Copyright © 2010 , Fawcett, J.

Critical Review: Examining the Reliability and Agreement of Listeners' Auditory-Perceptual Ratings of Dysarthric Speech Samples

Jenelle Fawcett
M.Cl.Sc. (SLP) Candidate
University of Western Ontario: School of Communication Sciences and Disorders

This critical review examines the literature measuring the reliability and agreement of listeners' auditory-perceptual ratings of dysarthric speech samples. Seven studies are reviewed with the following research designs: single group design, nonrandomized between groups design, two studies with a nonrandomized mixed design, and three studies with a within groups design. Overall, the results of these studies provide mixed results regarding the reliability and agreement of auditory-perceptual ratings. Clinical implications and future recommendations are discussed.

Introduction

Dysarthria is the term used to describe a group of speech disorders resulting from disturbances in muscular control over the speech mechanism due to a lesion in the central or peripheral nervous system (Darley, Aronson, & Brown, 1969). Dysarthia is typically classified as one of seven different types, including flaccid, spastic, ataxic, hyperkinetic chorea, hyperkinetic dystonia, hypokinetic, and mixed (flaccid-spastic), with each classification often related to a different underlying neurological condition. The different classifications of dysarthria depend on both the underlying neurological condition as well as the presence of deviant speech dimensions. These deviant speech dimensions were identified by Darley, Aronson, and Brown (1969a, 1969b). Overall, they identified 38 deviant perceptual dimensions that characterize the dysarthrias. They also identified which dimensions are typically present in each of the types of dysarthria, which allows for classification based on auditory-perceptual analysis of dysarthric speech.

Many speech-language pathologists (SLPs) who assess and treat patients with dysarthria routinely use auditory-perceptual analysis of the patient's speech and/or voice to rate the severity of the deviant dimensions present, as well as to identify the type of dysarthria. The importance of auditory- perceptual analysis in the diagnosis of dysarthria is highlighted when one considers the alternative approaches, such as physiologic and acoustic methods. These approaches can be expensive, may require specialized training and equipment to perform, and may have limited application (Zeplin & Kent, 1996). In addition, auditory-perceptual analysis does not require the patient to have a neurological diagnosis, which can often take a lengthy amount of time to receive.

Given the clinical reliance on auditory-perceptual analysis for the evaluation of dysarthria, it is critical to determine whether it is a reliable tool. A definition of 'reliability' and 'agreement' needs to be noted, as these terms are often incorrectly used interchangeably. Reliability is the "degree to which the ratings of different judges are proportional when expressed as deviations from their means" (Tinsley & Weiss, 1975) and is utilized when one is interested in the relative ordering of the ratings. Agreement is employed when the absolute value of the ratings matters, as it is the "extent to which the different judges tend to make exactly the same judgments about the rated subject."

Objectives

The primary objective of this paper is to critically evaluate the existing research literature regarding the reliability and/or agreement of auditory-perceptual procedures when they are used to identify types of dysarthria or features of dysarthria.

Methods

Search Strategy

Computerized databases including CINAHL, PsycINFO, and SCOPUS were searched using the following search criteria:

> (dysarthria) OR (motor speech disorder) AND (perceptual rating) OR (perceptual scaling) OR (auditory-perceptual) AND (reliability) OR (agreement)

Databases were searched for relevant articles and background information. In addition, the reference lists of the articles were also searched for other relevant articles. The search was limited to articles written in English. There was no limitation set on the date of the articles.

Selection Criteria

Studies selected for inclusion in this critical review paper were required to examine the reliability and/or

Copyright © 2010 , Fawcett, J.

agreement of listeners' auditory-perceptual ratings of the speech of patients diagnosed with dysarthria. The studies were required to use auditory-perceptual analysis alone to identify the type of dysarthria or to rate the deviant speech dimensions of dysarthria for inclusion in this review. In addition, the listeners were required to be SLPs or SLP students. No other limits were set on the demographics or linguistic profile of the research participants (speakers and listeners) or outcome measures.

Data Collection

Results of the literature search yielded the following types of articles congruent with the aforementioned selection criteria: nonrandomized between groups design (1), nonrandomized mixed design (2), single group design (1), and within groups design (3).

Results

In a pivotal study in the area of dysarthria, Darley, Aronson, & Brown (1969) conducted research to determine the speech patterns that are characteristic of seven neurological groups. In addition, they examined the reliability of expert SLPs' auditory-perceptual ratings of the dimensions of dysarthric speech. In their within groups research study, three expert judges (the authors) rated various dysarthric speech samples on a series of dimensions, one dimension at a time. They used a 7-pt severity rating scale, where 1 represented normal and 7 represented a very severe deviation from normal. The speech samples consisted of a standardized passage reading and, on some occasions, conversational speech. In very rare cases, they used sentences repeated by the patient after the examiner. The judges were aware of the neurologic type of each speech sample and rated only the dimensions considered relevant to that neurologic type. To determine intraobserver reliability, at least 30 patients were rated twice by the judges on all 38 dimensions.

In terms of intraobserver reliability, the overall average was 85%. For interobserver reliability, comparisons were made between the ratings of the three judges on 150 patients on 37 dimensions (total of 5550 sets of three ratings). The judges agreed on 84% of the samples as to whether they were normal or not. The judges agreed perfectly or within one scale value on 84% of the sets. This level of reliability was considered to be generally satisfactory.

This study was successful in demonstrating that expert SLPs were able to reliably use auditory-perceptual analysis to rate dimensions of dysarthric speech. A larger sample size of raters would have increased the generalizability of this study. It may seem a limitation that the raters were not blind to the neurologic conditions of the patients; however, given that the 38 perceptual dimensions had not yet been identified or attributed to a specific dysarthria type, the raters were blind in a different sense. Another point to note is that it seems the authors calculated agreement rather than reliability, although they use the term reliability.

In an attempt to replicate the findings of Darley, Aronson, & Brown, Zyski & Weisiger (1987) conducted a nonrandomized between-groups study to determine whether different groups of SLPs could use auditory-perceptual analysis to identify types of dysarthria. The speech samples were taken from the work of Darley et al. and contained a reading passage and syllable repetition. The listeners were split into three groups. Group 1 consisted of 17 SLPs with a minimum of five years experience with dysarthria. For this group, the number of dimensions rated was reduced from 38 to 16 by using only the dimensions with a mean scale value of 2.0 (as determined by Darley, Aronson, & Brown) and using only the dimensions that occurred in no more than four dysarthria types. This was to ensure greater differentiating power. Each listener reviewed the descriptions of each dimension and then listened to the samples. They were asked to record a check mark where they perceived the dimension as present. No limits were made on how many dimensions could be checked off in each sample. The responses were analyzed and the greatest number of dimensions checked off for each sample determined the dysarthria type, which was scored as accurate or inaccurate. The results established that this group identified only 19% of the samples correctly.

Listener Group 2 consisted of 11 SLPs with a minimum of five years of experience with dysarthria. They listened to the same speech samples in the same way as Listener Group 1, but the method of rating differed for this group. They were asked to list a maximum of three deviant dimensions present in each sample after listening. They were also asked to list the dysarthria type (or neurologic disease) of each speech sample. This group was able to identify 55% of the samples correctly.

Listener Group 3 was composed of 15 SLP graduate students who were in a motor speech disorders course. They had five hours of classroom training in auditory-perceptual analysis of dysarthria using the Audio Seminars tape (excluding the speech samples used in this study). They were given the same task as Listener Group 2 five days following training. The

Copyright © 2010 , Fawcett, J.

results for this group showed that they were able to identify 56% of the samples correctly.

Although this study did not include measures of reliability or agreement, which is a limitation, it was included in this review because it investigated SLPs' auditory-perceptual ratings of dysarthric speech and whether they could accurately identify dysarthria types. A weakness in this study was in the methods for Group 1. The forced choice method they used and classification of dysarthria type based on the quantity of dimensions checked off alone limited the listeners. Therefore, the 19% accuracy level achieved by this group may not have been reflective of their actual ability.

In another attempt to replicate the findings of Darley, Aronson, & Brown, a study was conducted by Zeplin & Kent (1996) using a within groups study design. This study involved five judges, two graduate students who had taken a course on dysarthria and three doctoral students who had a minimum of one year experience with dysarthria. The speech samples used in this study were again taken from the work of Darley et al. and consisted of a reading passage and syllable repetition. The judges reviewed the 38 perceptual dimensions, listened to the samples, and rated them using the same 7-pt scale as the one used in the Darley, Aronson, & Brown (1969a) study; however, pitch and loudness were scaled differently due to the fact that speech could vary between bipolar extremes. For these dimensions, the value 4 represented normal and the values 1 and 7 represented deviations.

Intrajudge reliability was estimated by obtaining two sets of ratings for one speaker from each dysarthric category (drawn randomly) and calculating discrepancy scores, which are the differences in scale values. The authors concluded that the judges repeated their ratings reasonably well. Interjudge reliability was estimated from standard deviations calculated for the five judge's ratings of each dimension for each dysarthric speaker. About half of the standard deviations across dysarthria types were less than 1.0 (1.0 was used as a cutoff for the most reliable ratings) and the majority were less than 2.0 (2.0 was used as a cutoff for the least reliable ratings). The authors interpreted this as satisfactory interjudge reliability; however they noted that the interjudge reliability was not as high as one might want for all of the dimensions.

This study did a good job of replicating the Darley, Aronson, & Brown (1969) study. Its' strength over the Zyski & Weisiger (1987) article were that the judges rated all 38 perceptual features. A limitation of this study was that the judges were chosen from a pool of volunteers who responded to an ad, which could create bias. In addition they did not explain how they were chosen from the pool and did not indicate whether they were SLP students. A greater number of judges would have contributed to the ability to generalize the results from this study. Also, the reliability was reported subjectively; results may have had more power had they been reported objectively by giving specific data.

Bunton, Duffy, Rosenbek, & Kent (2007) carried out another study using SLPs as judges. The purpose of their nonrandomized mixed study was to determine the intrarater and interrater agreement for auditory-perceptual ratings of patients with dysarthria. The speech samples they used were taken from 47 patients with different types of dysarthria, Hyperkinetic dysarthria was not included because there were not enough speakers in the database with that type. The samples contained conversational speech from an initial interview with an SLP; however they did not include any speech from the SLP or any leading information regarding diagnosis. The listeners in this study were split into two groups. The first group included 10 inexperienced SLPs who had just completed their master's degree, had taken a course in dysarthria, and had five hours of training using the Audio Seminars tape. The second group of listeners included 10 experienced SLPs with more than seven years of clinical experience and who regularly diagnosed and treated dysarthria. The listeners rated the samples (one dimension at a time) using the 7-pt scale, with pitch and loudness again scaled differently. Speaker order was randomized within each feature. Also, the order of the features was randomized across listeners. Each listener provided 1786 ratings and to control for listener fatigue, listening sessions were limited to one hour. Each listener therefore participated in six to eight one-hour listening sessions over the span of two weeks.

Intrarater agreement was determined by the listeners rating eight quasirandomly selected features for each speaker twice. No differences were found between the two groups so the data was collapsed across the groups. The mean differences between the two ratings were calculated. One feature had a mean difference of 0 (same both times they rated it), 29 features had a mean difference between 0 and 1, and eight features had a mean difference of 1 or greater than 1. In general, this suggests that listeners were reliable in their ratings. To determine interrater agreement, the frequency with which two listeners

Copyright © 2010 , Fawcett, J.

agreed with one another for speaker and feature was calculated. This provided 80,370 pairwise comparisons. The probability that two listeners would agree was calculated (0.14 for exact agreement and 0.39 within one scale value). The probability of observed outcomes was compared with the alpha level and the hypothesis was tested as two-tailed with an alpha level of 0.05. Analysis included only speakers with mean ratings between 2.5 and 5.5 on the scale to eliminate any artifact caused by high listener agreement at extreme ends of the scales. No differences were found between the two groups based on an analysis of variance (ANOVA), so the data was again collapsed across the two groups. The overall percent agreement across all features was 47.8% for exact agreement and 67% within 1 scale, which was judged to be reasonable.

This study used appropriate statistical methods (described in detail) with a large number of comparisons to support their findings, which was an advantage. In addition, they used new speech samples rather than the ones used by Darley et al. (1969), which demonstrated greater generalizability of the results; however, they did not include hyperkinetic speech samples. The authors note that a limitation of their study was that although no differences were found between the two listening groups, the differences may have been hidden by variability related to the rating task. Also, the training provided to the graduate students may have equalized the two groups and therefore, no differences were seen.

An additional two studies sought to determine the reliability and agreement of SLPs' auditory-perceptual ratings using only ataxic dysarthric speech samples. Kearns & Simmons (1988) conducted a single group study using speech samples (standardized reading passage) obtained from ten patients with ataxia. The judges were five experienced SLPs, with experience ranging from six months to nine years. All SLPs participated in three one-hour training sessions that reviewed the definitions of the dimensions used in this study, as well as practiced rating speech samples. The judges then independently rated the samples on a 7-pt severity rating scale. Each sample was played six times to allow the judges to rate each perceptual category separately.

An "agreement" was scored if the judges' severity ratings were within one scale value of one another and overall reliability was calculated on the basis of point to point agreement between judges. Pairwise comparisons were made for each of the 40 perceptual characteristics. Overall agreement between the judges on the 40 dimensions ranged between 60% and 90%, with a mean overall agreement level of 82%. These results suggest that judges can reliably rate speech dimensions of ataxic dysarthria after only minimal training.

The other study (Sheard, Adams, & Davis, 1991) implemented a within groups study design. The speech samples were collected from 15 patients with a diagnosis of cerebellar dysfunction and included a standardized reading passage. The judges were 15 SLPs with over 200 hours of experience with dysarthria. Using a 7-pt rating scale, they rated the speech samples on five deviant dimensions found in cerebellar dysfunction: imprecise consonants, excess and equal stress, irregular articulatory breakdown, distorted vowels, and harsh voice. One to two weeks later, the judges rated the samples again to determine intrarater agreement.

The results indicated that the average intrarater agreement across the speech dimensions was 86.4% (within one scale point). Interrater agreement was determined by calculating pairwise differences. A total of 3150 pairwise comparisons were made. The average interrater agreement was 70% (within one scale point). Interrater reliability was also calculated using an intraclass correlation, which implies a two-way ANOVA. The intraclass correlation coefficients were above 0.6 for imprecise consonants, excess and equal stress, and harsh voice, but below 0.6 for distorted vowels and below 0.5 for articulatory breakdown. These results suggest that experienced SLPs were overall moderately reliable in rating the deviant dimensions.

An obvious weakness to the preceding two studies was that they both used only ataxic dysarthric speech samples, which limits the generalizability of the results to other types of dysarthria. Another contributing factor that limited generalizability was the small samples sizes of judges included. Additionally, being aware of the dysarthria type at the time of rating may have influenced the judges' ratings by introducing bias. The study by Sheard, Adams, & Davis (1991) only rated five deviant features, which may have inflated the reliability. A strength of both studies were that they used appropriate statistical methods.

The final and most recent study included in this review was a nonrandomized mixed study conducted by Van der Graaf et al. (2009). The study included speech samples from eighteen patients with dysarthria and four healthy controls. The samples

Copyright © 2010 , Fawcett, J.

consisted of a standard reading passage and free speech. There were three groups of raters. Group 1 was eight neurologists, Group 2 consisted of eight neurology residents, and Group 3 included eight speech therapists. Each group rated the samples three times to determine whether clinical information would improve the score of each rater: the first time, they rated the samples and checked off the type of dysarthria, the second time, they were given some clinical information on each patient and rated the samples again, and the third time they rated the samples a week later.

Since this critical review is only interested in SLPs as raters, only their results will be reported. Group 3 (the SLPs) correctly identified 37% of the speech samples correctly in the first session, 31% in the second session, and 48% in the third session. Interrater agreement was κ=0.22 and intrarater agreement was κ=0.47 (calculated with the χ^2 test). These results demonstrated that correct identification based on auditory-perceptual analysis was poor and that interrater and intrarater agreement were fair to moderate, at best.

A key advantage to this study was that they included a control of healthy patients in their speech samples, which were not included by any of the other studies. This study also provided a detailed account of the statistical methods they used, which were suitable. The authors note that a limitation to their study was that while the speech samples were of good quality, the recordings did not allow for some sophisticated analyses, including breathing patterns. Also, the authors stated there was some uncertainty as to whether each of the dysarthria samples contained the whole set of characteristics used to determine a particular type.

Discussion

The seven studies reviewed above demonstrate mixed results regarding the reliability and agreement of listeners' auditory-perceptual ratings of dysarthric speech samples. Overall, five of the studies reported a reasonable or satisfactory level of reliability; however the other two studies reported poor reliability or poor accuracy in the identification of dysarthria types. The strength of evidence provided by these studies is below the "gold standard" because they all include research designs that are not truly randomized. Furthermore, there were other methodological limitations found within these studies which suggest the results should be interpreted with caution. As discussed above, the small sample sizes of listeners included in the studies may limit the generalizability of the findings. As well, some of the studies appeared to use the terms 'reliability' and 'agreement' interchangeably, which could have an impact on the weighting one puts on the results. Furthermore, using the terms interchangeably may be confusing to the reader.

The number of deviant features rated may also have affected the results, particularly in the studies with only ataxic dysarthric speech samples. As mentioned by Kearns & Simmons (1998), the small number of features rated may have caused an inflated level of reliability. On the other hand, Wertz & Rosenbek (1992) mention that experienced clinicians likely do not rate all 38 features when using auditory-perceptual ratings clinically. Instead, they suspect that they hone in on a few specific dimensions to make a diagnosis.

The nature of the speech tasks used in the studies is also something worth considering as it may have affected the findings. The type of speech task used may have had an impact on the ability of the listeners to identify the distinctive features of the dysarthic speech, as discussed by Zeplin & Kent (1996). For example, the speech tasks of 'syllable repetition' and 'prolonged vowels' make some features of dysarthric speech more easily identifiable. The speech tasks included in the above studies were variable and the type of speech task used in a particular study may have affected the overall accuracy and reliability and/or agreement of the listeners' ratings.

While there are some drawbacks to these studies, there are also some successes and important implications. They demonstrate that expert SLPs are able to use auditory-perceptual analysis reliably. Moreover, it appeared that SLPs who had recent training in auditory-perceptual analysis of dysarthric speech were more accurate and reliable in their ratings (Kearns & Simmons, 1998, and Zyski & Weisiger, 1987). This suggests that it may be worthwhile to provide auditory-perceptual training to clinicians, especially those with less experience.

Recommendations

It is recommended that further research on this topic be completed. In order to improve upon the evidence provided by the existing literature, future research should take the following into account:

a) Adequate sample sizes and random distribution of participants into experimental groups
b) Speech samples including healthy controls

Copyright © 2010 , Fawcett, J.

for comparison

c) A clear distinction between the terms 'reliability' and 'agreement'
d) Further investigation regarding the number of deviant features rated and how reliability and/or agreement is affected
e) The effect different speech tasks have on reliability and/or agreement
f) The effect of training SLPs to use auditory-perceptual analysis with dysarthria

Additionally, it is recommended that clinical researchers maintain close contact with clinicians when examining the reliability of auditory-perceptual ratings to ensure that valid methods are employed in research and can be effectively and reliably translated into clinical practice.

Clinical Implications

The studies reviewed indicate mixed support regarding the reliability and/or agreement of auditory-perceptual procedures when they are used to identify types of dysarthria or features of dysarthria. Overall, a suggestive level of evidence has been provided in the majority of cases, which tentatively promotes the use of auditory-perceptual analysis of dysarthric speech. It is vital that clinicians be cognizant of the shortcomings of the current research when making clinical decisions regarding auditory-perceptual analysis.

References

Bunton, K., Kent, R.D., Duffy, J.R., Rosenbek, J.C., & Kent, J. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language & Hearing Research, 50*(6), 1481-1495.

Darley, F., Aronson, A., & Brown, J. (1969a). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12*, 246-269.

Kearns, K. & Simmons, N. (1998). Interobserver reliability and perceptual ratings: More than meets the ear. *Journal of Speech and Hearing Research, 31*, 131-136.

Sheard, C., Adams, R.D., & Davis, P.J. (1991). Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. *Journal of Speech and Hearing Research, 34*, 285-293.

Tinsley, H.E.A., & Weiss, D.J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of CounselingPsychology, 22*, 358-376.

Van der Graaff, M., Kuiper, T., Zwinderman, A., Van de Warrenburg, B., Poels, P., Van der Kooi, A., Speelman, H., & De Visser, M. (2009). Clinical identification of dysarthria types among neurologists, residents in neurology, and speech therapists. *European Neurology, 61*, 295-300.

Wertz, R.T., & Rosenbek, J.C. (1992). Where the ear fits: A perceptual evaluation of motor speech disorders. *Seminars in Speech and Language, 13*, 39–54.

Zeplin, J., & Kent, R.D. (1996). *Reliability of auditory perceptual scaling of dysarthria*. In D. Robin, K. Yorkson, & D.R. Buekelman (Eds). Disorders of motor speech: Recent advances in assessment, treatment, and clinical characterization. Baltimore: Paul H. Brookes.

Zyski, B.J. & Weisiger, B.E. (1987). Identification of dysarthria types based on perceptual analysis. *Journal of Communication Disorders, 20*, 367-378.