The Multidisciplinary Feeding Profile: A Statistically Based Protocol for Assessment of Dependent Feeders

The Hugh MacMillan Medical Centre, Toronto, Ontario, Canada

Abstract. The development of the multidisciplinary feeding profile entailed a level of statistical analyses not commonly utilized in test development. This paper describes the statistical analyses and offers an explanation of why specific statistical tests were chosen. It also serves to identify where clinical knowledge and experience overrode specific statistical tests.

Key words: Feeding disorders, profile - Evaluation, statistical

The multidisciplinary feeding profile (MFP) is the first attempt to design a statistically based quantitative assessment protocol for children who are unable to feed themselves. The previous paper described the numerous factors that affect the oral preparatory phase of swallowing. In addition, current clinical assessment methods were outlined, the scope and rationale for the MFP developed, and the concept of content validity described. Finally, the complete MFP test form was provided [1].

A multidisciplinary team of 9 persons generated 198 test items or questions in 6 different categories based upon their combined knowledge (content) of assessment and treatment of children with feeding disorders. The 198 item test was applied by three raters to 8 children in a pilot study and each test item was analysed for interrater reliability and clinical value and then maintained, clarified, or discarded. The main study contained 146 items in 6 categories and was applied by 3 different raters to 18 children. Each test item had a number of choices or option cells. Each option cell was assigned a value from 1 to 5 based on the group evaluation of how important the option choice was to actual feeding behavior. Once again, individual test items were analysed for weakness and clarified or discarded. The MFP presented in the previous paper contains 136 test items chosen for additional analysis and 56 items that were reworded or altered but were included for analysis in future revisions. These 56 items were identified in the MFP as not analyzed (NA). Once individual items were selected, the six sections were tested for variance and tests of rater agreement and consistency were applied.

This paper presents the statistical analyses used and the results obtained and identifies where subjective decisions were made in their application. The next stages of development of the MFP will also be described.

Methods

Analysis and Selection of Test Items

Weighted kappa coefficients [2] were calculated for each pair of raters for each of the 198 test items of the pilot study and the 146 items of the main study as a measure of interrater agreement. The weighted kappa values identified items of disagreement among raters, so that these questions could be eliminated or reworded. The arbitrary threshold kappa value of 0.40 was chosen because it represents a moderate strength of rater agreement [3]. Items with kappa values greater than 0.40 were retained for the MFP. Items with weighted kappa values between 0.21 and 0.40, inclusively, were judged to have only a fair strength of rater agreement. Consequently, they were examined for their clinical/diagnostic value and were retained if judged to be of clinical value. Items that produced kappa values less than 0.20 for all combinations of rater pairs were removed from the MFP if judged to be of minimal clinical value. However, some items in this low-agreement category were judged to be clinically important and so were retained but not analyzed.
Section Analysis

A subject's score for each of the sections and his or her total score for all six sections were calculated by summation of the value assigned to each cell chosen by the rater. The distribution of the section scores and the total scores of the 18 subjects for each rater were examined for normality and homogeneity of variance to determine if the two-way analysis of variance was appropriate. This mean, variance, skewness, and kurtosis of each distribution of the section and total scores were calculated to see if transformation of the data was required. The assumptions of the analysis of variance (ANOVA) [4] were satisfied and two-way repeated measures ANOVAs were performed. The scores were partitioned into variation among subjects, variation among raters, and error.

The section scores and the total scores were subjected to a two-way repeated measures ANOVA to assess the relative magnitude of variation from different sources and to ascertain whether a particular part of the variation was greater than expected under the null hypothesis [4]. Two sets of hypotheses were tested in the ANOVA procedure. The first set of hypotheses considered each subject's abilities as a separate population. The null hypothesis that the population mean scores for all subjects in general are equal to one another was tested against the alternative hypothesis that at least one subject's abilities were different from the other subject's abilities. The second set of hypotheses considers each rater's scores assigned by all raters as a separate population. The null hypothesis that the scores given by all raters, in general, are equal to one another, was tested against the alternative hypothesis that at least one rater scored the subjects differently from other raters.

Two different forms of the intraclass correlation coefficient (ICC) were used to quantify reliability, as recommended by Shrout and Fleiss [5]. One measured rater agreement and the other measured rater consistency.

Results

Test Item Analysis

Weighted kappa coefficients were calculated for each of the 198 items of the pilot study and the 146 items of the main study for each pair of raters. The mean of the 436 weighted kappa coefficients of the main study was 0.36 with a standard deviation of 0.26. Approximately 13% of the kappa values were 0 or less while 50% of the kappa values were greater than 0.3. Kappa values of one, indicating perfect agreement between two raters, occurred in 14 items of 436, or about 3% of the time.

Table 1 shows an example of a high kappa value that indicated strong rater agreement. There was strong agreement between two raters on the assessment of a subject's ability to clear excess food from the lips with the tongue.

In some cases, such as the item that assessed trunk posture, the kappa value was 0. This indicated that rater agreement was due only to chance. When reviewed, the item was judged to be of significant clinical value so it was retained in the revised MFP, but not included in further analysis.

Table 1. Cross tabulation for weighted kappa

<table>
<thead>
<tr>
<th>Value assigned by Rater D</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Value assigned by Rater E</td>
<td>7.33</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Column total</td>
<td>11</td>
</tr>
</tbody>
</table>

An example of the cross tabulation used to calculate the weighted kappa value of item 6.1 h. The upper number in each cell of the body of the table is the actual number of subjects who were given a particular score by raters D and E. The middle number in each cell is the number of subjects expected to fall into that cell by chance, given the observed column and row totals. The lower numbers (bold) are the penalty scores for disagreement. The actual and observed numbers in the cells along the main (upper left to lower right) diagonal (penalty score of zero) indicate agreement: both raters had the same response. The extent to which the actual frequencies on the diagonal are greater or lesser than expected determines how large the kappa value will be. The column total and row total refer to the sum of the actual frequencies of agreement between raters D and E.

Section Analysis

The total scores for each section and the total score of all six sections of the MFP were calculated for each subject as judged by each of the raters. Table 2 shows the mean, standard deviation, minimum, maximum and skewness of the scores assigned by each of the raters. The standard deviation of the three rater's sectional and total scores are of almost equal magnitude. Thus, the homogeneity of variance assumption for the ANOVA was allowable. An examination of the skewness shows values that are either very small or are in the same direction for all three rater's. Sections 3 and 5 were the only exceptions, however; both of these sections had only a few test items. Thus, any lack of normality should not have a noticeable effect on the ANOVA [4].

Table 3 shows the significance levels of the F values used to test for differences among raters and among subjects. These results generally show