Dysphagia
https://doi.org/10.1007/s00455-018-9883-x

ORIGINAL ARTICLE

Reliability of Untrained and Experienced Raters on FEES: Rating Overall Residue is a Simple Task

Jessica M. Pisegna[1,2,3] · James C. Borders[1,3] · Asako Kaneoka[4] · Wendy J. Coster[5] · Rebecca Leonard[6] · Susan E. Langmore[1,2]

Received: 25 October 2017 / Accepted: 15 February 2018

Abstract

The purpose of this study was to investigate the reliability of residue ratings on Fiberoptic Endoscopic Evaluation of Swallowing (FEES). We also examined rating differences based on experience to determine if years of experience influenced residue ratings. A group of 44 raters watched 81 FEES videos representing a wide range of residue severities for thin liquid, applesauce, and cracker boluses. Raters were untrained on the rating scales and simply rated their overall impression of residue amount on a visual analog scale (VAS) and a five-point ordinal scale in a randomized fashion across two sessions. Intra-class correlation coefficients, kappa coefficients, and ANOVAs were used to analyze agreement and differences in ratings. Residue ratings on both the VAS and ordinal scales had acceptable inter- and intra-rater reliability. Inter-rater agreement was acceptable (ICC > 0.7) for all comparisons. Intra-rater agreement was excellent on the VAS scale ($r_c = 0.9$) and good on the ordinal scale ($k = 0.78$). There was no significant difference between expert ratings and other raters based on years of experience for cracker ratings ($p = 0.2119$) and applesauce ratings ($p = 0.2899$), but there was a significant difference between clinicians on thin liquid ratings ($p = 0.0005$). Without any specific training, raters demonstrated high reliability when rating the overall amount of residue on FEES. Years of experience with FEES did not influence residue ratings, suggesting that expert ratings of overall residue amount are not unique or specialized. Rating the overall amount of residue on FEES appears to be a simple visual-perceptual task for puree and cracker boluses.

Keywords Deglutition · FEES · Pharyngeal residue · Ratings · Reliability · Psychometrics

Portions of this manuscript were presented in poster format at the Dysphagia Research Society in Portland, Oregon on Saturday, March 4, 2017 and in a dissertation defense on December 7, 2016.

✉ Jessica M. Pisegna
jpisegna@bu.edu

James C. Borders
James.borders@bmc.org

Asako Kaneoka
kaneokaa-reh@h.u-tokyo.ac.jp

Wendy J. Coster
wjcoster@bu.edu

Rebecca Leonard
rjleonard@ucdavis.edu

Susan E. Langmore
langmore@bu.edu

[1] Department of Otolaryngology, Boston University School of Medicine, 820 Harrison Ave., FGH Building, 4th Floor, Boston, MA 02118, USA

[2] Sargent College of Health and Rehabilitation Sciences, Boston University, 635 Commonwealth Ave, Boston, MA 02215, USA

[3] Boston Medical Center, 830 Harrison Ave., Suite 1400, Boston, MA 02118, USA

[4] The University of Tokyo Hospital Rehabilitation Center, Tokyo, Japan

[5] Department of Occupational Therapy, Boston University College of Health and Rehabilitation Sciences: Sargent College, 635 Commonwealth Avenue, Boston, MA 02215, USA

[6] University of California at Davis, Davis, CA 95616, USA

Published online: 07 March 2018

Introduction

Rating pharyngeal residue on Fiberoptic Endoscopic Evaluation of Swallowing (FEES) is riddled with methodological challenges, one of which is quantifying how much residue is present. The few investigations that have studied residue ratings on FEES have relied on either reliability or consensus between raters or expert judges as the 'gold standard' to estimate how much residue is actually present [12]. While agreement is one way to estimate the amount of residue, it does not necessarily reflect the actual amount of residue present. This is because raters are judging the amount of residue by eye and not quantifying the amount precisely. The gold standard for residue measures on FEES remains elusive.

Given these challenges, it is not surprising that there are very few scales with established validity for rating pharyngeal residue on FEES. A recent systematic review found only seven scales that met the criterion for the review, and six of them had "*insufficient data to support their use as evidenced by methodological weakness*" [21]. For instance, only 4 of the scales reported on reliability measures (inter- or intra-rater) based on 2, 9, 15, and 2 expert raters, respectively [5, 14, 22, 38]. Other studies not included in that systematic review have also used expert consensus for rating residue on FEES as a substitute gold standard [12, 27, 33]. The use of expert raters as a substitute gold standard is concerning for several reasons: (1) it is a grossly subjective rating, based on a relatively small group of raters, and (2) no studies have compared expert ratings to other clinicians' ratings to determine if there truly is something unique about expert ratings. In the absence of an ability to determine true validity, we wondered about reliability on an untrained rating scale and how experts would compare to clinicians.

The purpose of this study was to further investigate psychometric aspects of residue ratings on FEES. Our aims were constructed in the context of a recent investigation that found differences between FEES residue ratings on a visual analog scale (VAS) and an ordinal rating scale [28]. We sought to determine inter- and intra-rater reliability on two different types of simple rating scales to determine if reliability can be acceptable in a mixed group of raters who are untrained on the scales. We also sought to examine rating differences based on clinician experience to determine if years of experience influenced residue ratings on the VAS. We hypothesized that more experienced clinicians would demonstrate better reliability.

Methods

Raters were asked to rate the overall amount of residue on FEES videos, twice, each time with a different rating method.

Participants

We aimed to recruit a range of speech pathologists and a group of laymen. The inclusion criteria for speech pathologists were clinicians, or students studying to be speech pathologists, who had at least heard of the procedure FEES. The exclusion criterion for clinicians was the inability to understand spoken or written English. The laymen, recruited by word of mouth and local ads, were participants who had to meet different criteria to ensure complete unfamiliarity with FEES and swallowing disorders. A convenience sample of laymen was invited to participate, consisting of people recruited from a local online ad, neighbors, and other local residents. Laymen were excluded if they worked in a medical setting, reported any familiarity with FEES or videofluoroscopic swallow studies (had heard of it at least once), reported any familiarity with swallowing disorders or head/neck anatomy, reported professional experience in making judgments from complex visual data (i.e., a computer graphics designer), or were color blind.

Videos

The FEES videos were prospectively collected from patients seen for a swallow evaluation in the outpatient clinic of an urban hospital. The videos are more thoroughly described elsewhere in a companion paper [28]. Videos were selected for use in the study if any of the following boluses were administered during FEES with two drops of green food dye: 5 mL thin liquid via spoon, 5 mL applesauce via spoon, ¼–½ saltine cracker.

The videos were categorized by consistency and residue severities until an adequate variety of residue presentations were collected to complete the following categories: 25 videos of 5 mL thin liquid, 25 videos of 5 mL applesauce, and 25 videos of ¼–½ of a saltine cracker. Within each bolus type, there were 5 videos demonstrating no residue, 5 demonstrating trace/coating, 5 demonstrating mild, 5 demonstrating moderate, and 5 demonstrating severe residue. To categorize the videos according to the aforementioned categories of residue severity, two experienced raters independently rated the overall residue severity using a previously published perceptual scale of *none, trace/coating, mild, moderate, severe* [13].

All videos were presented in the same exact format to raters: a 3-second title listing the bolus amount and consistency ("*5 mL applesauce*") followed by the clipped FEES video that included before, during, and after the swallow. The videos contained instruction titles to "*1. Score Now*" for the period of time after the first swallow and "*2. Score Now (clearing swallow)*" for the period of time after the very last clearing swallow(s). Sample videos can be seen in a companion paper [28]. Each video was numbered to correspond with a rating sheet in the provided packet (see Procedure).

Procedure

Participation occurred in small groups of ≤ 5 raters. They were not allowed to share impressions or to discuss the videos with each other. As the raters viewed each FEES video, they responded to a questionnaire that asked several questions described below. The only question pertaining to this investigation was, "*Overall, how much residue do you see?*" Residue was not defined and no operational definitions of severity were provided to the raters because this study aimed to compare the unprompted internalized scales of each clinician without any priming. There were 75 videos and 6 repeated videos shown within each session for intra-rater analyses, unbeknownst to the raters. The rating method for each sheet of paper was randomized to either ordinal or VAS. For the ordinal rating, choices were *none, trace/coating, mild, moderate,* or *severe* for the first question about overall amount of residue. On the VAS ratings, raters were asked to mark a slash (*/*) on the 100-mm line according to the impressions of residue severity. The companion paper includes a schema of the rating method presentation [28].

The rating method was planned such that each swallow was rated twice, but once on an ordinal and a VAS rating in a randomized fashion. In the first session, each rater viewed the 81 edited FEES videos and rated their impression of residue severity for each video. In the second session about 2 weeks later, they rated the same 81 videos. There were no dropouts. Both rating methods were presented within each session in a randomized order to avoid any habituation or repetitive answering effects. In the second session, the order was counterbalanced to change the rating method for each video. During the sessions, the videos were displayed on a bright 13-inch high retina full-screen computer display that was placed within 5 feet of the raters. Raters were allowed to watch the videos as many times as requested, as well as pausing at requested time points or using slow motion. Only the lead investigator was allowed to control the videos to allow for as much standardization in video presentation as possible. Sessions ranged from 45 min to 1.5 h and brief breaks were allowed as requested.

Statistics

To determine if one method yielded greater consensus among raters, intra-class correlation coefficients (ICC) were determined for ordinal ratings and VAS ratings. Many studies have demonstrated that ICC analyses can be used for continuous or ordinal data [32, 37]. Only clinician reliability was analyzed for generalizability to clinical practice; laymen's reliability was not of interest for this research question. The ordinal and continuous data were converted via logarithms with a base 10 to allow for a simple derivative value that would allow for comparison of ICC coefficients without altering the order of the data itself [9]. The log10 of the data for each clinician ($n = 33$) was entered into an SPSS computer program.

Reliability analyses for all raters for both ordinal and VAS ratings employed a two-way random model (2,1), and estimates were based on absolute agreement, not consistency. A 95% confidence interval was calculated. Pre-defined interpretation levels were assigned to the ICC coefficients: < 0.2 = poor, 0.21–0.4 = fair, 0.41–0.6 = moderate, 0.61–0.8 = good, and 0.81–1.0 = excellent. To assess reliability of experts independently ($n = 4$) and all other clinicians ($n = 29$) independently, two-way mixed models were used (3,1) due to the fixed set of raters. A 95% confidence interval was calculated and the interpretation levels were the same as defined above. It was hypothesized that ordinal ratings would demonstrate an ICC < 0.7 [12, 38] and VAS ratings would demonstrate an ICC > 0.7, given evidence from preliminary data [29].

To determine intra-rater reliability, Cicchetti–Allison weighted kappas were used for ordinal ratings [8] and Lin's concordance coefficient was used for VAS ratings [16]. A weighted kappa is a numeric index ranging from 0 to 1 that assesses agreement along an ordinal scale that gets progressively more severe and also accounts for agreement by chance. ICCs were not employed for intra-rater reliability measures due to a concern for homogeneity in ratings that would negatively interfere with the much-needed source of variance for ICC measures [32]. Lin's correlation is a numeric index ranging from 0 to 1 and assesses the degree of agreement between two continuous measures along the 45-degree slope line to account for concordance and discordance. Chance agreement is also factored into the statistic, making it a stronger analysis than other correlation analyses.

The ratings from the 6 repeated videos within each session were used for the intra-rater data analyses, 3 applesauce videos and 3 cracker videos. Thin liquid videos were not used for intra-rater data due to previously reported

variability in thin liquid ratings [25] and prioritization to reduce participant burden related to attention span and length of session. Not all of the clinicians rated the repeated videos due to an administrative error in the distributed packets. Pre-defined interpretation levels were assigned to the weighted kappa levels: < 0.2 = poor, 0.21–0.4 = fair, 0.41–0.6 = moderate, 0.61–0.8 = good, and 0.81–1.0 = excellent [1]. Pre-defined interpretation levels were assigned to the Lin's concordance coefficients, loosely structured around previously reported thresholds: < 0.6 = poor, 0.61–0.8 = good, 0.81–1.0 = excellent [20]. It was hypothesized that intra-rater agreement would be stronger for ordinal ratings than VAS ratings. SAS was used to calculate the kappa statistics of the ordinal variables; in cases where the data entry table was non-square, dummy weights were added to prevent errors in the calculations. An online site was used for Lin's calculations ("Statistical Calculations: Lin's Concordance," https://www.niwa.co.nz/node/104318/concordance).

To investigate differences in clinicians versus expert clinicians, only the VAS ratings were analyzed due to stronger statistical analyses over the ordinal ratings. The VAS ratings were grouped by clinician type into 5 groups: laymen, novice students and clinicians, proficient clinicians, advanced clinicians, and expert clinicians.

Every VAS rating was measured in mm from the left-hand side of the VAS line and rounded to the nearest 0.1 mm. A second data collector audited 20% of the data entry to ensure accuracy to the nearest mm. All data were entered into REDCap, an electronic data capture tool, and then were cleaned and sorted. Outlier data points, defined as $\pm$ 3 standard deviations from the mean, were removed to meet the assumptions of the desired ANOVA testing and a balanced statistical design; from the cracker videos, there were $n = 6$ outliers, from the applesauce videos, there were $n = 13$, and from the thin liquid videos, there were $n = 19$ outliers. Cracker, applesauce, and thin liquid ratings were analyzed independently, given the previously reported influence of bolus type [28]. SAS was used to calculate a multi-factor ANOVA for each of the three bolus types based on the following factors and levels: (1) clinician grouping by years of experience (5 levels), and (2) severity of VAS ratings (5 levels). Interaction terms were included in the model to investigate underlying effects. A p value of $p < 0.05$ was interpreted as significant for all analyses.

Results

A total of 44 participants rated videos. The participants were categorized into 5 groups: laymen (absolutely no familiarity with swallowing disorders, $n = 11$), novice students and clinicians (0–1 years of experience with FEES, $n = 10$), proficient clinicians (2–5.9 years of experience with FEES, $n = 8$), advanced (6–9 years of experience with FEES, $n = 11$), and experts ($\geq$ 15 years experience with FEES, $n = 4$). The mean number of days between sessions was 11.9 and the median was 14 (range 1–34 days). Participants rated every video but for 2 ordinal ratings, which were unusable due to difficulty in deciphering placement of the mark.

Inter-Rater Reliability

Both rating methods demonstrated acceptable inter-rater reliability. The intra-class correlation (ICC) coefficients for the inter-rater reliability are listed in Table 1. When all the participants' ratings across all boluses were combined, for both VAS and ordinal ratings, there was an acceptable reliability of > 0.80. There were lower ICCs for thin liquid ratings on both VAS and ordinal ratings (0.70–0.82), but they were still above what could be considered clinically acceptable (ICC > 0.7). Clinician ICCs were nearly identical to expert ICCs on most comparisons.

Intra-Rater Reliability

For intra-rater reliability, clinicians rated the same cracker and applesauce videos twice with each rating method. The reliability coefficients for the VAS ratings were higher than the coefficients for ordinal ratings. For all VAS repeated ratings (applesauce and cracker combined), the agreement was 0.90 and interpreted as 'excellent.' For all ordinal repeated ratings (applesauce and cracker combined), the agreement was 0.78 and interpreted as 'good.' The cracker and applesauce reliability indices are listed in Table 2, along with the lower and upper 95% confidence limits. Of note, these reliability coefficients and confidence limits cannot be numerically compared because of the use of disparate statistical testing procedures. In general, intra-rater reliability for just the applesauce ratings was excellent. For cracker ratings, the VAS ratings showed excellent agreement, while the ordinal ratings were only 'good.'

Cracker Ratings

There was no significant source of variance among the clinician experience groups (df = 4, $F = 1.46$, $p = 0.2119$). The overall model contained a significant source of variance (df = 24, $F = 54.84$, $p < 0.0001$), but the significant variance was due to the expected difference in severity levels of cracker residue ratings (df = 4, $F = 325.67$, $p < 0.0001$). There was no interaction effect of experience and severity ($p = 0.9602$), and removing the interaction term from the model did not alter the

Table 1 Inter-rater reliability coefficients (intra-class correlations; ICC) and their 95% confidence intervals (CI) for visual analog scale (VAS) and ordinal ratings across bolus types and clinician types

	All raters ICC ($n = 33$)	Clinicians only ICC ($n = 29$)	Experts only ICC ($n = 4$)
Inter-rater reliability for VAS ratings			
All cracker boluses ($n = 25$)	0.83 (0.74–0.90)	0.82 (0.73–0.90)	0.84 (0.72–0.92)
All applesauce boluses ($n = 25$)	0.87 (0.80–0.93)	0.87 (0.78–0.93)	0.87 (0.80–0.93)
All thin liquid boluses ($n = 25$)	0.73 (0.61–0.84)	0.73 (0.61–0.84)	0.70 (0.54–0.84)
All boluses ($n = 75$)	0.82 (0.76-0.86)	0.81 (0.76–0.86)	0.82 (0.75–0.87)
Inter-rater reliability for ordinal ratings			
All cracker boluses ($n = 25$)	0.87 (0.80–0.93)	0.86 (0.79–0.93)	0.89 (0.80–0.94)
All applesauce boluses ($n = 24$)	0.92 (0.87–0.96)	0.91 (0.995–0.998)	0.94 (0.89–0.97)
All thin liquid boluses ($n = 24$)	0.78 (0.70–0.88)	0.78 (0.67–0.87)	0.82 (0.71–0.91)
All boluses ($n = 73$)	0.86 (0.82–0.90)	0.86 (0.82–0.90)	0.89 (0.85–0.92)

Table 2 Intra-rater reliability coefficients for VAS (Lin's correlation coefficient, r_c) and ordinal ratings (weighted kappa, k). The lower and upper 95% confidence limits (CL) are provided

	Coefficient	Lower 95% CL	Upper 95% CL
Intra-rater reliability			
All cracker boluses ($n = 56$)			
VAS Ratings	$r_c = 0.86$	0.79	0.91
	EXCELLENT		
Ordinal ratings	$k = 0.73$	0.61	0.84
	GOOD		
All applesauce boluses ($n = 73$)			
VAS ratings	$r_c = 0.92$	0.87	0.95
	EXCELLENT		
Ordinal ratings	$k = 0.83$	0.75	0.91
	EXCELLENT		
All boluses (cracker and applesauce) ($n = 129$)			
VAS ratings	$r_c = 0.90$	0.86	0.93
	EXCELLENT		
Ordinal ratings	$k = 0.78$	0.71	0.85
	GOOD		

significance of other factors. Tukey's test was used for post hoc testing and all cracker severity levels were significantly different from one another ($p < 0.05$) except for none versus trace and mild versus moderate comparisons.

Applesauce Ratings

For applesauce videos, there was no significant source of variance among the clinician experience groups (df = 4, $F = 1.25$, $p = 0.2899$). The overall model contained a significant source of variance (df = 24, $F = 254.5$, $p < 0.0001$), but the significant variance was due to the expected difference in severity levels of applesauce residue rating (df = 4, $F = 1520$, $p < 0.0001$). There was no interaction effect of experience and severity ($p = 0.1164$), and removing the interaction term from the model did not alter the significance of the other factors. Tukey's test was used for post hoc testing and all applesauce severity levels were significantly different from one another ($p < 0.05$).

Thin Liquid Ratings

For thin liquid videos, the overall model contained a significant source of variance (df = 24, $F = 113.25$, $p < 0.0001$), and there was a significant source of variance among the clinician experience groups (df = 4, $F = 5.08$, $p = 0.0005$) and among thin liquid severity levels (df = 4, $F = 670.63$, $p < 0.0001$). There was no interaction effect of experience and severity ($p = 0.5207$), and removing the interaction term from the model did not alter the significance of the other factors. Tukey's test was used for post hoc testing. There was a significant difference between the laymen versus novice clinicians and laymen versus advanced clinicians. All thin liquid severity level ratings were significantly different from one another ($p < 0.05$).

All factors were pooled into a generalized linear model for an overall analysis of variance. There was a significant difference in the model (df = 10, $F = 700.08$, $p < 0.0001$). The Type III SS results were used due to an unbalanced design and there was a significant effect for experience (df = 4, $F = 2.70$, $p = 0.0290$), which was driven by only one significant comparison between the laymen and the advanced clinicians (all other comparisons were not significant). There was a significant interaction by severity (df = 4, $F = 1579.45$, $p < 0.0001$) and bolus type (df = 2, $F = 360.22$, $p < 0.0001$). All pairwise comparisons for severity and bolus groupings were significantly different at $p < 0.05$.

Discussion

This study investigated pharyngeal residue ratings on FEES, an understudied topic due to methodological and psychometric challenges. We wondered if the type of rating scale would influence reliability. We also investigated expert ratings to determine if their ratings, which are often referenced as the gold standard, were significantly different from other clinicians' ratings.

Inter-Rater Reliability

The visual analog scale (VAS) demonstrated slightly higher inter-rater reliability than ordinal ratings, although both could be interpreted as clinically acceptable (no ICC was lower than 0.7). It was hypothesized that categorical ratings would have demonstrated an ICC < 0.7 because of the previously reported variability in categorical ratings of residue on FEES [12, 13, 31]. Yet the current study demonstrated high agreement among participants. This is even more interesting considering that no training was provided and there was a wide range of residue severities presented across 75 different videos. In a different study with comparable design, raters were not given any training or prompting, and the reported reliability (ICC) was about 0.60–0.61 [12], much lower than the ordinal ICCs of 0.86–0.89 in the present study. Further, in that study, after providing 3 h of training on a new scale, the reliability between 4 raters across 63 videos increased to only ICC of 0.81 and 0.80 in two respective sessions, nearly equal to the present study consisting of a larger, untrained, and more experientially diverse group of clinicians.

One possible explanation of these different findings is that clinicians in the present study were given a prompt of when to judge the residue: immediately after the first swallow but before the clearing swallows. Such standardization may have increased the reliability. This is a valuable finding: if timing of residue rating is controlled for, clinicians can be reliable [6, 7]. Kaneoka et al. [12] did not distinguish a particular scoring time and included clearing swallows in the untrained overall impression. Another reason for high inter-rater reliability on categorical ratings could be that 5 categories, rather than 7, created greater chance agreement with fewer choices. However, this does not explain the high VAS reliability. A final consideration regarding the acceptable inter-rater reliability on both the categorical and VAS rating is that raters did not report on other factors, like rating location of residue, which likely increased the chances of clinician agreement on a more global impression. Perhaps, then, cueing clinicians when to judge overall amount of residue is more important for agreement than training clinicians how to judge residue amount.

Bolus consistency appears to make a small difference in inter-rater reliability while still yielding agreement levels that are acceptable. While cracker and applesauce reliability coefficients ranged from an ICC of 0.8 to 0.9, thin liquid reliability ranged between 0.70 and 0.82. This finding goes hand-in-hand with the difference in thin liquid ratings depending on years of experience. Thin liquid can be somewhat transparent or blend in with pre-existing secretions, making it difficult to detect and thus variably perceived from clinician to clinician. It was the only consistency that appeared to be influenced by years of experience, suggesting that a trained eye is more adept to detecting liquid residue. However, there was not an obvious and consistent pattern in the comparisons of experience levels, as only 2 comparisons of experience categories were significant. Therefore, the wide variability in thin liquid ratings may have also influenced this finding. Thin liquid fluidity appears to be difficult to assess, and is likely a reason why some scales for residue ratings have not used liquids in their development [22]. Some solutions to increase thin liquid rating reliability might be to report the worst location of thin liquid in addition to amount (to account for the fluidity), or to use strategies that have been shown to enhance visualization such as narrow-band imaging [23, 24] or food dye [19]. For research purposes, the lower reliability on thin liquid boluses should be considered if agreement is an important variable. For clinical practice, the influence of lower reliability on thin liquids remains unclear but should be noted when more than one clinician is treating a patient.

Intra-Rater Reliability

Interestingly, the VAS ratings had better within-clinician agreement (intra-rater reliability) than the ordinal ratings. Within-clinician agreement on VAS ratings was determined to be 'excellent' ($r_c = 0.9$), whereas agreement on ordinal ratings was determined to be 'good' ($k = 0.78$). This is an intriguing finding considering that there are countless choices on a 100-mm line and that the ratings of the same video were performed an average of 2 weeks apart. Further, there were no tick marks on the VAS line that could have biased placement. In a survey completed after the study, many clinicians commented that they felt their ratings on the VAS were unreliable, "*It was harder to tell if I was being consistent*," "*Difficult to replicate*," "*Not sure if I was consistent*," "*May not be as consistent each time*." However, the data demonstrated the opposite: clinicians were very consistent when using the VAS. About 67% of the repeated VAS ratings were within 10 mm of the first rating. In fact, 16 clinicians were surprisingly within

1 mm of their first rating. These results suggest that clinicians were indeed consistent when scoring on the VAS, more so than on the ordinal scale. It could be that with the greater precision allowed in rating amount on a VAS, clinicians were able to be more consistent in their ratings. Previous research on VAS measurement supports this finding; VAS ratings allow for more freedom in rating slight gradations, which results in better reliability [4, 10, 26].

Another point of discussion regarding high intra-rater reliability is that without any training or operational definitions, each clinician undoubtedly had a means of interpretation in their mind, which likely played a strong role in the consistency of ratings. This 'internalized scale' may have contributed to the high and steady within-clinician reliability that is reported here. Other studies of FEES residue have found similar results, i.e., intra-rater reliability was higher than inter-rater reliability of residue ratings on an ordinal scale [12, 13, 38]. The literature also reports that intra-rater reliability on ordinal ratings remains high but unchanged over time, even with training or repeated exposure [12, 22, 38]. Therefore, a clinician's internalized scale appears to be a strong factor that is difficult to influence, even with training.

Experience Does Not Influence Residue Ratings

Curiously, there has been no published study that directly compares expert ratings of residue on FEES to non-expert raters. The literature has assumed that there is something different or special about expert ratings without any empirical support. In fact, multiple studies have used experts as the reference standard in either a Delphi, a consensus, or an averaging method [11, 12, 18, 22]. The current investigation found no significant difference in ratings between experts and the other raters on cracker and applesauce ratings, including no difference in ratings from laymen who had absolutely no familiarity with head/neck anatomy or swallowing problems. This finding contradicts the hypothesis that years of FEES experience influences residue ratings. On thin liquid ratings, there was a significant difference between laymen and novice clinicians and also a significant difference between laymen and advanced clinicians. This could be due to poor visualization, location of residue, the movement of the residue, or a combination of these factors. It could also be due to years of experience in interpreting thin liquid residue, but if this were true, we would expect to see such an influence in other boluses. Rating the overall residue amount appears to be a simple, visual-spatial task that requires no training if using the VAS scale. Based on the present results, one should not expect expert ratings of global residue amount to be different from other clinician's ratings, when considered in the aggregate.

The lack of difference between laymen, novice, proficient, advanced, and expert ratings of residue is a provocative finding. Not only does it raise the question of whether residue ratings on FEES require an expertise, but it also provides an interesting platform for future studies. That is, if expert ratings are not unique, then what should be used as the reference standard to represent how much residue is present? Borrowing research from the field of non-verbal behavior seems appropriate as similar challenges are faced in measuring observations, for example, how much happiness is expressed [17]. What is striking is how similar that problem is to rating residue: dynamic events characterized by a swath of variables. Research with non-verbal behavior has shown that humans are perceptually very good at accurately discerning and synthesizing complex cues and perceiving in a gestalt-like manner [2, 3, 36]. In line with those conclusions, this body of work assumes that (1) the average of clinician ratings of residue will be near the truth, and (2) expert ratings are not very different from other clinicians' ratings. When future studies are looking to classify how much residue is present, having a group of clinicians rate the videos and then using the aggregate as the reference may be optimal, assuming that a gestalt rating judged from a video clip will be near the truth [34, 35]. Some may choose to continue to use experts as the reference standard, given their experience with other refined clinical decision-making skills like the importance of location of residue, management of residue, and diet recommendations. These important factors were not included in the present investigation. In summary, it is recommended to use numerous clinicians with a range of expertise to determine a reference standard of residue. This is a reasonable conclusion based on this study's findings, until other more valid measures of the amount of residue can be determined (Fig. 1).

The findings of the present study are not without limitations. Even though an average of nearly 2 weeks was provided between rating sessions and identifiers were removed from the videos, it is possible that clinicians recognized videos and their assigned rating, enabling higher intra-rater agreement. In the same way, the small groups could have created a setting where clinicians looked at each other's ratings, although clinicians were instructed to rate independently, discussion was not allowed, and the sessions were closely monitored. The intra-rater reliability testing consisted of only two exposures to videos, when ideally intra-rater reliability would include multiple exposures [32]. Additionally, we were unable to make direct comparisons of inter- and intra-rater reliability coefficients due to the necessary disparate statistical tests. Another limitation is that the group of experts was small. It would

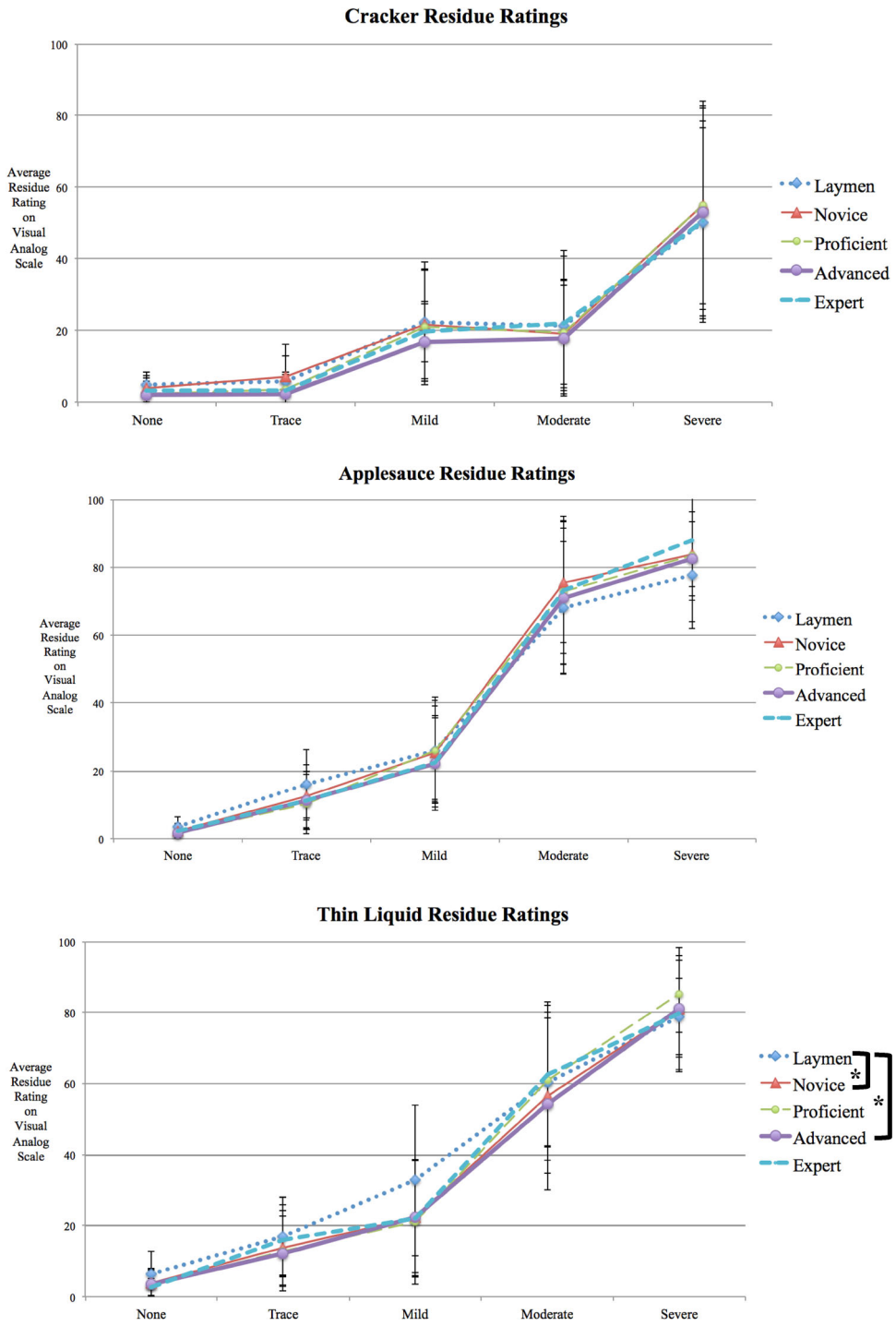

Fig. 1 Average visual analog scale ratings for all bolus severity types for cracker, applesauce, and thin liquid boluses. Raters are categorized as laymen, novice, proficient, advanced, and expert clinicians. Error bars are ± 1 standard deviation. *Indicates a significant difference

be interesting to see if a larger set of experts or an even-more specialized group with more than 20 or 25 years' experience would change the findings. The latter may not be possible, however, since FEES was first described in 1988 [15].

We also acknowledge that the present investigation analyzed the influence of experience on VAS ratings alone and not on the ordinal scale. This was intentional given our specific aims and intent to further study FEES residue ratings on a VAS. Differences in ratings between expert and non-expert ratings on an ordinal scale would be an interesting investigation for future research, although we hypothesize that there would continue to be no significant differences given a smaller set of choices and increased likelihood of agreement of an ordinal scale. Finally, the videos did vary in the number of clearing swallows, which

were spontaneous and not cued by the clinician. The frequencies were $n = 38$, $n = 19$, $n = 13$, and $n = 5$ for 0, 1, 2, and 3 clearing swallows. There were no videos containing more than 3 clearing swallows. We mention this because the number of clearing swallows could have influenced clinicians, although clinicians were asked to rate residue alone. Very little empirical data exist about clearing swallows and how many are considered normal versus abnormal, but recent analyses of the present's study data suggest that clinicians are reliable in their impressions of clearing swallows [29, 30].

Conclusion

When examining reliability of VAS and ordinal scales to measure residue on FEES, inter-rater reliability of both scales was excellent. However, intra-rater reliability was slightly better on the VAS. Thus, VAS provides optimal inter- and intra-rater reliability, likely because it affords the rater additional freedom to rate slight gradations that are not possible in ordinal scales. Overall, clinician experience with FEES did not influence ratings of residue when using a VAS scale. Expert ratings were not significantly different from other raters, even those completely unfamiliar with FEES. Training clinician judgment of overall amount of residue does not seem necessary to improve reliability if judgments are made immediately after the first swallow.

Acknowledgements The lead author would like to thank Dr. Elizabeth Hoover for her input and guidance in the development of this research. We are also grateful for the clinicians, raters, patients, and researchers who contributed to this work.

Compliance with Ethical Standards

Conflict of interest Salary and tuition support was provided to the first and last authors from the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Award Number R01DC012584 (PI: Kumar). The Department of Speech, Language, and Hearing of Sargent College (Boston University) also provided financial support for this research via the Dudley Allen Sargent Research Fund.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards (DEEMED EXEMPT BY IRB, as it did constitute research with human subjects, but ALSO meets the requirements of a defined low-risk category that is exempt from SOME (but not all) of the requirements governing human subjects research.).

Informed Consent The regulatory requirement for a consent form does not apply to Exempt research.

References

1. Altman DG. Practical statistics for medical research. 1st ed. London: Chapman and Hall; 1991.
2. Bernieri F. Coordinated movement and rapport in teacher-student interactions. J Nonverbal Behav. 1988;12:120–38.
3. Bernieri F, Resnick J, Rosenthal R. Synchrony, pseudo-synchrony, and dissynchrony: measuring the entrainments process in mother-infant dyads. J Pers Soc Psychol. 1988;54:243–53.
4. Brunier G, Graydon J. A comparison of two methods of measuring fatigue in patients on chronic haemodialysis: visual analogue versus Likert scale. Int J Nurs Stud. 1996;33(3):338–48.
5. Donzelli J, Brady S, Wesling M, Craney M. Predictive value of accumulated oropharyngeal secretions for aspiration during video nasal endoscopic evaluation of the swallow. Ann Otol Rhinol Laryngol. 2003;112(5):469–75. https://doi.org/10.1177/000348940311200515.
6. Farneti D. Pooling score: an endoscopic model for evaluating severity of dysphagia. Acta Otorhinolaryngol Ital. 2008;28(3):135–40.
7. Farneti D, Fattori B, Nacci A, Mancini V, Simonelli M, Ruoppolo G, Genovese E. The pooling-score (P-score): inter- and intra-rater reliability in endoscopic assessment of the severity of dysphagia. Acta Otorhinolaryngol Ital. 2014;34(2):105–10.
8. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problem of two paradoxes. J Clin Epidemiol. 1990;43:543–9.
9. Hildebrand DK, Laing JD, Rosenthal R. Analysis of ordinal data: quantitative applications in the social sciences. Newbury Park: SAGE Publications Inc; 1977.
10. Holmes S, Dickerson JW. Malignant disease: nutritional implications of disease and treatment. Cancer Metastasis Rev. 1987;6(3):357–81.
11. Hutcheson KA, Barrow MP, Barringer DA, Knott JK, Lin HY, Weber RS, Lewin JS. Dynamic imaging grade of swallowing toxicity (DIGEST): scale development and validation. Cancer. 2017;123(1):62–70. https://doi.org/10.1002/cncr.30283.
12. Kaneoka AS, Langmore SE, Krisciunas GP, Field K, Scheel R, McNally E, Cabral H. The boston residue and clearance scale: preliminary reliability and validity testing. Folia Phoniatr Logop. 2013;65(6):312–7. https://doi.org/10.1159/000365006.
13. Kelly AM, Leslie P, Beale T, Payten C, Drinnan MJ. Fibreoptic endoscopic evaluation of swallowing and videofluoroscopy: does examination type influence perception of pharyngeal residue severity? Clin Otolaryngol. 2006;31(5):425–32. https://doi.org/10.1111/j.1749-4486.2006.01292.x.
14. Kelly AM, Macfarlane K, Ghufoor K, Drinnan MJ, Lew-Gor S. Pharyngeal residue across the lifespan: a first look at what's normal. Clin Otolaryngol. 2008;33(4):348–51. https://doi.org/10.1111/j.1749-4486.2008.01755.x.
15. Langmore SE, Schatz K, Olsen N. Fiberoptic endoscopic examination of swallowing safety: a new procedure. Dysphagia. 1988;2(4):216–9.
16. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255–68.
17. Lyons KD, Tickle-Degnen L. Reliability and validity of a videotape method to describe expressive behavior in persons with Parkinson's disease. Am J Occup Ther. 2005;59(1):41–9.
18. Martin-Harris B, Brodsky MB, Michel Y, Castell DO, Schleicher M, Sandidge J, Blair J. MBS measurement tool for swallow impairment–MBSImp: establishing a standard. Dysphagia. 2008;23(4):392–405. https://doi.org/10.1007/s00455-008-9185-9.
19. Marvin S, Gustafson S, Thibeault S. Detecting aspiration and penetration using FEES With and without food dye. Dysphagia. 2016;31(4):498–504. https://doi.org/10.1007/s00455-016-9703-0.

20. McBride G. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. Hamilton, New Zealand: National Institute of Water and Atmospheric Research Ltd. 2005. http://www.medcalc.org/download/pdf/McBride2005.pdf.
21. Neubauer PD, Hersey DP, Leder SB. Pharyngeal residue severity rating scales based on fiberoptic endoscopic evaluation of swallowing: a systematic review. Dysphagia. 2016;31(3):352–9. https://doi.org/10.1007/s00455-015-9682-6.
22. Neubauer PD, Rademaker AW, Leder SB. The yale pharyngeal residue severity rating scale: an anatomically defined and image-based tool. Dysphagia. 2015;30(5):521–8. https://doi.org/10.1007/s00455-015-9631-4.
23. Nienstedt JC, Muller F, Niessen A, Fleischer S, Koseki JC, Flugel T, Pflug C. Narrow band imaging enhances the detection rate of penetration and aspiration in FEES. Dysphagia. 2017;32(3):443–8. https://doi.org/10.1007/s00455-017-9784-4.
24. Niessen A, Nienstedt J, Pflug C. Methodic background of narrow band imaging (NBI) in Dysphagia diagnostic-proposing a high sensitivity FEES. Paper presented at the Dysphagia Research Society, Barcelona. 2017.
25. Park JM, Yong SY, Kim JH, Jung HS, Chang SJ, Kim KY, Kim H. Cutoff value of pharyngeal residue in prognosis prediction after neuromuscular electrical stimulation therapy for Dysphagia in subacute stroke patients. Ann Rehabil Med. 2014;38(5):612–9. https://doi.org/10.5535/arm.2014.38.5.612.
26. Pfennings L, Cohen L, van der Ploeg H. Preconditions for sensitivity in measuring change: visual analogue scales compared to rating scales in a Likert format. Psychol Rep. 1995;77(2):475–80. https://doi.org/10.2466/pr0.1995.77.2.475.
27. Pilz W, Baijens LW, Passos VL, Verdonschot R, Wesseling F, Roodenburg N, Kremer B. Swallowing assessment in myotonic dystrophy type 1 using fiberoptic endoscopic evaluation of swallowing (FEES). Neuromuscul Disord. 2014;24(12):1054–62. https://doi.org/10.1016/j.nmd.2014.06.002.
28. Pisegna JM, Kaneoka A, Leonard R, Langmore SE. Rethinking residue: determining the perceptual continuum of residue on FEES to enable better measurement. Dysphagia. 2017. https://doi.org/10.1007/s00455-017-9838-7.
29. Pisegna JM, and Langmore S. Measuring residue: categorical ratings versus a visual analog scale. Paper presented at the Dysphagia Research Society Annual Convention, Chicago. 2015.
30. Pisegna JM, and Langmore S. Clinician judgments of clearing swallows are reliable without training. Paper presented at the American Speech Language and Hearing Association, Los Angeles. 2017.
31. Pisegna JM, Langmore SE. Parameters of instrumental swallowing evaluations: describing a diagnostic dilemma. Dysphagia. 2016;31(3):462–72. https://doi.org/10.1007/s00455-016-9700-3.
32. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. 3rd ed. Upper Saddle River: Pearson/Prentice Hall; 2015.
33. Rommel N, Borgers C, Van Beckevoort D, Goeleven A, Dejaeger E, Omari TI. Bolus residue scale: an easy-to-use and reliable videofluoroscopic analysis tool to score bolus residue in patients with Dysphagia. Int J Otolaryngol. 2015;2015:780197. https://doi.org/10.1155/2015/780197.
34. Rosenthal R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts; 1966.
35. Rosenthal R. Conducting judgment studies: Some methodological issues. In: The new handbook of methods in nonverbal behavior research. Oxford University Press; 2005. p. 199–234.
36. Tickle-Degnen L, Rosenthal R. The nature of rapport and its nonverbal correlates. Psychol Inq. 1990;1(4):285–93.
37. Tinsley H, Weiss D. Interrater reliability and agreement of subjective judgments. J Couns Psychol. 1975;22:358–76.
38. Tohara H, Nakane A, Murata S, Mikushi S, Ouchi Y, Wakasugi Y, Uematsu H. Inter- and intra-rater reliability in fibroptic endoscopic evaluation of swallowing. J Oral Rehabil. 2010;37(12):884–91. https://doi.org/10.1111/j.1365-2842.2010.02116.x.

Jessica M. Pisegna PhD, MS-CCC-SLP, MEd

James C. Borders MS-CCC-SLP

Asako Kaneoka PhD

Wendy J. Coster PhD, OTR/L, FAOTA

Rebecca Leonard PhD

Susan E. Langmore PhD, CCC-SLP, BCS-S