Sensory Integration Therapy: Affect or Effect

(research, practice, developmental disabilities, learning disabilities)

Kenneth Ottenbacher

The results of studies examining the effectiveness of sensory integration therapy were reviewed, using recently developed quantitative methods that treat the literature review process as a unique type of research. Forty-nine studies were located initially. Eight of these studies met the following criteria: (a) they investigated the effect of sensory integration therapy; (b) they included dependent measures of academic achievement, motor or reflex performance, and/or language function; (c) they included a comparison between at least two groups; and (d) they reported quantitative results of the effect of sensory integration therapy. The 8 studies contained a total of 47 statistical hypothesis tests that evaluated the effectiveness of sensory integration therapy. An analysis of these tests, using quantitative reviewing methods, revealed that subjects participating in sensory integration therapy performed significantly better than members in the control groups who did not receive sensory integration therapy.

The application of sensory integration therapy to various client populations is discussed in relation to the existing empirical support revealed in the studies reviewed. The advantages of quantitative reviewing procedures are discussed, and use of the procedures with the developing occupational therapy research literature is recommended.

Sensory integration therapy is popular with occupational therapists and has been applied to various clinical populations. Price (1) observed that "sensory integrative therapy has been extensively applied to patients with neuromuscular dysfunction, as well as learning disability, retardation, and sensory impairment. Though most visible in pediatrics, it has a firm base and wide application in acquired brain dysfunction, specifically cerebrovascular accidents, usually at the upper end of the age continuum... Even more promising and challenging is the increasing role of the sensory integration approach in the treatment of psychiatric disorders, as preventive therapy with children, as well as remedial therapy with children, adolescents, and adults." (1, pp 287-288)

This widespread application of sensory integration therapy has not gone uncriticized. For example, Lerer (2) stated that "there is no valid, convincing proof in the limited studies conducted thus far on sensory integrative therapy to indicate that this treatment has directly remediated or helped anyone with a learning disorder of any kind" (2, p 4) and went on to recommend that sensory integration therapy only be "provided within the framework of experimental studies." (2, p 4) Other authorities have also raised concerns regarding the lack of empirical evidence substantiating the efficacy of sensory integration therapy for various populations (3, 4).

Sensory integration theory as developed by Ayres (5) has generated considerable research in occupational therapy and related areas; however, most of this research has been concerned with identifying particular neuropsychological characteristics associated with sensory integrative dysfunction (6-13). These studies resulted in a substantial body of evidence that some learning-disabled and neurologically impaired children exhibit neuropsychological dysfunctions, particularly in relation to the processing of vestibular-related sensory information. These studies have not, however, addressed in a controlled manner the question of whether such dysfunctions can be remediated by a particular therapeutic procedure. Hightower-Vandamm (14) commented on therapists' difficulties in establishing and providing sensory integration therapy. Some of these difficulties are undoubtedly related to the perception by physicians, educators,
and others that sensory integration therapy has not received adequate empirical testing.

The empirical legitimacy of any therapeutic regimen in the behavioral sciences is generally established through the statistical testing of hypotheses under controlled conditions. The hypothesis-testing model established by Fisher (15) in the 1930s and 1940s requires the researcher to formulate a null hypothesis complementary to the hypothesis the researcher is interested in supporting. Through the application of statistical tests the researcher demonstrates support or rejection of the null hypothesis at a predetermined probability level. The efficiency of the hypothesis-testing model has been enhanced over the years by the development of increasingly sophisticated experimental design procedures (16).

When a sufficient number of hypothesis tests on a particular topic (of interest) have been conducted, a review of the literature is carried out to determine what the aggregate studies reveal. Traditionally, these reviews consisted of a narrative description of the studies and their outcomes followed by an attempt to state a general conclusion. Although stringent criteria have been developed for the conduct of primary research studies, these same studies are often reviewed in a subjective manner with little or no objective criteria (17). The subjective and judgmental nature of traditional literature reviews is unfortunate because it is often these reviews of particular topic areas that are instrumental in establishing or refuting the empirical legitimacy of a finding.

Recently, techniques that synthesize and integrate bodies of research evidence in a quantitative manner have been developed and refined (17-19). These procedures treat the literature review process as a unique type of research and allow the investigator to quantitatively compare research studies and make consensus judgments based on the results. Such techniques facilitate the systematic investigation of variation across studies, including differences in sampling procedures and research design characteristics, and the use of multiple types of dependent variables. In addition, information can be obtained from comparing studies that would not be available in one large well-conducted study (20). These techniques have yet to be applied to occupational therapy research.

The purposes of this paper are: 1. to synthesize the existing evidence on the efficacy of sensory integration therapy as applied to various clinical populations using quantitative reviewing techniques, and 2. to expose consumers of research in occupational therapy to the methods and procedures of quantitative reviewing.

Methods

Potentially relevant studies were obtained through searches of Psychological Abstracts, Current Index to Journals in Education, Resources in Education (ERIC), and Dissertation Abstracts International. An online computer search was also conducted by using the Index Medicus database. In addition to the manual and computer searches, an examination of the bibliographies of retrieved studies (citation tracking) resulted in the location of additional information.

Relevance Decision Criteria. The search procedures yielded a total of 49 nonoverlapping research report titles that were broadly construed as potentially relevant to the topic. The abstracts and full reports of these studies were then judged for relevance on several specific criteria. The first criterion for inclusion in the review was that the study had to investigate the effects of sensory integration therapy, the independent variable. An operational definition adapted from Ayres (21), which delimited sensory integration therapy to only those procedures that employed sensory stimulation and adaptive responses involving total body movement, was used to determine whether the therapy used in a study met the criteria. The sensory stimulation procedures had to include some combination of vestibular, proprioceptive, or tactile stimulation and be based on the theoretical work of Ayres (5). In order to be classified as sensory integration therapy the treatment could not involve desk activities, speech training, reading lessons, or specific perceptual or motor skills training. The stated or implied goal of the therapy was to improve the way the brain processed and organized sensation, not to teach specific academic or motor skills. This first criterion served to differentiate sensory integration therapy from other forms of sensorimotor or perceptual-motor therapy. Studies that provided only controlled vestibular or tactile stimulation did not meet this criterion and therefore were not included in the review.

The second criterion for inclusion in the review was related to the type of dependent variable(s) employed in the study. One advantage of the quantitative reviewing technique is that it permits the use of broad, dependent variables. The purpose of this review was to evaluate the literature on the effectiveness of sensory integration therapy (i.e., to determine whether such therapy results in "developmental" gains). Improvement in development was
broadly defined by performance on any measure that evaluated (a) academic achievement, (b) improvement in motor and/or reflex integration, or (c) improvement in language function.

The last two criteria were related to the study's design and method of analysis. The study had to report a comparison between at least two groups—one that received sensory integration therapy and one that did not. In some cases, when a within-subjects experimental design was used, the comparison or control group was the same as the experimental group. Finally, the study had to report findings and results in a manner that allowed quantitative analysis.

Because they did not meet the above criteria, 19 of the 49 studies were eliminated after available abstracts and titles were reviewed. Another 20 studies were eliminated for similar reasons after the full report was scrutinized. The remaining ten studies met the criteria outlined previously. However, in two instances reports were found that analyzed the same data from the same sample of subjects (22-25); only two of these four studies were included in the review. In both cases the study with the more complete or more recent presentation of the data was included in the analysis. The deletion of the two repeated studies left a total of eight studies that met the criteria.

Some studies made comparisons between more than two groups. For example, Montgomery and Richter (26) tested the effect of a program of sensory integration therapy on 75 mentally retarded children divided into three groups. One experimental group received sensory integration therapy, whereas a second experimental group received a program of developmental physical education activities. The control group remained in regular physical education class but received no special training. For purposes of quantitative analysis, only the data from the sensory integration therapy group and the control group were analyzed. The comparisons involving additional experimental groups receiving other types of therapy were not considered in the analysis for any of the studies included in the review. Thus, all of the comparisons analyzed involved a group of subjects receiving sensory integration therapy and a control group not receiving therapy. In all, 47 statistical hypothesis tests of such two-group comparisons were contained in the 8 studies.

Characteristics of Individual Studies. With the selection criteria for the studies defined and the general boundaries of the review determined, the next step was to identify aspects of the studies that might be related to study outcomes. These variables fell into four general categories: 1. subject characteristics (i.e., number of subjects and their mean ages); 2. design characteristics, including (a) how the subjects were selected, (b) the type of assignment used (i.e., random or matched), (c) the amount of time a group spent receiving therapy, and (d) the type of dependent variable used to measure improvement (i.e., academic achievement, motor reflex abilities, or language function); 3. aspects of the study's outcome such as statistical test used, test value reported, accompanying probability level, z-score corresponding to the probability level, and degrees of freedom; and 4. retrieval characteristics (i.e., the year of publication, whether published in a journal or another source, and how the study was located). Information on the categories was coded for each individual study and subjected to computer analysis.

Quantifying Outcomes. Two primary methods—combining probabilities by adding z-scores and explaining variation in study effect sizes—were used to quantitatively compare the results of the 47 statistical hypothesis tests included in the 8 studies. The first method of aggregation, referred to as the Stouffer Method (27), is easy to compute when probability levels are reported. Probability values and z-scores are computed for each hypothesis test, and the z-scores are summed and then divided by the square root of the number of hypothesis tests. The result is itself a z-score that can be interpreted as gauging the probability that the set of study results could have been generated by chance. The method of obtaining the overall or combined z, referred to as the Zma, is described by Rosenthal (28).

In the present application, if a study reported a nonsignificant result and no p-level was given, a p-level of .50 and accompanying z-score of 0.00 were assumed.

The method of combining probabilities has several inadequacies. First, studies with significant p-levels are more likely to be published than are studies with nonsignificant p-levels (29). Rosenthal (30) and Cooper (31) present equivalent procedures for estimating the potential impact of relevant but unretrieved null studies.

A second shortcoming is that the technique does not tap the wealth of information contained in the variation in results from one study to another. Procedures capable of uncovering systematic variation in study results have been pioneered by Glass (17, 18, 32). Glass (17) refers to the synthesis of quantitative outcomes across studies as “meta-
analysis," which he defines as the "statistical analysis of a large collection of analyses results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature." (p 3) Glass' procedures involve the calculation of study effect sizes (33) and the correlation of these with study characteristics. O'Brien and Shapiro (34) point out that the usual criterion for gauging the importance of a study—its level of significance—is strongly influenced by the size of the sample involved. Significance testing, then, which compares an observed relation to the chance of no relation, becomes less informative as evidence supporting a phenomenon accumulates. The question turns from whether an effect exists to how much of an effect exists. Effect size measures that are free of sample-size influence can play a vital role in determining the degree to which a treatment exerts an effect on a population.

Effect Size Estimation. An effect size was calculated for each of the 47 hypothesis tests. The effect size used was the d-index (33). The d-index gauges the difference between two group means in terms of their common (average) standard deviation. If \( d = .3 \), it means that 3/10 of a standard deviation separates the average person in the two groups. This effect size transforms the results from any two-group comparison into a common standardized metric regardless of the original measurement scales. There are effect size measures appropriate for any type of research design or statistical analysis (33). Cohen presents a description of various effect sizes and tables to compute power, sample size, and significance level (33). The d-index was used because it is simple to compute, it is "scale free," and it is applicable to a plurality of studies in occupational therapy.

Effect sizes can be computed from \( t \) and \( F \) ratios when means and standard deviations are not reported in an article. Friedman (35) has provided formulas and a rationale for transforming \( t \) and \( F \) values to \( d \)-indexes. In instances where \( t \) and \( F \) ratios are not reported, they may be estimated from the significance level and the sample size (18). When nonparametric statistics or percentages are reported, effect sizes can be computed using procedures described by Glass (18, 32). Also, Hedges (36) has noted that \( d \)-indexes may be biased as the sample size becomes smaller (< 50), so Hedges' correction factors should be employed to adjust for potentially inflated effect sizes. Finally, in this analysis, when a study reported a nonsignificant result but not enough information to determine the effect size, an effect size of 0.00 was assumed.

Cohen (33) presents several measures of distribution overlap meant to enhance the interpretability of effect size indexes. The overlap measure employed in this review, called \( U_3 \), tells the percentage of the population with the smaller mean that is exceeded by the average person in the population with the larger mean. A table for converting the \( d \)-index to \( U_3 \) is presented by Cohen (33, p 22).

### Results

**Description of the Literature.** A total of 317 subjects participated in the 47 hypothesis tests found in the 8 studies reviewed. The subjects ranged in age from about 4 to 62 years. Eighty-nine of the subjects had a diagnosis of mental retardation, 191 were diagnosed as learn-
ing disabled, 18 were diagnosed as aphasic, and 19 were diagnosed as "at risk" for learning (reading) disorder.

*Type of Assignment.* Some form of random assignment was used in four of the studies, three studies used a combination of random assignment and matching on selected variables, and one study involved pre-existing groups with no experimental manipulation in terms of assignment.

*Description of Quantitative Analysis.* Effect size analysis: Table 1 lists the eight studies included in the review, the mean $d$-index for that study, the associated $U_3$ value, and the number of hypothesis tests. The works by Glass (17, 18, 32), and Cohen (33), Cooper (31), and Friedman (35) contain specific instructions concerning the calculation of $d$-indexes and other effect sizes and examples of how effect sizes are translated into $U_3$ values.

The mean $d$-index for the 47 total hypothesis tests was .79 (SD = .29). Cohen proposed labels for various effect sizes to describe their relative magnitude (33). In the behavioral sciences Cohen (33) interprets a $d$-index of .20 to .50 as small, .50 to .80 as medium, and greater than .80 as large. According to Cohen, the effect size boundaries are relative (33). The $U_3$ associated with a $d$-index of .79 is 78.8. This means that the average performance of subjects in the experimental groups receiving sensory integration therapy was better than 78.8 percent of the subjects in the control groups not receiving sensory integration therapy (see Figure 1).

The hypothesis tests were then broken down according to the type of dependent variable used. Nineteen tests used dependent variables that measured motor or reflex performance. Twelve of the 19 hypothesis tests found a significant effect ($p < .05$) in favor of the treatment groups, whereas 7 found no difference. The $d$-index for the motor/reflex dependent variable was 1.03 (SD = .67). The $U_3$ of 84.8 meant that the average subject receiving sensory integration therapy performed better than 84.8 percent of the subjects in the control groups when a measure of motor/reflex performance was used to evaluate improvement.

Seventeen of the 47 hypothesis tests employed a dependent variable of academic performance or achievement to measure outcome. Nine found a significant effect ($p < .05$) in favor of the experimental groups, whereas 8 found no difference. The $d$-index for the dependent variable of academic achievement was .75 (SD = .39) and the $U_3$, 77.3. This $U_3$ indicates that the average subject in the experimental groups receiving sensory integration therapy performed better than 77.3 percent of the subjects in the control groups when the study used a dependent variable of academic achievement.

Finally, 11 of the 47 hypothesis tests used a dependent variable of language function. Six of the tests found no difference between the experimental and control groups, whereas four tests found significant effects ($p < .05$) for the treatment groups. One test found a significant effect in favor of the control group. The $d$-index for the 11 hypothesis tests using language measures as dependent variables was .43 (SD = .57). The $U_3$ was 66.6, indicating that the average subject receiving sensory integration therapy performed better than 66.6 percent of the subjects in the control groups when the dependent variable was a measure of language function.

The effect size analysis by dependent variable revealed that sensory integration therapy had its greatest
effect when the dependent measure was some type of motor or reflex evaluation and its least effect when the dependent variable was some type of language measure.

Effect sizes were next analyzed according to diagnostic category. Seventeen hypothesis tests involved subjects diagnosed as mentally retarded. The d-index for hypothesis tests that involved mentally retarded subjects was \( d = .65 \) \((SD = .49)\), whereas the \( U_3 \) was 69.8. This means that, for studies involving mentally retarded individuals, the average subject in the experimental groups receiving sensory integration therapy performed better than 69.8 percent of the mentally retarded subjects in the control groups.

Fifteen hypothesis tests were found in which learning-disabled subjects were used. The d-index for hypothesis tests using learning-disabled subjects was \( d = .68 \) \((SD = .34)\), and the \( U_3 \) was 75.2, indicating that the average learning-disabled student receiving sensory integration therapy performed better than 75.2 percent of the learning-disabled subjects not receiving sensory integration therapy.

The remaining 15 hypothesis tests used subjects diagnosed as “at risk” for learning disorder or aphasic. The d-index for “at risk” or aphasic subjects was \( d = 1.20 \) \((SD = .65)\), and the \( U_3 \) was 88.5. This indicates that, for subjects diagnosed as “at risk” or aphasic, the average subject receiving sensory integration therapy scored better than 88.5 percent of the subjects in the control groups.

The analysis by diagnostic category suggests that sensory integration therapy has its greatest effect on subjects diagnosed as aphasic or “at risk,” and the least effect on subjects diagnosed as mentally retarded. It is interesting to note that the diagnostic category of “at risk” or aphasic represented the youngest age group, whereas the mentally retarded category represented the oldest category.

A possible confound in these results exists between the type of dependent measure used and the diagnostic category. Since it was demonstrated previously that the largest effect for dependent variables was found for motor/reflex measures, it is possible that if a majority of the hypothesis tests with “at risk” or aphasic subjects used a dependent measure of motor/reflex performance a confound may exist.

To test this possibility a chi-square (with Yates correction) was computed between the type of dependent variable (motor/reflex vs. academic/achievement vs. language function) and the diagnostic category (mentally retarded vs. learning disabled vs. “at risk”/aphasic) for the 47 hypothesis tests. The result, \( X^2 = 20.3 \) \((p < .01; df = 4)\), indicates that a confound may exist between the variables of type of dependent measure and diagnostic category.

This confound makes it more difficult to interpret the effect of sensory integration therapy for the various diagnostic categories.

**Combined probability analysis:**

The overall \( Z \) for the analysis was 11.65 \((p < .001, n = 47)\). This indicates a highly significant effect for the combined experimental groups receiving sensory integration therapy when compared with the combined control groups not receiving therapy. A fail-safe \( N \) \((N_{fs} = 2305)\) was computed using procedures described by Cooper (31). The fail-safe \( N \) indicates the number of additional hypothesis tests with summed null findings that would be needed to increase the probability of the \( Z \) above .05. For the overall \( Z \) the \( N_{fs} \) was 2305. This means that 2305 tests with null results would be needed to raise the overall \( Z \) above .05, in other words, to make it nonsignificant.

Cooper described a limitation of the fail-safe \( N \). “It is an appropriate guide for the reader only if the assumption of a summed null relation in undiscovered studies is acceptable. It is always possible that a smaller number of studies exist that have a summed \( Z \) score of equal but negative value to the sum of those reviewed. The plausibility of this alternative also should be considered by the reader.” (31, p 135)

The breakdown according to type of dependent measure revealed a \( Z \) of 9.01 \((p < .001, n = 19, N_{fs} = 550)\) for dependent variables of motor/reflex function; a \( Z \) of 7.12 \((p < .001, n = 17, N_{fs} = 301)\) for measures of academic achievement; and a \( Z \) of 3.42 \((p < .01, n = 11, N_{fs} = 37)\) for measures of language function. The breakdown by diagnostic category revealed the following \( Z \) values:

- Mentally retarded subjects—\( Z = 7.09 \) \((p < .001, n = 17, N_{fs} = 146)\).
- Learning-disabled subjects—\( Z = 7.23 \) \((p < .001, n = 15, N_{fs} = 275)\).
- “At risk”/aphasic subjects—\( Z = 7.98 \) \((p < .001, n = 15, N_{fs} = 337)\).

The interpretation of these findings is confounded by the relationship between the dependent measures and diagnostic category as discussed earlier.

**Discussion and Conclusion**

The decision to include in the review only studies that met certain criteria eliminated a larger number of studies. Several of these studies used sensorimotor procedures very similar to those employed in sensory integration therapy (37) or used sensory integration procedures in conjunction with other forms of sensorimotor therapy (38). For example, Morrison and Pothier (38) em-
ployed a combined, modified version of Kephart (39) and Ayres' (5) therapeutic procedures with 27 mentally retarded children. They found statistically significant results in favor of the sensorimotor group on measures of overall development, gross motor ability, and language development. The inclusion of studies such as that by Morrison and Pothier (38) would have strengthened the outcome of the analysis, but would have introduced a possible confound with the independent variable: sensory integration therapy.

**Advantages and Limitations of Quantitative Reviewing.** As the empirical base of occupational therapy expands, therapists needing summaries of accumulated research evidence will find themselves relying heavily on the efforts of literature reviewers. Time constraints will make it impossible to keep abreast of primary data reports except within a few specialized areas of particular interest. Occupational therapy is in a unique position to begin integrating separate research projects using procedures that involve scientific inferences as central to the validity of empirical knowledge as inferences involved in primary data interpretation. The validity of literature review conclusions cannot be taken for granted. A reviewer performing a traditional narrative literature review makes numerous decisions bearing on the outcome of the review and each choice contains threats to the validity of the outcome. If occupational therapy is to take advantage of the use of the procedures illustrated here to synthesize empirical research results, then reviewers must be required to use the same rigorous methods that are required of primary researchers. Although substantial attention has been paid to validity issues in primary research (16), only recently have guidelines been developed for systematically evaluating the research review process (17, 18). These guidelines promise to help researchers accumulate and interpret scientific knowledge. As bodies of research in areas of occupational therapy practice expand, it is hoped that the use of quantitative reviewing techniques will help to establish the efficacy of occupational therapy procedures.

Quantitative reviewing procedures are only appropriate for a particular type of review. From among the four primary types of reviews identified by Jackson—(a) those that size up new substantive and/or methodological developments, (b) those that verify existing theories or develop new ones, (c) those that synthesize knowledge from different lines of research, and (d) those that infer generalizations about substantive issues from a set of studies directly bearing on those issues (40, p 438)—quantitative methods are appropriate only for the last type of review.

While quantitative reviewing procedures can clarify some problems of conceptualization and methodological artifact, they cannot resolve these problems. As in evaluating a single study, alternate conceptualizations of included independent variables may rival the one offered. Also, since multiple results are derived from one study, the ensuing data cannot be considered independent and may lead to inflated or unreliable estimates derived from inferential statistical procedures. The question of whether or not to employ inferential procedures in the analysis is complex; Glass and colleagues noted: "a facile solution to this problem of non-independence would be to average all findings within a study up to the level of the study and proceed with a meta-analysis with 'studies' as the unit of analysis. No doubt there will be instances in which this resolution of the problem will be satisfactory. But in most instances, it is likely to obscure many important questions that can only be addressed at the 'within study' level." (41, p 229)

**Implications.** The results of this research have demonstrated the effect of sensory integration therapy in the studies reviewed; however, the limitations of this review should be emphasized. In spite of the widespread application of sensory integration therapy, only eight studies that met the pre-established criteria for inclusion in the review were found. The criteria for inclusion in the review were specifically developed to meet empirical standards commensurate with traditional inferential research in the behavioral sciences. Clearly more studies are needed to resolve some of the questions of interpretation.

The effect of sensory integration therapy applied to the representative populations appears to have empirical support. However, the justification for some applications of sensory integration therapy may be more affect than demonstrated effect. No studies meeting the criteria were found that evaluated the effect of sensory integration therapy applied to client populations diagnosed as having emotional or psychiatric disorders. Nor were any studies meeting the criteria retrieved that evaluated the effect of sensory integration therapy with a geriatric or physically disabled (i.e., CVA) population. The argument could be made that the empirical base for applying sensory integration ther-
apy to individuals with these disorders does not yet exist in spite of Price's (1) observation that sensory integration therapy has a "firm base and wide application" in such populations. There is obvious professional and personal risk in applying any treatment procedure that has not been demonstrated to be therapeutically effective by the traditionally accepted empirical standards. This is particularly true if such a procedure is being advocated for particular client populations does not yet exist in spite of "preventive therapy." The possibility of premature application of sensory integration therapy to particular client populations should be carefully considered by therapists using sensory integrative procedures.

REFERENCES
5. Ayres AJ: Sensory Integration and Learning Disorders, Los Angeles, CA: Western Psychological Services, 1972

Phys Occup Ther Pediatr 1: 31-38, 1980
27. Montgomery P, Richter E: Effect of sensory integrative therapy on the neuro-motor development of retarded chil-