Journal of Speech, Language, and Hearing Research, Volume 40, 83–94, February 1997

Stuttering and Communicative Suitability of Speech

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

Renée van Bezooijen
Department of General Linguistics and Dialectology
University of Nijmegen
The Netherlands

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

The purpose of the present study was to develop and evaluate an instrument for assessing the communicative suitability of speech (i.e., the speaking situation-dependent adequacy of speech as judged by listeners). Listeners judged the suitability of speech of people who stutter ($N = 10$) at three stages of treatment (before, immediately after, and 6 months after) and that of people who do not stutter ($N = 10$, the latter serving as a reference). The listeners rated the suitability of the speech, using a 10-point scale, for 10 speaking situations that supposedly make different demands, with listeners consisting of three groups: unsophisticated listeners ($N = 17$), clinicians specializing in the treatment of stuttering ($N = 17$), and stuttering listeners ($N = 17$). Results indicate that the rating instrument can be scored reliably. Analysis of variance for the ratings of the reference speakers showed that the factor "situation" had a significant effect on the suitability ratings, with more demanding situations receiving lower suitability scores than the less demanding ones. Also, the speech of the people who stutter was judged significantly less suitable than the speech of the reference speakers. Furthermore, unsophisticated listeners were considerably less tolerant in their judgments than clinicians and stuttering listeners. Findings suggest that communicative suitability is a promising criterion to further investigate, especially as it may apply to the objective evaluation of treatment outcome for stuttering.

KEY WORDS: stuttering, speech fluency, speech quality, evaluation of speech, evaluation of stuttering

One of the main reasons that a person who stutters seeks treatment is that listeners hear him or her speak in a non-normally fluent fashion. Ideally, a treatment for stuttering helps a person who stutters learn to speak with more normally fluent speech, that is, in a manner that cannot be readily distinguished from average speakers. In order to measure the extent to which speakers produce a normally fluent speech quality, a naturalness scale (Martin, Haroldson, & Triden, 1984) has been introduced. Several studies have demonstrated that spontaneously produced, (nearly) stutter-free posttreatment speech sounds relatively unnatural (Franken, Boves, Peters, & Webster, 1992; Ingham, Gow, & Costello, 1985; Ingham, Martin, Haroldson, Onslow, & Leney, 1985; Onslow, Hayes, Hutchins, & Newman, 1992; Runyan, Bell, & Prosek, 1990). Franken et al. (1992) found that extemporaneous posttreatment speech, resulting from fluency-shaping treatment, sounded as unnatural as pretreatment speech. Follow-up speech, recorded 6 months after treatment, showed a slight improvement in naturalness. Furthermore, using reading passages as speech samples, Kalinowski, Noble, Armson, and Stuart (1994) found that the speech of people who stutter was judged even less natural after fluency-shaping treatment than before. Thus, evaluation outcome studies of various treatments for

1092-4388/97/4001-0083

stuttering suggest that more often than not the quality of the posttreatment speech falls short of the "ideal" (i.e., speech that is not easily distinguished from speech produced by average speakers).

The naturalness scales used in the studies referred to above were all 7- or 9-point equal-appearing interval scales with one extreme defined as "highly natural" and the other extreme defined as "highly unnatural." From a psychometric point of view, equal-appearing interval scales rely on relatively abstract anchors, provided by the meaning of the terms used to label the extremes. It is well known that the actual meaning of the scale terms is determined to a considerable extent by the most (un)natural stimuli in the set to be rated (Boves, 1984). Often, raters are provided with a number of training stimuli at the beginning of the experiment that double as indicators of the range of qualities within the experimental stimuli to be judged. Ratings on equal-appearing interval scales allow one to position stimuli relative to each other and therefore to conclude that stimulus "x" is more "natural" than stimulus "y." However, when evaluating treatment outcome, one would like to be able to go a step further: One would also like to be able to determine whether the speech resulting from a treatment is *sufficiently natural* so that it can be considered to fall somewhere in the distribution of normal speakers.

When requested to assess whether some stimulus is sufficiently "x," the scale extremes are no longer the major anchors, as with equal appearing interval scales. Instead, the scale position indicating the caesura from insufficient to sufficient becomes the anchor point. In addition, when making judgments of sufficiency, the judge has to know: "Sufficient for what?" For instance, a term paper that would deserve an A in high school might be worth a D in graduate school. In the case of judging the sufficiency of speech quality, the judge has to know under what or for which speaking condition or situation "sufficiency" must be judged.

Sociolinguistic research in the United States (Labov, 1972), Britain (Trudgill, 1974), and the Netherlands (Brouwer, 1989) indicates that norms applied to speech are stricter as the speaking situation becomes more formal. Also, when speaking situations emphasize information transmission, speech should be more precise and more standard. Apparently, different situations place different demands on the speech. That is, the type of speech that may be acceptable in the privacy of one's home differs from what is allowed or expected in the public domain with listeners unknown to the speaker. Thus, one might reasonably assume that perceptual judgments of suitability of speech resulting from a treatment will also depend on the specific speaking situation in which the speech is to be used.

Moreover, the 7- or 9-point naturalness scale is not only a relative and abstract evaluation scale, it is also a very global scale (Conture & Wolk, 1990) that is sensitive to many different kinds of deviations from an ideal standard. Thus, the naturalness scale is also sensitive to deviations that do not seem relevant when evaluating speech resulting from stuttering treatment. For instance, Onslow, Adams, and Ingham (1992) pointed out that regional accents or dialectal coloring of speech samples may act as a possible confounding factor in judging the naturalness of speech. Language errors may also confound naturalness judgments. Yet, speech with a regional accent coloring and with some nongrammatical constructions may be perfectly suitable for communicating in many normal daily life situations (Costello Ingham, personal communication, August 1994).

In attempts to minimize the psychometric limitations of the naturalness scale when evaluating results of speech treatment, we propose the concept of communicative suitability that we will define, for the purposes of this study, as the adequacy of the speech relative to the speaking situation as judged by listeners. We attempted to operationalize this concept in the form of the 10-point scale that is commonly used in the Dutch school system. For Dutch adults this scale has clear meanings for all eight intermediate values. A set of situations was selected, ranging from low communicative demands to highly demanding, by choosing 10 specific speaking situations in which the setting (private or public), the number of listeners (single or multiple), and the relation between speaker and listener (known or unknown) were varied (Biber, 1995). Listeners were asked to judge the suitability of speech samples for use in those 10 speaking situations. For instance, listeners rated the suitability of a speech sample for asking directions from a stranger.

The main purpose of our study was to develop and evaluate a rating instrument to measure the communicative suitability of the speech of people who stutter, before and after treatment. A new rating instrument cannot be appropriately used until a number of basic features have been established. For the "suitability scale" to be readily used, it must first of all be proved that ratings are reliable. Furthermore, we would have little trust in the instrument (and maybe even the concept of communicative suitability that it intends to measure) if it did not show that speech samples rated as marginally suitable for low demanding situations are even less suitable for more demanding situations. Finally, we would require that the instrument differentiates atypical fluent speech from typical fluent speech.

The communicative suitability of speech can be judged by many different listener groups. In the case of speech produced by stutterers who have received treatment, three groups seem to be especially relevant: the unsophisticated persons who are likely to confront the

clients in daily life, clinicians specializing in stuttering, and people who stutter. Judgments of clinicians specializing in stuttering should be considered, since their assessments of stutterers' speech play an essential role in the treatment process. If people who stutter judge other stutterers' posttreatment speech insufficiently suitable for their own communicative needs, then one must seriously reckon with the possibility that they will decline the opportunity to use such treated speech in their daily lives. Unsophisticated listeners act as a reference group, unaffected by specialized knowledge of speech disfluencies or specific emotional reactions to disfluencies. If it would appear that unsophisticated listeners judge the speech of people who stutter(ed) differently from clinicians or stuttering listeners, that information should be accounted for in assessing treatment results.

In summary, this study has two major goals. First, we want to investigate the viability of the newly proposed "communicative suitability" scale. To accomplish this goal we will investigate the reliability of the scores, the difference in scores between low-demand and high-demand communicative situations as well as the difference in scores for pretreatment, posttreatment, and follow-up treatment speech of people who stutter(ed) and nonstuttering reference speakers. The second goal is to investigate whether suitability scores depend on the group of judges who rate speech samples. Specifically, we want to investigate whether unsophisticated listeners behave differently from clinicians specialized in stuttering and people who stutter.

Method

Speakers

Speakers were 10 men who stutter (M = 23.7 years, ranging from 15.8 to 39.3 years) and 10 men who do not stutter (M = 27.2 years, ranging from 17.2 to 40.3 years). The 20 speakers were selected from a larger group of 32 stutterers and 20 nonstutterers previously described (Franken et al., 1992). The 10 persons who stutter were selected randomly. Most had a regional accent from the southeast or south of the Netherlands. All men who stutter participated in the Dutch adaptation of the Precision Fluency Shaping Program (PFSP) developed by Webster (1974, 1979, 1980a, 1980b). PFSP can be characterized as a fluency-shaping treatment (Peters & Guitar, 1991). PFSP represents a tightly structured speech motor training program requiring about 120 treatment hours of client participation, from beginning to end. The 10 men who do not stutter were matched with the 10 stutterers for sex, age, educational background, and regional accent. The 10 men who stutter were recorded three times: pretreatment, immediately following treatment ("posttreatment") and 6 months after treatment ("follow-up treatment"). The 10 men who do not stutter were recorded during a single session. The nonstutterers served both as "distractors" in the judgment experiment (i.e., their samples were used to reduce the chance that the listeners would notice the repeated presentation of the same stutterers) and as reference speakers.

Speech

Stimuli for the judgment experiment were selected from a semi-spontaneous speech task in which speakers summarized and commented upon a newspaper article for about 5 minutes. Stutterers were recorded in the clinic pretreatment, posttreatment, and 6 months follow-up treatment, each time commenting on a different topic. Within a condition (e.g., pretreatment) different topics were used for different subjects. Nonstutterers were recorded commenting on two topics in a quiet room (most of them also in the clinic, a few in the school they attended). The signal/noise ratio for all audio-recordings was perceptually acceptable and there was no audible difference between the recordings made in the clinic or the school.

Stimulus material selected consisted of fragments of about 45 seconds following the first 30 seconds of a recording, starting with a new utterance. The fluency of all speech samples used in this experiment was evaluated by the first author. In doing so, normal speech disfluencies were distinguished from stutter-like disfluencies. For the 10 stuttering subjects the mean percentage of all speech disfluencies (normal plus stutter-like) was 22.5 (SD = 9.4) pretreatment, 5.6 (SD = 2.8) posttreatment, and 12.7 (SD = 10.0) 6 months follow-up treatment. The mean percentages of stutter-like disfluencies for the 10 stutterers are 20.0 (SD = 10.7) pretreatment, 3.4 (SD = 3.0) posttreatment, and 8.6 (SD = 6.0) 6 months follow-up treatment. The mean percentages of all speech disfluencies for the nonstutterers were 2.2 (SD = 1.8; first recording) and 2.1 (SD = 2.6; second recording). The mean percentage of stutter-like disfluencies for the nonstutterers of the first recording was 1.0 (SD = 1.0) and of the second recording 0.9 (SD = .9). From these data it can be seen that on the average the two groups show equal proportions of normal disfluencies; also, it can be seen that nonstutterers do occasionally produce disfluencies that would be considered as stutter-like (e.g., a short block introducing a sentence revision).

So, the total number of stimuli judged was 50 (10 stutterers × 3 recordings + 10 nonstutterers × 2 recordings). The 50 stimuli were placed in two random orders, separated by 5 s interstimulus intervals. The experimental stimuli were preceded by five practice stimuli.

Journal of Speech, Language, and Hearing Research, Volume 40, 83–94, February 1997

In subsequent data analysis of communicative suitability and acceptability ratings, only one randomly selected recording of each nonstutterer was used, which means that the ratings of 40 stimuli (10 stutterers × 3 recordings + 10 nonstutterers) were analyzed in order to simplify statistical processing of the data.

Listeners

Stimuli were presented freefield via a Revox A77 tape recorder and a pair of good quality loudspeakers to three groups of listeners ($N = 17$ per group). Within each group, each listener was randomly assigned to one of the two stimulus orders. The three listener groups were (a) a sample of adult (8 men, 9 women) unsophisticated listeners, that is, 17 members of a chorus, who do not stutter, and varied in educational background; (b) "trained" listeners, that is, clinicians specializing in stuttering (6 men, 11 women) who were members of the Dutch Organization for Stuttering Therapy; and (c) "stuttering" listeners, that is, adults who stutter (12 men, 5 women) and who were all involved in a type of group treatment that can be characterized as a stuttering modification treatment (Peters & Guitar, 1991) named "Stichting Stottertherapie Doetinchemse Methode." All listeners served as unpaid volunteers.

Rating Scales

The 51 listeners judged the communicative suitability of the stimuli on 10 scales that refer to specific speaking situations that supposedly make different demands. A point of departure in creating a set of situations that should differ in terms of communicative demands was formed by the results of sociolinguistic research (Biber, 1995). In sociolinguistic theory a number of factors (dimensions) are distinguished that affect the demands or the formality of a communication situation. The three most important factors are: (a) the setting where the communication takes place (private vs. public domain), (b) the number of persons spoken to (single conversational partner vs. multiple conversational partners), and (c) the relation to the person(s) spoken to (known conversational partner[s] vs. unknown conversational partner[s]). The third factor includes aspects of emotional bonding, that is, the likelihood that the speaker has some kind of emotional bond with a known person is rather high.

The true dimensionality of the communication situation space is not known; although the three factors mentioned above are believed to be the most important ones, they may not cover the complete space. At the same time, it is not guaranteed that the factors we included are orthogonal, let alone orthonormal (in other words, we do not know whether the factors are really independent and equally important). Yet, we decided that it was possible to define five global situations that supposedly span a continuum from least demanding to most demanding. Communication in a private environment with a single person who is known to the speaker was considered least demanding. Speaking in public to a large number of unknown listeners was considered most demanding. The intermediate situations were defined by varying the position along the three dimensions. Since we wanted to use two specific communicative contexts (one stressing the social function of speech and the other stressing information transfer) in each of the five global situations, we ended up with 10 communicative contexts. This was the maximum number of contexts we considered feasible for rating in this experiment. Table 1 shows the five pairs of speaking situations that were included in the experiment.

Listeners judged the suitability of each speech sample for use in each of the 10 speaking situations on the 10-point suitability scale, keeping in mind the grading scale that is commonly used in the Dutch educational system. To prevent any possible confusion, the meaning of all 10 points was explicitly explained in the scoring instructions (1: very bad; 2: bad; 3: moderate; 4: insufficient; 5: just insufficient; 6: just sufficient; 7: amply sufficient; 8: good; 9: very good; 10: excellent). Therefore, the scale can be considered an anchored 10-point scale, allowing interpretation of the judgments of

Table 1. Five pairs of speaking situations combining different +/− values of private, single, and known. Situations #1 and #2 are Low Demanding, and situations #9 and #10 are Highly Demanding.

Low demanding
+ private, + single, + known
1. talking about everyday events with a friend
2. telling a housemate about one's new job
+ private, − single, + known
3. chatting with housemates during a party game
4. giving a speech at a family celebration
− private, + single, + known
5. making conversation with a friend in the train
6. ordering bread from the baker around the corner
− private, + single, − known
7. getting into contact with a stranger on the bus
8. asking a passerby for directions
− private, − single, − known
9. instructing a group at a dancing school
10. giving a lecture to a newly founded professional association
Highly demanding

communicative suitability in absolute terms. Order of presentation of the 10 situations to be rated was random and changed every 10 stimuli.

Scoring Procedure

Listeners rated the scales while listening to the 45-second stimuli, so, on the average, listeners had about 4 seconds to rate each scale. The total experiment took about one hour. Before the scoring session started the 10 situations were explained to the listeners. In order to do that, they were asked to rate the suitability of their own speech for each of the situations. Moreover, listeners were instructed to pay attention only to *how* things were said, not to *what* was said. In this way we intended to focus the assessment on speech quality and guide it way from other, possibly interfering linguistic characteristics that might affect communicative suitability.

Results

From previous experiments with ratings of large numbers of speech stimuli it has appeared that effects of offering the stimuli in different orders are negligible. Therefore, we decided to collapse the scores of the two orders of presentation from the outset.

Table 2 shows the mean ratings and standard deviations for the 10 speaking situations, the four speaker groups (10 stuttering speakers pre-, post-, and 6 months follow-up treatment and 10 reference speakers) and the three listener groups (unsophisticated listeners, clinicians, and stuttering listeners). The table should allow the reader to obtain an impression of the ranges spanned by the scores.

Table 2 illustrates the differences between the speaking situations, the listener groups, and the speaker groups. For instance, posttreatment speech is judged

Table 2. Means and (in italics) standard deviations of suitability ratings for three listener groups (unsophisticated listeners, clinicians, and stuttering listeners) and four speaker groups (people who stutter pre-, post-, and 6 months follow-up treatment, and reference speakers), for each of 10 speaking situations separately (number in first column) as well as averaged over the 10 situations (av in first column). For a detailed description see text; for the meaning of the numbers of the speaking situations, see Table 1.

	Unsophisticated listeners				Clinicians				Stuttering listeners			
	pre	pos	fol	ref	pre	pos	fol	ref	pre	pos	fol	ref
1	4.4	4.9	5.6	7.4	5.9	6.7	6.7	8.3	5.9	6.6	6.8	8.2
	2.0	*1.7*	*2.0*	*1.2*	*2.0*	*1.9*	*1.8*	*1.4*	*2.1*	*1.6*	*2.0*	*1.3*
2	4.2	4.6	5.4	7.0	5.6	6.3	6.4	7.9	5.6	6.5	6.5	8.2
	1.8	*1.7*	*2.0*	*1.4*	*2.0*	*1.8*	*1.8*	*1.5*	*2.1*	*1.8*	*2.0*	*1.4*
3	4.1	4.6	5.4	7.0	5.6	6.2	6.4	7.9	5.5	6.4	6.6	8.1
	1.8	*1.6*	*1.9*	*1.5*	*2.0*	*1.8*	*1.9*	*1.5*	*2.2*	*1.8*	*2.0*	*1.2*
4	2.4	3.1	3.6	5.9	3.1	4.0	4.2	6.6	4.0	5.1	5.3	7.5
	1.7	*1.8*	*2.3*	*2.1*	*1.8*	*1.8*	*2.0*	*2.2*	*2.2*	*1.9*	*2.2*	*1.7*
5	3.9	4.6	5.2	7.0	5.4	6.2	6.3	7.9	5.2	6.2	6.5	8.1
	1.8	*1.7*	*2.0*	*1.3*	*1.8*	*1.7*	*1.7*	*1.4*	*2.0*	*1.5*	*1.9*	*1.2*
6	3.9	4.8	5.4	7.1	5.1	6.4	6.3	8.0	4.5	5.7	5.9	7.7
	2.0	*1.8*	*2.2*	*1.4*	*2.1*	*1.6*	*1.9*	*1.4*	*2.2*	*1.6*	*2.1*	*1.3*
7	3.1	3.8	4.4	6.6	4.6	5.6	5.6	7.6	4.1	5.2	5.4	7.6
	1.7	*1.6*	*2.1*	*1.5*	*1.8*	*1.7*	*1.8*	*1.6*	*2.2*	*1.6*	*2.2*	*1.4*
8	3.4	4.3	5.0	6.8	4.9	6.2	6.0	7.8	4.3	5.6	5.7	7.8
	2.0	*1.8*	*2.1*	*1.5*	*2.1*	*1.7*	*1.9*	*1.5*	*2.0*	*1.7*	*2.2*	*1.4*
9	1.8	2.1	2.9	4.8	2.4	3.0	3.4	5.4	3.0	3.9	4.2	6.9
	1.3	*1.5*	*2.3*	*2.2*	*1.7*	*1.9*	*2.1*	*2.5*	*2.0*	*1.8*	*2.4*	*1.7*
10	1.9	2.1	3.1	4.9	2.3	3.1	3.2	5.6	3.1	4.0	4.4	6.9
	1.4	*1.5*	*2.3*	*2.3*	*1.6*	*1.8*	*2.1*	*2.4*	*1.9*	*1.7*	*2.3*	*1.9*
av	3.3	3.9	4.6	6.4	4.5	5.4	5.4	7.3	4.5	5.5	5.7	7.7
	2.0	*2.0*	*2.4*	*1.9*	*2.3*	*2.2*	*2.3*	*2.0*	*2.3*	*1.9*	*2.3*	*1.5*

sufficiently suitable for talking about everyday events with a friend (situation #1) by the clinicians and stuttering listeners, but not by the unsophisticated listeners. Posttreatment speech is definitely insufficient for instructing a group at a dancing school (situation #9) for all three listener groups. Overall ratings for the people who stutter increase, going from pretreatment, to posttreatment, to 6 months follow-up treatment speech, but they remain much lower than the ratings for the reference speakers.

We will now address the research questions formulated in the Introduction. First, we will investigate the basic psychometric properties of the new instrument. We will do that by addressing a number of issues, including the reliability of the scores, the differences in scores between putative low- and high-demand situations and the difference between the scores for the men who stutter at the 3 moments in time compared to the scores for the reference speakers. Finally, we will investigate differences between ratings by the three listener groups.

Reliability of Listener Judgments

Data analysis was based on the judgments of 40 stimuli, namely 30 recordings of 10 men who stutter at three different stages of treatment and 10 recordings of 10 men who do not stutter. Reliability of listener judgments was assessed separately for the 10 scales and the three listener groups by means of Cronbach's alpha, which can be considered as the average correlation between the judgments of listeners, taking into account the number of raters. All alphas exceeded .95, which is extremely high (Rietveld & Van Hout, 1993). To investigate the extent to which the alphas were inflated by the large between-speaker differences, the alphas were also computed separately for the 10 scales and the four speaker groups. All these alphas exceeded .93. From these results it is safe to conclude that the scores are sufficiently reliable to warrant further statistical processing.

Ranking of the 10 Communication Situations

In order to check whether the 10 communication situations used in our instrument do indeed span a continuum from low demand to high demand we analyzed the scores for the reference speakers. The hypothesis behind this procedure is that speech samples produced by an average group of normal speakers should obtain lower average ratings as the demands grow. To answer this question an analysis of variance (using BMDP8V—general mixed model) was applied to the ratings for the reference speakers, with the fixed factors situation (10 levels) and listener group (three levels), plus the random factor speaker (10 levels). The factor situation had a significant effect on the ratings, $F(9, 81) = 87.45$, $p < 0.001$, eta^2 37%. In addition to the main effect for "situation" there was a significant interaction between situation and listener group, $F(18, 162) = 19.56$, $p < 0.001$, eta^2 4%.

Table 3 shows the mean communicative suitability ratings for the reference speakers as a function of speaking situation.

It can be seen that the speech was judged least suitable for the most demanding situations #9 and #10 (speaking in public to a group of unknown listeners) and most suitable for the least demanding situations #1 and #2 (speaking to a friend in a private environment). The remaining pairs of situations received intermediate ratings. To investigate which differences among the 10 speaking situations were significant, post hoc analyses using Tukey's HSD test (Jaccard, Becker, & Wood, 1984) were carried out. They revealed that a substantial number of situations did not differ significantly from each other. Homogeneous subsets were: #9 and #10, and #4 on its own. Overlapping homogeneous subsets were: #7, #8, and #6; #8, #6, #3, #5, and #2; and, finally, #3, #5, #2, and #1.

Figure 1 shows the significant interaction between speaking situation and listener group regarding the reference speakers. In essence, the clinicians were relatively more strict in their judgments in the most demanding situations (i.e., situations #9, #10, and #4). In addition, the stuttering listeners judged relatively less strictly in these situations.

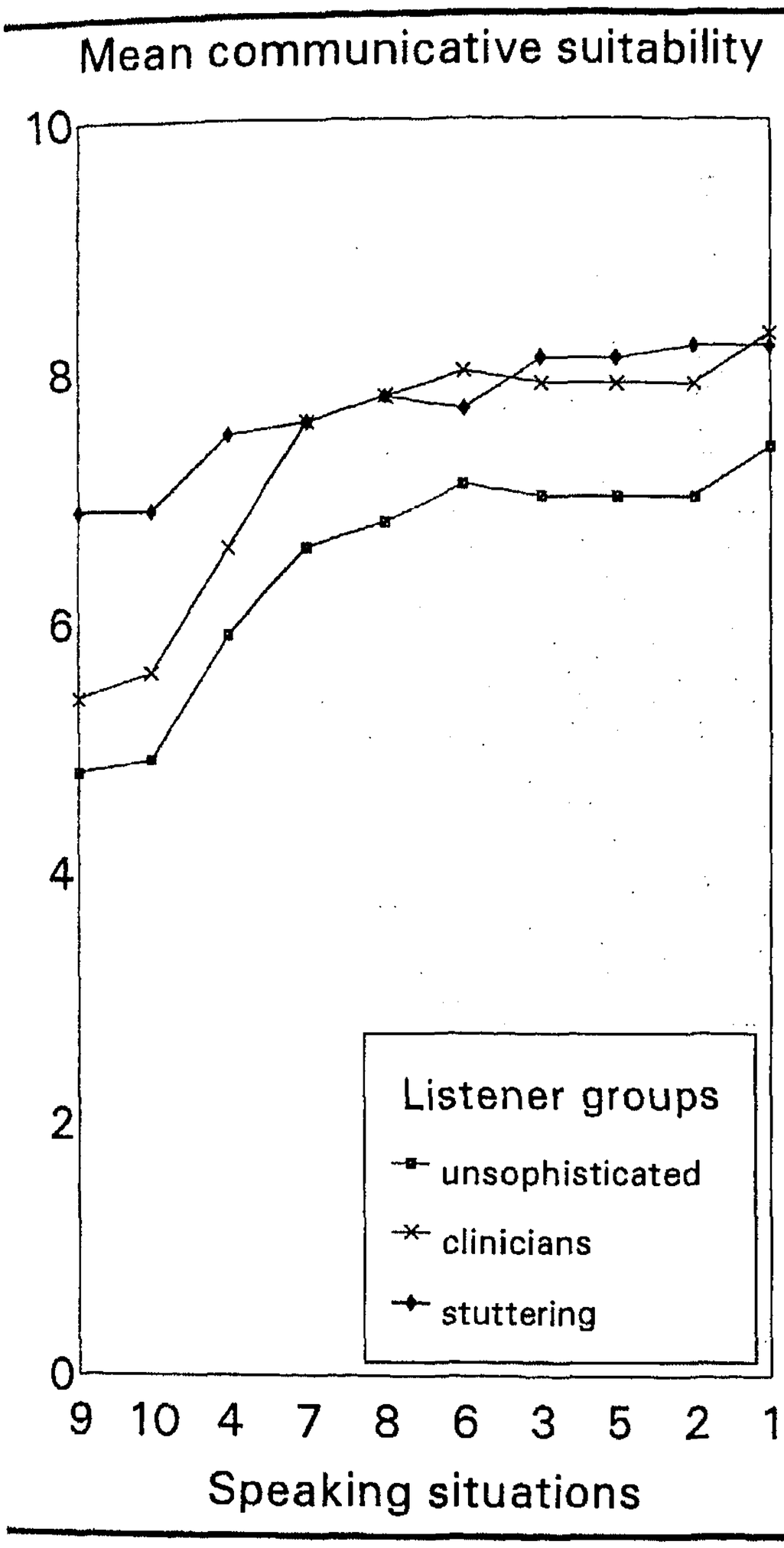
Does the Rating Instrument Discriminate Atypically From Typically Fluent Speech?

For a rating instrument to be valid we would require that it is at least able to show significant differences

Table 3. Mean communicative suitability ratings for reference speakers for 10 speaking situations (10-point scales), ordered from lowest to highest suitability. For the meaning of the situation numbers, see Table 1.

Speaking situation	Mean	SD
9	5.7	2.2
10	5.8	2.2
4	6.7	2.0
7	7.3	1.6
8	7.5	1.5
6	7.7	1.4
3	7.7	1.4
5	7.7	1.3
2	7.7	1.4
1	8.0	1.3

Figure 1. Interaction between speaking situation and listener group (unsophisticated listeners, clinicians, and stuttering listeners) for the speech of the reference speakers. Suitability ratings (10-point scales) for the 10 speaking situations. The 10 situations are ordered from low to highly demanding. See the means for the reference speakers in Table 3. For the meaning of the speaking situations, see Table 1.

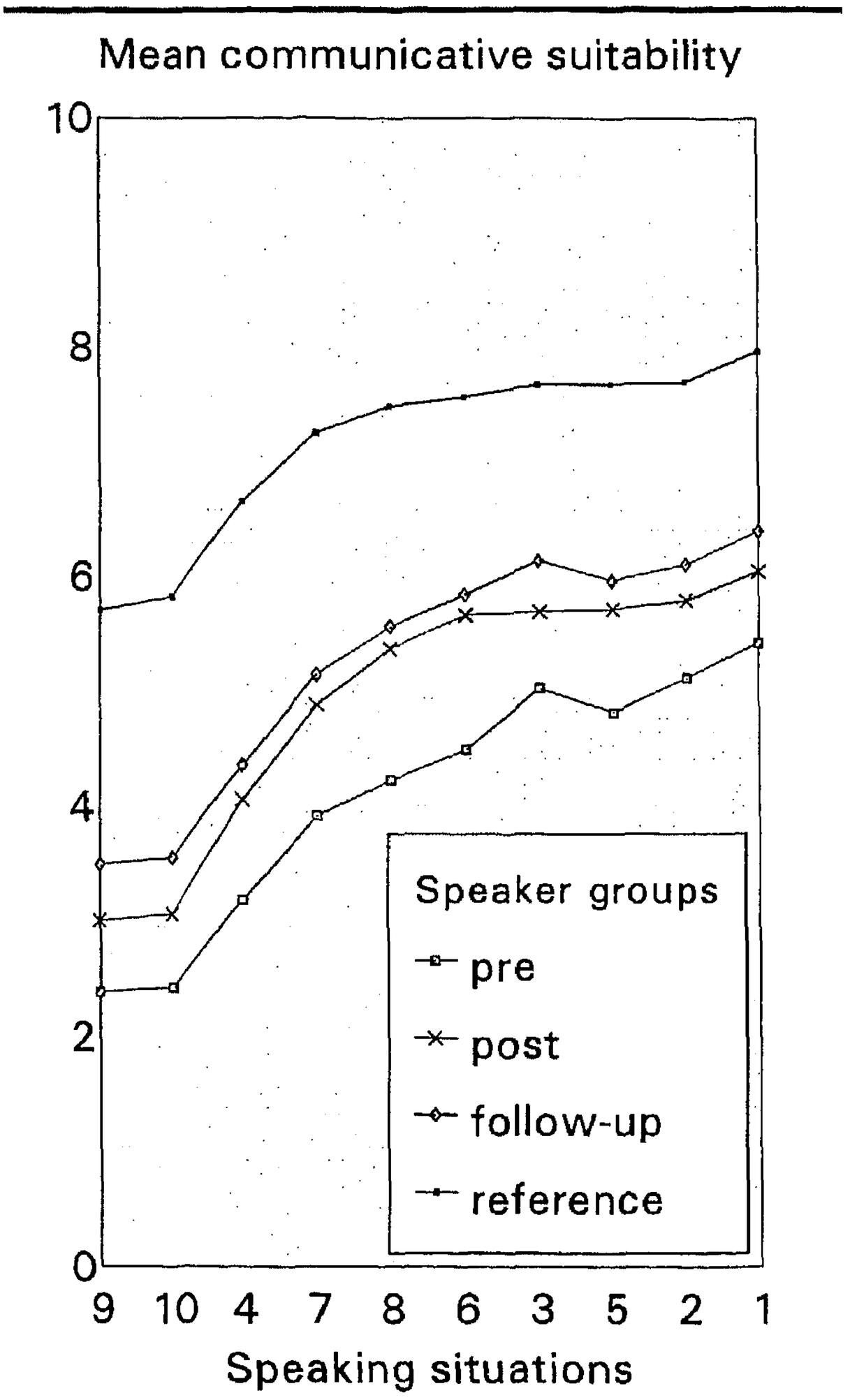


between untreated stuttered speech and the speech of reference speakers. Since previous research has invariably shown that also post- and follow-up treatment speech differs quite audibly from reference speech we also expect the instrument to be able to distinguish the speech in these situations from reference speech. Lastly, we would expect that the instrument would bring to light differences between the three conditions for the speech of the persons who stutter.

To investigate these issues three analyses of variance were carried out. The design for these analyses comprised the fixed factors situation (10 levels), speaker group (two levels), and listener group (three levels), plus the random factor speaker (10 levels). The results show that ratings for the speakers who stutter were significantly different from ratings for the reference speakers for pretreatment, $F(1, 18) = 41.21$, $p < 0.001$, eta^2 50%, posttreatment, $F(1, 18) = 37.51$, $p < 0.001$, eta^2 42%, and 6 months follow-up treatment, $F(1, 18) = 15.04$, $p < 0.001$, eta^2 27%. These findings confirm that raters judged samples from speakers who stutter as significantly less suitable than those samples produced by the reference speakers.

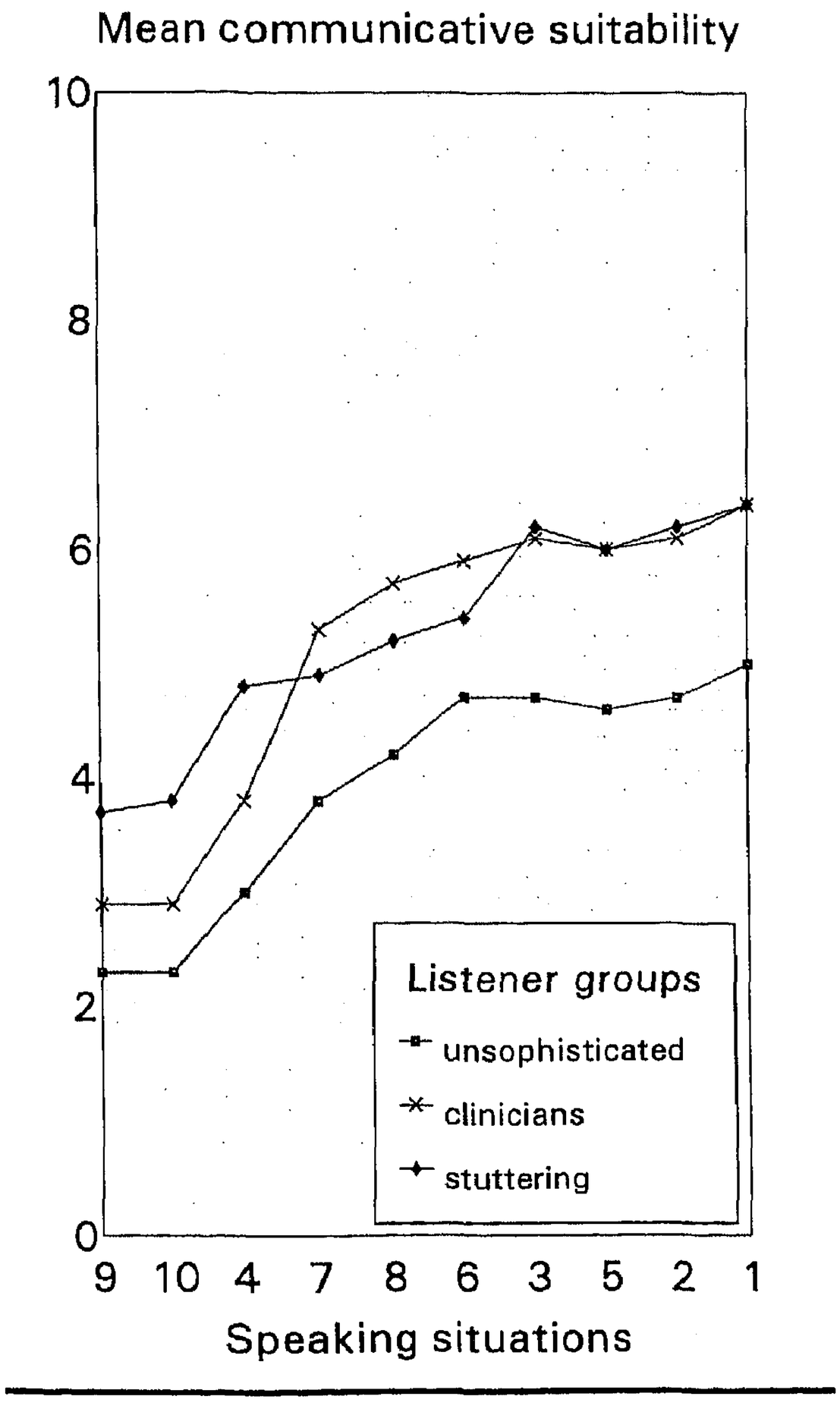
The mean scores for the pretreatment, posttreatment, and 6 months follow-up treatment speakers, averaged over the 10 speaking situations and the three listener groups, were 4.1, 4.9, and 5.2, respectively. The mean score for the reference speakers was 7.2. Figure 2

Figure 2. Mean suitability ratings for the 10 speaking situations separately for the three moments of measurement of the stuttering speakers and for the reference speakers. The 10 situations are ordered from low to highly demanding. See the means for the reference speakers in Table 3. For the meaning of the speaking situations, see Table 1.



shows the suitability ratings for the 10 speaking situations separately for the three moments of measurement of the stuttering speakers and for the reference speakers. It can be seen that the pretreatment speech of the persons who stutter is rated as insufficiently suitable for each of the 10 speaking situations, with their posttreatment and follow-up treatment speech both being judged to be slightly more suitable. Listeners appear to consider stuttering speakers' posttreatment and follow-up treatment speech (almost) suitable for low and medium demanding situations, but reject it for highly demanding situations.

Figure 3. Interaction between situation and listener group (unsophisticated listeners, clinicians, and stuttering listeners) for the speech of the speakers who stutter (including their pretreatment, posttreatment, and 6 months follow-up treatment measures). Suitability ratings (10-point scales) for the 10 speaking situations. The 10 situations are ordered from low to highly demanding. See the means for the reference speakers in Table 3. For the meaning of the speaking situations, see Table 1.



Do Judgments of Communicative Suitability Differ Significantly Between Unsophisticated Listeners, Clinicians, and Stuttering Listeners?

To compare the judgments for the three moments of measurement for the stuttering speakers (pretreatment, posttreatment, and 6 months follow-up treatment) an analysis of variance was carried out, using a repeated measurements design with the fixed factors situation (10 levels), moment of measurement (three levels), and listener group (three levels), plus the random factor speaker (10 levels). For the stuttering speakers the factor listener group was highly significant, $F(2, 18) = 107.58$, $p < 0.001$, eta^2 11%. The mean ratings of the unsophisticated listeners, clinicians, and stuttering listeners for the stuttering speakers only were 3.9, 5.1, and 5.3, respectively. Post hoc analyses using a Tukey HSD test revealed significant differences between the unsophisticated listeners and the clinicians and between the unsophisticated listeners and the stuttering listeners. The clinicians and stuttering listeners did not differ significantly.

In addition to the main effect of listener group, there was a significant interaction between listener group and situation, $F(18, 162) = 39.38$, $p < 0.001$, eta^2 1.8%. Figure 3 graphically depicts this interaction for the speech of the stuttering speakers. A similar deviation from parallelism as in Figure 1 (for the reference speakers) can be observed: The clinicians were relatively more strict in most demanding situations (i.e., situations #9, #10, and #4).

Thus, although clinicians and stuttering listeners seem to behave as a single group, the clinicians judge more strictly in the most demanding situations, both for the speech of the reference speakers and the speech of the stuttering speakers. Moreover, the speech of the reference speakers is judged relatively mildly by the stuttering listeners in the most demanding situations.

Discussion

The aim of the present study was to assess whether a communicative suitability rating instrument could be used to meaningfully measure the suitability of the speech resulting from stuttering treatment for several speaking situations differing in demands. The present study measured the communicative suitability of the speech of people who stutter, before and after treatment, as well as the speech of people who do not stutter for use in 10 typical conversational situations.

Reliability of Listener Judgments

In this study evaluation reliability was computed using Cronbach's alpha. All alphas exceeded .93,

including those computed separately for the four speaker groups. Based on these findings, it seems safe to conclude that the listeners agreed to a high extent on the suitability of the experimental speech samples for various speaking situations. Thus, the listeners were able to carry out the rating task in a reliable and meaningful way. For the judgments of the clinicians it would be interesting to know whether the ratings are also stable over time. An experiment to investigate this issue is under preparation.

Speech Material

The present study was based on excerpts from monologues in which speakers summarized and commented on newspaper articles they had read. The speaking situations for which the suitability of the speech samples was to be judged were varied in that they asked for (spontaneous) dialogues as well as (prepared) monologues, pertaining to a wide variety of topics, ranging from talking about one's job to ordering bread and giving a lecture. It is not clear to what extent this less-than-ideal matching affected listener judgments. However, it is also not clear how the method can be improved.

The "ideal" study in which situations and speech samples are exactly matched would require 10 speech samples for each speaker, one for each of the speaking situations under judgment. However, such a study would incur enormous methodological problems: firstly, it would be very expensive to make all necessary recordings; secondly, the samples are likely to differ substantially in many aspects, making comparison in terms of measurable characteristics of the speech very difficult, if not impossible. In the instructions for the listeners it was carefully explained that they should imagine the situation and try to decide how suitable the type of speech represented by the samples would be in that situation. Previous studies have shown that ratings of speech quality are not affected significantly by the situation in which the speech has been produced (Onslow, Hayes, Hutchins, & Newman, 1992).

Effect of Speaking Situations

Listeners' judgments of the samples of the reference speakers support the basic assumption that there is a continuum of speaking situations that range from low demand on speech quality to (very) high demands: A speech quality that is just good enough for a low-demanding situation is judged increasingly less suitable as the situation becomes more demanding. Thus, the idea of measuring the concept of "communicative suitability" is basically sound. The differences between the speaking situations correspond to differences in the strictness of norm application. The characteristics typical of formal speaking situations (e.g., importance of the message, listeners unfamiliar with the speaker [and his or her speech style], large physical distance between listeners and speaker, background noise, etc.) require speech that is sufficiently clearly and loudly enunciated, without deviant and unpredictable properties. Our findings support the notion that listeners expect conformation to standard norms with regard to outward appearance in formal situations. In a home situation, when speaking with friends, one can do as one pleases; however, outside the home, in the public domain, one has to act and talk according to generally accepted rules.

Current findings show that five homogeneous subsets of situations (three of which were overlapping) could be discriminated for the reference speakers. These subsets of increasing level of speech demands do represent an underlying low to highly demanding continuum. Taking the overlap into account, our results suggest that one can meaningfully measure communicative suitability by using four levels of demands: (a) low demanding: "nonformal"; (b) medium demanding: "somewhat formal"; (c) highly demanding: "formal"; and (d) very highly demanding: "very formal."

The first, low demanding, nonformal level, comprises the subset of situations #1, #2, #5, and #3. The second level, somewhat formal, comprises the subset of situations #7, #8, and #6. The third or formal level describes situation #4. The most demanding, fourth or very formal level, describes the homogeneous subset of situations #9 and #10. Such an arrangement would eliminate the overlap between three (out of five) homogeneous subsets.

Although our efforts to design a rating instrument for measuring communicative suitability of speech show promising results, we think that the current version of the instrument can be improved. First, it is not yet clear how many different levels of demands should be distinguished in treatment outcome evaluation; neither is it completely clear what the optimal dimensions of the continuum should be. In the present study, three aspects of speaking situation were varied: setting (private vs. public), number of persons spoken to (single vs. multiple), and relation to the person spoken to (known vs. unknown). However, other aspects, such as age and sex of the speaker and the emotional relation between speaker and listener (i.e., factors other than the degree of formality of the situation) might also constitute relevant dimensions. To clarify the situation additional research is necessary.

For the instrument in its present form (emphasizing formality) one may question whether highly demanding (formal) levels should be included in an instrument

that is aimed at measuring quality of speech of people who stutter before and after treatment. We decided to include these situations in the present study in order to test the hypothesis that more demanding situations do indeed give rise to lower suitability scores, which clearly happened to be the case. This supports the soundness of the underlying concept. One should be very cautious, however, with including (very) formal situations in a "suitability instrument" intended for routine applications, because that might convey the (inappropriate) suggestion that the quality of posttreatment speech should be sufficient for these situations. Indeed, our findings show that even reference speakers are not always judged sufficiently suitable for very formal, public speaking situations (mean ratings 5.7 and 5.8 in situations #9 and #10, respectively). Speech fluency alone, although necessary, is not sufficient in (very) formal speaking situations. The speaker's verbal and cognitive abilities are probably also taken into account in making judgments whether somebody's speech is "sufficiently suitable." Stuttering treatments are aimed at improving fluency but typically neglect verbal and/or cognitive skills and development. It is beyond the scope of a stuttering treatment to demand that the posttreatment speech of all people who stutter should be sufficiently suitable for (very) formal situations.

Differences Between Atypically and Typically Fluent Speech

Listeners judged the samples from the speakers who stutter significantly less suitable than the samples of the reference speakers. The largest difference in judgment existed between the reference speech and the pretreatment speech. The mean judgments for the reference speech show that the listeners considered all reference speakers sufficiently suitable for nonformal and somewhat formal speaking situations. The judgments for the pretreatment speech of people who stutter, however, show that the untreated speech was judged as essentially unsuitable for all types of communicative functioning for all communication situations in our instrument. Thus, the data show that the current instrument can discriminate untreated, atypically fluent speech from typically fluent speech.

The judgments for the treated speech are of special interest. The mean suitability values for posttreatment and follow-up treatment speech are higher than for pretreatment speech, but the ratings of the speech in these two conditions are still significantly different from the reference speech. This difference cannot satisfactorily be explained by the fact that not all clients were completely fluent in the post- and follow-up treatment conditions. The percentage of stutter-like disfluencies is 3% in posttreatment samples, a proportion that does not differ from the proportion of stutter-like disfluencies in the speech of the control speakers, and 9% in 6 months follow-up. It is extremely unlikely that the percentage of stutter-like disfluencies dominates the overall suitability judgments, because the mean suitability rating in the posttreatment condition was 4.9, whereas 6 months follow-up treatment the mean rating was 5.2. Thus, when the percentages of stutter-like disfluencies are relatively low, it seems more likely that other aspects of speech quality determine the suitability judgments. This seems the more so because the presence of stutter-like disfluencies in the speech of the reference speakers did not cause them to be judged as unsuitable for the less demanding communicative situations.

Naturalness Versus Suitability

In the Introduction we argued that the concept of "communicative suitability" is less global than naturalness. Can we use the present dataset to demonstrate that suitability and naturalness are conceptually different? The 10 speakers who stuttered in the present study were a subset of the 32 speakers from Franken et al. (1992). In that study a significant shift in speech naturalness between post- and 6 months follow-up treatment speech was established. On the other hand, posttreatment speech was judged equally natural as pretreatment speech. The mean naturalness ratings of the 10 speakers in this study are: pretreatment 3.1 (SD = 0.72), posttreatment 3.3 (SD = 0.66), and follow-up treatment 4.0 (SD = 0.85). Thus, the major shift in speech naturalness occurs between posttreatment and follow-up treatment. The data of the present study show that on the average the posttreatment speech was judged as more suitable than the pretreatment speech, and that follow-up treatment the suitability was better than posttreatment. The improvement in the suitability judgments for the treated speech is concentrated in the step from pre- to posttreatment. This trend is especially obvious in the scores of the clinicians and stuttering listeners. Listeners seem to consider the controlled fluency and artificial speech technique rather unnatural (Franken et al., 1992). However, at the same time, they tend to accept it for nonformal speaking situations and almost accept it for somewhat formal speaking situations. On the other hand, listeners do consider the uncontrolled, severely stuttered speech unsuitable, regardless of the formality of the speaking situation. It is not yet completely clear why listeners judge the speech of stutterers who control their fluency by means of an audible, artificial speech technique more suitable, but not more natural than pretreatment speech. Perhaps listeners feel more at ease with a speaker who controls his or her speech performance than with a speaker who stutters unpredictably.

Effect of Listener Group

The third main factor, listener group, had a significant and meaningful effect on the suitability ratings for the stuttering speakers. The judgments of the unsophisticated listeners are (relatively) more severe than clinicians and stuttering listeners. Apparently, unsophisticated, nonstuttering listeners place higher demands on the quality of speech than clinicians specialized in stuttering and persons who stutter. On the average, the difference is more than one point on the 10-point scale. This implies that persons who stutter and clinicians may judge a specific speech sample as communicatively suitable, whereas the "person in the street" has a different opinion.

It should be noted that the treatments received by the speakers in the study and the stuttering judges are different. The judges in this study were taking part in a stuttering modification treatment, whereas the speakers under judgment had followed a fluency-shaping treatment. It would be interesting to investigate whether judges who followed a fluency shaping treatment would react differently to the speech in the post- and follow-up treatment conditions. Different reactions toward pretreatment and reference speech seem quite unlikely.

We assume that the difference in rating behavior between the unsophisticated listeners and clinicians plus stuttering listeners can, at least partly, be explained by a difference in familiarity with atypically fluent speech. As a result of repeated exposure, speech of people who stutter may sound more familiar to clinicians specialized in stuttering and people who stutter than to unsophisticated listeners who do not stutter. Moreover, the milder judgment of the former two groups may be related to the fact that they know by experience how difficult it is to achieve and establish fluent speech production. Finally, both groups have learned that some speech disfluencies are normal. More research is needed to establish the importance of the discrepancies between the judgments of clinicians, stuttering listeners, and unsophisticated listeners. If the "man in the street" does indeed feel that clinicians and persons who stutter judge posttreatment speech too positive, this should eventually have consequences for the decision to consider a treatment completed and successful.

There was also a significant interaction between listener group and speaking situation: The clinicians judge relatively strictly in the most demanding situations, both for the speech of the reference speakers and the speech of the stuttering speakers. In the most demanding situations, the clinicians move closer to the unsophisticated listeners. It could be that for the public use of speech, clinicians take the point of view of outsiders into account. Another aspect of the interaction relates to the judgment of the reference speech by the stuttering listeners in the most demanding situations: This judgment appears to be relatively mild. This is encouraging, since it suggests that the unnatural character of post- and follow-up treatment speech may not be a very important factor for people who stutter to learn and actually employ artificial fluency-enhancing techniques.

Conclusions

The present study aimed to evaluate an instrument to perceptually measure the adequacy of speech quality relative to the speaking situation, that is, communicative suitability. The following conclusions can be drawn.

1. Communicative suitability ratings are reliable. Unsophisticated listeners, clinicians, and stuttering listeners showed good agreement in judging the communicative suitability of speech samples from persons who stutter (recorded pretreatment, posttreatment, and 6 months follow-up treatment) and from nonstuttering reference speakers.
2. The concept of measuring listener judgments of "communicative suitability" seems basically sound. The data showed that speech qualities that are just good enough for low-demanding situations are judged unsuitable for highly demanding situations.
3. The instrument to measure communicative suitability herein described can discriminate speech of people who stutter from speech of nonstuttering, reference speakers.
4. Distinguishing between nonformal, somewhat formal, formal, and very formal speaking situations seems relevant when judging communicative suitability.
5. Our data suggest that "the man in the street" is less tolerant in his judgments than clinicians specializing in stuttering and listeners who stutter. But in their judgments for the most demanding speaking situations the clinicians are relatively strict too; stuttering listeners are milder in their judgments than clinicians and unsophisticated listeners.

Acknowledgments

The contribution by the second author to this research has been made possible by a fellowship from the Royal Netherlands Academy of Arts and Sciences. The authors wish to thank Ed Conture for his many insightful editorial comments on an earlier version of this manuscript, and Anneke Olierook and Margot Osse-Spanhof for their assistance with data collection.

References

Biber, D. (1995). An analytical framework for register studies. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 31–56). New York: Oxford University Press.

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

Brouwer, D. (1989). *Gender variation in Dutch. A sociolinguistic study of Amsterdam speech.* Dordrecht, Holland/ Providence, RI: Foris Publications.

Conture, E. G., & Wolk, L. (1990). Stuttering. *Seminars in Speech and Language, 11*(3), 200–210.

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,* 223–242.

Ingham, R. J., Gow, M., & Costello, J. M. (1985). Stuttering and speech naturalness: Some additional data. *Journal of Speech and Hearing Disorders, 50,* 217–219.

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

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin, 96,* 589–596.

Kalinowski, J., Noble, S., Armson, J., & Stuart, A. (1994). Pretreatment and posttreatment speech naturalness ratings of adults with mild and severe stuttering. *American Journal of Speech-Language Pathology, 3*(2), 61–66.

Labov, W. (1972). *Sociolinguistic patterns.* Philadelphia: University of Pennsylvania Press.

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

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.

Peters, T. J., & Guitar, B. (1991). *Stuttering. An integrated approach to its nature and treatment.* Baltimore: Williams & Wilkins.

Rietveld, T., & Van Hout, R. (1993). *Statistical techniques for the study of language and language behaviour.* Berlin/ New York: Mouton de Gruyter.

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.

Trudgill, P. (1974). *The social differentiation of speech in Norwich.* Cambridge, MA: Cambridge University Press.

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. 17–61). New York: Wiley.

Webster, R. L. (1979). Empirical considerations regarding stuttering therapy. In H. H. Gregory (Ed.), *Controversies about stuttering therapy* (pp. 209–239). Baltimore: University Park Press.

Webster R. L. (1980a). *The precision fluency shaping program: Speech reconstruction for stutterers. Clinician's program guide.* Roanoke, VA: Communications Development Corporation.

Webster R. L. (1980b). Evolution of a target-based behavioral therapy for stuttering. *Journal of Fluency Disorders, 5,* 303–320.

Received January 29, 1996

Accepted July 9, 1996

Contact author: Marie-Christine Franken, Academic Hospital Rotterdam, Sophia Children's Hospital, Department of Hearing and Speech, Dr. Molewaterplein 60, 3015 GJ Rotterdam, The Netherlands. Email: franken@knos.azr.nl