Journal of Speech and Hearing Research, Volume 38, 280–288, April 1995

Perceptual Rating Instrument for Speech Evaluation of Stuttering Treatment

Marie-Christine Franken
Department of Voice and Speech Pathology
University of Nijmegen
Nijmegen, The Netherlands

Louis Boves
Department of Language and Speech
University of Nijmegen
Nijmegen, The Netherlands

Herman F. M. Peters
Department of Voice and Speech Pathology
University of Nijmegen
Nijmegen, The Netherlands

Ronald L. Webster
Hollins Communications Research Institute
Hollins College
Roanoke, VA

A rating instrument is described that can be used to assess the results of stuttering treatments. The instrument is designed for use with naive listeners. It yields a comprehensive and detailed description of the speech quality in terms of articulation, phonation, pitch, and loudness; in addition, it includes a naturalness scale. Analysis of ratings obtained with the instrument show that naturalness is a multidimensional characteristic. Moreover, the speech characteristics that determine the naturalness ratings appear to be different pretreatment, posttreatment, and at follow-up treatment.

The psychometric characteristics of the instrument are analyzed in detail. It is concluded that mixing of samples of stutterers and nonstutterers in one rating experiment may artificially inflate the reliability of the ratings. Also, ratings on equal-appearing interval scales cannot be interpreted in an absolute sense. Solutions for this methodological problem are suggested.

KEY WORDS: speech evaluation, stuttering treatment, perceptual rating of speech, rating instruments

---

The importance of evaluating the outcome of speech treatment for scientific, economic, and social reasons is widely recognized. Given the large body of literature on the evaluation of stuttering treatments, stuttering seems to be one of the disorders where the quality of existing and newly proposed treatments can easily and reliably be assessed; for existing treatments by looking up their evaluation results in the literature, and for new treatments by subjecting them to some well-established test. Moreover, the literature suggests that existing stuttering treatments live up to their expectations: "substantial improvement, as defined in these studies, typically occurs as a result of almost any kind of treatment in about 60 to 80 percent of cases" (Bloodstein, 1987, p. 399).

However, on closer inspection it appears that the quotation from Bloodstein probably overestimates the real success of stuttering treatment. First, few evaluation studies seem to have been carried out by independent investigators; clinicians/researchers evaluating their own treatment might be tempted to stress positive outcomes and ignore less favorable aspects of their favorite treatment. And even if the evaluation was objective, one must be aware that the measures for success of a stuttering treatment have conventionally been limited to the proportion of disfluencies and/or speech rate just after treatment completion. The validity of these measures can be questioned, because there is ample evidence that even naive listeners are able to discriminate posttreatment speech from the speech of nonstutterers (Ingham & Packman, 1978; Runyan & Adams, 1979; Franken, 1987; Runyan, Bell, & Prosek, 1990; Franken, Boves, Peters, & Webster, 1991, 1992).

Apparently, there is more to normal sounding speech than absence of disfluencies and a normal speech rate. This is especially true for treatments that teach special techniques to induce fluent speech (like prolongation of speech sounds and/or

0022-4685/95/3802-0280

continuous phonation). Thus, if one wants to assess the extent to which a treatment succeeds in restoring the ability of stutterers to speak normally, a more comprehensive evaluation of posttreatment speech is necessary.

A number of recent papers have proposed a naturalness scale to complement speech rate and stuttering frequency measures (Martin, Haroldson, & Triden, 1984; Ingham, Gow, & Costello, 1985; Ingham, Martin, Haroldson, Onslow, & Leney, 1985; Ingham & Onslow, 1985; Ingham, Costello-Ingham, Onslow, & Finn, 1989; Onslow, Hayes, Hutchins, & Newman, 1992). However, in the research mentioned above the naturalness scale was primarily used as an instrument in a treatment: clients are urged to speak more naturally if the clinician feels that the client's speech sounds unnatural. A study by Ingham, Gow, and Costello (1985) showed that two out of three stutterers who had passed the establishment phase of a prolonged speech treatment could modify their speech towards more natural sounding speech when instructed to do so. However, Ingham et al. (1989), reporting on three different stutterers who were part way through a treatment program using prolonged speech or rate control, found that instructing the stutterers to speak more naturally did not result in improved naturalness ratings by the experimenters or independent judges. The stutterers themselves, however, thought that improvements in naturalness were achieved. Another study (Ingham et al., 1985) with six stutterers who had not recently received treatment showed that less severe stutterers could improve the naturalness of their speech as an effect of regular feedback of listener-judged speech naturalness. But feedback about the naturalness of their speech did not help the more severe stutterers to improve their speech quality. Perhaps the limited efficacy of instructing stutterers to speak more naturally should be expected. It has been shown that naturalness is a multifaceted variable that is related to a number of other perceptual characteristics of a speech sample (Franken, 1987; Franken et al., 1992). Thus, speech may fail to sound natural for a number of reasons. Because of the multidimensional nature of speech naturalness, the probability that a clinician can help a stutterer to improve the overall speech quality becomes higher if he or she can diagnose the dimensions that are most deviant. And it is quite likely that a clinician who knows why the speech still does not sound natural is more effective in giving detailed instructions on how to improve the speech.

The naturalness scale has also been used in posttreatment assessment (Metz, Schiavetti, & Sacco, 1990; Onslow et al., 1992). Metz et al. were especially concerned with the psychometric properties of the naturalness scale, as well as with the extent to which naturalness ratings can be predicted by relatively simple acoustic measurements like voice onset time, vowel duration, and speech rate. They concluded that naturalness ratings can be obtained by means of equal-appearing interval scales. Voice onset time and speech rate (expressed in terms of sentence duration) appeared to be the most powerful—but still relatively weak—acoustic predictors of rated naturalness.

In summary, we can say that a naturalness scale is an important addition to the conventional measures of percentage of stuttered syllables and speech rate, as an instrument in treatment, and as a measure for the evaluation of posttreatment speech. Naturalness can be rated with the use of popular 7- or 9-point equal-appearing interval scales (cf. Martin et al., 1984). It has been shown that naturalness ratings during treatment and in treatment outcome evaluation can be made reliably, both by trained and untrained raters. However, most of the research has addressed naturalness ratings as used during treatment; less attention has been given to naturalness ratings in treatment outcome evaluation. Moreover, Kreiman, Gerrat, Kempster, Erman, and Berke (1993) emphasize that reliability measures appropriate for scales to be used by a single clinician are different from the measures when the reliability of averages obtained from many raters is concerned. Finally, the real meaning of speech naturalness remains somewhat unclear; naturalness may depend on several, possibly independent, perceptual and physical characteristics of a speech sample.

In this study we investigated the reliability of average ratings obtained from groups of judges as well as details of the meaning of the speech naturalness scale applied to the speech of stutterers before and after treatment, and to the speech of nonstutterers. To that end, we have developed an instrument that yields a comprehensive and detailed formal and technical description of the quality of speech samples. In addition, the instrument includes a naturalness scale. The instrument has been developed with treatment evaluation in mind.

Our speech quality instrument consists of 14 bipolar (equal-appearing interval) rating scales, defined by contrastive terms that label extremes analogous to the Semantic Differential introduced by Osgood, Suci, and Tannenbaum (1957). With this instrument we investigated in what ways "naturalness" relates to other characteristics of speech signals, like loudness, pitch, voice dynamics, and so forth. We also investigated whether the meaning of rated speech naturalness (in terms of its relation to other speech characteristics) depends on the point in time when the speech was evaluated (pre-, post-, or follow-up treatment). This should help in grasping the precise nature of speech naturalness.

Although equal-appearing interval scales are seemingly easy to use, they have a number of possibly less desirable properties, one of which is that they have no fixed anchors. Even if the scale positions on, for example, a 7-point scale ("fast–slow") are explained to the raters in terms of "extremely fast," "rather fast," "somewhat fast," "not fast nor slow," "somewhat slow," "rather slow," "extremely slow," it remains unclear what part of the range of this scale is considered "acceptable," "normal," or "pleasant." Speech that is too slow is probably not desirable, but extremely fast speech may be judged as undesirable too. Also, it has been shown (Boves, 1984; Kreiman et al., 1993) that even on scales that have undisputable positive and negative labels, like "not at all rough–extremely rough," the exact scale value of a specific speech sample is determined by its quality relative to the least rough and the most rough samples. Subjects tend to use the whole scale, so that the anchor positions at the scale's extremes are essentially defined by the characteristics of the extreme samples in the set under judgment. Thus, in the context of the naturalness scale, a sample that gets a scale value in the range labeled

"somewhat natural" might still be considered as actually below the threshold of what is normal or acceptable. We will investigate this problem by means of a comparison of our findings with previously reported data.

In our experiment "naive" judges were used because we feel that their judgments should be the reference. What counts most for a stutterer when he or she assesses the result of a treatment are the reactions and opinions of the persons with whom he or she must communicate in normal daily life—not the judgments of experts.

Method

The Speakers

Speech samples were provided by 32 male stutterers and 20 male nonstutterers. Mean age of the stutterers was 25.3; the youngest subject was 15.1, the oldest 46.3. Subjects were recorded immediately before they started a Dutch adaptation of Webster's Precision Fluency Shaping Program (PFSP) (Webster, 1974, 1979, 1980; Peters & Kooijman, 1985). The stutterers were also recorded immediately after completion of the treatment (posttreatment) and 6 months after that date (follow-up). The PFSP is a tightly structured, systematic speech motor training program that aims to reconstruct behavior details involving respiration, voicing, and articulation (Webster, 1980). The program starts with a form of slow-motion speech behavior, to establish a framework into which specific, corrected speech movements can be transferred in order for the stutterer to feel complete control of the speech production system. The Dutch version of this program follows the guidelines given by Webster as precisely as possible. The complete treatment program takes about 120 clinical hours.

In the Dutch health care system only confirmed severe stutterers are allowed to enroll in this residential, and therefore relatively expensive, treatment. Moreover, in order to be accepted the clients must show positive proof that their speech problem is primarily motoric in nature; clients who have considerable (concomitant) psychological problems are not admitted. The primacy of psychological problems is confirmed with the help of a number of proven personality scales. The presence of severe motoric problems is established by experienced speech-language pathologists (Franken et al., 1992). In effect, the subject group consisted of all clients who completed the PFSP program in the years 1987 to 1989 and who were willing to participate in an evaluation study that required them to return to the clinic each semester during a period of at least 2 years. It should be clear that, because of these selection criteria, the treatment results quoted in this paper cannot be construed as an evaluation of the PFSP as such.

The 20 nonstuttering males had no known speech problems; they were matched to the stutterers on age, level of education, and extent of deviation of the standard pronunciation. The last aspect is important to prevent a confounding of differences between the groups due to speech characteristics related to the disorder *and* speech characteristics that are known to affect perceptual ratings but that are independent of stuttering problems. Research in sociolinguistics and social psychology has shown that speech approximating standard pronunciation is judged more favorably than speech characterized by regional accents, everything else being equal (Giles & Powesland, 1975). Onslow et al. (1992) have also pointed out that a dialectal coloring of speech samples under judgment may act as a possibly confounding factor in judging the speech of (treated) stutterers and nonstutterers. In our experiment the degree of deviation from standard pronunciation of the stutterers and nonstutterers was scored by the first author, using a 4-point scale; she was also responsible for the matching of stutterers and nonstutterers. The nonstutterers were recorded twice; both recordings were used as samples in the experiment.

The Speech Samples

The speech stimuli were obtained in a task in which the speakers had to summarize and comment upon a recent newspaper article. In most cases the experimenter had to ask questions about the event described in the text in order to elicit speech beyond a short summary. These conversations were kept going for about 5 minutes. The stutterers were recorded in the clinic pre-, post-, and 6 months after treatment, each time on a different topic. The nonstutterers were recorded on two topics. Both recordings were made on the same day. The signal/noise ratio of all the recordings was acceptable.

In total, $3 \times 32 + 2 \times 20$ (136) recordings were used. Fragments of about 45 seconds following the first 30 seconds of a recording, starting with a new utterance, were selected for the experimental sample. Table 1 shows the median and range of the percentage of Stuttered Syllables (% SS) and the mean and standard deviation of the speech rate in Syllables/sec for the three conditions under which the stutterers were recorded as well as for the nonstutterers. For the nonstutterers, only a single value is given, that is, the average over the two recordings.

The fluency of the samples was scored on the basis of an elaborate protocol. According to this protocol sound-, syllable-, and word repetitions, as well as silent and filled blocks, prolongations, and interjections (speech and nonspeech sounds) must be scored as disfluencies. Judges are encouraged to listen repeatedly to the recorded samples before scoring the disfluencies. In the eventual measure, % SS, all categories of disfluencies are collapsed. To compute intra- and interjudge agreement for the % SS, 20% of the speech samples were selected randomly. These samples were rerated by the first author and also by a trained clinical research assistant. The total percentage intraobserver and interobserver agreement (Kearns, 1990) turned out to be 98% and 94%, respectively.

To compute speech rate, the duration of each sample (including pauses) was measured twice using an electronic stopwatch. The two duration measures were then averaged to obtain the "true" speaking time. Only linguistically relevant syllables were counted to establish speech rate in Syllables/sec; thus, a monosyllabic word that was repeated five times was counted as a single syllable.

TABLE 1. Group medians and ranges of percentage stuttered syllables (% SS) and average speech rate and standard deviations in Syllables per second (Syll/sec) for the speech samples of stutterers in three conditions (pretreatment, posttreatment, and at 6 months follow-up) and for the speech samples of nonstutterers. Values for nonstutterers are means over two recordings on the same day.

	Stutterers (*N* = 32)			Nonstutterers (*N* = 20)
	Pre	Post	Follow-up	
% SS				
median	20.5	4.5	11.8	2.6
range	7.1–75.0	0–18.5	2.3–47.1	0–7.8
Syll/sec				
average	2.1	2.1	2.3	3.9
SD	1.0	0.7	1.0	0.5

TABLE 2. Perceptual rating scales for untrained listeners, with their Ru-coefficients, based on the judgments of two groups of respectively 24 (Ru-1) and 20 listeners (Ru-2).

Scale	Ru-1	Ru-2
Low Pitch–High Pitch	.95	.94
Slow–Quick	.95	.96
Slovenly–Polished	.90	.90
Flat–Expressive	.95	.94
Shrill–Deep	.93	.91
Soft–Loud	.84	.85
Monotonous–Melodious	.96	.95
Tense–Relaxed	.96	.95
Weak Acc.–Strong Accentuation	.90	.90
Unpleasant–Pleasant	.95	.95
Slurred–Precise	.94	.91
Fluent–Halting	.96	.96
Weak–Powerful	.88	.88
Unnatural–Natural	.97	.97

Naturalness in a Speech Quality Measurement Instrument

An instrument for the evaluation of a speech treatment can be considered to have ecological validity only if it takes into account the judgments of "naive" listeners, that is, the persons who are most likely to be addressed by the treated stutterer in daily life. Naive listeners cannot be expected to be able to use technical terms in their evaluation of speech quality. At best they can be requested to express global ratings. But, clever combinations of global, associative ratings have been shown to yield highly useful and informative data (Osgood et al., 1957). To investigate the global judgments of untrained listeners judging the speech of treated stutterers we modified the Semantic Differential type instrument proposed by Fagel, van Herpt, and Boves (1983). In its original form that instrument consists of 14 bipolar point scales, chosen to yield comprehensive descriptions of the quality of the speech of "normal" speakers. For reasons of efficiency we did not want to construct an instrument that includes more than 14 scales. Thus, adding scales addressing characteristics of stutterer's speech would require that scales considered less important for characterizing the results of a stuttering treatment be sacrificed. In doing so, we deleted the scales "Ugly–Beautiful," "Husky–Not Husky," "Dull–Clear," and "Broad–Cultured." We considered these scales, which address general evaluation, voice quality, and pronunciation quality, respectively, as less important for the evaluation of speech before and after a stuttering treatment. They were replaced by the scales "Tense–Relaxed," "Weak Accentuation–Strong Accentuation," "Slurred–Precise," and "Halting–Fluent." Finally, the scale "Dragging–Brisk" was removed from the original instrument, because we suspected that it might acquire an undesirable attitudinal loading in the context of stutterer's speech. This scale was replaced by the scale "Unnatural–Natural." Table 2 gives an overview of the scales; the Ru-coefficients in the table are explained in the Result section below.

On the forms given to the raters, the meaning of the seven intervals on the scales is explained by explicitly pointing out that each rating scale is supposed to cover the range from one extreme to the other. This is done by example, using the familiar scale "slow–quick," giving the following definitions: (1) very slow, (2) rather slow, (3) somewhat slow, (4) not slow nor quick, (5) slightly quick, (6) rather quick, and (7) very quick.

Rating Procedure and Listeners

The 136, 45" stimuli were copied onto two tapes with the same number of samples from stutterers and nonstutterers on each tape. Each tape was presented to a different group of listeners; 20 speech samples were present on both tapes (viz. the pre-, post-, and follow-up samples of 4 stutterers, plus both samples of 4 nonstutterers). Thus the total number of stimuli on each tape was 78. The order of the stimuli was randomized under the restriction that two samples of the same speaker had to be separated by at least three samples of other speakers. The ratings were organized in the form of classroom sessions. The sample recordings were played back on a Revox A77 recorder through a pair of high quality loudspeakers. Listeners were not informed about the origin of the speech samples. Rating sessions lasted about 2 hours, including two pauses of about 15 minutes.

To familiarize the listeners with the scales and the rating procedure and to give them an impression of the stimuli to be judged, 10 training stimuli were presented at the start of the experiment. To help ensure that listeners were judging seriously at sample 11, they were told that the first three samples (instead of the first 10) were training stimuli.

The first tape was judged by 24 listeners, the second by 20. All listeners were students of logopedics who were in their first semester; thus, the judges can be considered as essentially naive with respect to the formal and technical aspects and terminology of speech science.

Results

Preparatory Data Processing

Based on previous research (Fagel et al., 1983; Boves, 1984), listener ratings were treated as interval data. Table 2

TABLE 3. Principal components analysis of transformed and combined scores on 14 rating scales by two groups of 24 and 20 untrained listeners. Factor loadings on the rotated Factor Matrix.

Rating scales	F#1	F#2	F#3
Low Pitch–High Pitch	.23	−.05	.94
Slow–Quick	.70	.30	.43
Slovenly–Polished	.45	.75	−.09
Flat–Expressive	.84	.32	.32
Shrill–Deep	−.19	.14	−.95
Soft–Loud	.80	.02	−.03
Monotonous–Melodious	.84	.39	.26
Tense–Relaxed	−.03	.94	−.14
Weak Accentuation–Strong Accent.	.86	.31	.27
Unpleasant–Pleasant	.53	.81	−.03
Slurred–Precise	.34	.83	.04
Fluent–Halting	−.12	−.94	−.03
Weak–Powerful	.87	.30	.17
Unnatural–Natural	.52	.80	.06

reports Ru-coefficients (Winer, 1971) for the ratings in the speech quality experiment. The coefficient Ru gives the unadjusted reliability, that is, no adjustment is made for possible bias of individual raters towards high or low ratings. In essence, Ru expresses the ratio of true between-item variance and within-item variance, which is considered as error. The Ru-values range from .84 to .97. These values are high enough to warrant the replacement of the ratings of the individual listeners by average ratings in all subsequent analyses.

In order to arrive at a single set of scores the ratings by the two groups were combined in the following way. First, correlation coefficients were calculated between the average scores of the 20 samples rated by both groups. For all 14 scales we found $r = >.95$ in both experiments. Next, linear regression coefficients were computed for the scores of group 2 on the scores of group 1, separately for all scales. These coefficients were used to transform the scores of group 2 to the same reference as group 1. Finally, the original scores of group 1 and the transformed scores of group 2 were merged. For the items rated by both groups the average of the ratings of the groups after transformation was taken.

Factorial Components of the Speech Scale Ratings

In order to grasp the dimensional structure of the 14 speech rating scales a principal components analysis was carried out using the FACTOR program in SPSS (1990) on the transformed scores. This analysis resulted in a three-factor solution, which is shown in Table 3.

The first factor can be described as a Voice Dynamics factor; it has high loadings from the scales: Weak–Powerful, Weak Accentuation–Strong Accentuation, Flat–Expressive, Monotonous–Melodious, Soft–Loud, and Slow–Quick. The second factor can be described as an Articulation Quality factor. It has high loadings from the scales Fluent–Halting, Tense–Relaxed, Slurred–Precise, and Slovenly–Polished; in addition, this factor has high loadings from the scales Unpleasant–Pleasant and Unnatural–Natural. The third and last factor is a Pitch factor, with high loadings from the scales Shrill–Deep, and Low Pitch–High Pitch. The variance in the judgments of the listeners explained by this factor solution amounts to 86.6%, with contributions of 57.4%, 21.2%, and 7.9%, respectively, from the individual factors.

The results correspond essentially with the finding of Fagel et al. (1983) indicating that semantic differential ratings of the speech of normal talkers on five "speech" dimensions can be distinguished in addition to, or confounded with, the General Evaluation and Potency dimensions (viz. Melodiousness, Articulation Quality, Voice Quality, Pitch, and Tempo). The last factor did not appear in the present factor solution; the only scale that could represent this factor, Slow–Quick, loads highly on the Voice Dynamics factor. The instrument does not contain other scales that measure speech rate. It is well known that it is quite unusual for solitary scales to crop up as a factor. The failure of the present analysis to isolate the Voice Quality factor is because the voice quality scales proposed by Fagel et al. (1983) were sacrificed in favor of scales referring more directly to specific aspects of the speech of stutterers, like Weak Accentuation–Strong Accentuation and Unnatural–Natural.

Our factor solution confirms previous findings that General Evaluation and Potency tend to associate with the Melodiousness and Articulation Quality factors (Fagel et al., 1983; Boves, 1984; Franken, 1987). In our factor solution Potency seems to associate almost exclusively with Melodiousness; that is the reason why we prefer the label "Voice Dynamics" for the first factor. It strikes the eye that most of the scales have one-dimensional loadings: they load highly on one factor and have negligible loadings on the remaining factors. Unnatural–Natural and Unpleasant–Pleasant are the most notable exceptions; they divide their loadings between the first two factors, even if there seems to be a preference for the second, Articulation Quality, factor. This shows that naturalness and pleasantness of speech samples are at least two-dimensional concepts—concepts that relate to other, more technical aspects of the speech quality in complicated ways.

Thus, it appears that the judgment space for speech samples obtained from a mix of normal speakers and

TABLE 4. Principal components analysis of transformed and combined scores on 14 rating scales by two groups of 24 and 20 untrained listeners, for the pretreatment condition only. Factor loadings on the rotated Factor Matrix.

Rating scales	F#1	F#2	F#3
Low Pitch–High Pitch	−.10	.15	.93
Slow–Quick	.36	.56	.33
Slovenly–Polished	.76	.24	−.36
Flat–Expressive	.13	.81	.42
Shrill–Deep	.20	−.11	−.92
Soft–Loud	.07	.66	−.28
Monotonous–Melodious	.36	.76	.41
Tense–Relaxed	.93	−.02	.00
Weak Accentuation–Strong Accent.	.15	.78	.25
Unpleasant–Pleasant	.93	.21	−.13
Slurred–Precise	.82	.13	−.03
Fluent–Halting	−.90	−.13	−.24
Weak–Powerful	.18	.87	−.12
Unnatural–Natural	.91	.21	−.18

(treated and untreated) stutterers has essentially the same factor structure as the space for normal speech. This suggests that naive listeners do not switch to a special "stutter-mode" when judging speech of stutterers, despite the fact that the stimulus ensemble in our experiment contained a number of items that were unmistakably severely stuttered speech. The two bi-dimensional scales, (Un)Natural and (Un)Pleasant, appear to be essentially generally evaluative in nature. In fact, their close relation is confirmed by the extremely high Pearson correlation between these scales ($r = 0.96$).

Factor Solutions for Individual Conditions

It may well be the case that the bi-dimensional nature of the general evaluative scales is mainly due to the fact that their behavior, as that of all other scales, of course, has been "averaged" over four conditions: stutterers pre, post-, and follow-up treatment and normal speakers. Their behavior within individual conditions might be substantially different. In order to investigate this we carried out four separate principal component analyses, one for each of the four conditions. Below, we will present only the results for the pre- and posttreatment conditions in detail. The two remaining conditions confirmed the major findings that are illustrated for the pre and post conditions: If there is some salient and unusual aspect that (virtually) all stimuli have in common, then this aspect will attract most of the meaning of the (un)naturalness scale. If such a dominant feature is absent (as is the case in the speech of the normal subjects) (un)naturalness is likely to lose its discriminating power. When used with a sample consisting of normal speakers only, that might well result in the scores on this scale being very unreliable (Fagel et al., 1983; Boves, 1984).

Table 4 shows the factor solution for the pretreatment condition. Once again, the three factors—Articulation Quality, Voice Dynamics, and Pitch—are extracted, but the proportion of the variance explained is somewhat lower than in the solution for the total material: 79.4% for the three factors together, with individual contributions of 43.0%, 24.5%, and 11.9%, respectively. Contrary to the solution for the complete set of stimuli, however, in the pretreatment condition the Articulation Quality factor is much more important than the Voice Dynamics factor. Also, the behavior of the general evaluation scales (Un)Natural and (Un)Pleasant is quite different; they load on the Articulation Quality factor exclusively. This finding is easy to explain, since the speech samples in the pretreatment condition are mainly characterized by the fact that most are severely stuttered; it is the stuttering that determines the naturalness and pleasantness of the stimuli, almost on its own.

The factor solution for the posttreatment condition, shown in Table 5, is quite different from what we have seen before. The first, and by far the most powerful factor (explaining 48.5% of the total variance, as compared to 19.3% for the second) is a very broad and diffuse amalgam of Dynamics and Pronunciation scales. The fact that the Pitch factor appears as second here is most probably because the posttreatment speech is significantly lower and more voiced than the speech in all remaining conditions. This seems a direct consequence of the "gentle voice onset" target behavior. The third factor, which explains no more than 11.0% of the total variance, has high loadings of the scales Tense–Relaxed and Fluent–Halting. Thus, it seems that this factor attracts what remains of pretreatment stuttering behavior. The proportion of variance explained by the first three factors with an eigenvalue greater than 1.0 together amounts to 78.7%. The overwhelming contribution of the first factor, Voice Dynamics/Expressiveness, is easy to understand. The posttreatment speech is strikingly monotonous in all possible respects; pitch, loudness, and tempo variations are virtually absent. In such a situation it is not surprising to see that the general evaluation scales (Un)Natural and (Un)Pleasant associate with the scales Flat–Expressive, Monotonous–Melodious, and Weak Accentuation–Strong Accentuation.

Discussion

In this discussion we will not reiterate the results of the factor analyses; nor will we try to add to their interpretation.

TABLE 5. Principal components analysis of transformed scores on 14 rating scales by two groups of 24 and 20 untrained listeners, for the posttreatment condition only. Factor loadings on the rotated Factor Matrix.

Rating scales	F#1	F#2	F#3	F#4
Low Pitch–High Pitch	.01	.95	−.16	−.04
Slow–Quick	.57	.56	.05	.03
Slovenly–Polished	.83	−.05	.16	.01
Flat–Expressive	.96	.16	−.06	−.04
Shrill–Deep	.02	−.95	.15	−.04
Soft–Loud	.04	−.09	−.14	.92
Monotonous–Melodious	.95	.11	.01	.10
Tense–Relaxed	.01	−.20	.93	−.00
Weak Accentuation–Strong Accent.	.91	.11	−.09	.26
Unpleasant–Pleasant	.95	−.10	.15	.07
Slurred–Precise	.77	.01	.22	.25
Fluent–Halting	−.21	.08	−.91	.15
Weak–Powerful	.58	.29	.03	.63
Unnatural–Natural	.95	.02	.12	−.00

There is ample room, however, for a discussion of the psychometric characteristics of our rating instrument.

From a psychometric point of view it is interesting to try to understand why the reliability of the ratings was so very high. After all, essentially naive subjects judged a very complex type of stimuli with which they were not familiar. Under such circumstances one would expect at best mediocre reliability. We suspect that the very high Ru values are to a large extent due to the very large differences between the speech samples of the stutterers and the nonstutterers. This "inflated" between-stimulus variance must lead to somewhat inflated estimates of the reliability of the ratings, if reliability is expressed as the ratio of the between-item variance and the error variance. This interpretation is corroborated by inspection of the scatter plot of the correlation between the scores on the Naturalness and Pleasantness scales (not shown). Despite the overall correlation of .96 the plot clearly shows four subsets of the data, which spread around the diagonal. The cluster made up by the scores of the normal speakers particularly shows a high degree of scattering, but at the same time remains clearly separated from the other clusters. This shows that the raters agree that the overall position of the normal speakers should be near the positive end of the scales, but that they do not agree on the precise rank of the normal speakers on the very limited part of the naturalness scale that remains if their distance to the stutterers is duly expressed. This effectively confirms the conjecture made earlier that, in the absence of a salient cause of unnaturalness, the Naturalness scale loses most of its meaning (or, at least, its discriminating power).

In Franken et al. (1992) average scale values are given for the stutterers pre-, posttreatment, and at follow-up as well as for the nonstutterers on the naturalness scale; these values are 2.74, 2.83, 3.35, and 5.06, respectively, on an unanchored 7-point scale. From these data it cannot be concluded that, on average, stutterers following treatment speak sufficiently naturally because the average value of 3.35 is below the midpoint of the scale (which is at 3.5). Apart from the fact that semantic differential scales are essentially unanchored, one must also take into account that the ratings of the stutterers are based on just one sample of spontaneous speech per speaker and per condition. Metz et al. (1990) found a correlation of .80 and .84 between naturalness ratings using equal-appearing interval and direct estimation scales, respectively, of read and spontaneous speech samples provided by 20 posttreatment stutterers and 20 nonstutterers. Raters were undergraduate speech-language pathology students. From these data it becomes apparent that speech naturalness, as measured by a rating instrument, is inevitably a characteristic of a sample, rather than a characteristic of a speaker. In a similar vein, in rating a complex concept like speech naturalness one should expect at least some interaction between aspects of the stimuli and idiosyncrasies of the raters.

Probably it is unwise to try to interpret absolute scale values. In any case, one must keep in mind that the scale value of a speech sample is inevitably affected by the requirement to fit stimuli with a very large range of (multidimensional) qualities onto a finite-width one-dimensional scale. Metz et al. (1990) have shown that going from a 7- or 10-point scale to a continuous or to a 10-point scale will not improve the rating accuracy. Samples belonging at the two extremes of a scale will always attract "correct" scores, but it will always be difficult to establish the "true" scale values of the less extreme samples (cf. also Kreiman et al., 1993). The scales cannot be broadened by omitting the normal samples either, because they serve as essential anchoring samples. In the absence of these anchoring samples, naive judges will tend to use the high end of the scale for the "best" (most natural) stuttered samples, thereby making it impossible to interpret the scores as (un)satisfactory in any absolute sense. This seems to be a problem inherent in the use of naive listeners who are required to rate speech samples on what are essentially global semantic scales. However, the problem can perhaps be circumvented by the use of a standardized set of calibration samples, that is, samples that are expressly selected to span the continuum from extremely positive on the one hand and either extremely negative or just above sufficient on the other. Mixing these anchoring samples with the test samples to be assessed might yield judgment scores that can be interpreted in absolute terms. In a way we have already used a precursor of the concept of calibration samples in repeating 20 samples from tape I on tape II, and using the scores on these

samples to transform the scores on tape II to the same reference as tape I.

Clinical Use of the Rating Instruments

Another way of circumventing the anchoring problem might be to use trained raters instead of naive judges. Our research focused on the use of the rating instruments in treatment outcome evaluation. To that end we consistently used relatively large groups of untrained judges. There is a substantial body of literature on the clinical use of the naturalness scale, from which it appears that (most) clinicians do provide reliable naturalness ratings (e.g., Metz et al., 1990; Kreiman et al., 1993). It is also clear, however, that ratings by trained clinicians are usually not more reliable than ratings by untrained judges. Similar results were found by Kreiman et al. (1993) for voice quality ratings. This raises the question whether the instruments described in this paper can at all safely be used by clinicians in their daily practice.

Many recent papers (Kreiman et al., 1993; Onslow et al., 1992, to name just a few) have proposed that the reliability and agreement measures of clinical ratings can be improved by means of specific training procedures, in which all scale positions of all rating scales are "defined" by means of calibrated reference stimuli. Judges should be obliged to listen to these reference stimuli regularly—for example, once before each major rating session. This proposal is not new. It can be found in Laver (1980); his book comes with a tape of examples illustrating most scale values on a large set of speech quality rating scales. Our speech quality scales, although different in shape and definition from the scales proposed by Laver, are sufficiently similar to his scales to instill confidence that we can develop an effective calibration tape for our instrument. We have started work in that direction. Work in our lab has shown that these calibration samples do not need to be developed for many individual languages. Raters trained with the English tape coming with Laver's book were able to rate Dutch material reliably (van Bezooijen, 1988).

Clinicians trained in the use of the speech quality instrument proposed in this paper may employ it to make decisions in planning stuttering treatment. If the speech still sounds unnatural, the clinician can use the scaling instrument to pinpoint the aspects of the speech behavior that need additional improvement and shaping. Eventually, it may be even the other way round. When ratings on calibrated speech quality scales yield results that are less than sufficient, they can be considered to be a trustworthy indication that the speech is still not natural.

References

Bezooijen, R. van (1988). The relative importance of pronunciation, prosody, and voice quality for the attribution of social status and personality characteristics. In R. van Hout & U. Knops (Eds.), *Language attitudes in the Dutch language area* (pp. 85–103). Dordrecht, Holland: Cinnaminson.

Bloodstein, O. (1987). *A handbook on stuttering*. Chicago: National Easter Seal Society.

Boves, L. (1984). *The phonetic basis of perceptual ratings of running speech.* Dordrecht, Holland: Cinnaminson.

Fagel, W. P. F., Van Herpt, L. W. A., & Boves, L. (1983). Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation. *Speech Communication, 2,* 315–326.

Franken, M. C. (1987). Perceptual and acoustic evaluation of stuttering therapy. In H. F. M. Peters & W. Hulstijn (Eds.), *Speech motor dynamics and stuttering* (pp. 285–294). Wien/New York: Springer Verlag.

Franken, M. C., Boves, L., Peters, H. F. M., & Webster, R. L. (1991). Prosodic features in the speech of post-therapy stutterers compared with the speech of nonstutterers. In H. F. M. Peters, W. Hulstijn, & C. W. Starkweather (Eds.), *Speech motor control and stuttering* (pp. 527–535). Amsterdam: Elsevier Science Publishers.

Franken, M. C., Boves, L., Peters, H. F. M., & Webster, R. L. (1992). Perceptual evaluation of the speech before and after fluency shaping stuttering therapy. *Journal of Fluency Disorders, 17,* 223–241.

Giles, H., & Powesland, P. F. (Eds.). (1975). *Speech style and social evaluation.* London/New York/San Francisco: Academic Press.

Ingham, R. L., Costello-Ingham, J. M., Onslow, M., & Finn, P. (1989). Stutterers' self-ratings of speech naturalness: Assessing effects and reliability. *Journal of Speech and Hearing Research, 32,* 419–431.

Ingham, R. J., Gow, M., & Costello, J. M. (1985). Stuttering and speech naturalness: Some additional data. *Journal of Speech and Hearing Disorders, 50,* 217–220.

Ingham, R., Martin, R., Haroldson, S., Onslow, M., & Leney, M. (1985). Modification of listener judged naturalness in the speech of stutterers. *Journal of Speech and Hearing Research, 28,* 495–504.

Ingham, R. J., & Onslow, M. (1985). Measurement and modification of speech naturalness during stuttering therapy. *Journal of Speech and Hearing Disorders, 50,* 261–281.

Ingham, R. J., & Packman, A. C. (1978). Perceptual assessment of normalcy of speech following stuttering therapy. *Journal of Speech and Hearing Research, 21,* 63–73.

Kearns, K. J. (1990). Reliability of procedures and measures. In L. B. Olswang, C. K. Thompson, S. F. Warren, & N. J. Mingetti (Eds.), *Treatment efficacy research in communication disorders.* Rockville, MD: American Speech-Language-Hearing Foundation.

Kreiman, J., Gerrat, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21–40.

Laver, J. (1980). *The phonetic description of voice quality.* Cambridge: Cambridge University Press.

Martin, R., Haroldson, S., & Triden, K. (1984). Stuttering and speech naturalness, *Journal of Speech and Hearing Disorders, 49,* 53–58.

Metz, D. E., Schiavetti, N., & Sacco, P. R. (1990). Acoustic and perceptual dimensions of the perceived speech naturalness of nonstutterers and posttreatment stutterers. *Journal of Speech and Hearing Disorders, 55,* 516–525.

Onslow, M., Adams, R., & Ingham, R. (1992). Reliability of speech naturalness ratings of stuttered speech during treatment. *Journal of Speech and Hearing Research, 35,* 994–1001.

Onslow, M., Hayes, B., Hutchins, L., & Newman, D. (1992). Speech naturalness and prolonged-speech treatments for stuttering: Further variables and data. *Journal of Speech and Hearing Research, 35,* 274–282.

Onslow, M., & Ingham, R. J. (1987). Speech quality measurement and the management of stuttering. *Journal of Speech and Hearing Disorders, 51,* 2–17.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning.* Urbana: University of Illinois Press.

Peters, H. F. M., & Kooijman, P. (1985). *Systematische spraakreconstructie voor stotteraars.* Nijmegen: Associatie Systematische Spraakreconstructie voor Stotteraars.

Runyan, C. M., & Adams, M. R. (1978). Perceptual study of the

speech of "successfully therapeutized" stutterers. *Journal of Fluency Disorders, 3,* 25–39.
Runyan, C. M., & Adams, M. R. (1979). Unsophisticated judges' perceptual evaluations of the speech of "successfully treated" stutterers. *Journal of Fluency Disorders, 4,* 29–38.
Runyan, C. M., Bell, J. N., & Prosek, R. A. (1990). Speech naturalness ratings of treated stutterers. *Journal of Speech and Hearing Disorders, 55,* 434–438.
Runyan, C. M., Hames, P. E., & Prosek, R. A. (1982). A perceptual analysis between paired stimulus and single stimulus methods of presentation of the fluent utterances of stutterers. *Journal of Fluency Disorders, 7,* 71–77.
SPSS Inc. (1990). *Statistical data analysis.* Chicago: Author.
Webster, R. L. (1974). A behavioral analysis of stuttering: Treatment and theory. In K. S. Calhoun, H. E. Adams, & K. M. Mitchell (Eds.), *Innovative treatment methods in psychopathology* (pp. 17–67). New York: Wiley.
Webster, R. L. (1979). Empirical considerations regarding stuttering therapy. In H. H. Gregory (Ed.), *Controversies about stuttering therapy* (pp. 209–240). Baltimore: University Park Press.
Webster, R. L. (1980). *The precision fluency shaping program: Speech reconstruction for stutterers* (Clinician's Program Guide). Roanoke, VA: Communications Development Corporation.
Winer, B. J. (1971). *Statistical principles in experimental design.* New York: McGraw-Hill.

Received November 11, 1993
Accepted September 15, 1994

Contact author: Marie-Christine Franken, Department of Hearing and Speech, University Hospital Rotterdam, Sophia Children's Hospital, Dr. Molenaterplein 60, 3015 GJ Rotterdam, The Netherlands.