Interrater Reliability of Sensory Integration and Praxis Tests (SIPT) Score Interpretation

Asha V. Asher, L. Diane Parham, Susan Knox

OBJECTIVE. This study examined interrater reliability of score interpretation on the Sensory Integration and Praxis Tests (SIPT).

METHOD. Using SIPT scores of two complex cases, 20 trained participants independently rated each case for presence of sensory integrative dysfunction and for relevance of specific patterns of dysfunction. They also provided comments to justify their ratings.

RESULTS. Agreement on the presence of sensory integrative dysfunction was 70% for Case A and 100% for Case B. Reliability was more variable for dysfunctional pattern ratings, ranging from 50% to 100% agreement for Cases A and B, respectively. Participants consistently appeared to use configural decision-making strategies to guide their ratings.

CONCLUSION. Interrater reliability was moderate to high for interpretation of the presence of sensory integrative dysfunction using SIPT scores. Less agreement was apparent regarding specific patterns of dysfunction. Results suggest that additional clinical information, such as clinical observations and case history, may be needed to make reliable distinctions among dysfunctional patterns.

Sensory integration is a clinical specialty area in occupational therapy that may be integrated into occupation-centered practice (Parham, 2002). When conducting a clinical assessment using the sensory integration frame of reference, the clinician integrates data from standardized test scores with other sources of information, including clinical observation, occupational profile, case history, and parent or teacher interview addressing goals and concerns. These data are used to determine whether sensory integration difficulties are likely to be a significant factor in the child’s presenting problems and to formulate recommendations (Parham & Mailloux, 2005).

The Sensory Integration and Praxis Tests (SIPT; Ayres, 1989) is the gold standard for assessing sensory integration functioning. The SIPT is a battery of 17 standardized tests designed to assess praxis, various aspects of sensory processing, and the integration of sensory inputs (Ayres, 1989). In clinical assessment, the evaluator uses SIPT scores, supplemented with other sources of information, to determine whether sensory integrative functioning is compromised and, if so, the types of sensory integration problems that affect the child.

The SIPT have been carefully standardized, and each of the 17 tests has strong evidence of both test–retest and interrater reliability (Ayres, 1989). However, interrater reliability of judgments as to whether dysfunction is present, using SIPT score profiles for cases, has not been previously assessed. An estimate of the reliability of the interpretation of SIPT score profiles was used to evaluate the trustworthiness of this test battery. In this study, we investigated the extent to which a group of trained clinicians could agree on the presence and type of dysfunction for the two cases.
evaluators, independently analyzing only SIPT test scores of unfamiliar children, each arrived at the same conclusions regarding the presence of sensory integrative dysfunction and the relevance of particular dysfunctional patterns for each case.

**Literature Review**

**Clinical Judgment in Assessment**

The construction of the SIPT followed traditional psychometric procedures for standardized test development, which emanated from the discipline of psychology. In psychological assessment, the assessor integrates data about a client and assigns value to an attribute or to a behavior to make predictions about a client’s future behavior, to draw inferences about causes of a problem, to select a treatment program for the client, and to evaluate the effectiveness of that treatment program (Haynes & O’Brien, 2000). Many factors affect the quality of assessment data, including validity of the methods used to gather data and the inferences that the clinician draws from that data (Garb, 1998; Haynes & O’Brien, 2000; Murphy & Davidshofer, 1988). On the basis of an extensive review of research on clinical judgments in psychological assessments, Garb (1998) suggested that the validity of the judgments made by clinicians is not related to their total clinical experience but rather to the training they received in the specific assessment procedure. In addition, studies have shown that the confidence evaluators have in their clinical judgment is not related to the actual validity of their judgments (Dana, Cocking, & Dana, 1970; Garb, 1998).

Clinical judgments can be based on linear models, in which predictions are based on a linear combination of each of the available cues, or on a configural model, in which the weighting of an item of information varies according to the nature of other information available (Hoffman, 1960; Slovic & Lichtenstein, 1971). Haynes and O’Brien (2000) claimed that reliability and validity of clinical judgments can be enhanced with quantitatively aided decision-making strategies and quantitative criteria. They added that clinicians often use oversimplification strategies when required to make clinical judgments in the context of complex arrays of assessment information about a client.

Clinical reasoning in occupational therapy has been studied to gain an understanding of how therapists make sense of clinical situations and how they decide to proceed in therapy (e.g., Crabtree, 1998; Leicht & Dickerson, 2001; Mattingly, 1991a, 1991b). However, the inferences made by occupational therapists when interpreting standardized test scores have not been researched previously. This study was designed to investigate not only the reliability of test interpretation but also the reasoning process and inferences made by occupational therapists when examining SIPT standard scores.

**Interpretation of SIPT Scores**

Several informal and formal assessments are used to identify problems in sensory integration and their impact on activities and participation, to document current level of function, and to appraise change based on expected outcomes (Windsor, Roley, & Szklut, 2001). The SIPT (Ayres, 1989) is standardized on a North American normative sample of nearly 2,000 children ages 4 years to 8 years, 11 months. Ayres (1989) described the content of these tests as falling under four main categories: form and space perception, somatosensory and vestibular sensory processing, praxis, and bilateral integration and sequencing. A child’s z score on each of the 17 tests is generated by computer, and the profile of the child’s SIPT scores is compared by means of computer analysis to the normative sample and presented graphically. This graph compares the child’s pattern of scores with six diagnostic groups derived from a cluster analysis of scores from a subset of the normative sample plus a group of clinically referred children (Ayres, 1989). These cluster groups are as follows: low-average bilateral integration and sequencing, generalized sensory integrative dysfunction, visuo- and somatodyspraxia, low-average sensory integration and praxis, dyspraxia on verbal command, and high-average sensory integration and praxis.

Comprehensive evidence regarding validity and reliability of the individual tests are provided in the SIPT manual (Ayres, 1989). Generally, validity and reliability data exceed minimum standards of acceptability. Because validity and reliability of test procedures is a critical factor affecting assessment conclusions (Garb, 1998), the individual tests of the SIPT provide a sound foundation on which clinical decisions can be built. Interrater reliability of the major scores obtained on the individual SIPT tests range from .94 to .99. In addition to the major SIPT scores, some SIPT tests provide partial scores that reflect specific aspects of performance, such as severity of errors, speed of performance, or right-versus left-hand skills (e.g., Design Copying, Manual Form Perception, Constructional Praxis). Interrater reliability for most of the scores is above .90, and all have reliabilities of .85 or above, with the exception of the Figure–Ground time score and some of the Design Copying atypical approach parameters (Ayres, 1989).

Two separate organizations, Sensory Integration International and the University of Southern California in partnership with Western Psychological Services, provide training in the use of the SIPT through continuing education courses for therapists. These courses provide information on the theoretical foundations of sensory integration, including
its role in child development and its contributions to atypical behavior; in-depth training in administration of the SIPT; interpretation of the data obtained from the SIPT battery; and clinical reasoning strategies to design and implement intervention using a sensory integration frame of reference. Demonstration of competency in the administration of this test battery by trainees is included in the course structure offered by the University of Southern California–Western Psychological Services program, whereas Sensory Integration International trainees undergo a separate observation of test administration and must successfully complete a competency examination to be certified in the use of the SIPT. For the final interpretation of the scores on the SIPT, instructors in both programs recommend the use of worksheets to facilitate the inspection of score patterns that portray the sensory integrative functioning of the child. The first part of the worksheet guides the therapists in reviewing these scores with respect to the foundational abilities of the child tested, whereas the second part addresses the dysfunctional patterns indicated by the scores. Therapists are also instructed to check whether additional clinical information (such as history or nonstandardized observations) supports or contradicts test results and whether the presence of extraneous factors, such as environmental distractions during testing, could affect reliability of the test scores.

SIPT training appears to develop and strengthen clinical judgment used by therapists to interpret SIPT scores in a clinical evaluation. Therapists undergo rigorous training in the use of the specific assessment procedures, as recommended by Garb (1998). Configural decision-making strategies are explicitly taught for interpretation of test results (Hoffman, 1960; Slovic & Lichtenstein, 1971). Therapists are taught to consider how different combinations of low scores indicate different types of problems. For example, a combination of low scores on the Oral Praxis and Bilateral Motor Coordination tests is suggestive of the cluster group bilateral integration and sequencing. However, if a low score on Oral Praxis appears in combination with low scores on Postural Praxis and other tests of praxis, this score configuration is a strong indicator of somatodyspraxia. Interpretation worksheets are designed to allow therapists to use configural decision-making strategies. Although these strategies are not the linear, quantitative strategies that Haynes and O’Brien (2000) recommended, it is reasonable to expect that they should enhance the reliability of clinical judgments made by therapists undergoing SIPT interpretation training.

In clinical practice, therapists almost always use additional information beyond SIPT test scores to complete the clinical assessment of a child. For example, scores from parent or teacher rating scales, such as the Sensory Profile (Dunn, 1999) or Sensory Processing Measure (Parham, Ecker, Kuhaneck, Henry, & Glennon, 2007), may be used to augment assessment, particularly in relation to assessment of sensory modulation. In addition, nonstandardized data, such as referring problems, developmental history, occupational profile, and clinical observations, provide critical supplementary information (Ayres, 1989; Parham & Mailloux, 2005). For example, clinical observations of motor responses, such as co-contraction (the simultaneous contraction of all the muscles around a joint to stabilize a position) and equilibrium reactions (compensatory movements of body parts that help to maintain or regain the center of gravity; Parham & Mailloux, 2005), furnish important information regarding vestibular–proprioceptive processing.

In her discussion of the use of tests in clinical assessment, Anastasi (1986) suggested that spurious evidence of test validity might occur when judges have obtained cues from nontest sources, such as conversation with the client, case history, or additional assessment procedures. Anastasi recommended that this spurious validity could be controlled by blind analysis, in which a scorer who has no contact with the examinee or any other information beyond that contained in the test protocol interprets the test record.

This study was designed to follow Anastasi’s (1986) recommendation by using blind analyses of SIPT scores. Participants in this study were given only the SIPT score profiles of anonymous children so that estimates of interpretation validity reflected the SIPT scores alone and were not biased by additional case information. Because reliability is a critical element of validity (Anastasi, 1986), this study contributes to knowledge regarding the validity of the use of SIPT scores in clinical assessment.

**Research Questions**

Although the reliability and the validity of the individual SIPT tests are extensively documented, interrater reliability data are lacking for the interpretation of the entire group of tests. If a group of clinicians independently analyzed the computer-generated SIPT score profiles of children, would they arrive at the same conclusions regarding the presence of sensory integrative dysfunction for each child? Accurate assessment is a prerequisite to appropriate and efficient treatment and is vital to the respectable image of the profession. A demonstration of interrater reliability of test score interpretation, without additional case information, would be valuable because it would provide evidence as to whether the test is a worthwhile component of clinical assessment. If test score profiles cannot be reliably interpreted without additional assessment information, the value of using the test in clinical practice is limited.

Three questions were specifically addressed in this study. The first two questions addressed the interrater reliability of
trained clinicians independently interpreting the presence of sensory integration dysfunction using SIPT score profiles of anonymous cases.

The first research question inquired about determination of dysfunction: To what extent do different clinicians agree on whether a child’s SIPT test score profile is indicative of dysfunction?

The second research question addressed foundational abilities and dysfunctional patterns measured by the particular patterns of scores on the 17 SIPT tests: To what extent do clinicians agree on the relevance of specific foundational abilities and dysfunctional patterns to a particular child’s profile?

The third research question attempted to understand the reasoning on which the clinicians based their judgments: What are the strategies that therapists consciously use when interpreting test scores?

In the study, clinicians were asked to state the factors they considered in reaching the decisions made in response to Questions 1 and 2. This research question was designed to elucidate the reasoning process clinicians use to draw conclusions from test data.

Method

To evaluate the interrater reliability of the interpretation of the SIPT on the basis of the score profiles alone, a group of clinicians were asked to rate the SIPT profiles of two children on the presence of sensory integrative dysfunction and the relevance of subpatterns on the SIPT scores. Their ratings were quantitatively analyzed to assess interrater agreement. In addition, content and thematic analyses (Smith, 1992) of their comments regarding the decision-making process were completed. In keeping with the recommendation of Anastasi (1986), the clinicians had no personal contact with the anonymous cases, nor did they have access to any identifying background or diagnostic information related to the cases.

Participants

A convenience sample of 20 occupational therapists participated in the study. The participants’ experience was variable enough to minimize bias that might result from including only exceptionally good or poor clinicians. The occupational therapists were selected from different locations within the Los Angeles, California, area and were all female, between 26 and 60 years old, with 1.5 to 20 years of clinical experience in pediatric occupational therapy. All had attended postprofessional courses on the theoretical background of sensory integration, SIPT administration, and interpretation of SIPT scores. One participant had written materials for these courses. Participants’ estimates of the number of clinical reports they had written that contained SIPT interpretations ranged from 1 to 60 ($M = 13.80$).

Clinical Cases

Each participant was given only the computerized SIPT score profiles of two anonymous children (see Figures 1 and 2). In view of the time required to review each profile, the number of profiles presented was restricted to two to ensure that participants would complete all ratings. A therapist certified in the use of the SIPT, but not participating as a participant in the study, selected the cases, which were chosen for their complexity. In Case A, the computer-generated graphical representation of the child’s scores had not likened the profile to any of the diagnostic groups previously identified through cluster analysis of SIPT normative data (Ayres, 1989). In Case B, the profile was likened to three diagnostic clusters, generalized sensory integrative dysfunction, visuo-and somatospraxia, and low-average bilateral integration and sequencing. The complexity of the SIPT score patterns of these two cases was intended to permit a rigorous evaluation of reliability. Had profiles been likened to only one cluster group, a high interrater reliability might result because of the relative ease of interpretation.

Instrumentation

A questionnaire was developed for this study. It was pilot tested to confirm the clarity of the instructions and its appropriateness to obtain the desired information. This instrument asked the participants to rate on a dichotomous (yes–no) scale the likelihood of the presence of sensory integrative dysfunction on the basis of the SIPT scores alone. In addition, the participants were asked to rate each child’s score profile on the relevance of several foundational abilities and dysfunctional patterns that were specified as relevant to SIPT score interpretation in the SIPT manual (Ayres, 1989) and in the SIPT training process that the participants had completed. The ratings were defined as follows: not relevant—evidence is generally incompatible with this interpretation; possibly relevant—evidence is suggestive but not strong; probably relevant—evidence is strong but not conclusively so; definitely relevant—evidence very strongly fits this interpretation. Finally, in an attempt to gain some insight into the reasoning used, the participants were asked to state the factors they had considered in deciding their ratings for these questions.

Procedures

After the two cases were selected, the names of the children were removed and the SIPT score profiles were provided to the researcher. The 20 participants were presented the case...
material, which consisted of the graphical representation of the SIPT scores for each case and the questionnaire. The order of presentation of the cases was counterbalanced across participants. The investigator explained to each participant that, for the purpose of the study, best judgments should be used to rate each child with the SIPT scores alone. To assist the process of interpretation, a worksheet that had been used in SIPT training courses was provided for each case. However, participants were not required to use this worksheet; it was provided simply as an aid for participants who wished to use it.

Data Analysis

To answer the first research question, which addressed the presence or absence of sensory integrative dysfunction on the basis of the interpretation of SIPT score profile, interrater reliability was estimated through calculation of the percentage of agreement among the participants for each case.

The second research question addressed the extent to which the participants agreed on the relevance of foundational abilities and dysfunctional patterns to a particular child’s profile. Ratings given by the participants on the 4-point rating scale were assigned numerical values from 1 to 4 to aid statistical calculations: 1 = not relevant, 2 = possibly relevant, 3 = probably relevant, and 4 = definitely relevant. These ratings were ordinal with respect to increasing relevance; however, it must be remembered that the values are not equidistant. Using these ordinal values, the average rating received by each child for each pattern was determined. The deviation of the rating assigned by each participant from the average was calculated. Moreover, the average absolute deviation for each pattern was determined for the two children. Average absolute deviation is a measure of dispersion that tells us the average distance that a piece of data is from the mean (Johnson, 1984). The average absolute deviation was calculated by considering each deviation from the mean as positive.
and further determining the average of these deviations. The
dysfunctional patterns being rated could then be arranged in
order of decreasing agreement among the participants.

Some of the participants commented on their difficulty
in differentiating between probably and definitely relevant
and possibly and not relevant because the cases involved were
not clear cut. Therefore, ratings were collapsed to allow for
an analysis of percentage of agreement. Probably relevant
was collapsed with definitely relevant, and possibly relevant
was collapsed with not relevant to reduce the number of
categories from four to two (relevant and not relevant).
Numerical values were assigned to these categories (1 = not
relevant and 2 = relevant). From the resulting dichotomized
ratings, percentage of agreement was calculated for each of
the dysfunctional patterns.

For the third research question, which inquired about
the strategies that therapists use when interpreting test scores,
the comments made by the participants on the questionnaire
were transcribed and categorized according to score patterns
and types of problem-solving strategies reported.

Results

Reliability for Case A

Of the 20 participants, 6 rated Case A as having sensory
integrative dysfunction, whereas 14 participants rated dys-
function as not present. Therefore, percentage of agreement
in ruling out dysfunction was 70%. The average ratings by
the participants for each pattern and the average absolute
deviations of these ratings are presented in Table 1. Percentages of agreement on dichotomized ratings of pat-
terns are shown in Table 2.

All the participants completed their ratings for bilateral integration and sequencing, and they were in
almost total agreement in their ratings for dyspraxia on verbal
command. In order of decreasing agreement (as indicated by average absolute deviations of ratings) were visual form and space deficit, visuodypraxia and somatodypraxia, somatosensory dysfunction, and vestibular dysfunction. Percentages of agreement for dichotomized ratings were 100% for several dysfunctional patterns, but they were unacceptably low for vestibular dysfunction (63%) and somatosensory dysfunction (58%).

The 6 participants who deemed that this child had sensory integrative dysfunction agreed that visual form and space deficit, bilateral integration and sequencing, and dyspraxia on verbal command were not relevant to this child’s profile. A similarity was seen in the extent of agreement of these participants between somatosensory dysfunction and vestibular dysfunction, whereas visuodypraxia and somatodypraxia showed greater disagreement.

The 14 participants who decided that sensory integrative dysfunction was not present in this case agreed entirely only in that bilateral integration and sequencing were not relevant to this child’s profile and were almost unanimous in ruling out the relevance of dyspraxia on verbal command. In order of decreasing relevance were visuodypraxia and somatodypraxia, then visual form and space deficit, and somatosensory dysfunction. Least agreement was seen on vestibular dysfunction.

Reliability for Case B

All 20 participants agreed that Case B had definite sensory integrative dysfunction, yielding 100% agreement. However, examination of the ratings that participants assigned for each pattern revealed greater variability compared with ratings given for Case A. The averages of the ratings given by the participants for each pattern and the average absolute deviations of these ratings are presented in Table 3. Table 4 depicts percentages of agreement for dichotomized judgments regarding the relevance of each of the patterns of foundational and dysfunctional patterns.

Most agreement was seen on visuodypraxia and somatodypraxia (100% agreement with a small average absolute deviation). Least agreement was seen on bilateral integration and sequencing. The other four patterns showed a similar amount of deviation from the average score for that pattern. Percentages of agreement for dichotomized ratings were unacceptably low for somatosensory dysfunction (55%) and bilateral integration and sequencing (50%). All other percentages of agreement were high, ranging from 80% to 100%.

Content Analysis of Participant Comments on Reasoning

Participants had been asked to state the factors they considered in reaching their decisions regarding the presence or absence of sensory integration dysfunction and the relevance of particular foundational abilities and dysfunctional patterns. In their comments, participants consistently referred to scores that were indicators or contradicts of the particular foundational abilities and dysfunctional patterns. For example, for Case A, 1 participant commented that the average scores on Space Visualization and Figure Ground rule out Form and Space Deficit; below average scores on Standing and Walking Balance, Finger Identification, and Localization of Tactile Stimuli Left, Kinesthesia and Graphesthesis suggest possible somatosensory dysfunction, and but Standing and Walking Balance with eyes closed [is within normal limits], Right Kinesthesia, Graphesthesis, Localization of Tactile Stimuli are average, Manual Form Perception part II is above average.

Analysis of Comments for Case A. The score profile of Case A had not been likened to any cluster groupings. Three participants found that the scores obtained by this child did not provide enough evidence to conclude that any one dysfunctional pattern was present.

<table>
<thead>
<tr>
<th>Dysfunctional Patterns</th>
<th>Average Rating</th>
<th>Average Absolute Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visuodypraxia</td>
<td>3.65</td>
<td>0.45</td>
</tr>
<tr>
<td>Dyspraxia on verbal</td>
<td>1.70</td>
<td>0.70</td>
</tr>
<tr>
<td>Somatosensory dysfunction</td>
<td>2.45</td>
<td>0.70</td>
</tr>
<tr>
<td>Visual form and space deficit</td>
<td>1.60</td>
<td>0.72</td>
</tr>
<tr>
<td>Vestibular dysfunction</td>
<td>2.00</td>
<td>0.75</td>
</tr>
<tr>
<td>Bilateral integration and sequencing</td>
<td>2.50</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 1. Average Ratings and the Average Absolute Deviations of the Ratings for Case A

<table>
<thead>
<tr>
<th>Dysfunctional Patterns</th>
<th>Average Rating</th>
<th>Average Absolute Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilateral integration and sequencing</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Dyspraxia on verbal command</td>
<td>1.05</td>
<td>0.10</td>
</tr>
<tr>
<td>Visual form and space deficit</td>
<td>1.20</td>
<td>0.32</td>
</tr>
<tr>
<td>Visuodypraxia and somatodypraxia</td>
<td>2.00</td>
<td>0.42</td>
</tr>
<tr>
<td>Somatosensory dysfunction</td>
<td>2.47</td>
<td>0.55</td>
</tr>
<tr>
<td>Vestibular dysfunction</td>
<td>2.05</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 2. Agreement of Participants in Judging the Relevance of Dysfunctional Patterns for Case A

<table>
<thead>
<tr>
<th>Dysfunctional Patterns</th>
<th>% in Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilateral integration and sequencing</td>
<td>100</td>
</tr>
<tr>
<td>Dyspraxia on verbal command</td>
<td>100</td>
</tr>
<tr>
<td>Visual form and space deficit</td>
<td>100</td>
</tr>
<tr>
<td>Visuodypraxia and somatodypraxia</td>
<td>79</td>
</tr>
<tr>
<td>Vestibular dysfunction</td>
<td>63</td>
</tr>
<tr>
<td>Somatosensory dysfunction</td>
<td>58</td>
</tr>
</tbody>
</table>

Table 3. Average Ratings and the Average Absolute Deviations of the Ratings for Case B
For this case, 9 of 20 participants specifically asked for additional information, although they had been asked to make their best estimates using the SIPT scores alone. Of these 9 participants, 7 specified the need for clinical observations for appropriate interpretations of the test scores. For example, Participant 12, who judged that sensory integrative dysfunction was not present in Case A, wrote, based on these scores alone I could not say that there is sensory integrative dysfunction present in this child. I would want more information regarding medical–developmental, sensory history, clinical observations, etc., to support the presence of sensory integrative dysfunction.

Interestingly, of the 6 participants who concluded that Case A was dysfunctional, only 1 asked for additional information. In other words, the participants who commented on the desirability of additional information nevertheless tended to be in strong agreement that Case A scores did not provide evidence of sensory integrative dysfunction.

Some participants commented that they needed additional information specifically in relation to interpreting foundational abilities and dysfunctional patterns for Case A. For example, Participant 4 noted, I’m questioning the finger identification score since it falls so far below all other scores. LTS [Localization of Tactile Stimuli] is really a borderline score—I would need clinical observations before I could put check [on] the possibly relevant category.

Of the 20 participants, only 3 reported their consideration of and rejection of the bilateral integration and sequencing and dyspraxia on verbal command patterns for Case A. These participants had systematically commented on every grouping presented in the worksheet and rejected these patterns based on average or high scores on key identifying tests.

Case A’s low scores on Constructional Praxis, Manual Form Perception (Part I), Design Copying (Part II), and low Preferred Hand Use on the Motor Accuracy test influenced 5 participants to support the presence of the dysfunctional pattern visuo- and somatosensory. Of these, 2 participants reported that they considered the low score on Preferred Hand Use on the Motor Accuracy test in reaching their decision. One participant offered a rationale for the fact that scores expected to be low in somatodyspraxia were elevated into the low-average range. She explained that average or high-average scores on tests having visual–perceptual or proprioceptive components might have positively influenced some of the other scores expected to be low in the somatodyspraxia group. However, 3 other participants refuted the presence of this pattern because the major somatopraxia tests (Postural Praxis, Oral Praxis, and Sequencing Praxis) were within the average range.

There appeared to be stronger agreement that this child had problems with processing of somatosensory input. Nine of the 20 participants said that low-average to low scores on the tactile tests (Finger Identification, Localization of Tactile Stimuli, and others) pointed to a deficit in somatosensory processing, although the evidence was relatively weak. Two other participants considered this possibility but thought that scores on other tests refuted it.

**Table 4. Agreement of Participants in Judging the Relevance of Dysfunctional Patterns for Case B**

<table>
<thead>
<tr>
<th>Dysfunctional Patterns</th>
<th>% in Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visuodyspraxia and somatodyspraxia</td>
<td>100</td>
</tr>
<tr>
<td>Dyspraxia on verbal command</td>
<td>85</td>
</tr>
<tr>
<td>Visual form and space deficit</td>
<td>85</td>
</tr>
<tr>
<td>Vestibular dysfunction</td>
<td>80</td>
</tr>
<tr>
<td>Somatosensory dysfunction</td>
<td>55</td>
</tr>
<tr>
<td>Bilateral integration and sequencing</td>
<td>50</td>
</tr>
</tbody>
</table>

Analysis of Comments for Case B. The score profile of Case B had been likened on the graph to three clusters, generalized sensory integrative dysfunction, visuo- and somatodyspraxia, and low-average bilateral integration and sequencing. Of the 20 participants, 18 did not mention consideration of generalized dysfunction at all. Of the 2 who discussed this grouping, 1 participant thought that it was not present because there were not enough low scores on the visual and tactile tests. The other participant described the case as generalized dysfunction with dyspraxia as a major problem. This participant suggested that other patterns of dysfunction could not be conclusively determined on the basis of scores but did not give a further explanation of her reasoning.

The presence of a deficit in bilateral integration and sequencing was supported by 7 participants, mainly on the basis of major identifying tests that included Oral Praxis, Standing and Walking Balance, Sequencing Praxis, and Bilateral Motor Coordination. Of these participants, 2 reported that low Preferred Hand Use and borderline Contralateral Hand Use scores offered additional support for this pattern. The 5 participants who rejected this pattern appear to have considered low scores obtained on other tests, which should have been within the average range for this pattern. They reported that scores obtained on tests other than those mentioned previously supported the presence of other patterns of dysfunction. Eight participants did not report consideration of this group in spite of the computer analysis likening Case B’s profile to this group.

The 2 participants who considered the presence of dyspraxia on verbal command indicated that the child’s apparent problem with this type of functioning was caused by postural demands rather than language comprehension. However, 8 other participants reported that although praxis...
appeared to be generally affected in this child, there was not enough contrast between the scores on the test Praxis on Verbal Command, which should have been lower, and Postrotary Nystagmus, which should have been higher, to classify this child as having this particular deficit. One participant commented that, with regard to this pattern, the Graphesthesia score should not have been so low.

Visuo- and somatodypraxia was the third grouping to which the computer had likened the profile of this child. The 14 participants who supported the presence of this deficit justified their decision with the low scores on all the components of dyspraxia, including Oral Praxis, Sequencing Praxis, Design Copying, and Constructional Praxis. The low score on Graphesthesia was assumed to be adequate to justify the somatosensory component of somatodypraxia. One participant did not reach a conclusive decision regarding presence of this deficit on the basis of the scores alone but did not clarify her reasoning.

In the analysis of Case B, participants did not mention the need for clinical observations or other assessment data as frequently as for Case A. However, it did appear occasionally in the analysis of Case B comments. For example, Participant 5 commented on this case, “It would be necessary to observe this child in the clinic for further vestibular conclusions.”

**Thematic Analysis of Decision-Making Strategies**

The comments offered by the participants were further analyzed thematically to identify common strategies used to reach decisions. Comments were coded according to the type of strategies they suggested. The decision-making strategies were grouped under several common themes that emerged from the data. These are presented in Tables 5 and 6 together with the number of participants describing each. For Case A, the participants appeared to have used a somewhat wider range of strategies to reach their decisions, compared with Case B.

Most participants appeared to have focused on comparison of the different dysfunctional patterns. In general, participants appeared to agree more in ruling out relevance of the patterns rather than in confirming the presence of the patterns. Two participants noted that they found the interpretation worksheets more useful than the graph in reaching their decisions.

Results suggest that participants tended to rely on configural strategies, where the weighting of an item of information varies according to the nature of other information available (Hoffman, 1960; Slovic & Lichtenstein, 1971). For example, Participant 4 commented on Case A, “Although the SWB [Standing and Walking Balance] is low [indicating a vestibular deficit], the high PRN [Postrotary Nystagmus] tends to make me cautious in even considering vestibular dysfunction.” When elaborating on the factors considered in reaching their decisions regarding Case A, 14 participants reported considering the interplay of four or more factors.

It was apparent that even when a particular score is thought to reflect a specific dysfunction, participants tended to interpret it in view of the other test scores. For instance, Participant 6 wrote with reference to Case A,

(1) A typically high PRN is a sign of insufficient CNS [central nervous system] inhibition of the lower nystagmatic reflex, not insufficient processing of vestibular sensory input. (2) Low score on SWB [Standing and Walking Balance] with high score on PRN is suggestive of neuromuscular incoordination.

This participant considered the high PRN by itself, then compared it with other data before concluding that the child did not appear to have a problem with vestibular processing. This conclusion, in turn, was pivotal in the participant’s decision that Case A did not have sensory integrative dysfunction.

**Discussion**

This study was intentionally limited to an examination of interrater reliability of SIPT interpretation using standard
scores only. Had we included other assessment data, such as clinical observations, along with the SIPT scores, we would not have been able to discern the extent to which results reflected agreement on interpretation of SIPT scores because participants’ judgments would be affected by the additional assessment data. The finding that participants were able to generally agree on their conclusions regarding the presence of sensory integrative dysfunction using scores alone (agreement on the presence of dysfunction, 70% for Case A and 100% for Case B) indicates that the SIPT provides meaningful and trustworthy information to trained users.

When the ratings on the relevance of six specific patterns of dysfunction were analyzed for the two cases, wide variability was seen. Agreement was acceptable at more than 79% for most ratings of dysfunctional patterns, and perfect agreement was achieved for ratings of some patterns. However, several ratings of dysfunctional patterns fell unacceptably low (percentages of agreement between 50% and 63%—levels that could be accounted for by chance). This finding will not be surprising to some readers, considering that Ayres (1989) had recommended that additional information, such as case history, functional capacity of the child, and clinical observations, should be considered in making decisions regarding presence or absence of types of dysfunction. In this study, such additional information was not available to the participants. This information was withheld to pursue the purpose of this study, which was to examine the reliability of SIPT score interpretation. The weak levels of agreement for interpreting some types of dysfunctional patterns indicate that SIPT scores should not be solely relied on to make more nuanced interpretations regarding specific types of dysfunction.

It is plausible that provision of additional information, such as clinical observations, would have enabled participants to more reliably organize the data in relation to particular foundational abilities and dysfunctional patterns. Some participants explicitly indicated that having access to additional information might have made a difference in their responses. However, if additional information had been provided in this study, reliability estimates would have reflected the influence of the additional information, and consequently it would have been impossible for us to elucidate the reliability of SIPT score profile interpretation.

The comments provided by the participants were analyzed for clues to the strategies they used in decision making. In discussing psychological assessments, Haynes and O’Brien (2000) and Slovic and Lichtenstein (1971) reported that clinicians use simplified decision strategies to reduce cognitive strain when presented with complex data. However, participants in this study did not appear to identify sensory integrative dysfunction using a simple algorithm of test scores. Instead, the interplay between the patterns of low and normal or high scores on the various tests was carefully weighed by the participants to reach conclusions. Participants appeared to use these complex, configurational strategies reliably in reaching conclusions regarding the presence or absence of sensory integrative dysfunction for the two cases.

The prior formal postprofessional training of the participants likely played a critical role in their reliability. Sensory integration postprofessional certificate programs provide training in SIPT administration and interpretation, including application of specific interpretation strategies. Such training is thought to increase the validity of professionals’ judgments (Garb, 1998). In addition, research suggests that training clinicians to use specific decision-making strategies enables them to make decisions more reliably and effectively (Slovic & Lichtenstein, 1971). Explicit guidelines are given in SIPT interpretation courses, which recommend that the results from the individual tests composing the SIPT should be inspected with reference to foundational abilities and dysfunctional patterns using configurational decision-making strategies. The participants in this study appeared to follow these guidelines, contributing to the reliability of the interpretation of the SIPT.

For each of the cases, a graph was provided that depicted the child’s main standard SIPT scores, as well as any diagnostic cluster groups to which the child’s profile was similar, as determined previously by computer analysis. The computer analysis of Case A’s score profile had not identified it as similar to any of the diagnostic groupings, a result that could partially explain why participants tended to agree on the nonrelevance of the dysfunctional patterns. It could also account for the lower percentage of agreement (70%) on judgments as to whether Case A’s profile indicated the presence of sensory integration dysfunction. By contrast, Case B’s computer analysis had likened the score profile of this child to three diagnostic clusters. This score profile may account for the 100% agreement regarding the presence of dysfunction, but it also could contribute to the high level of variability among participants regarding the relevance of specific types of dysfunctional patterns. Clearly, the computer-generated diagnostic cluster groupings did not dictate the conclusions reached by the participants regarding presence or type of sensory integrative dysfunction. For example, with Case B, 8 participants did not report considering a deficit in bilateral integration and sequencing, even though the computer analysis had likened Case B’s profile to this cluster group. Participants’ comments indicated that they used their own configurational strategies to interpret scores rather than relying on the computer-identified groupings.
The weak reliability for several dysfunctional patterns raises concerns about how these patterns are interpreted in clinical assessment. In particular, somatosensory dysfunction and vestibular dysfunction yielded percentages of agreement ranging from 50% to 63%. Bilateral integration and sequencing had 100% agreement for Case A, but for Case B, agreement was unacceptably low (50%). Perhaps these results are particular to the two difficult cases that were examined in this study, or it may be that decisions regarding specific dysfunctional patterns cannot be reliably made without supplemental assessment data from sources beyond the SIPT. Even so, results of this study suggest that training programs on SIPT interpretation may need to reexamine instructional strategies for teaching interpretation of somatosensory dysfunction, vestibular dysfunction, and bilateral integration and sequencing to bolster reliability of interpretation of these patterns.

Limitations

Although the SIPT provides the most rigorous standardized assessment of sensory integration available, it does not measure sensory modulation, an important domain of sensory integration. SIPT scores primarily measure sensory discrimination, perceptual–motor functioning, and praxis. Therefore, it should be kept in mind that participants’ interpretations of sensory integration dysfunction in this study did not include sensory modulation problems, such as overresponsivity, underresponsivity, and sensory seeking.

The process of diagnostic decision making was studied only through content analysis of comments that participants chose to write on the rating forms. It is possible that important decision-making considerations and strategies were not documented by the participants and therefore were not detected in this study.

Results suggested that the participants tended to agree more in ruling out relevance of dysfunctional patterns rather than confirming the presence of the patterns. However, because only two cases were analyzed by participants, this finding cannot be generalized to all cases. It may be that in more straightforward, less ambiguous case profiles, the presence of patterns is more easily agreed on.

An additional limitation is posed by the geographical area from which the participants were recruited. The study was conducted in Southern California, where advanced training and expert mentoring in SIPT interpretation is easily accessible. The existence of a local community of clinicians well versed in SIPT usage may have resulted in stronger reliability estimates than might be found in geographic areas where such expertise is not available.

Conclusions and Implications for Future Research

This study showed that trained clinicians can reliably draw conclusions regarding the presence of sensory integration dysfunction using SIPT score profiles only. Reliability is less consistent with respect to particular dysfunctional sub-patterns of SIPT scores. It appears likely that more consistently reliable interpretations of sub-patterns would require the integration of SIPT scores with additional assessment information, such as clinical observations and case history. Therefore, we recommend strongly that when using the SIPT to clinically assess an individual child, additional clinical information be integrated with SIPT scores. Such clinical information would be particularly important when assessment data are being used to make clinical decisions about interventions.

Future research could address interrater reliability of a more comprehensive assessment process that involves provision of SIPT scores along with additional clinical information on each case. We hypothesize that the additional information would enable participants to make finer differentiations required to tease out the relevance of different patterns of dysfunction in a reliable manner. For example, clinical observations of proximal stability and equilibrium reactions might substantially improve the reliability of judgments regarding vestibular dysfunction. The addition of sensory history questionnaires, such as the Sensory Profile (Dunn, 1999) and the Sensory Processing Measure (Parham et al., 2007) would permit the reliability judgments to address sensory modulation issues as well as the perceptual–motor and praxis functions measured by the SIPT. Inclusion of a larger number of cases with SIPT profiles of varying complexity, and participants from diverse geographical locations, would strengthen generalizability of results.

Findings showed that participants relied on configural decision-making strategies and not on the diagnostic clusters. After completion of this study, the results were used to update SIPT interpretation content in the University of Southern California–Western Psychological Services program. Specifically, the computer-generated clusters were deemphasized, and greater emphasis was placed on examples of configural reasoning strategies. Future course revisions should address techniques to strengthen the reliability of SIPT score interpretation regarding the dysfunctional patterns that were found to have the weakest reliability in this study (vestibular dysfunction, somatosensory dysfunction, and bilateral integration and sequencing). ▲
Acknowledgments

The authors are grateful to the clinicians (participants) who generously donated their time to participate in the study. This study was conducted in partial fulfillment of Asha V. Asher’s requirements for the master of arts degree at the University of Southern California. The involvement of L. Diane Parham was supported in part by a grant from Maternal and Child Health (Grant #MCJ 009048).

This study was presented as a research poster at the American Occupational Therapy Association Annual Conference & Expo, April 2006, in Charlotte, North Carolina.

References


