Research Article

The Efficacy of Stuttering Measurement Training: Evaluating Two Training Programs

Lauren A. Bainbridge,^a Candace Stavros,^a Mineh Ebrahimian,^a
Yuedong Wang,^a and Roger J. Ingham^a

Purpose: Two stuttering measurement training programs currently used for training clinicians were evaluated for their efficacy in improving the accuracy of total stuttering event counting.

Method: Four groups, each with 12 randomly allocated participants, completed a pretest–posttest design training study. They were evaluated by their counts of stuttering events on eight 3-min audiovisual speech samples from adults and children who stutter. Stuttering judgment training involved use of either the Stuttering Measurement System (SMS), Stuttering Measurement Assessment and Training (SMAAT) programs, or no training. To test for the reliability of any training effect, SMS training was repeated with the 4th group.

Results: Both SMS-trained groups produced approximately 34% improvement, significantly better than no training or the SMAAT program. The SMAAT program produced a mixed result.

Conclusions: The SMS program was shown to produce a “medium” effect size improvement in the accuracy of stuttering event counts, and this improvement was almost perfectly replicated in a 2nd group. Half of the SMAAT judges produced a 36% improvement in accuracy, but the other half showed no improvement. Additional studies are needed to demonstrate the durability of the reported improvements, but these positive effects justify the importance of stuttering measurement training.

Wendell Johnson and his colleagues were the first to quantify the speech of persons who stutter (PWS) in terms of the frequency of “moments of stuttering” (Johnson & Knott, 1936). Indeed, despite Johnson’s (1948) confusing assertion that “stuttering is what the speaker does while trying not to stutter again” (p. 240), the “moment of stuttering” (or a “stuttering event”) became the dependent variable that ultimately supported an enormous body of stuttering research and treatment. However, although research based on this measure has continued to flourish, the reliability with which this perceptually based dependent variable could be identified or measured has remained problematic (Ingham, 1990).

Cordes and Ingham (1994a) critically reviewed studies assessing the reliability with which moments of stuttering or stuttering events are measured. They noted that at that time most current measurement systems favored

quantification of overt speech behaviors rather than emotional or physiological aspects of the disorder. Consequently, stuttering measurement methods rely heavily on the judgments of human observers; some definitions of stuttering, in fact, define the disorder and its behavioral characteristics almost solely in terms of human observers’ judgments. . . . Obviously, then, it is as critical to stuttering measurement as it is to other direct behavior observation methods that human observers provide reliable data; direct-observation recording systems cannot function as sources of valid and reliable data if they are affected by differences among observers. (p. 279)

Those observations came in the wake of a series of studies that showed remarkably poor levels of agreement among total stuttering event counts made by experienced judges from different laboratories (Kully & Boberg, 1988), from different academic training programs (Ingham & Cordes, 1992), and by prominent stuttering treatment investigators (Cordes & Ingham, 1995). In fact, it was a situation that had remained virtually unchanged since 1940, when Tuthill demonstrated huge levels of disagreement among judges when they identified events perceived as “stuttering”

^aUniversity of California, Santa Barbara

Correspondence to Roger J. Ingham: rjingham@speech.ucsb.edu

Editor: Jody Kreiman

Associate Editor: Hans-Georg Bosshardt

Received July 24, 2014

Revision received October 27, 2014

Accepted January 16, 2015

DOI: 10.1044/2015_JSLHR-S-14-0200

Disclosure: The authors have declared that no competing interests existed at the time of publication.

during speech samples. Given the implications of these findings, there seemed to be very little concern within the treatment or research community to find a resolution—at least until the 1990s.

Measurement validity and measurement reliability are inextricably connected components of measurement (Cronbach, 1951). Thus, one creative approach toward solving the stuttering measurement problem was introduced in the 1980s by Perkins (1983, 1984, 1990; Moore & Perkins, 1990). He argued that the PWS rather than an observer is in the best position to make a valid judgment as to the occurrence of a stuttering event. What distinguishes that event, he argued, is a sensation that the speaker experiences during a “real” moment of stuttering. That experience is a sense of loss of control over the “ability to move forward in the execution of linguistically formulated speech” (Perkins, 1984, p. 431), and it is a sensation that only the speaker—and not an observer—can perceive. Unhappily, the shift from an observer-judged event to a self-judged sensation only served to exacerbate the stuttering measurement reliability problem. This was exemplified in a study by Martin and Haroldson (1986), who reported vivid instances of events that observers clearly recognized to be stuttered but were not speaker-judged to be accompanied by a sensation of “loss of control.” The stuttering measurement problem had additional implications: It highlighted the embarrassing absence of standardized training systems for teaching students, clinicians, and researchers to measure stuttering events in real time, or from recordings, with adequate accuracy and reliability. That problem persists and continues to be understudied. Perhaps the most problematic example of this issue pertains to the widely used Stuttering Severity Instrument—Third Edition (Riley, 1994) and Stuttering Severity Instrument—Fourth Edition (Riley, 2009). It is puzzling, for example, to find that these instruments are employed without requiring clinicians or judges to be trained to judge stuttering events.

Beginning in the early 1990s, a series of studies initiated at the University of California, Santa Barbara, led to the development of two computerized programs that were designed to improve the reliability of stuttering event measurement and to train clinicians to measure stuttering in real time or from recordings. The first of these was the Stuttering Measurement System (SMS; Ingham, Bakker, Ingham, Kilgo, & Moglia, 1999). This computer-based program actually evolved from an earlier videotape-recorded program (Fowler & Ingham, 1986; Ingham & Ingham, 1987). It accompanies a training program that is composed of 1- to 3-min audiovisual recordings of children and adults who stutter. Transcripts of the samples were marked to show within-, between-, and across-word stuttering events plus syllable counts and an overall speech-naturalness rating using the 9-point scale designed by Martin, Haroldson, and Triden (1984). The original stuttering events, syllable markings, and speech naturalness ratings were made independently by five highly experienced judges. Subsequently, five additional and equally experienced judges repeated these judgments—also independently. Transcript comparisons provided stuttering event markings across judges that

agreed on 80% or more instances. Therefore, only those instances that were judged as stuttered or not stuttered with at least 80% (eight of 10 judges) agreement constituted the master transcript record.

The second computerized program is known as the Stuttering Measurement Assessment and Training (SMAAT; Ingham, Cordes, Kilgo, & Moglia, 1998). SMAAT evolved from a series of studies on the use of interval-based judgments of stuttered speech as a way to improve agreement among judges in identifying stuttering events in connected speech (Cordes & Ingham, 1994a, 1994b, 1995, 1999; Cordes, Ingham, Frank, & Ingham, 1992; Ingham & Cordes, 1997). The SMAAT program requires judges to identify stuttering events from audiovisual recordings by pressing a mouse button for the duration of each moment of stuttering. All recordings had been previously subdivided into 5-s intervals. Recognized authorities on stuttering had then independently judged a randomized presentation (on videodisc) of all intervals as either stuttered (including a stuttering event) or not stuttered with greater than 80% agreement among the judgments. Those with less than 80% agreement were classified as disagreed or ambiguous. Thus, a judge’s mouse button press while observing a sample would be automatically recognized within a nonvisible record of superimposed 5-s “stuttered” or “not stuttered” intervals. During training sessions, an incorrect judgment (button press) would stop the playing of the sample and the 5-s interval involved would be replayed and labeled as either “stuttered” or “not stuttered.” The efficacy of the SMAAT training program in improving stuttering judgments was investigated by Cordes and Ingham (1999). They demonstrated that across two groups ($n = 10$ per group) of essentially naive judges, training did improve a number of factors, including “correspondence between interval-based analyses of the judges’ stuttering judgments and the previously developed standards” (1999, p. 862) as well as intra-judge and interjudge agreement. However, the study was partially confounded by a “ceiling effect,” that is, almost all judges achieved at least 80% correct interval judgments without training (thereby also demonstrating the accuracy of interval-based stuttering judgments). Furthermore, it was not clear from this study if the SMAAT program would also improve the accuracy of stuttering event counts within a sample. Subsequently, the SMAAT program has been used with samples from Icelandic preschool children who stutter to assist clinicians identify stuttering events by preschool children who stutter (Einarsdóttir & Ingham, 2008). The program’s principles have also been used to evaluate parents’ judgments of children who stutter (Einarsdóttir & Ingham, 2009).

The aim of the present study was to comparatively evaluate the efficacy of the SMS and SMAAT programs in improving stuttering event count judgments in real time. In addition, this study aimed to compare the effects of these programs with a “no training” control group. A pretraining–posttraining study using participants’ change scores (from pre- to posttraining) was designed to answer the following four questions:

1. Are the mean change scores from two SMS groups significantly different?
2. Are the mean change scores from the SMS groups and the SMAAT group significantly different?
3. Are the mean change scores from the SMS groups significantly different from the mean change scores in the control group?
4. Are the mean change scores from the SMAAT group significantly different from the mean change scores in the control group?

Method

Participants

Four groups of undergraduate students (19 to 23 years old) were paid volunteer participants in this study. Each group contained 12 participants ($n = 44$ women, four men) who were majoring in biology, psychology, linguistics, or communication studies at the University of California, Santa Barbara. None had any previous experience in judging stuttering; they represented the target audience for these training programs of preservice students or early career professionals in need of training. All participants agreed to attend the laboratory where the study was conducted on 3 consecutive days. Any participant who could not fulfill the 3 consecutive day requirement was replaced (this was the case for three judges). Participants completed their experimental trials individually and in a sound-treated booth in R. J. Ingham's laboratory. By their own report, and as demonstrated by their conduct during this experiment, no judge had any hearing, vision, or motor disability that would interfere with the task of watching and listening to recordings of speech and making button-press responses. This study was preapproved by the Human Subjects Committee at the University of California, Santa Barbara.

Materials

All phases of this study used a part of the SMS program. That part (Part 5 of the SMS training program) is composed of eight 3-min audiovisual spontaneous speech samples of PWS (six men, one woman, and one boy). The samples were originally assessed by 10 highly experienced clinicians and researchers specializing in developmental stuttering. Those 10 individuals were provided with a transcript of the samples, and they underlined words or intervals between words they judged were stuttered. No definition was provided; judges made a perceptual judgment that there was a stuttering event. The stuttered syllable or word count for the eight samples was based on words, syllables, or intervals between words where eight of 10 judges agreed there was stuttering.

For the experimental groups, the participants completed sections of either the SMS (Ingham, Bakker, Ingham, Kilgo, & Moglia, 1999) or SMAAT (Ingham et al., 1998)

program.¹ Those sections are designed to train judges to count stuttering events to a criterion level of accuracy.

The stuttering judgment training for the SMS and SMAAT programs differ in some important ways. The SMS Part 2 or stuttering event count training requires the subject to count events using a computer mouse during 15 different speech samples ranging in duration from 1 to 3 min. A successful trial required a total count within 5% of the pre-judged and prescribed number for the sample. No sample or speaker included in SMS Part 2 appeared among the evaluation samples (SMS Part 5) nor among the samples used in the SMAAT program. The judges were given up to five opportunities to complete a successful trial with a sample.

The SMAAT stuttering judgment training involved nine 2-min audiovisually recorded spontaneous speech samples from nine different adults who stutter (seven men and two women). As mentioned above, these samples had been previously divided into 5-s intervals that were classified as either stuttered, nonstuttered, or ambiguous. Consequently, any judge's mouse button press during an interval was immediately judged as "correct" or "incorrect" (if it either did or did not occur during an ambiguous interval, it was judged "correct"). The first minute of each sample was used for training, and the second minute was a probe section used to evaluate training. During the first minute of each 2-min sample, the program immediately pauses and replays the 5-s interval that is prejudged to be "incorrect." The second minute served as a probe sample; incorrect judgments during that sample were not replayed so as to determine if the training on the first minute of a sample generalized to improved judgment skill with the same speaker. The probe sample was replayed and training continued only with samples for which the judge failed to reach the 90% correct criterion. Training ceased when the judge achieved 90% or better correct judgments on all nine 1-min probe samples or failed to do so after five additional attempts.

Neither data recording sheets nor written transcripts were used with the SMAAT training; thus, the judges did not have access to the target or correct judgments unless they were provided with that information in the form of performance-contingent feedback.

The SMS training involved 15 different speech samples (24 min), and the SMAAT training involved nine different samples (18 min). Consequently, the average training time for the two training programs might be expected to differ. That was the case but not in the expected direction; the average time for the SMS group ranged from 40 to 75 min, and for the SMAAT group, it ranged from 60 to 90 min.

Experimental Design

As mentioned above, the SMS and SMAAT training was evaluated by the judge's performance on eight 3-min

¹The SMS program is available as a free download for either a PC or Mac computer at <http://sms.id.ucsb.edu/>. The SMAAT program is available, also freely, by e-mailing R. J. Ingham.

samples. This was also true for the control group. The experimental format used to evaluate the programs is shown in Table 1. With the exception of control group participants, each participant completed the tasks within each phase of the experimental format on 3 consecutive days. In the case of the control group, there was no training task on Day 2 (participants completed tasks only on Days 1 and 3). In order to test for the reliability or consistency of any training effect, SMS training was repeated with a fourth group of judges.

Data Analysis

The basic data used to evaluate this experiment were each judge's pretraining versus posttraining change scores. These were derived from each judge's absolute deviation scores, relative to the sample's stutter-count target range, for each of the eight pretest–posttest samples. This derivation is illustrated in Figure 1 and described below. Figure 1 shows that for one particular sample, the stutter-count target range was 78–86 stutters. This range incorporates the mean stutter count of the 10 experienced judges. In this case, their mean score was 82; thus, $\pm 5\%$ of that score (rounded to the nearest whole number) then formed the target range (78–86) for that sample. Thus, Figure 1 shows that on Day 1, Judge 1 counted 72 stutters on the sample, which was 6 units outside of the target range, whereas Judge 2 counted 90 stutters on that sample, which was 4 units outside of the target range. All units were treated as positive numbers so as to record judgment accuracy with respect to the target range; a deviation score of 0 would mean that the judge's count was within the range and considered to be accurate. In the example provided in Figure 1, an improvement in judgment accuracy occurred on Day 3 when Judge 1 counted 88 stutters because that was only 2 units beyond the target range, an improvement by 4 units from the deviation score for Day 1. Therefore, Judge 1's change score on the sample was 4. However, Judge 2 showed a deterioration in accuracy after training that produced a change score of -5 . The data from all participants were, therefore, suitable for change-score analyses. A positive change score meant that there had been an improvement in judgment accuracy from Day 1 to Day 3, whereas a negative change score meant that judgment accuracy had deteriorated.

Table 1. Experimental design.

Group	Day 1	Day 2	Day 3
SMS 1	Criterion test	SMS training	Criterion test
SMS 2	Criterion test	SMS training	Criterion test
SMAAT	Criterion test	SMAAT training	Criterion test
Control	Criterion test	No training	Criterion test

Note. For all groups, $n = 12$. SMS = Stuttering Measurement System; SMAAT = Stuttering Measurement Assessment and Training.

Statistical Analysis

The two-sample Hotelling's T^2 test (Winer, Brown, & Michels, 1991) was applied to investigate differences in the vector of mean change scores from eight samples between two generic groups. Specifically, let $X_1, \dots, X_m$ be change scores from eight samples for each of the m judges in the first group, and let $Y_1, \dots, Y_n$ be change scores from eight samples for each of the n judges in the second group. We assume that $X_1, \dots, X_m$ are independent, identically distributed multivariate normal with mean μ_1 and covariance matrix Σ_1 , $Y_1, \dots, Y_n$ are independent, identically distributed multivariate normal with mean μ_2 and covariance matrix Σ_2 , and X and Y samples are mutually independent. Hotelling's T^2 was applied to test the hypothesis that $\mu_1 = \mu_2$, that is, mean change scores between the two groups are the same for each of the eight samples. If this overall test is significant at the 5% level, the difference in change scores between the two groups was investigated for each sample using the t test. The normal distribution was checked by normal quantile–quantile plots, and there was no sign of violation. The following combination of groups were investigated: (a) SMS 1 versus SMS 2, (b) SMS versus SMAAT, (c) SMS versus control, and (d) SMAAT versus control.

Results

Mean absolute error scores (stutter counts outside of the target range), with their standard deviations, are shown for each group in Table 2 and illustrated in Figure 2. These data show that the two SMS training groups, SMS 1 and SMS 2, displayed a similar level of improvement (approximately 35%) in the accuracy of their stutter counts (reduced error scores or stutter counts outside of the target range) across the eight 3-min samples that were used to evaluate the training. By contrast, there was very little improvement in the mean accuracy of total stutter counts from pretraining (Day 1) to posttraining (Day 3) for judgments by the SMAAT-trained group or the control group. The mean for each of the eight samples is shown in Table 2. These scores also provided the change scores on each sample for each participant and were used in all subsequent analyses that addressed this study's four questions.

Hotelling's T^2 test was employed to test the hypothesis that there is no significant difference between the groups' change scores. If this overall test was significant at the 5% level, the difference between the two groups was investigated for each sample using the t test. Four questions were addressed systematically by the data analyses. They are listed below along with the findings.

Question 1: Are the mean change scores from the two SMS groups significantly different? Hotelling's T^2 test yields $F = 1.190$ with null distribution $F(8, 15)$, $p = .367$. It was concluded that there was no difference between the SMS 1 and SMS 2 groups. Consequently, they were combined for the following questions.

Question 2: Are the mean change scores from the SMS groups and the SMAAT group significantly different?

Figure 1. An illustration of how the change scores were calculated in the present study. Stutter-count scores are displayed for two of the 12 judges during pretraining and posttraining on one of the eight speech samples used for evaluating the effects of training. The predetermined target range for that sample was 78–86 stutter counts. During pretraining, Judge 1 counted 72 stuttering events, which is 6 units below the target range, but after training, she counted 88 events, which is 2 units outside of the target range. Therefore, this judge improved by 4 units, which was the change score used in the data analysis. Judge 2 counted 90 events during pretraining, which is 4 units outside of the target range, but during posttraining, she counted 69 events, which is 9 units outside of the target range. Therefore, this judge's performance deteriorated by 5 units. Hence, a –5 change score was entered into the data analysis for this judge.

Table 2. Means and standard deviations of stuttering event count errors (counts outside target range) for the four groups before (Day 1) and after training (Day 3); the control group received no training. The mean percentage change scores are also shown.

Group	Sample no.	Day 1		Day 3		M change	M % change
		M	SD	M	SD		
SMS 1	1	20.58	13.43	19.25	10.28	1.33	6.48
	2	5.75	2.09	2.42	2.15	3.33	57.97
	3	14.17	7.48	7.92	5.68	6.25	44.12
	4	21.08	9.89	13.58	7.22	7.50	35.57
	5	9.67	10.15	4.67	5.90	5.00	51.72
	6	2.08	1.88	2.67	2.61	–0.58	–28.00
	7	9.58	8.99	1.33	2.10	8.25	86.09
	8	17.83	10.08	13.33	10.25	4.50	25.23
SMS 2	Mean	12.59	6.98	8.15	6.56	4.45	35.26
	1	30.58	16.36	22.17	11.58	8.42	27.52
	2	6.67	3.52	3.17	2.95	3.50	52.50
	3	17.25	9.23	11.33	7.13	5.92	34.30
	4	22.92	8.01	12.25	7.10	10.67	46.55
	5	12.75	13.86	8.50	7.00	4.25	33.33
	6	6.08	9.51	2.42	4.21	3.67	60.27
	7	11.42	10.60	4.83	7.79	6.58	57.66
SMAAT	8	18.58	12.72	18.17	7.94	0.42	2.24
	Mean	15.78	8.33	10.35	7.11	5.43	34.41
	1	28.67	13.99	24.50	17.41	4.17	14.53
	2	6.33	3.17	5.58	3.03	0.75	11.84
	3	19.50	7.93	15.33	10.29	4.17	21.37
	4	29.83	6.10	26.83	7.15	3.00	10.06
	5	15.50	6.65	14.75	7.96	0.75	4.84
	6	2.08	2.11	3.83	3.90	–1.75	–28.00
Control	7	6.33	6.77	8.75	8.51	–2.42	–38.16
	8	24.00	8.89	22.17	12.10	1.83	7.64
	Mean	16.53	10.73	15.22	8.73	1.31	7.92
	1	39.17	19.82	33.50	23.36	5.67	14.47
	2	6.33	3.80	6.25	3.31	0.08	1.32
	3	18.83	8.19	19.42	8.47	0.58	–3.10
	4	24.67	10.50	26.17	11.23	1.50	–6.08
	5	18.75	10.32	17.58	11.80	1.17	6.22
	6	3.17	2.12	3.75	2.42	–0.58	–18.42
	7	10.75	7.70	9.50	6.96	1.25	11.63
	8	26.67	13.24	26.08	13.11	0.58	2.19
	Mean	18.54	11.80	17.78	10.61	0.76	4.09

Figure 2. Mean pre- and posttraining incorrect stutter counts (plus one standard deviation) for each group relative to the target range across the eight samples. For each judge in each group, the pretraining assessment on the eight samples occurred on Day 1, and posttraining assessment occurred on Day 3 with the same eight samples. Subsequent data analyses showed that there was no difference between the amounts of improvement in judgment accuracy after training (Trg) by the two different SMS-trained groups. Their accuracy was also improved significantly more than the SMAAT and no training groups' accuracy. The accuracy of either of the latter two groups' stutter-count scores was not significantly improved from Day 1 to Day 3. SMS = Stuttering Measurement System; SMAAT = Stuttering Measurement Assessment and Training.

Hotelling's T^2 test yields $F = 2.733$ with null distribution $F(8, 27)$, $p = .024$. Therefore, it was concluded that there is a significant difference between the SMS and SMAAT groups' mean change scores. The overall effect size ($d = .670$) was considered to be of "medium" magnitude (Cohen, 1988). A t test was applied to compare difference in mean change scores between the SMS and SMAAT for each of eight samples (S1–S8) and produced the following results: S1: $t = 0.116$ with null distribution $t(20.020)$, $p = .909$; S2: $t = 2.467$ ($df = 22.869$), $p = .022$; S3: $t = 0.558$ ($df = 20.657$), $p = .583$; S4: $t = 1.965$ ($df = 28.961$), $p = .059$; S5: $t = 1.082$ ($df = 25.201$), $p = .290$; S6: $t = 1.586$ ($df = 33.674$), $p = .122$; S7: $t = 2.799$ ($df = 26.715$), $p = .009$; S8: $t = 0.158$ ($df = 27.997$), $p = .877$. In other words, SMS was determined to be more effective than SMAAT in producing significantly improved judgment accuracy for S2 and S7.

Question 3: Are the mean change scores from the SMS groups significantly different from the mean change scores in the control group? Hotelling's T^2 test yields $F = 3.115$ with null distribution $F(8, 27)$, $p = .013$. Therefore, it was concluded that the SMS training program does produce changes greater than those that would occur without training. In order to locate which particular samples were significantly improved by SMS training, t tests were applied once more and produced the following results: S1: $t = -0.144$ ($df = 23.101$), $p = .886$; S2: $t = 3.373$ ($df = 26.091$), $p = .002$; S3: $t = 2.977$ ($df = 33.868$), $p = .005$; S4: $t = 2.715$ ($df = 20.788$), $p = .013$; S5: $t = 1.159$ ($df = 32.517$), $p = .255$; S6: $t = 1.162$ ($df = 31.359$), $p = .254$; S7: $t = 1.873$ ($df = 29.434$), $p = .071$; S8: $t = 0.525$ ($df = 32.417$), $p = .603$. Thus, SMS training

was determined to be more effective than no training in producing significantly improved judgment accuracy for samples S2, S3, and S4.

Question 4: Are the mean change scores from the SMAAT group significantly different from the mean change scores in the control group? Hotelling's T^2 test yields $F = 0.722$ with null distribution $F(8, 15)$, $p = .654$. Hence, it was concluded that SMAAT did not produce a significant improvement in stutter-count accuracy relative to no training.

One issue that could undermine the validity of the findings is the effect of differences between the groups' pretraining stutter-count accuracy. Examination of Table 2 shows that there were some rather obvious differences among the Day 1 means that might confound the findings. If it is concluded that the SMS groups were significantly more accurate than the SMAAT or control groups, then this could simply be a result of the SMS groups being relatively more accurate before training. This seems to be the case when the Day 1 SMS 1 group's mean stuttering event count errors (counts outside target range) is compared with the SMAAT group's mean stuttering event count errors. However, this confounding issue is avoided, rather fortuitously, by the presence of the SMS 2 group data. A comparison between the Day 1 accuracy scores for the SMS 2 and SMAAT groups across all eight speech samples showed there was no significant difference between their Day 1 scores. It follows, therefore, that the findings favoring the SMS program over the SMAAT program could not be attributed to pretraining levels of accuracy.

Finally, the accuracy of the absolute stutter counts per group was checked by determining how many judges made counts that were within the target ranges for the samples. The results for the four groups before and after training across the eight speech samples are shown in the Appendix. This shows that the counts within target range data still favored the SMS program. After training, both SMS groups more than doubled the number of judges who had total stutter counts on samples that were within the target range. There was essentially no change in target range stutter counts for the SMAAT or control groups. Hence, these findings serve to complement the pattern of results reported above.

Discussion

The findings of this study appear to be quite straightforward, but they may also be somewhat misleading. It certainly appears as though the SMS training program (albeit, the program's stuttering event training module only) is superior to the SMAAT training program in improving the accuracy of stutter judgment counts. It also appears that SMS's stuttering event judgment training is superior to the absence of training. However, there are at least four important points to take into account when interpreting the results of this study. The first is that these were judgments of the accuracy of the total stuttering event counts; they do not necessarily reflect the accuracy with which individual

stuttering events were identified. The second is that these findings do not indicate if the judges achieved sustained improvements in their stuttering event count accuracy; that is obviously an important matter for further research. It should also be noted that, in practice, a judge's stuttering judgment training could continue until a required "satisfactory" level of judgment accuracy is achieved. The third is that these are group data and do not identify individual differences among the judges in their response to training or the judges who failed to respond to training. There were, however, some potentially important differences between the variety of samples and amount of training time for SMS and SMAAT training (see Method), differences that may have produced differential training effects.

The fourth point is related to the third but carries some more complex implications. The current findings not only suggest that the SMAAT training program was ineffectual, but they are also in direct conflict with the results of a number of previous studies showing that the SMAAT training program does improve judgment performance (Cordes & Ingham, 1996, 1999; Einarsdóttir & Ingham, 2008). In some important respects, the present study highlights the classic problem of translating the results of a group study to the behaviors of individuals or, in this case, the judgment behavior of individuals. Inspection of the change scores for the 12 judges who received the SMAAT program shows that they could be essentially divided into two very different subgroups. One subgroup of six judges produced mean pre- and posttraining accuracy scores of 16.19 and 10.36, respectively, whereas the other six judges produced mean pre- and posttraining scores of 14.02 and 17.40, respectively. In other words, the first subgroup had a 36.61% improvement (remarkably close to the SMS groups' 34.84% improvement), whereas the second subgroup had a 24.11% deterioration in judgment accuracy. At the very least, this bimodal response within the SMAAT group suggests that the training was not completely ineffective as would be suggested by the group data. It is impossible to know just why one subgroup performed so poorly, but it was obvious to the observing research assistants that the SMAAT program appeared to be a more challenging task for the SMAAT-trained judges than the SMS program was for the SMS-trained judges. It is not surprising, perhaps, that having to learn to identify individual moments of stuttering is also more difficult than learning to count a total number of events rather than specific stuttering events.

It is important to note that neither the SMS nor the SMAAT programs draw attention to different types of stuttering events. They are primarily focused on perceptual judgments and presume that the use of exemplars will assist judges to use what Martin and Haroldson (1981) described as perceptual threshold identification. It remains an empirical question as to whether guiding judges' attention toward exemplars of self-judged moments of stuttering or perhaps parent-identified stuttering events would have improved their judgment skills.

One additional feature of the present study was the inclusion of an additional SMS training group. The original

design of this study was to try to replicate any gains produced by training with a second and independent group. This proved to be fortuitous because it clearly demonstrated the reliability and/or replicability of the training effect produced by SMS training. Indeed, the similarities between the treatment effects achieved by both SMS-trained groups—approximately 34%—is a surprisingly impressive finding. A second SMAAT training group was not employed because the group results suggested that the training had not been effective. Ultimately, however, there may well be considerable value in replicating the SMAAT training section of the study because it might elucidate the rather puzzling bimodal effect described above.

Finally, the data analysis on the eight speech samples used to evaluate the effects of training revealed some useful, albeit only partially useful, findings. They showed that not more than four of eight samples were significantly associated with the SMS training effect. It is not clear exactly how that finding should be interpreted and implemented. It is common in evaluation procedures to use some questions or tasks that are not functionally important. It could be argued that this may reduce the demands of the overall task and make it somewhat more user-friendly. Conversely, it could suggest that other more "functional" samples would help differentiate individual training effects. Both options have benefits that might be explored in later studies of this stuttering measurement training instrument.²

The present findings of this essentially exploratory study strongly suggest that the SMS's stuttering judgment training section does make a functional contribution to improving the accuracy of stuttering event measurement. That is not a trivial finding because it illustrates the contribution that an exemplar-based approach to stuttering measurement can make. For that reason, this also appears to be a very important contribution toward training students, clinicians, and researchers to measure the frequency of stuttering events with reference to a standard. In that respect, it is interesting to take note of the conclusions reached by Jani, Huckvale, and Howell (2013, p. 860), who cite the SMS program in their recent review as essentially the only program that advocates procedures that should underpin stuttering measurement:

Stuttering assessment instruments should provide training packages that include practice samples with baseline syllables, stuttered syllables, and duration counts for comparing results and monitoring progress according to different procedures. (Ingham, Bakker, Ingham, Kilgo, & Moglia, 1999)

²An important consideration in evaluating the findings of this study concerns their external validity. When used clinically or in a customary manner, rating with the SMS program requires the judge not only to count stutter events during a speech sample (including the eight used in the present study) but to concurrently count syllables in real time and make speech naturalness ratings at predetermined intervals. This may mean that in practice the accuracy of stuttering event counts would have been affected by the SMS's multitask requirements.

The present study serves to validate the efficacy of at least part of that program's training package. The clinical significance of these findings is that they demonstrate that stuttering judgments, so necessary for delivering and evaluating many stuttering treatments, are able to be improved by exemplar-based training. The SMS procedure may also provide a standard method that researchers can use as a basis for establishing the credibility of perceptually based measures of stuttering within their research.

Acknowledgments

The study was supported by National Institute on Deafness and Other Communication Disorders Grant DC007893, awarded to R. J. Ingham, and a grant from the Undergraduate Research Committee at the University of California, Santa Barbara, awarded to C. Stavros and M. Ebrahimián. Special thanks are offered to Anne Bothe for her comments and edits on earlier versions of this article. We also thank Dan Orias for assistance with various parts of this study.

References

- Cohen, J.** (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cordes, A. K., & Ingham, R. J.** (1994a). The reliability of observational data: II. Issues in the identification and measurement of stuttering events. *Journal of Speech and Hearing Research*, 37, 279–294.
- Cordes, A. K., & Ingham, R. J.** (1994b). Time interval measurement of stuttering: Effects of training with highly agreed or poorly agreed exemplars. *Journal of Speech and Hearing Research*, 37, 1295–1307.
- Cordes, A. K., & Ingham, R. J.** (1995). Judgments of stuttered and nonstuttered intervals by recognized authorities in stuttering research. *Journal of Speech and Hearing Research*, 38, 33–41.
- Cordes, A. K., & Ingham, R. J.** (1996). Time-interval measurement of stuttering: Establishing and modifying judgment accuracy. *Journal of Speech and Hearing Research*, 39, 298–310.
- Cordes, A. K., & Ingham, R. J.** (1999). Effects of time-interval judgment training on real-time measurement of stuttering. *Journal of Speech, Language, and Hearing Research*, 42, 862–879.
- Cordes, A. K., Ingham, R. J., Frank, P., & Ingham, J. C.** (1992). Time interval analysis of interjudge and intrajudge agreement for stuttering event judgments. *Journal of Speech and Hearing Research*, 35, 495–511.
- Cronbach, L.** (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Einarsdóttir, J., & Ingham, R. J.** (2008). The effect of stuttering measurement training on judging stuttering occurrence in pre-school children who stutter. *Journal of Fluency Disorders*, 33, 167–179.
- Einarsdóttir, J., & Ingham, R. J.** (2009). Accuracy of parent identification of stuttering occurrence. *International Journal of Language and Communication Disorders*, 44, 847–863.
- Fowler, S. C., & Ingham, R. J.** (1986). Stuttering Treatment Rating Recorder [STRR software]. Retrieved July 16, 2005, from <http://www.speech.ucsb.edu/>
- Ingham, J. C., & Ingham, R. J.** (1987). Stuttering Measurement Training [Instructional video and manual]. University of California, Santa Barbara.
- Ingham, R. J.** (1990). Commentary on Perkins (1990) and Moore and Perkins (1990): On the valid role of reliability in identifying “what is stuttering?” *Journal of Speech and Hearing Disorders*, 55, 394–397.
- Ingham, R. J., Bakker, K., Ingham, J. C., Kilgo, M., & Moglia, R.** (1999). Stuttering Measurement System [SMS software]. Retrieved from <http://sms.id.ucsb.edu/>
- Ingham, R. J., & Cordes, A. K.** (1992). Interclinic differences in stuttering event counts. *Journal of Fluency Disorders*, 17, 171–176.
- Ingham, R. J., & Cordes, A. K.** (1997). Identifying the authoritative judgments of stuttering: Comparisons of self-judgments and observer judgments. *Journal of Speech, Language, and Hearing Research*, 40, 581–594.
- Ingham, R. J., Cordes, A. K., Kilgo, M., & Moglia, R.** (1998). Stuttering Measurement Assessment and Training (SMAAT) [Software and videodisc]. University of California, Santa Barbara.
- Jani, L., Huckvale, M., & Howell, P.** (2013). Procedures used for assessment of stuttering frequency and stuttering duration. *Clinical Linguistics & Phonetics*, 27, 853–861.
- Johnson, W.** (1948). Stuttering. In W. Johnson, S. F. Brown, J. F. Curtis, C. W. Edney, & J. Keaster (Eds.), *Speech handicapped school children* (pp. 229–329). New York, NY: Harper & Row.
- Johnson, W., & Knott, J. R.** (1936). The moment of stuttering. *Journal of Genetic Psychology*, 48, 475–479.
- Kully, D., & Boberg, E.** (1988). An investigation of interclinic agreement in the identification of fluent and stuttered syllables. *Journal of Fluency Disorders*, 13, 309–318.
- Martin, R. R., & Haroldson, S. K.** (1981). Stuttering identification: Standard definition and moment of stuttering. *Journal of Speech and Hearing Research*, 24, 59–63.
- Martin, R. R., & Haroldson, S. K.** (1986). Stuttering as involuntary loss of speech control: Barking up a new tree. *Journal of Speech and Hearing Disorders*, 51, 187–190.
- Martin, R. R., Haroldson, S. K., & Triden, K. A.** (1984). Stuttering and speech naturalness. *Journal of Speech and Hearing Disorders*, 49, 53–58.
- Moore, S. E., & Perkins, W. H.** (1990). Validity and reliability of judgments of authentic and simulated stuttering. *Journal of Speech and Hearing Disorders*, 55, 383–391.
- Perkins, W. H.** (1983). Onset of stuttering: The case of the missing block. In D. Prins & R. J. Ingham (Eds.), *Treatment of stuttering in early childhood: Methods and issues* (pp. 1–20). San Diego, CA: College-Hill Press.
- Perkins, W. H.** (1984). Stuttering as a categorical event: Barking up the wrong tree—Reply to Wingate. *Journal of Speech and Hearing Disorders*, 49, 431–434.
- Perkins, W. H.** (1990). What is stuttering? *Journal of Speech and Hearing Disorders*, 55, 370–382.
- Riley, G. D.** (1994). *Stuttering Severity Instrument for Children and Adults—Third Edition*. Austin, TX: Pro-Ed.
- Riley, G. D.** (2009). *Stuttering Severity Instrument—Fourth Edition*. Austin, TX: Pro-Ed.
- Tuthill, C. E.** (1940). A quantitative study of extensional meaning with special reference to stuttering. *Journal of Speech Disorders*, 5, 189–191.
- Winer, B. J., Brown, D. R., & Michels, K. M.** (1991). *Statistical principles in experimental design* (3rd ed.). New York, NY: McGraw-Hill.

Appendix

Results for the Four Groups Before and After Training Across the Eight Speech Samples

This table shows the mean stutter event counts per group for each of the eight speech samples during pre- and posttraining. Also shown are the number of participants (or judges) whose stutter counts per sample were within each sample's stutter-count target range (TR). These scores complement the findings based on the error counts shown in Table 2. SMS = Stuttering Measurement System; SMAAT = Stuttering Measurement Assessment and Training.

Group	Sample no.	Pretraining mean count	No. judges in TR (max = 12)	Posttraining mean count	No. judges in TR (max = 12)	TR
SMS 1	1	62.4	1	59.3	1	78–86
	2	6.3	0	9.7	2	12–14
	3	18.9	1	24.0	2	31–35
	4	21.9	0	29.6	1	43–47
	5	39.3	2	39.4	4	44–48
	6	11.2	3	11.4	4	10–12
	7	44.9	0	39.8	7	36–40
	8	27.8	1	32.7	0	45–49
	<i>n</i> =		8		21	
SMS 2	1	51.4	1	55.8	0	78–86
	2	9.5	0	10.9	4	12–14
	3	22.1	0	22.7	1	31–35
	4	23.3	0	33.1	0	43–47
	5	41.1	2	38.7	1	44–48
	6	14.6	2	12.6	6	10–12
	7	40.6	2	40.5	5	36–40
	8	29.7	1	30.0	0	45–49
	<i>n</i> =		8		17	
SMAAT	1	49.4	1	61.7	0	78–86
	2	5.6	0	6.5	1	12–14
	3	11.5	0	16.3	0	31–35
	4	13.2	0	16.2	0	43–47
	5	28.5	0	33.3	1	44–48
	6	8.7	4	11.4	4	10–12
	7	29.9	3	28.9	2	36–40
	8	21.0	0	22.9	1	45–49
	<i>n</i> =		8		9	
Control	1	38.8	0	45.7	0	78–86
	2	5.7	0	5.8	0	12–14
	3	12.4	1	12.1	0	31–35
	4	15.3	0	16.8	1	43–47
	5	26.1	0	29.5	1	44–48
	6	9.2	2	10.1	0	10–12
	7	26.7	2	28.3	1	36–40
	8	18.3	0	19.0	1	45–49
	<i>n</i> =		5		4	