Original Article

Measuring Public Attitudes Toward Stuttering: Test-retest Reliability Revisited

Kenneth O. St. Louis[1], Elizabeth Fisher Aliveto[2], Mohyeddin Teimouri Sangani[3], Salman Abdi[4], Hossein Rezai[5], Fauzia Abdalla[6], Aneta Przepiórka[7], Agata Błachnio[7], Katarzyna Węsierska[8], Lejla Junuzović-Žunić[9], Sarah Eisert[10], Chelsea Kuhn Roche[11], Isabella Reichel[12], Ann Beste-Guldborg[13], Timothy Flynn[14], Benjamin Bolton-Grant[15], Sheryl Gottwald[16], Sara Spears[17], Daniel Hudock[18], Jessica Hartley[19], Mariswamy Pushpavathi[20], Kia N. Johnson[21]

[1]Department of Communication Sciences & Disorders, West Virginia University, Morgantown, West Virginia; [2]Speech-Language Therapy, Loudoun County Public Schools, Ashburn, Virginia, United States; [3]Department of Speech Therapy, School of Rehabilitation, Mashhad University of Medical Sciences, Mashhad; [4]Neuromuscular Rehabilitation Research Center, Semnan University of Medical Sciences, Semnan; [5]Department of Speech Therapy, Faculty of Rehabilitation Sciences, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran; [6]Department of Communication Disorders Sciences, Kuwait University, Kuwait City, Kuwait; [7]Institute of Psychology, The John Paul II Catholic University of Lublin, Lublin; [8]Department of Humanities, Institute of Linguistics, University of Silesia, Katowice, Poland; [9]Department for Speech and Language Pathology, University of Tuzla, Tuzla, Bosnia & Herzegovina; [10]Children's Therapy of Woodinville, Woodinville, Washington; [11]Advanced Dysphagia Diagnostics LLC, Grand Junction, Colorado; [12]School of Health Sciences, Touro University, New York City, New York, United States; [13]Department of Communication Sciences & Disorders, Minot State University, Minot, North Dakota; [14]Arlington Public Schools, Arlington, Virginia, United States; [15]Department of Speech and Language Sciences, Leeds Beckett University, Leeds; also Leeds Community Health Care, United Kingdom; [16]Department of Communication Sciences and Disorders, University of New Hampshire, Durham, New Hampshire; [17]Speech and Language Services, Oceanside Unified School District, Oceanside, California; [18]Department of Communication Sciences and Disorders, Idaho State University, Pocatello, Idaho; [19]Boothby Therapy Services, Laconia, New Hampshire, United States; [20]Department of Audiology and Speech Language Pathology, All India Institute of Speech and Hearing (AIISH), Mysuru, India; [21]Arthur M. Blank Center for Stuttering Education and Research, Atlanta, Georgia, United States

Purpose: Previous studies of the *Public Opinion Survey of Human Attributes-Stuttering* (*PO-SHA-S*), using test and retest designs in modest-sized samples, have reported satisfactory test-retest reliability, i.e., correlations of about 0.80. Simultaneously, lower but moderate correlations between different first and second test respondents were observed and hypothesized to represent unspecified "societal" influences on stuttering attitudes. This study sought to clarify this and other potential relationships between first and second tests with the *PO-SHA-S* in a large, geographically and linguistically diverse sample.

Methods: *POSHA-S* Overall Stuttering Scores (OSSs) of 345 respondents from 12 test-retest samples from four countries and languages, with no intervening interventions, were analyzed with correlations and by grouping respondents according to whose stuttering attitudes improved, remained the same, or worsened from test to retest.

Results: Test and retest OSSs generally conformed to normal distributions and were not significantly different. Correlations between first versus second tests replicated earlier research. However, when the degree and direction of change from test to retest was considered, both in other correlations and in sorts of respondents, unexpected results emerged. Respondents with intermediate attitudes changed minimally, while those with most and least positive attitudes at the first test changed in opposite directions past the overall mean at second test.

Conclusions: While demonstrating adequate test-retest reliability correlations on the *POSHA-S*, public attitudes were found to be less stable than previously assumed.

Keywords: Stuttering, Public attitudes, Reliability, POSHA-S

INTRODUCTION

Psychometric requirements of measures of attitudes

In order to be confident that any scale accurately measures its intended constructs its psychometric properties must be assessed [1]. Among other measures, these include aspects of validity (e.g., face validity, construct validity, and concurrent validity) and internal consistency. A satisfactory scale also must also be reliable, that is, generating consistent results in similar contexts. This study addresses reliability, or more specifically, test-retest reliability.

Sources of variance in test–retest reliability

Revelle and Condon [2] identified four sources of variance in measurements of individuals or groups, i.e., trait, state, specific, and random error. They wrote, "All tests are assumed to measure something stable over time (trait like), something that varies over time (reflecting the current state), some specific variance that is stable but does not measure our trait of interest, and some residual, random error" (p. 4). In other words, state variance refers to changes in one's responses based on moment-to-moment or short-term influences, while trait variance identifies one's long-term changes in responses. Specific and error variance components are latent constructs that cannot be observed. Revelle and Condon give the following example of a test item, "I enjoy a lively party." One's response is an unknown mixture of extraversion (trait), positive emotion (state), and wording of the item or how one might interpret "lively" or "party" (specific). All of these are sources of variability in test-retest reliability. Tests, scales, or surveys typically intend to measure traits, or, for the purpose of this paper, the attitudes the public holds toward stuttering and people who stutter.

The standard way to assess test-retest reliability of a self-reported opinion or perception in the social sciences is to administer the scale, survey, or test on two occasions with no intervening influence other than time between the two administrations, but the time not so short that respondents will clearly remember their responses to individual items on the instrument [1]. Ideally, if trait attitude scores remain similar in the same respondents from test to retest, while state and specific components are kept to a minimum, moderate to high test-retest correlations indicate that test-retest reliability is satisfactory. While Revelle and Condon [2] regard the single test-retest correlation as insufficient in some cases, this "classic" approach has served to determine which measures of attitudes generate results that can be regarded as "repeatable" in similar environments, but not necessarily valid, which is another important psychometric dimension not considered here.

Another way to assess test-retest reliability is to determine the percentage of agreement of 1st and 2nd responses that are identical, differ by one unit, differ by two units, differ by three units, and so on. The larger the percentages at the identical and small difference categories, the greater the test-retest reliability.

Test–retest reliability of one measure of stuttering attitudes

The focus of this paper is the *Public Opinion Survey of Human Attributes-Stuttering* (*POSHA-S*), a well-documented measure of public attitudes toward stuttering [3]. Forty-five *POSHA-S* items are combined into 11 components, three subscores, and an Overall Stuttering Score (OSS) (which is the mean of two stuttering subscores, that is, Beliefs and Self Reactions regarding people who stutter). The Beliefs subscore is derived from stuttering items that tap judgments that are external to the respondents in that they do not consciously consider their own behavior or emotions on their answers, such as beliefs about causes. Self Reactions require internal judgments of respondents' own real or potential actions, feelings, or personal knowledge, such as whether they would tell a stuttering person to "Slow down" or experience pity. A comparative general section compares one's overall impression, desire to be, amount known, and persons known who (a) stutter, (b) are obese, (c) are mentally ill, (d) are left handed, and (e) are intelligent. Items and components related to obesity and mental illness make up the Obesity/Mental Illness subscore, the purpose of which is to consider stuttering within a comparative disability perspective. Notably, public attitudes toward stuttering are, overall, less positive than attitudes toward obesity but more positive than those for mental illness [4]. The two other non-stuttering "anchor" attributes, intelligence (positive) and left handedness (neutral), are not included in standard *POSHA-S* summary ratings. All scaled ratings are converted to a -100 to +100 scale, with 0 being neutral, and ratings for some items are inverted so that, consistently, higher ratings reflect more positive attitudes while lower ratings reflect more negative attitudes.

The instrument contains a demographic section with questions about age, education, gender, parental and marital status, income, languages known, race, and religion. Additionally, it includes self-ratings for health, abilities, and 12 different life priorities that relate to a variety of personality traits.

The *POSHA-S*, and its earlier experimental version, have

been shown in numerous publications to be user-friendly, internally consistent, valid, reliable, translatable, and unaffected by different modes and settings of administration [5-13].

In terms of test-retest reliability, the *POSHA-S*, and its near-final experimental version (then termed the *POSHA-E2*) had test-retest correlations in the stuttering and general ratings in four different paper and online samples averaged +0.80 (range =+0.69 to +0.86), reflecting an acceptably high correspondence between mean values for tests (1st tests) and retests (2nd tests). Absolute percent agreement in ratings was also satisfactory. A later test-retest reliability investigation in Iran yielded a slightly lower correlation of +0.70 [14].

Table 1 summarizes these results in the stuttering and general sections for the paired, same respondents. The table also summarizes a finding reported by St. Louis [7] and, earlier, by St. Louis et al. [9], that 1st test scores of individual respondents were also correlated by +0.57 (range =+0.39 to +0.71), not with their own 2nd test scores, but with those of other respondents. At the time, the authors surmised that stuttering attitudes were affected by both "individual" and "societal" influences. They explained that, given correlations of about +0.80 for the *same* respondents, if the correlations between *different* respondents had been close to zero, then unique individual attitudes would have been responsible for the +0.80 correlations. If the correlations between *different* respondents had

been as high as +0.80, widely held societal influences would have superseded unique individual influences. The fact that *different* respondent correlations averaged +0.57 and the *same* respondent correlations averaged +0.80 indicated that common societal influences accounted for most of the similarities in 1st versus 2nd administrations (i.e., +0.57) and that individual influences accounted for the difference between the two correlations (i.e., +0.80 minus +0.57). St. Louis et al. [9] and St. Louis [7] concluded that stuttering attitudes are strongly influenced by shared societal values and opinions, sometimes almost as much or even more than unique individual views. To our knowledge, a standard protocol for determining the contributions of societal and individual contributions to standard test-retest reliability scores has not been advanced.

Since the above-mentioned test-retest reliability reports, a few studies have documented some atypical results in a few *POSHA-S* studies. Abdalla and St. Louis [15] found that practicing teachers were unaffected by a video intervention while pre-service education students improved their stuttering attitudes substantially after watching the video. Also, unlike several previous studies that showed improvements in *POSHA-S* measured stuttering attitudes after various interventions, two studies did not. Węsierska et al. [17] found that neither a video nor a presentation on stuttering changed stuttering attitudes of Polish high school or university students, and Kuhn and St.

Table 1. Pearson product-moment correlation coefficients and percent agreement for test-retest reliability comparisons of the *POSHA-S*

Sample	Number	Paired correlations: same		Paired correlations: different		Paired percentages: same		
		General section: correlations	Stuttering section: correlations	General section: mean individual correlations	Stuttering section: mean individual correlations	Percent identical agreement	Percent±1 agreement	Percent±2 agreement
Test-retest *POSHA-E2* paper version: test versus retest[a]	32	+0.83	+0.80	+0.70	+0.53	51	75	86
Test-retest *POSHA-E2* online version: test versus retest[b]	21	+0.86	+0.82	+0.71	+0.60	60	75	87
Test-retest *POSHA-S* paper version: test versus retest[b]	25	+0.85	+0.69	+0.53	+0.39	75	91	99
Test-retest *POSHA-S* online version: test versus retest[b]	27	+0.78	+0.78	+0.64	+0.48	78	95	99+

[a]St. Louis, et al. [9].
[b]St. Louis, et al. [16].
This table was adapted from Table 3 in St. Louis [7].
"Paired: Same" columns compare the same individuals in the standard way. "Paired: Different" columns compare different individuals in an unconventional way.

Louis [18] reported a similar failure to improve the attitudes of American middle school students after watching a commercial video on stuttering. These were not test-retest reliability studies; however, in the Węsierska et al. study, both intervention and non-intervention control groups had 1st versus 2nd OSS means that were essentially identical. Similarly, the 2nd means of the Kuhn and St. Louis [18] intervention sample also showed essentially no change. In an attempt to explain these puzzling findings, individual respondent profiles from 1st to 2nd administrations were explored rather than considering only the means of each administration. Correlations were generated and revealed a surprising negative correlation between respondents' 1st OSS and the difference between their 1st and 2nd OSSs (explained in detail in the Method section). This meant that high scorers on the 1st test were low scorers on the 2nd test, and *vice versa*. The surprising extent of this idiosyncratic pattern led us to wonder if this was simply a characteristic of these Polish and American student samples or was a phenomenon that characterized respondents to the *POSHA-S* in general that could be observed in a larger and more diverse sample.

A search of the literature revealed no previous reports of participants responding differently as individuals than what would be assumed from their mean results. However, a few reports identified different responses based on participants' prior biases. For example, Dholakia and Morwitz [19] showed that a 10-min telephone survey of approximately half of 2,000 customers of a large financial services company were substantially more likely to open accounts or defect to another company one year later than the other half who were not surveyed, even though the survey dealt with neither option, and no further contact occurred. In the realm of attitudes, Boysen and Vogel [20] explored perceived persuasiveness and perceived effectiveness of written descriptions of schizophrenia or alcohol addiction as a function of measured attitudes toward each condition in advance. Sorting college student respondents according to positive versus negative pre-intervention attitudes, the authors found that those attitudes significantly affected perceptions of both persuasiveness and effectiveness of the descriptions. Respondents with positive pre-intervention attitudes held more positive perceptions in both persuasiveness and effectiveness, and those with negative pre-intervention attitudes held more negative perceptions. This, incidentally, was the only study in our careful search that sorted respondents by more versus less positive attitudes from a pre attitude measure rather than by a typical demographic variable, such as sex, age, or education.

The purpose of the current study was to revisit test-retest reliability of the *POSHA-S* a decade later with a large and diverse sample. Research questions addressed were (a) to confirm and better explain the individual and societal correlations using a larger, more diverse sample and (b) to extend the reliability analysis to individual respondent performance in 1st and 2nd tests by comparing those who were stable in attitudes from 1st to 2nd test versus those who improved or worsened. For research question (a), we had two hypotheses. First, we expected that overall test-retest correlations in the larger sample would be similar to previous results with the *POSHA-S*. Second, we hypothesized that the negative correlations between 1st tests and changes between 1st and 2nd tests in the Węsierska et al. [17] and Kuhn and St. Louis [18] samples were uncharacteristic and would not be observed in the larger sample. For research question (b) we hypothesized that any changes—or lack of changes—for the large majority of individual respondents would align with the means for the test and retest conditions.

METHODS

Respondents

The samples evaluated in this study were taken from the *POSHA-S* database. Following a strategy of permitting responsible researchers who obtained human subject clearance at their respective institutions and who agreed to share copies of their raw *POSHA-S* data, the instrument's author has developed and maintained a growing database on attitudes. By January, 2022, nearly 21,000 respondents from 230 different public and professional samples representing 48 countries and translations to 30 different languages, had contributed to the database. About 84% of the database comprised respondents who filled out the instrument once, while 16% (or 3,277 respondents from 55 different 1st versus 2nd comparisons) were those who filled out the *POSHA-S* two or more times. Of the latter, most were carried out in studies that introduced interventions designed to improve public attitudes toward stuttering. The remainder of these pre-post studies were carried out as non-intervention control groups in a few of the intervention studies or studies of test-retest reliability of the *POSHA-S* instrument. This total non-intervention component used in the current study included 12 samples of 345 respondents and included those featured earlier in Table 1. Two studies used an experimental edition of the instrument (i.e., the *POSHA-E2*) [16]; the remainder used either complete or slightly modified versions of the final version of the *POSHA-S*) [7,12,14,15,17,18,21-25]. Respondents

were obtained utilizing samples of convenience.

The samples were from the USA, Kuwait, Poland, and Iran. Respondents' mean age was 28.5 years, and mean years of schooling was 13.8 years. More females than males were involved, 65% and 35%, respectively. Forty-four percent were parents, and 61% were married. Their mean relative income, which is a weighted value on a -100 to +100 scale that is derived from scaled ratings of one's income relative to (a) one's family and friends and (b) all the people in one's country was +10. This is higher than the median of relative income means in the *POSHA-S* database, which is essentially neutral, or +1 on the -100 to +100 scale. The mean percentage of self-reported stuttering among respondents was 1% (which is the expected percentage [26]), and the mean percentage reporting knowing no one who stuttered was 21%. Mean response times to fill out the *POSHA-S* were 11.3 minutes for 1st tests and 10.0 minutes for 2nd tests. The duration of time between 1st and 2nd tests was 3-5 days for one sample, two weeks for 10 samples, and one month for one sample.

Test and retest distributions of the OSS

The OSS, which is the mean of the Beliefs and Self Reactions subscores, served as the dependent variable in this study as it has been used to compare samples in numerous *POSHA-S* investigations, many of which were summarized by St. Louis [27]. In both the test and retest distributions, skewness (1st = -0.277; 2nd = -0.236) and kurtosis (1st = -2.80; 2nd = -0.272) were satisfactory [28]. Even so, the distributions were not completely normal (Shapiro-Wilk's test of normality (1st: $W[345] = 0.990$, $p = 0.019$; 2nd: W[345] = 0.991, $p = 0.042$). Homogeneity of variance was confirmed for both OSS values (Levene's $W = 0.26$ [1,688], $p = 0.607$). The two OSS means (1st = 10.24; 2nd = 10.96) were not significantly different using a dependent, paired t-test (t [344] = -0.932; $p = 0.352$). More simply, the 1st- and 2nd test distributions were essentially no different.

Correlations between test and retest *POSHA-Ss*

Pearson product-moment correlations were run between pairs of 1st and 2nd OSS scores for the same 345 respondents, and, as in the earlier reports [7,9], for different pairs of the 345 respondents. Additionally, correlations were run between 1st test OSSs and the amount of improvement or worsening in attitudes from 1st test to 2nd test (or 2nd-minus-1st OSSs). Finally, parallel correlations were carried out between 2nd test versus 2nd-minus-1st OSSs.

Additionally, in order to explore the conditions that might explain the previously unexpected negative correlations between the 1st versus 2nd-minus-1st values in Kuhn and St. Louis [18] and Węsierska et al. [17] we carried out a series of exploratory correlations using both actual and random values (Microsoft Excel random number generator) as well as various amounts added to or subtracted from the 1st values. The goal was to generate 1st versus 2nd, 1st versus 2nd-minus-1st, and 2nd versus 2nd-minus-1st correlations that would approximate the actual correlations.

Sorting individual respondents according to changes in *POSHA–S* summary scores from test to retest

To identify those respondents who improved, got worse, or essentially stayed the same, a sort was performed on all the respondents using the following criteria. On the -100 to +100 scale, those with a change from 1st to 2nd OSSs between -5 and +5 units (a 10-unit spread) were operationally considered to have minimal change or stay the same. Based on previous test-retest studies [7], we assumed that the 1st and 2nd test means would be nearly identical, but we wanted to allow for a reasonable amount of variability in minimum changers yet, also to identify the outliers who changed positively or negatively. Thus, those considered to have made a positive change from 1st test to 2nd test improved their OSSs by any amount greater than +5 units, and those with negative change worsened theirs by more than -5 units. In other words, the 1st test OSS value were subtracted from the 2nd test OSS value for each respondent. If the 2nd-minus-1st difference value was greater than +5, he or she was deemed to have improved stuttering attitudes or was a positive changer. If the difference was less than -5, the person was regarded to manifest worse attitudes or a negative changer. And, if the difference value was within ±5 units, the respondent was considered to have changed minimally or remained essentially the same. Subsequently, based on the "positive change," "minimal change," or "negative change" grouping, the mean 1st, 2nd, and 2nd-minus-1st values were calculated for these three change groups.

RESULTS

Correlation analyses

As shown in row 1 of Table 2, the correlation between the OSSs for all actual 1st versus 2nd OSS values was +0.79, which was nearly identical to the mean test-retest correlation reported by St. Louis [7]. As in the unexpected correlations described above in two samples that motivated this analysis [17,18], the

Table 2. Correlations between test (1st) versus retest (2nd) OSS values using actual and randomly generated values generated to approximate the actual 1st and 2nd correlations shown in row 1

Row	POSHA-S OSS rating	1st versus 2nd	1st versus 2nd-minus-1st difference	2nd versus 2nd-minus-1st difference
1	Actual 1st vs. actual 2nd	+0.79	-0.37	+0.28
2	Actual 1st vs. actual 2nd but shifted to pair with different 1st scores within each sample	+0.42	-0.56	+0.52
3	Actual 1st vs. actual 2nd sorted randomly for the total respondents	-0.06	-0.73	+0.73
4	Random 1st (±100) vs. random 2nd (±100)	-0.05	-0.71	+0.71
5	Random 1st (±25) vs. random 2nd (±25)	-0.04	-0.72	+0.72
6	Random 1st (±100) vs. random 2nd (±10) subtracted from 1st	+0.99	+0.02	+0.13
7	Random 1st (±100) vs. random 2nd (±25) subtracted from 1st	+0.97	-0.02	+0.24
8	Random 1st (±100) vs. random 2nd (±50) subtracted from 1st	+0.88	+0.02	+0.49
9	Random 1st (±100) vs. random 2nd (±75) subtracted from 1st	+0.78	+0.01	+0.62
10	Actual 1st vs. random 2nd (1-30) added to 1st	+0.94	+0.06	+0.40
11	Actual 1st vs. if 1st >0, random 2nd (1-30) added to 1st; if 1st ≤0, random (1-30) subtracted from 1st	+0.68	-0.70	+0.05
12	Actual 1st vs. if 1st >10, random 2nd (1-30) added to 1st; if 1st ≤10, random (1-30) subtracted from 1st	+0.60	-0.70	+0.15
13	Actual 1st vs. if 1st >10, random 2nd (1-35) added to 1st; if 1st ≤10, random (1-35) subtracted from 1st	+0.48	-0.70	+0.29
14	Actual 1st vs. if 1st >10, random 2nd (1-40) added to 1st; if 1st ≤10, random (1-40) subtracted from 1st	+0.32	-0.73	+0.42
15	Actual 1st vs. if 1st >10, random 2nd (1-37) added to 1st; if 1st ≤10, random (1-37) subtracted from 1st	+0.42	-0.72	+0.34

All correlations beyond ±0.105 are significant at $p \leq 0.05$.

2nd-minus-1st difference versus the 1st test correlation was negatively correlated ($r = 0$-0.37). The correlations between 2nd versus 2nd-minus-1st values was +0.28. Row 2 shows correlations when, *within each of the 12 samples*, the 2nd OSSs were shifted to pair with different 1st OSSs. In this case, the 1st versus 2nd correlation was +0.42, which is moderate and similar to a "societal" contribution reported by St. Louis [7]. However, when the 2nd scores of all 345 respondents were randomized to compare different respondents (row 3), the 1st versus 2nd correlation was near zero ($r = $-0.06), clearly showing no "societal" influence. The 2nd-minus-1st correlations for different 1st and 2nd respondents versus the 1st scores were -0.73 and +0.73 versus the 2nd scores.

Given the non-intuitive negative correlation between the 1st values and the difference values between those and the 2nd values, we then sought to determine if our *actual* correlations and pattern (+0.79, -0.37, and +0.28 in row 1) could be replicated with other numbers. This would be analogous to building a model to explain a natural phenomenon. Numerous trial analyses were

carried out using actual and random numbers in a 345-person sample. Generating random numbers between -100 and +100 in both 1st and 2nd OSS values (row 4) produced no correlation (-0.05) between 1st versus 2nd, -0.71 between 1st versus 2nd-minus-1st, and +0.71 between 2nd versus 2nd-minus-1st, or virtually the same as randomizing the actual 2nd scores in row 3. Reducing the random number range from -25 to +25 (row 5) yielded essentially the same result. Next, a series of correlations were run between randomly generated 1st scores from -100 to +100 and 2nd values between ±10 and ±75 added to (or subtracted from) the 1st values. 1st versus 2nd, 1st versus 2nd-minus-1st, and 2nd versus 2nd-minus-1st correlations at four levels were, respectively (in rows 6-9) as follow: ±10: +0.99, +0.02, and +0.13; ±25: +0.97, -0.02, and +0.24; ±50: +0.88, +0.02, and +0.49; and ±75: +0.78, +0.01, and +0.62. These essentially revealed 1st and 2nd values that were highly correlated but with no association between difference scores and 1st values. Increasing positive relationships between difference and 2nd values occurred as added or subtracted values increased in magnitude.

Table 3. Overall stuttering score (OSS) ratings for sorts of respondents according to change from 1st to 2nd on the -100 to +100 scale

	OSS All respondents 1st	OSS All respondents 2nd	Mean OSS change 1st to 2nd	OSS positive change 1st	OSS positive change 2nd	OSS positive change difference	OSS minimal change 1st	OSS minimal change 2nd	OSS minimal change difference	OSS negative change 1st	OSS negative change 2nd	OSS negative change difference	OSS percent positive change	OSS percent minimal change	OSS percent negative change
Column	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	10	11	+1	1	17	+15	11	12	0	19	4	-16	35.7%	34.8%	29.6%

2nd-minus-1st difference categories were: negative change (≤-5 units), Minimal Change (>-5 but <+5 units), and Positive Change (≥+5 units). Shown are 1st, 2nd, and difference (2nd-minus-1st) values as well as the percentage of respondents in each sample group.

Next in row 10, to serve as a comparison for the next five trials, the actual values were used as the 1st test, while for the 2nd test, random values from 1 to 30 were added to the actual 1st test values. This resulted in values similar to the random arrays in rows 6-9 and. In row 11, a similar procedure was used except if the 1st test score was negative (<0 in the -100 to +100 scale), then a random number from +1 to +30 was generated for the 2nd test score. Alternatively, if the 1st-OSS was positive (>0), a random number from -1 to -30 became the 2nd-OSS. This generated a +0.68 for 1st versus 2nd, -0.70 for 1st versus 2nd-minus-1st, and +0.05 for 2nd versus 2nd-minus-1st. The remaining trials (rows 12-15) were carried out to approximate as closely as possible the pattern of the actual correlations in row 1. The closest approximation (row 15) occurred when the cutoff was 10 (which was the average OSS in the entire sample) and to that cutoff, random values between 1 and 37 were either added to or subtracted. The three values were: 1st versus 2nd=+0.42, 1st versus 2nd-minus-1st=-0.72, and 2nd versus 2nd-minus-1st=+0.34.

These results confirmed that the negative 1st versus 2nd-minus-1st correlations observed in two intervention studies that showed no positive change in attitudes [17,18] did not support our hypothesis that those results were atypical or uncharacteristic of test-retest studies with the *POSHA-S*. The Table 2 results also showed that the negative correlations were not a statistical anomaly.

Amount, direction, and distribution of change of positive, minimal, and negative changers

The analysis of the change groups involved (a) determining the *magnitude and direction* of changes in the 1st OSSs of individual respondents who improved, worsened, or changed minimally on the 2nd OSSs as well as (b) determining the *percentages* of the total number of respondents who changed positively, minimally, and negatively. The third column of Table 3 shows the *mean change* in OSS of the samples, which serves as a comparison to values for the three change groups. Mean 1st,

2nd, and 2nd-minus-1st difference OSS values sorted according positive, minimal, or negative change are shown in columns 4-12. It should be noted that plus (+) values for 2nd-minus-1st difference scores reflect positive changes, and minus (-) values reflect negative changes. Columns 13-15 display the percentages of respondents within each change group.

It was hypothesized that 1st and 2nd OSS values for the respondents would cluster around the mean values. Further, because none of these respondents were exposed to any intervention, it was assumed that virtually all of them would be in the minimal change group in both the 1st and 2nd *POSHA-S* administrations, with only a few outliers in the positive or negative change groups. Moreover, most of these outliers were hypothesized to remain in their respective change groups. Finally, based on the fact that the 1st and 2nd test distributions were essentially equal, it was hypothesized that the large majority of respondents would show little or no change from 1st to 2nd test values.

Table 3, however, tells an entirely different story. The numbers showing the magnitude of rating changes, indicate that the positive change group began with the lowest 1st OSS scores of +1 but improved 15 units to +17 (rounded). Minimal changers started at +11, ended at +12, with a 0-unit (rounded) change. By contrast, the negative change group began with the highest 1st OSS of +19 but worsened by 16 units (rounded) to +4. We have labeled this pattern as a "crossover" effect 1. The respondent percentages also were counter to the hypothesized profile that virtually all of the respondents would be in the minimal change group. Fully 36% percent of the respondents improved and 30% worsened, while only 35% changed minimally. Each of these percentages approximate one-third of the total sample. Finally, very few outliers remained in their 1st test groups at the 2nd test.

DISCUSSION

Correlational analysis

This study replicated the expected high 1st versus 2nd correlation of about +0.80 for the OSS, which closely approximated

earlier test-retest reliability of all the items of the *POSHA-S* [7,9,14]. In addition, the earlier hypothesized "societal" component of to .57 between 1st and 2nd scores of different respondents [7] was approximated at +0.42 in the current sample, but only so long as different 1st and 2nd scores were compared *within* their original samples. Different respondent pairs were not correlated when all 12 samples were randomized together. These two procedures appear to support the existence of a "societal" hypothesis, given that attitudes of respondents *only within* each of the samples were apparently similar enough to generate a moderate correlation among different people. However, when *different* population samples are combined, this moderate "societal" correlation disappears. As noted, we are unaware of any standard protocol for identifying and quantifying "societal" and "individual" components of test-retest reliability. However, the procedure carried out here provides a guide: compare the same and different respondents *within* several individual samples and then compare the same respondents as an aggregate.

Considering this finding from Revelle and Condon's [2] perspective, trait, state, and specific components would likely contribute to some extent to both the "societal" component and "individual" components of the +0.80 correlation. Speculating, it might be reasonable to assume that the trait component would be strongest in the "societal" component. For example, shared influences on stuttering attitudes have been found to vary from culture-to-culture or geographic region-to-region [29,30]. A variety of influences in the specific component (e.g., memories of interacting with a person stuttering) might well predominate to generate approximate 0.30 to 0.40 "individual" "additions" to the "societal" component. The large differences in the positive and negative changers in the 2nd tests would most likely be due to the state component. Perhaps, upon thinking about a previous affirmative response to the item, "If I were talking to a person who stutters, I would tell them to 'slow down' or 'relax,'" a respondent might respond less affirmatively based on an overall feeling of greater empathy toward a stuttering person. Only careful research designs could tease out these and other possibilities.

In contrast to the test-retest correlations, to our knowledge, the moderate negative correlation ($r = -0.37$) between 1st-OSS versus 2nd-minus-1st OSS differences has not been reported before. At first glance, this was thought to be either an error or perhaps a statistical anomaly. However, after calculating numerous similar 1st and 2nd scores with combinations of actual and randomly generated numbers, a pattern of 1st and

2nd scores was generated that mirrored the pattern of the actual correlations. For reasons that are not apparent, when the average score on the 1st-OSS (i.e., +10) was set as the point at which randomly generated 2nd test numbers between 1 and 37 that were opposite in sign of actual 1st test values, similar correlations to the actual data emerged.

Direction and profiles of change in individual respondents' attitudes

It must be reiterated that, prior to this report, all of the investigators who carried out these reliability or control group studies had assumed that measured attitudes of the vast majority of the respondents were generally stable from 1st to 2nd *POSHA-S* ratings. Yet, contrary to these assumptions and expectations, only about one-third of respondents who held close to average attitudes of the combined sample were stable in their attitudes on the second *POSHA-S*. And, as noted, approximately one-third of respondents with the best ratings on the 1st test had the worst ratings on the 2nd test while the remaining third with the worst 1st test ratings had the best 2nd test ratings. In light of these results, the hypotheses that most of the 1st and 2nd scores would be stable from 1st test to 2nd test and that the overwhelming majority would be in the minimal change category were not supported. The sorting results were, however, consistent with the negative correlation between 1st test and 2nd-minus-1st differences.

How might the "crossover" effect be interpreted? It arguably reflects a "paradigm shift" [31] in understanding aspects of measured stuttering attitudes. It could be regarded as encouraging news in that about one-third of the public holding attitudes toward stuttering that are worse than average will acquire more accurate and sensitive beliefs and reactions simply by virtue of filling out an attitude scale a second time or thinking about what they discerned on the first administration. What is not encouraging—and even more puzzling—was our parallel finding that respondents with the most positive attitudes at the 1st test held the worst attitudes after retaking the scale.

It must be emphasized that these positive and negative changes in individuals were both large and consistent. Those who changed in the positive direction gained 15 units on the OSS, and those who changed in the negative direction on the 2nd test reduced by 16 units (Minimal changers had a mean difference value of 0 units). From numerous previous *POSHA-S* studies, mean OSS sample differences of 10 units, considerably smaller than either those for the positive and negative

changers, have been interpreted as substantial and important [7,11,25,27,29,32].

Especially intriguing, how could the test-retest reliability respondents have shown virtually no *mean* change in OSS from 1st to 2nd *POSHA-S*s, both in the original samples and as a combined group while, simultaneously, two-thirds of them changed dramatically in the 2nd test? The answer is that the positive and negative change groups had roughly equal magnitude changes in opposite directions and roughly equal percentages in the two groups, which resulted in one canceling the other out. For example, re-analysis in this study of the high school students from Poland [17], who were controls in an intervention study, had an OSS mean change from 1st to 2nd of -1 unit. Of this sample, 33% increased by +15 units, 29% decreased by -24 units, and 38% minimal changers remained unchanged at +1 unit.

Regression to the mean

The correlation trials in Table 2 using combinations of random and actual numbers for 1st and 2nd OSS values convinced us and two statistical consultants that the correlations were not the result of a statistical anomaly. Instead, as noted, they were clearly confirmed by the positive, minimal, and negative change group results. To our knowledge, no relevant test-retest study of attitudes has documented a "crossover" pattern such as was observed here. Nevertheless, it would be expected that the so-called "regression to the mean" phenomenon [33,34] could have affected our results. This phenomenon indicates that, when sorted according to high or low scores on some 1st test variable, an intervention-induced change will "move" the mean of either group closer to the overall 1st test mean in 2nd test values. A typical example is a baseball player who has an exceptionally high batting average one year will typically have a lower but respectable average the next year, and the player with the worst average will show some improvement the next year. Notwithstanding that none of the respondents reported here were exposed to a planned intervention, if present, regression to the mean would predict that mean 2nd test scores in the respondents with the highest scores would be somewhat *lower* than their 1st test scores, but still well above the 1st test mean. Conversely, the mean 1st test scores of respondents with the lowest scores would generate somewhat *higher* 2nd test mean scores, yet again, well below the 1st test mean.

Statistically, regression to the mean is a function of the extent to which 1st and 2nd test scores are correlated. There is no regression to the mean if the 1st versus 2nd test correlation equals 1.0 but complete regression to the mean if the correlation equals 0. The formula to calculate the percentage regression (movement) in the direction of the mean is $100 \times (1$ minus 1st versus 2nd correlation) [35]. Consider a hypothetical case wherein a 1st test mean of 20 is increased in a 2nd test mean to 30 among the one-half of the participants with the higher 1st test ratings. Also consider that all 1st and 2nd test ratings are correlated at +0.50. In this case, 50% of the change from the 1st to 2nd test was due to regression to the mean, so the actual change from 1st to 2nd would be 5 units rather than 10 units.

In the current study, the mean correlations between the OSS 1st versus 2nd ratings of all individual respondents were quite high as follows: $r = +0.79$ (Table 2). The mean differences between 1st and 2nd scores were 1 unit (0.73 rounded). Using the above formula, regression to the mean adjustment would be only 0.15 unit (0.73×0.21) or essentially no such regression.

"Crossover" effect

The impact of this "crossover" effect cannot be underestimated even though it was not recognized in investigations when they were being carried out. The method of measuring attitudes—and many other things—does have an effect on the results, as illustrated by Dholakia and Morwitz [19] and Boysen and Vogel [20]. In retrospect, it should not be surprising that the act of filling out a *POSHA-S* constitutes a significant treatment effect on any subsequent *POSHA-S*. While repeated surveys can be affected by remembering earlier responses or an overall learning effect, this study suggests the respondents were affected by influences beyond such well known biases.

For the positive changers who initially had decidedly negative beliefs about—and self reactions to—stuttering, the act of simply being obliged to think again about the *POSHA-S* items apparently challenged some of the negativity. Arguably, they then decided that stuttering is "not so bad." Of course, this was true of only about one-third of the respondents. For the third with initially quite positive attitudes who changed in the opposite direction, simply thinking about the *POSHA-S* items again apparently had an opposite effect. To them, thinking about stuttering likely changed their minds to stuttering being regarded as "worse than they thought." These two effects may be related to the well-known placebo effect and its much less recognized opposite, the nocebo effect. According to Raypole [36], "The placebo effect demonstrates how positive thinking can improve treatment outcomes. The nocebo effect suggests

that negative thinking may have the opposite effect."

Implications

The results of this study have identified a heretofore unreported characteristic in measured public attitudes toward stuttering, that is, their lack of stability in test-retest situations. Beyond the aforementioned speculations that might have impacted the positive and negative changers, our data do not offer further explication of the origin of the "crossover" effect. A careful search of the literature yielded no reports of the phenomenon. It is possible that "crossover" in test-retest reliability studies is a characteristic only of public attitudes toward stuttering. It is also possible that "crossover" is unique to items aggregated in the *POSHA-S*. However, a post-hoc analysis of a subset of data from St. Louis et al. [12] suggests it is not. Thirty-four college students twice filled out another stuttering attitude scale, a 1-7 semantic differential (or bipolar adjective) scale [37]. Their correlations were similar to those in the current study: 1st versus 2nd=+0.039, 1st versus 2nd-minus-1st=-0.74, and 2nd versus 2nd-minus-1st=0.33. It is also plausible that the "crossover" effect is a characteristic of other attitudes or measured phenomena in social science or education experiments, but not yet documented. Only future research using carefully designed studies could further illuminate these possibilities.

Another way to think about this is that public attitudes toward stuttering, unlike attitudes toward many other things, are probably not well established. Consider beliefs about—and reactions to—stuttering compared to beliefs and opinions about current controversial issues. Whereas stuttering attitudes are fluid, perhaps more similar to attitudes about consumer products [38], many other attitudes, such as those related to global warming [39] or university health care [40] are typically much more rigid. While experts in stuttering attitude research might be concerned that stuttering stereotypes and other negative attitudes are not as stable as was thought [41], this cloud has a silver lining. Opinions that are not well established are arguably easier to change than those that are not. With interventions that are maximally appropriate to the segment of the nonstuttering majority that is targeted, and with intrinsically or extrinsically motivated participants, success in changing relatively fluid attitudes is likely [32].

Strengths and limitations

A strength of this study is the number and diversity of samples that were evaluated on the same attitude measure. The total sample of 12 individual samples contained nearly 350 respondents representing four different countries filling out the *POSHA-S* in four different languages. This diversity among samples could, of course, be viewed as a limitation. More convincing, in our view, however, is that, given differences in demographics, geography, interventions, and differences in early versions of the *POSHA-S* for some early samples, the robustness of positive, minimal, and negative change groupings with a similar "crossover" pattern occurring across all samples yields findings that are more robust or generalizable than those that could be derived from a more uniform population.

Relatedly, because the original samples assigned to each category were from different populations, and also because a few samples utilized an earlier version of the *POSHA-S* without one later-added item, the 1st and 2nd OSS values from sample to sample were not perfectly comparable. This is clearly a limitation. Nevertheless, the mean *difference* ratings are entirely comparable because every respondent in every sample—and, therefore, in the combined sample—filled out exactly the same *POSHA-S* in the test and retest conditions.

Another potential limitation of the study was that the criteria to determine the positive, minimal, or negative change groups undoubtably affected the results. Different values would generate different percentages in the change groups. It is likely, however, that the "crossover" effect would still be an overarching finding, as would be predicted by the negative correlations in Table 2, but in slightly different magnitudes and percentages. In retrospect, the magnitude of the 1st test means of the high versus low scores obliquely validates our quite conservative >5-unit and <5-unit criteria because the 1st test means of the positive changes were 9 units *lower* than the overall mean and the means of the negative changers were 9 units *higher*.

Some also might regard as a limitation the fact that regression to the mean is involved anytime high or low scorers are chosen for analysis. Given the miniscule 0.15-unit correction calculated and, more importantly, the atypical "crossover" pattern far past the mean retest scores, it is doubtful that the typical regression to the mean adversely influenced the results.

Suggested future research

Future research implications are legion. As a first step, the results of this study point clearly to the need to explore the stability of measured public attitudes toward stuttering over time. To accomplish this, the *POSHA-S* could be administered monthly three to five times with no intervening interventions and in

such a way that individual respondents could be matched anonymously from administration-to-administration. The OSSs from the first *POSHA-S* could then be rank-ordered and divided into the top, middle, and bottom thirds according to the criteria utilized in this study. Those respondents could then be tracked in each successive *POSHA-S* to assess stability of individual item ratings and summary scores. More importantly, the "crossover" pattern observed in the current study could be confirmed or disconfirmed in early and/or later administrations. Our results would predict that initial stability would be the greatest for the middle third of the respondents. The proposed study would also provide needed evidence as to whether or not the top and bottom thirds would become more stable with repeated exposure to thinking about and evaluating beliefs and reactions to stuttering.

Similar studies could be carried out with other measures. For example, it would be useful to carry out sorts of similar 2nd minus-1st difference ratings and compare them with 1st ratings on both stuttering and non-stuttering attitude measures when no interventions are introduced.

CONCLUSION

Twelve samples consisting of 345 persons from four countries filled out the *POSHA-S* on two occasions with no intervening treatment or intervention between administrations. Group means for 1st and 2nd OSSs were essentially the same. Also, confirming earlier reports [7], the standard test-retest correlation procedure for assessing test-retest reliability yielded a correlation very close to +0.80, which is generally interpreted as an acceptable level. By contrast, further correlation analyses and sorting of those respondents who improved, worsened, or stayed the same yielded entirely unexpected results. Nearly one-third of the 345 respondents with the lowest (worst) OSSs improved to the highest (best) OSSs. Conversely, one-third with the highest (best) 1st test scores worsened to the lowest (worst) 2nd test scores. Only about one-third, with intermediate OSS values had very similar scores in the 1st and 2nd tests.

ACKNOWLEDGEMENTS

REFERENCES

1. Babbie E. The practice of social research, 13th ed. Belmont, CA: Wadsworth Centage Learning; 2013.
2. Revelle W, Condon DM. Reliability from alpha to omega: a tutorial; 2018.
3. St. Louis KO. The public opinion survey of human attributes-stuttering (POSHA-S): summary framework and empirical comparisons. Journal of Fluency Disorders. 2011;36(4):256-261.
4. St. Louis KO. Comparing and predicting public attitudes toward stuttering, obesity, and mental illness. American Journal of Speech-Language Pathology. 2020;29:2023-2038.
5. Al-Khaledi M, Lincoln M, McCabe P, Packman A, Alshatti T. The attitudes, knowledge and beliefs of Arab parents in Kuwait about stuttering. Journal of Fluency Disorders. 2009;34(1):44-59.
6. St. Louis KO. POSHA-S public attitudes toward stuttering: online versus paper surveys. Canadian Journal of Speech-Language Pathology and Audiology. 2012;36(2):116-122.
7. St. Louis KO. Research and development for a public attitude instrument for stuttering. Journal of Communication Disorders. 2012;45:129-146.
8. St. Louis KO, Lubker BB, Yaruss JS, Adkins TA, Pill JC. Development of a prototype questionnaire to survey public attitudes toward stuttering: principles and methodologies in the first prototype. The Internet Journal of Epidemiology. 2008;5(2).
9. St. Louis KO, Lubker BB, Yaruss JS, Aliveto EF. Development of a prototype questionnaire to survey public attitudes toward stuttering: reliability of the second prototype. Contemporary Issues in Communication Sciences and Disorders. 2009;36:101-107.
10. St. Louis KO, Roberts PM. Measuring attitudes toward stuttering: English-to-French translations in Canada and Cameroon. Journal of Communication Disorders. 2010;43:361-377.
11. St. Louis KO, Reichel I, Yaruss JS, Lubker BB. Construct and concurrent validity of a prototype questionnaire to survey public attitudes toward stuttering. Journal of Fluency Disorders. 2009;34:11-28.
12. St. Louis KO, Williams MJ, Ware MB, Guendouzi J, Reichel I. The Public Opinion Survey of Human Attributes-Stuttering (POSHA-S) and Bipolar Adjective Scale (BAS): aspects of validity. Journal of Communication Disorders. 2014;50:36-50.
13. Valente ARS, Jesus LMT, Roberto MT, Leahy MM, St. Louis KO. Cross-cultural adaptation of the Public Opinion Survey of Human Attributes-Stuttering (POSHA-S): the European Portuguese (EP) translation and adaptation process as a model. In D. Santiago (Ed). Stuttering: Risk factors, public attitudes and impact on psychological well-being. Hauppaguge, NY: Nova Science Publishers; 2016;93-110.
14. Abdi S, Jalaei Sh, Teimouri Sangani M, Pourmohammad A. Development and evaluation of psychometric properties of the Persian version of "Public Opinion Survey of Human Attributes-Stuttering". New Scientific Research Journal of Rehabilitation. 2014;9(5):87-92.
15. Abdalla F, St. Louis KO. Modifying attitudes of arab school teach-

ers toward stuttering. Language, Speech, and Hearing Services in the Schools. 2014;45:14-25.

16. St. Louis KO, Remley CL, Hancock BR. POSHA: test-retest reliability of the final version. Poster presented at the Annual Convention of the American Speech-Language-Hearing Association, Philadelphia, PA; 2010.

17. Wesierska K, Błachnio A, Przepiorka A, St. Louis KO. Zmiana postaw wobec jakania w polsce-wstepne doniesienia z badan [Changing attitudes towards stuttering in Poland: preliminary study report]. Paper presented at the International Conference on Speech-Language Therapy: Modern Trends in Logopaedic Diagnosis and Therapy. Chorzow, Poland; 2015.

18. Kuhn CD, St. Louis KO. Attitudes toward stuttering of middle school students before & after a stuttering video. Poster presented at the Annual Convention of the American Speech-Language-Hearing Association. Denver, CO; 2015.

19. Dholakia PM, Morwitz VG. How surveys influence customers. Harvard Business Review. https://hbr.org/2002/05/how-surveys-influence-customers; 2002.

20. Boysen GA, Vogel DL. Education and mental health stigma: the effects of attribution, biased assimilation, and attitude polarization. Journal of Social and Clinical Psychology. 2008;27:447-470.

21. Beste-Guldborg A, St. Louis KO, Shorts JE. Face-to-face interviews of people who stutter: effects on attitudes of SLP students. Poster presented at the Annual Convention of the American Speech-Language-Hearing Association. Denver, CO; 2015.

22. Holcombe K, Eisert S. Understanding the impediment: investigating the nature of people's attitudes toward stutterers. Unpublished capstone paper. Pacific Lutheran University, Tacoma, WA; 2012.

23. Junuzovic-Žunic L, Weidner ME, Reichel IK, Cook S, St. Louis KO, Ware MB. Effects of fluency disorders coursework on students' stuttering attitudes in two countries. In KO St. Louis (Ed.), Stuttering meets stereotype, stigma, and discrimination: an overview of attitude research. West Virginia University Press. 2015;226-242.

24. St. Louis KO, Enoch MB. Stuttering coursework influences on undergraduate student attitudes. Poster presented at the Annual Convention of the American Speech-Language-Hearing Association. Atlanta, GA; 2012.

25. St. Louis KO, Przepiórka AM, Beste-Guldborg A, Williams MJ, Błachnio A, Guendouzi J, et al. Stuttering attitudes of students: professional, intracultural, and international comparisons. Journal of Fluency Disorders. 2014;39:34-50.

26. Bloodstein O, Ratner NB, Brundage SB. A handbook on stuttering (7th ed.) Plural Publishing. Inc., ISBN. 2021;13:978-1-63550-317-3.

27. St. Louis KO. Epidemiology of public attitudes toward stuttering. In KO St. Louis (Ed.), Stuttering meets stereotype, stigma, and discrimination: an overview of attitude research. 2015;7-42. West Virginia University Press.

28. West SG, Finch JF, Curran PJ. Structural equation models with nonnormal variables: problems and remedies. In RH Hoyle (Ed.), Structural equation modeling: concepts, issues, and applications. Sage Publications, Inc; 1995;56-75.

29. St. Louis KO, Wesierska K, Polewczyk I. Improving polish stuttering attitudes: an experimental study of teachers and university students. American Journal of Speech-Language Pathology. 2018; 27:1195-1210.

30. Üstün-Yavuz MS, Warmington M, Gerlach H, St. Louis KO. Cultural difference in attitudes towards stuttering among British, Arab and Chinese students: considering home and host cultures. International Journal of Language and Communication Disorders. 2021;56:609-619.

31. Kuhn TS. The structure of scientific revolutions. Chicago: University of Chicago Press; 1962.

32. St. Louis KO, Wesierska K, Przepiorka A, Błachnio A, Beucher C, Abdalla F, et al. Success in changing stuttering attitudes: a retrospective study of 29 intervention samples. Journal of Communication Disorders. 2020;84:1-18. 105972.

33. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology. 2005;34:215-220.

34. Zhang X, Tomblin JB. Explaining and controlling regression to the mean in longitudinal research designs. Journal of Speech, Language, and Hearing Research. 2004;46:1340-1351.

35. Trochim WMK. Regression to the mean. Web Center for Social Research Methods. 2024. https://conjointly.com/kb/regression-to-the-mean/.

36. Raypole C. What is the nocebo effect? Healthline.com; 2024. https:// www.healthline.com/health/nocebo-effect.

37. Woods CL, Williams DE. Speech clinicians' conceptions of boys and men who stutter. Journal of Speech and Hearing Disorders. 1971;36:225-234.

38. Duhachek A, Zhang S, Krishnan S. Anticipated group interaction: coping with valence asymmetries in attitude shift. Journal of Consumer Research. 2007;34:395-405.

39. Howell RA, Capstick S, Whitmarsh L. Impacts of adaptation and responsibility framings on attitudes toward climate change mitigation. Climatic Change. 2016;136:445-461.

40. Catapano R, Tomala ZL, Rucker DD. Perspective taking and self-persuasion: why "putting yourself in their shoes" reduces openness to attitude change. Psychological Science. 2019;30:424-435.

41. Hughes S, Strugalla EA. Recognizing positive aspects of stuttering: a survey of the general public. In KO St. Louis (Ed.), stuttering meets stereotype, stigma, and discrimination: an overview of attitude research. West Virginia University Press. 2015;176-186.