3 Pearson *r* correlations between test and retest for the subtest and COMP raw and standard scores on the SCAN test. As indicated in this table, poor and nonsignificant test-retest correlations emerged for both grades on the FW and AFG raw and standard scores (for all, $r < 0.35$). Test-retest correlations for the CW and COMP raw and standard scores, however, were moderately strong and significant ($p < 0.01$) for both grades ($0.70 \leq r \leq 0.78$).

To examine potential patterns of change in performance, scatterplots of test and retest scores were generated for the two measures with the strongest correlations (CW and COMP). Figure 3 depicts two scatterplots of Test 1 (T1) and Test 2 (T2) for the CW and COMP standard scores for all participants. Rather than raw scores, standard scores are depicted, because they are most likely to be used when evaluating individual performance and because the nonsignificant effect of grade on these scores permitted the collapse of data across all participants (47 total). Included in Figure 3 are solid lines representing linear regression (with dotted lines representing 95% confidence ranges) and dashed lines representing equivalent test-retest performance. On the CW subtest, 36 children (76.6%) had scores such that T2 > T1 (improved on retest), 5 children (10.6%) had scores such that T1 = T2, and 6 children (12.8%) had scores such that T1 > T2. For the COMP score, 38 children (80.9%) had scores such that T2 > T1, 5 children (10.6%) had scores such that T1 = T2, and 4 children (8.5%) had scores such that T1 > T2. Hence, the significant effect of test ($p < .01$) observed previously for the group data, in which the mean CW and COMP scores improved from test to retest, is apparent in the individual data for about 80% of the participants.

In a similar vein, an appendix of the SCAN manual (Keith, 1986) provides values to establish various confidence intervals (e.g., 80%, 85%, 90%, and 95%) around subtest and COMP standard scores according to an individual child's age. Thus, the respective 95% confidence ranges were computed for the CW and COMP standard scores for all participants. On the CW subtest, 16 children (34.0%) had T2 scores outside of the established 95% confidence intervals (i.e., a significant difference between T1 and T2). For the COMP score, 23 children

Figure 3. Scatterplots of Test 1 and Test 2 for competing word (CW) and composite (COMP) standard score outcomes for all participants. Fewer than 47 data points appear in each panel because of some identical individual scores.



(48.9%) had T2 scores outside of the established 95% confidence intervals. In all cases (for CW and COMP) where retest scores fell outside of the 95% confidence range, Test 2 performance was better than Test 1. Thus, one third to one half of the individual CW and COMP standard scores improved beyond the 95% confidence intervals for these scores when retested after 6 to 7 weeks.

Also depicted in Figure 3, the relationship of the best-fit linear regression line to the diagonal would indicate a potential pattern of change in individual performance from T1 to T2. A regression line within the 95% confidence range that is parallel to the diagonal would suggest equivalent improvement in performance from test to retest for both low and high scorers. In Figure 3, convergence of the lines for both the CW and COMP outcomes noticeably occurs in the high-standard-score range. A ceiling effect resulting from maximum possible CW and COMP standard scores of 17 and 135, respectively, however, could lead to the observed convergence of these two lines at high scores. Thus, a single linear regression line may not be a true "best fit." Rather, two lines, one horizontal and at the maximum possible score, with the other sloping and parallel to the diagonal, would appear to be as appropriate as a single line. Regardless, the CW and COMP standard-score test-retest correlations were only moderately strong ($0.70 \leq r \leq 0.73$), making it difficult to reliably predict an average amount of improvement for low- or high-scoring children from test to retest. The primary point here is that the individual data are consistent with trends in the group data and indicate higher CW and COMP scores on retest of SCAN.

Factor Analyses

Finally, to examine potential redundancy within SCAN, principal-components factor analyses were performed on the standard scores for the three subtests (FW, AFG, and CW). Because no grade effect emerged in previous standard-score analyses, all the data were pooled for the factor analyses. Results were very similar for both Test 1 and Test 2 scores. The three subtests were found to be associated with one independent, underlying factor, and each of the three subtests (components) contributed similar weighting to this single factor. Table 6 reports the principal-component weightings of the single factor identified for Test 1 and Test 2 standard scores for all participants. The independent factor, identified as a general auditory-processing ability, accounted for 43% and 50% of the variance for Tests 1 and 2, respectively. Thus, because the three SCAN subtests are all related to the same underlying factor, an examiner could potentially use just one subtest (e.g., CW) to screen a child's general auditory-processing ability. However, because maximal test-retest correlation is desired, and because stronger reliability is generally attained with a greater number of test items, it would seem most prudent to administer all three subtests and only use the COMP score for recommendations. Additionally, results of these factor analyses suggest that attempts to differentially interpret individual subtest outcomes are not warranted given that the three subtests all appear to be equally related to the same underlying auditory-processing factor.

Table 6. Principal-component weightings of the single factor identified for Test 1 and Test 2 standard scores for all participants on SCAN.

SCAN variable		Factor
Test 1	FW	0.58
	AFG	0.70
	CW	0.66
Test 2	FW	0.61
	AFG	0.72
	CW	0.78

Note. FW = filtered words subtest; AFG = auditory figure ground subtest; CW = competing words subtest.

Discussion

Published studies have indicated that SCAN has been used by both researchers and clinicians despite test-retest reliability concerns. As noted in the introductory remarks, previously published reliability data for first- and third-grade children (Keith, 1986) used a 6-month test-retest interval, and the average Pearson *r* test-retest correlation across all scores was about $r = 0.40$ (Keith, 1986), which is well below the generally targeted range of $r > 0.80$ (Nunnally, 1959). A lack of additional reliability data prompted the present investigation, which examined the stability of SCAN outcomes using a 6- to 7-week retest interval for normal-hearing, Caucasian, first- and third-grade children (ages 6 to 9 years). To minimize potential confounding, both time of day and examiner were held constant for each child from Test 1 to Test 2.

Raw score, standard score, percentile rank, and age-equivalent outcomes of this study indicated that both first and third graders did not perform maximally when taking SCAN for the first time. Specifically, both raw and standard scores improved significantly from Test 1 to Test 2 for two of the three SCAN subtests (FW and CW) and the COMP score. The AFG was the only subtest for which a significant test effect did not emerge. Further, the COMP percentile rank and age equivalence (AE) outcomes also demonstrated significant improvement from test to retest for both grades. This result is

not surprising, however, because COMP precentile rank and AE are based on the COMP standard and raw scores, respectively, which both demonstrated a significant effect of test. Also, although not as important from a reliability standpoint, third graders performed significantly better than first graders on the FW raw scores and marginally better on the COMP raw scores and AE, but, as would be expected, no grade effects emerged following conversion to age-based standard scores and percentile ranks.

Given the observed improvements in performance from Test 1 to Test 2, Pearson *r* test-retest correlations and scatterplots were generated to examine the potential to reliably predict the expected amount of improvement in scores at retest. If reliable, this could eliminate the need to re-administer SCAN in order to determine an individual's true performance (e.g., it may be adequate just to add 10 points to the initial COMP standard score). However, poor and nonsignificant test-retest correlations emerged for both grades on the FW and AFG scores ($r < 0.35$), and significant moderately strong positive correlations emerged for both grades on the CW and COMP scores ($0.70 \leq r \leq 0.78$). It is quite possible that the better CW and COMP score test-retest correlations may result from the greater number of test items contributing to these scores (i.e., 100 and 180, respectively, vs. 40 each for the FW and AFG subtests). Despite the significant CW and COMP test-retest correlations, it should be noted that $r > 0.8$ frequently is considered to be a minimally acceptable test-retest correlation and $r \geq 0.9$ most desirable. None of the SCAN scores for either grade, however, demonstrated test-retest correlations of $r \geq 0.8$. Thus, although scatterplots may have suggested that all participants generally improved equally (see Figure 3) on the CW and COMP scores, the correlations were only moderately strong, making it difficult to predict amount of improvement on retest for a specific child.

From a reliability standpoint, the group and correlational data of this study necessarily create a dilemma for examiners. On one hand, it appears necessary to administer SCAN a second time to obtain an estimate of an individual child's best performance. This may be even more useful for minorities, non-native English speakers, and children of ages 3 to 5 years, for whom the test developer (Keith, 1986) recommended that outcomes be interpreted with caution. However, the norms in the SCAN test manual (Keith, 1986) were not established on the basis of second administration of the test. Therefore, it would be inappropriate for examiners to compare retest scores for specific individuals to the published norms in the test manual. Thus, particularly for initially poor or marginal performers for whom additional testing, followup, and remediation would likely be recommended, examiners are left to choose from the following: (a) Administer SCAN a second time to obtain a better estimate of an individual child's best performance, but have no established "second-score" norms for comparison. or (b) Use an individual child's initial test scores and the published norms (Keith, 1986) with the knowledge that the reliability of such scores has been shown to be highly questionable. Neither choice is desirable, and it is clear that additional research must be conducted to use and interpret SCAN effectively. It is not clear, for example, that "second-score" SCAN results reveal asymptotic or "best" performance nor that such second-score SCAN results will necessarily be more reliable than single administration, although this would be expected.

Of importance is the fact that this investigation used a retest interval of 6 to 7 weeks. Thus, although some examiners may retest children to obtain a better estimate of maximal performance, the effects of immediate retest (within a few days) and the effects of multiple administrations (more than two) are unknown. It is apparent that future research should seek to establish normative data for different test-retest intervals and multiple test administrations. In addition, an alternative approach to address the apparent learning effect may be to increase the number of practice items. SCAN currently provides two practice items per ear per subtest. Such items could be replayed for children who miss one or more, or could be increased in number to provide additional practice before each subtest. Further, perhaps subtests could be administered but not scored initially, then later re-administered and scored for the first time. Again, however, the potential effects of such variations in practice would need to be investigated to provide examiners with normative data.

It may also be consequential that the children who participated in this investigation essentially were developing normally, with no evidence of CAPD. As with the present study, it is important to examine the test-retest reliability of a CAPD test on a population with no damage to the central nervous system (Musiek & Chermak, 1994). However, it may be just as important to determine the reliability of such a test on a population with the disorder it is intended to measure (Cacace & McFarland, 1995). Examiners should be cognizant that the present discussion is based on data from essentially normal children and therefore may or may not be applicable to children demonstrating CAPD characteristics. Future research should seek to address the present issues in both normal and disordered populations (Cacace & McFarland, 1995; Musiek & Chermak, 1994).

It must also be noted that this study examined the reliability and not the validity of SCAN. The validity of a test is a separate matter from its reliability. Validity addresses whether a test truly measures what it was

designed to measure (e.g., central auditory processing ability; for reviews, see Humes, 1996, and McFarland & Cacace, 1995). Even if the validity of SCAN is viewed by most researchers and clinicians as acceptable, use of a test that is potentially unreliable could result in grave inaccuracies in research conclusions and client management.

In addition to examining the stability of SCAN outcomes in this investigation, principal-components factor analyses were performed on the standard scores for the three subtests (FW, AFG, and CW). The results revealed one independent, underlying factor to which the three subtests (components) contributed similar weighting. Thus, the SCAN subtests were all found to be similarly related to the same underlying factor, which could be identified as a general auditory-processing or speech-understanding ability factor. Although this suggests that an examiner could potentially use just one of the subtests to screen a child's general auditory-processing ability, such use is not recommended. Because stronger reliability generally is attained with a greater number of test items, it would be prudent for examiners to administer all three subtests of SCAN and use only the COMP score for recommendations. As indicated in the test manual, SCAN is a screening instrument, and interpretation of the COMP standard score is of primary interest for screening (Keith, 1986).

Furthermore, the SCAN test manual suggests that an individual's performance on particular subtests allows for comparison of specific aspects of auditory processing and that differences may help indicate the most appropriate direction for further diagnostic testing or remediation (Keith, 1986). Specific implications are discussed in the manual. The factor analyses conducted in this study, however, suggest that attempts to differentially interpret individual subtest outcomes are not warranted given that the three subtests were found to be equally related to a single underlying auditory-processing or speech-understanding factor. Again, it appears that the primary value of the particular subtests rests in their contribution to the COMP score, not in an ability to indicate specific direction(s) for further diagnostic testing or remediation (Keith, 1986). As stated previously, the data from this study suggest that recommendations for additional testing or remediation should be based only on the COMP score.

Finally, approximately 80% of the children in this study demonstrated higher (improved) COMP standard scores on retest following a 6- to 7-week test-retest interval. Approximately 50% of the children had retest COMP standard scores that fell outside of the established 95% confidence ranges (Keith, 1986), and all of such children demonstrated better performance on retest. Although reliability data for children with CAPD were not assessed, the present results with normally developing children tentatively suggest that use of SCAN to monitor effectiveness of treatment of CAPD in children is highly questionable. That is, simply by allowing 6 to 7 weeks to pass between test administrations the overwhelming majority of children (80%) in this study had COMP standard scores that were higher on retest than on the initial test, and half of the children had improvements that exceeded the established 95% confidence intervals.

Acknowledgments

The authors would like to thank Kevin Caudill and Anne Summers, as well as the teachers, parents, and children of Lakeview Elementary School, Bloomingon, IN, for their assistance and cooperation in this project. Also, the authors give special thanks to the reviewers for their insightful comments.

References

American National Standards Institute. (1989). *Specifications for audiometers* (ANSI S3.6-1989). New York: Author.

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Cacace, A. T., & McFarland, D. J. (1995). Opening Pandora's box: The reliability of CAPD tests. *American Journal of Audiology, 4*(2), 62–62.

Chermak, G. D., & Musiek, F. E. (1992). Managing central auditory processing disorder in children and youth. *American Journal of Audiology, 1*(3), 61–65.

Chermak, G. D., Styer, S. A., & Seikel, J. A. (1995). Comparison of the Selective Auditory Attention test and the SCAN administered to boys with histories of otitis media. *Hearing Journal, 48*(5), 29–34.

Dietrich, K. N., Succop, P. A., Berger, O. G., & Keith, R. W. (1992). Lead exposure and the central auditory processing abilities and cognitive development of urban children: The Cincinnati lead study cohort at age 5 years. *Neurotoxicology and Teratology, 14,* 51–56.

Emerson, M. F., Crandall, K. K., Seikel, J. A., & Chermak, G. D. (1997). Observations on the use of SCAN to identify children at risk for central auditory processing disorder. *Language, Speech, and Hearing Services in Schools, 28,* 43–49.

Humes, L. E. (1996). Speech understanding in the elderly. *Journal of the American Academy of Audiology, 7,* 161–167.

Humes, L. E., Amos, N. A., & Wynne, M. (in press). Issues in the assessment of central auditory processing disorders. *Proceedings of the Fourth International Symposium on Childhood Deafness.* Nashville, TN: Bill Wilkerson Press.

Keith, R. W. (1986). *SCAN: A Screening Test for Auditory Processing Disorders.* San Antonio, TX: The Psychological Corporation, Harcourt, Brace, Jovanovich, Inc.

Keith, R. W., Rudy, J., Donahue, P. A., & Katbamna, B.

(1989). Comparison of SCAN results with other auditory and language measures in a clinical population. *Ear and Hearing, 10*, 382–386.

McCartney, J. S., Fried, P. A., & Watkinson, B. (1994). Central auditory processing in school-age children prenatally exposed to cigarette smoke. *Neurotoxicology and Teratology, 16*, 269–276.

McFarland, D. J., & Cacace, A. T. (1995). Modality specificity as a criterion for diagnosing central auditory processing disorders. *American Journal of Audiology, 4*(3), 36–48.

Musiek, F. E., & Chermak, G. D. (1994). Three commonly asked questions about central auditory processing disorders: Assessment. *American Journal of Audiology, 3*(3), 23–27.

Nunnally, J. C., Jr. (1959). *Tests and measurements: Assessment and prediction*. New York: McGraw-Hill.

Task Force on Central Auditory Processing Consensus Development. (1996). Central auditory processing: Current status of research and implications for clinical practice. *American Journal of Audiology, 5*(2), 41–54.

Willeford, J. A. (1974). *Central auditory function in children with learning disabilities*. Paper presented at the American Speech-Language-Hearing Association Annual Convention, Las Vegas, NV.

Received August 4, 1997

Accepted January 26, 1998

Contact author: Nathan E. Amos, Speech and Hearing Sciences, Indiana University, 200 South Jordan, Bloomington, IN 47405. Email: namos@indiana.edu