Interrater and test-retest reliability of the Behavioral Assessment Scale of Oral Functions in Feeding were computed for two separate samples of severely and profoundly handicapped students having oral-motor and feeding disorders. Reliability was assessed using the intraclass correlation-generalizability theory approach and resulted in coefficients ranging from .68 to .84. These values were considered only marginally acceptable by current standards of instrument development advocated in the behavioral sciences. This paper discusses the difficulty in developing psychometrically satisfactory assessment instruments for use with the severely and profoundly developmentally disabled and gives implications for practice and future research.

Oral-motor facilitation techniques are widely employed by occupational therapists to improve functional feeding skills in the developmentally disabled (1-3). For example, Stratton (4) recently observed that “therapeutic feeding techniques are often the most realistic and practical treatment approaches for the profoundly retarded, multihandicapped adolescent or adult” (p 719). Treatment procedures that employ sensory-motor facilitation procedures designed to reduce abnormal oral reflex activity and normalize the feeding process are often the most realistic and practical treatment approaches for the profoundly retarded, multihandicapped adolescent or adult. (p 719). Treatment procedures that employ sensory-motor facilitation procedures designed to reduce abnormal oral reflex activity and normalize the feeding process are widely employed by therapists (5-7). Despite the popularity of oral facilitation treatment strategies in programs for the developmentally disabled, there is little evidence indicating that the procedures are effective in normalizing feeding patterns. Snell (8) recently noted that there have been very few empirical tests of intervention procedures recommended for use with oral-motor dysfunction.

One of the obvious impediments to research in the area of oral-motor habilitation is the lack of evaluation or assessment instruments with reported reliability and validity. For instance, Ottenbacher, Scoggins, and Wayland (2) reported on the efficacy of a program of oral-motor therapy based on principles outlined by Mueller (9) and Rood as reported in Stockmeyer (10) and presented by Gallender (5). Analysis of the data obtained from 20 severely and profoundly retarded subjects using a pretest-posttest control group design revealed no statistically significant difference in body weight gain or development of specific feeding behaviors over the nine-week treatment period. In discussing some of the limitations of their study, Ottenbacher and others (2) reported that the relatively low in-
terrater reliability of the informal evaluation they employed “points out another area in need of further research: the development of a comprehensive standardized evaluation of oral-motor and feeding function” (p 157).

Stratton (4) recently presented and described the Behavioral Assessment Scale of Oral Functions in Feeding, which “provides an objective and graded method of documenting the major aspects of oral function in relation to feeding difficulties” (p 719). This scale measures variations of nine feeding-related behaviors, including jaw closure, lip closure, three variations of swallowing, tongue control, chewing skills, and the ability to sip liquids. Each behavior is rated on a scale from 0 to 5. A score of 0 indicates a totally passive response to the feeding process, while a score of 5 represents a functionally normal response. [See Stratton (4) for more detailed information.]

The purpose of the present study was to provide initial data on the psychometric properties of Stratton’s scale. Specifically, the study was designed to provide information on interrater reliability, test-retest reliability, and the internal consistency of the assessment.

Method

The scale was administered to 46 developmentally delayed students according to the instructions and guidelines presented by Stratton (4). The students ranged in age from 10 months to 258 months (mean = 124 mo; SD = 43.4) and were all residents of a state-supported facility located in the upper midwest. All students were identified as severely and profoundly retarded by a multidisciplinary habilitation team and were dependent in most daily living skills, including feeding. Residents in the institutional setting were routinely fed by aids.

The scale was administered by four registered occupational therapists who were experienced in the habilitation of the developmentally disabled. These therapists had a combined total of more than 30 years of experience in working with the severely and profoundly handicapped.

Two pairs of therapists assessed the degree of interrater and test-retest reliability for two separate samples of residents. The therapists were selected based on their interest in the study and were assigned to a unit within the institution that contained appropriate subjects for the study. The first pair of observers (therapists 1 and 2) administered the scale to 23 students (group A). Both examiners independently scored the student’s “feeding behaviors” as the student was fed by his or her regular attendant. The students were fed in their natural environments by the person who was usually responsible for this task. No unusual procedures or positioning were initiated during the assessment. The examiners did not confer with each other or with the feeders during the evaluation. In describing guidelines for administering the scale, Stratton (4) states, “the assessment of eating skills is most effective if completed during the actual feeding session with the client in the position and fed in the manner that most closely represents the average meal” (p 719).

The second pair of examiners (therapists 3 and 4) also administered the scale to 23 students (group B) in the manner described above. This procedure resulted in two separate interrater reliability studies being completed on a total of 46 severely and profoundly handicapped students. Each student was rated by either therapists 1 and 2 or therapists 3 and 4. Table 1 gives descriptive information for the students evaluated by each pair of therapists.

Within ten days after the initial administration of the scale, one of the therapists from the original pair of examiners returned and readministered the assessment to each student (groups A and B). Thus, interrater and retest data were collected for the two separate samples of students residing in the same institutional facility. Every attempt was made to administer the retest at approximately the same time of day (usually the noon meal) and with the same attendant feeding the student. The data collected were transcribed and subjected to statistical analysis.

Results

The ages of students in the two groups differed significantly (t = 8.50; p < .01; df = 44): the students in group B were significantly older than those in group A. The two groups of students were selected from different units within the residential setting. The units were formed on the basis of age; hence the significant difference in age between the two groups was expected. The ratio of males to females in each sample was relatively

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Descriptive Information for the Two Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group A</td>
</tr>
<tr>
<td>Total N</td>
<td>23</td>
</tr>
<tr>
<td>Males</td>
<td>13</td>
</tr>
<tr>
<td>Females</td>
<td>10</td>
</tr>
<tr>
<td>Age, mo</td>
<td>80.37</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>26.99</td>
</tr>
<tr>
<td>No. of days between first and second test</td>
<td>6.50</td>
</tr>
<tr>
<td>Mean</td>
<td>1.21</td>
</tr>
</tbody>
</table>
equal, with slightly more males in group B (see Table 1). The overall scores for the two groups were very similar. The mean scores for students in group A were 35.90 (therapist 1) and 34.71 (therapist 2), respectively ($t < 1.0; p = NS$). The mean scores for students in group B were 36.91 (therapist 3) and 36.03 (therapist 4), respectively ($t < 1.0; p = NS$). These overall scores indicate that the oral-motor and feeding abilities of the two samples, as measured by the Behavioral Assessment Scale of Oral Functions in Feeding, were very similar both across therapists and across groups.

A measure of internal consistency was computed for each group of students. This measure was Cronbach's (11) coefficient alpha. Coefficient alphas were calculated for the two samples for which the therapist completed the evaluation. They appear in Table 2. The interpretation of coefficient alpha is closely related to the interpretations given for reliability estimates based on split-half methods. Coefficient alpha can be considered a unique estimate of the expected correlation of one test with an alternate form containing the same number of items (12). The coefficient alpha is generally considered a better measure of internal consistency than are the alternate forms or split-half methods (13, 14).

Interrater and test-retest reliability coefficients were computed using the intraclass correlation-generalizability theory approach. Berk (15) has observed that the computation of interrater reliability "requires a methodology capable of treating quantitative as well as categorical data with equal precision" (p 462). He reviewed several commonly used statistical procedures and concluded that the intraclass correlation method provided the precision, comprehensiveness, and flexibility required to deal with the complexities associated with reliability assessment (15). The theory of generalizability enables investigators to make a detailed analysis, which is not possible when more conventional procedures are used. The components of reliability are determined by the specific conditions under which it is obtained, such as time, setting, and observer. The contributions of these conditions are determined by analysis of variance components, intraclass correlation coefficients, and measurement error statistics. The large variance components associated with a dimension related to reliability serve as a warning that the generalizability of the observations may be limited. For example, a large variance component associated with occasions of observation (test and retest) suggests that observations or measurements taken at a different time will produce different results. Traditional procedures, such as chi-square or the Pearson product-moment correlation, have several limitations as measures of reliability or, more precisely, agreement (16, 17). Chi-square and the product-moment correlation are both measures of association rather than agreement. Departures from chance in the direction of disagreement affect chi-square as much as departures in the direction of agreement. A related problem for the product-moment correlation is additive bias. A perfect correlation of 1.0 can be obtained when judges or examiners never agree exactly, as long as ratings are proportional (i.e., when a second examiner consistently scores a student on a level or step below the first examiner).

In view of these limitations, Tinley and Weiss (18) argue, "At the present time the intraclass correlation is the most appropriate measure of interrater reliability" (p 363). Actually, there are several versions of the intraclass correlation coefficient (16, 19). The particular case of the intraclass correlation coefficient used in this study corresponds to a random-effects analysis of variance (ANOVA) model described by Berk (15) in which subjects and observers serve as independent factors.

### Interrater Reliability

A subjects-by-observers random effects ANOVA provided the estimates of variance components from which the various generalizability coefficients were derived. Two types of coefficients were computed for each sample: a) the generalizability of a single examiner, which estimates the average between one examiner and another, and b) the generalizability of the examiners, which estimates the average agreement between the sample of examiners employed in the study and a hypothetical set of other examiners from the same population (15). For group A, the intraclass correlation coefficient for agreement between therapists 1 and 2 was .72, whereas the intraclass correlation estimating the average agreement between the two examiners and a theoretical set of other randomly selected examiners was .81. For group B, the intraclass

---

**Table 2**

Coefficient Alpha for Each of the Administrations of the "Behavioral Assessment Scale of Oral Functions in Feeding"

<table>
<thead>
<tr>
<th>Therapist</th>
<th>Group A</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>($N = 23$)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>($N = 23$)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>($N = 23$)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>($N = 23$)</td>
<td></td>
</tr>
</tbody>
</table>
correlation coefficient for agreement between therapists 3 and 4 was .76, whereas the intraclass correlation estimating the average agreement between the two therapists and a theoretical set of other similar examiners was .84.

The estimated between-examiner variance (examiner bias), which is computed as part of the formula, indicated that examiner bias was a negligible source of error in the computations for both groups.

Test-retest reliability

Reliability coefficients for test-retest were computed using the intraclass correlation method described above. A modification was made in the design model, so that a random effect model with subjects by occasions (test and retest) was used. Only one intraclass coefficient was computed for each group, indicating the agreement for a particular examiner over time. For group A, the coefficient was .68; for group B, the coefficient was .79.

Most treatments of the intraclass correlation for interrater agreement indicate that skewed distributions of ratings, with their concomitant reduction in variance, should be considered a result of rater bias. That is, it is assumed that the underlying ability or trait being measured is distributed normally and that significant departures from normality and reduced variance in obtained ratings occur because the sample of examiners or raters is predisposed to use one end of the measurement scale. This assumption may not be justified in the present study. Oral-motor functions may represent a unique case in that a skewed distribution of ratings, with the subsequent reduced variance, may be the result of the underlying nature of oral-motor pathology rather than examiner bias. If the Behavioral Assessment Scale of Oral Functions in Feeding scores for severely handicapped students are represented by a positively skewed distribution in which most scores fall at the lower end of the rating scale, then it will be difficult to obtain high intraclass correlations for interrater agreement. Finn (20) proposed a reliability index that corrects for the reduced variance associated with a skewed distribution by comparing the obtained error variance with that which would have occurred had the ratings been randomly assigned. To determine if such a correction was indicated for the data collected in this study, skewness (third-moment) coefficients were computed for the response recorded from both groups of students. These values ranged from .32 to .81, which suggests that the correction advocated by Finn (20) for skewed distributions was not indicated.

Discussion and Conclusions

The importance of developing methods of evaluating the efficacy of oral-motor therapy programs for developmentally disabled persons has received considerable attention in recent professional literature (1, 21, 22). The Behavioral Assessment Scale of Oral Functions in Feeding is one of the first published attempts by an occupational therapist to develop a systematic method of measuring oral function and feeding skills in the developmentally disabled. Stratton (4) states that the evaluation provides "an objective, graded format that can be used in documenting levels of skill/deficit as they relate to feeding difficulties" (p. 721). The extent to which this scale is capable of achieving the above goal is directly related to how reliable the assessment is. The results of our investigation suggest that the reliability of the evaluation is marginal. The individual interrater reliability coefficients for the two separate samples were .72 and .76.

When an assessment device relies on the ability of a therapist to observe and record patient responses, Berk and DeGangi (23) state, "It is crucial that the observations of one therapist be replicable with other therapists. That is, ideally, regardless of who administers the scale to an individual, the observed performance should be virtually identical" (p. 243). These researchers go on to argue that the interrater reliability coefficients should exceed .80 for scales measuring motor-based performance. Other authorities, notably Nunnally (13), have observed that if an instrument is to be used for diagnostic purposes, that is, to determine whether a person will qualify for services, then the interrater reliability should be at least .90.

The test-retest reliabilities were also found to be only marginally acceptable by traditional psychometric standards. The test-retest for the two samples (groups A and B) were .68 and .79, respectively. Although test-retest correlations represent an intuitively appealing procedure to assess reliability, they have been found to have several limitations (14). For example, test-retest coefficients may reflect changes in performance that occur between test administrations. Reactivity may also be a problem with test-retest data. Because the interval between tests in this study was relatively short (less than 11 days) and because the test-retest coefficients for both samples were similar, the estimates of the stability of the assessment procedure are probably relatively accurate.

Another reliability-related issue concerns the expertise and experience of the therapists administering the assessment. The four ther-
therapists involved in collecting data in this investigation were all experienced in providing services to severely disabled patients and had specific expertise in the area of oral-motor habilitation. The examiners' experience probably "inflated" the coefficients that were obtained. It is likely that the values would have been lower if the assessment had been administered by inexperienced therapists. However, this is only speculation, and additional research will be required to resolve this issue.

The problem of what is a satisfactory level of reliability is a complex one with no obvious answer. The difficulty of measuring performance in severely and profoundly handicapped populations is widely acknowledged (24). The low incidence of the disorders, the heterogeneity of the populations involved, the restricted range of functional performance combine to make the development of a psychometrically sophisticated assessment device for this population a very difficult task.

As a rule, educators and psychologists established .80 as the minimum for estimates of interrater and test-retest reliability. At that level, correlations are attenuated by relatively little random measurement error. It should be noted, however, that behavioral scientists do not completely agree on this level. Thus, the coefficients reported in this investigation are only marginally acceptable. However, we agree with Carmines and Zeller (14) who said, "the most important thing ... is to report the reliability of the scale and how it was calculated. Then other researchers can determine for themselves whether it is adequate for any particular purpose" (p. 51).

The construction of any evaluation or assessment is an evolutionary and dynamic process. We hope that other investigators and test developers will continue the work initiated by Stratton (4). Perhaps an expansion of the assessment or a refinement of the item pool will produce an evaluation instrument that has reliabilities more closely approaching those traditionally accepted as adequate in behavioral sciences research. As Gillette (25) recently stated, "The almost total lack of standardized evaluation tools on which to base outcome research is a strong deterrent to an expanded data base showing outcomes of occupational therapy intervention" (p. 501). This observation is particularly pertinent to the development of oral-motor habilitation programs for the severely and profoundly handicapped.

ACKNOWLEDGMENTS

The authors thank the administration, the staff, and the students of the Central Wisconsin Center for the Developmentally Disabled for their assistance and cooperation. This study was partially supported by grants from the American Occupational Therapy Foundation and the Hilldale Trust Fund (School of Allied Health Professions, University of Wisconsin–Madison).

REFERENCES