Studies in Second Language Acquisition (2026), 1–26
doi:10.1017/S0272263126101582

METHODS FORUM

Post hoc score category collapsing for L2 pronunciation research

Taichi Yamashita

Faculty of Foreign Language Studies, Kansai University, Osaka, Japan
Email: t.yamash@kansai-u.ac.jp

(Received 16 August 2025; Revised 01 January 2026; Accepted 18 January 2026)

Abstract

Second language (L2) pronunciation research has measured speech comprehensibility by asking listeners to assess L2 learners' speaking performance with rating scales. While some studies have provided validity evidence for these rating scales, few studies have examined the extent to which those scales effectively distinguish among L2 speakers. To fill this gap, the present study examines the 9-point scale used in Saito et al. (2020: *Annual Review of Applied Linguistics, 40*, 9–25.) and the 100-point scale in Huensch and Nagle (2023: *Studies in Second Language Acquisition, 45*(2), 571–585.) from a Rasch measurement perspective and showcases post hoc score category collapsing as a potential countermeasure against suboptimal rating scale functioning. Findings suggested that different score categories represented the same ability level and were therefore interchangeable. Collapsing these score categories yielded shorter but more functional scales without compromising the psychometric qualities of the original scales. These findings suggest that researchers need to empirically refine their scale lengths rather than uncritically following their conventional measurement practices.

Keywords: comprehensibility; pronunciation; rating scale; Rasch measurement; validation

Introduction

Second language (L2) pronunciation research has examined a range of constructs that manifest in learners' speaking performance, such as accentedness, comprehensibility, and intelligibility (Munro & Derwing, 2015). These studies have provided collective evidence of the interrelationships among different constructs, suggesting that, for example, accented speech is not necessarily incomprehensible (Chau & Huensch, 2025) and thereby that achieving nativelikeness should not be the ultimate goal for many L2 learners (Levis, 2005). In response to this paradigm shift, comprehensibility—defined as "perceived degree of difficulty experienced by the listener in understanding speech" (Munro & Derwing, 2015, p. 14)—has emerged as one of the key constructs

in L2 pronunciation research. Comprehensibility has often been measured by asking listeners to assess L2 learners' speaking performance with rating scales, such as 9-point (e.g., Saito et al., 2016; Suzuki & Kormos, 2020; Thorpe et al., 2025), 100-point (e.g., Huensch & Nagle, 2021; Nagle et al., 2022; Trofimovich et al., 2020), and 1,000-point scales[1] (e.g., Bergeron & Trofimovich, 2017; Crowther et al., 2015; Saito et al., 2017). For example, Saito et al. (2020) asked listeners to choose a score category[2] on a 9-point scale ranging from a score of 1 that represented *very difficult to understand* to a score of 9 that represented *very easy to understand*, while Huensch and Nagle (2023) used a 100-point scale where *very difficult to understand* and *very easy to understand* represented incomprehensible speech and comprehensible speech, respectively.

Validity evidence for these rating scales has been collected primarily by examining Cronbach's alpha and intraclass correlation that indicate interrater consistency among listeners (Kostromitina et al., 2025) and by examining the relationship between linguistic features and comprehensibility ratings (e.g., Isaacs & Trofimovich, 2012; Saito et al., 2016; Suzuki & Kormos, 2020; Tsunemoto & Trofimovich, 2024). Meanwhile, other studies have challenged the use of these rating scales because such scales present a substantially large number of score categories, which sometimes do not come with numerical values, and consequently cause raters' inconsistent scale use and suboptimal scale functioning (Isaacs & Thomson, 2013; Isbell, 2018; Kermad, 2024). Additionally, the presence of many score categories does not guarantee that scores can be treated as continuous variables because obtained scores are in principle ordinal unless equal intervals among adjacent scores are empirically observed (Isbell, 2018). Given that previous studies have provided both positive and negative evidence regarding long rating scales, it is reasonable to claim that further validation studies are still needed, rather than accepting these rating scales as *validated* instruments.

However, many L2 pronunciation studies still use rating scales with many score categories, justifying their scale length simply because previous studies used the same scale length without referring to validity evidence (e.g., Ali, 2023; Bergeron & Trofimovich, 2017; Crowther et al., 2015; Thorpe et al., 2025). This methodological convention can be concerning because researchers' arbitrary choice of scale length may influence their research findings. For example, Chau and Huensch (2025) reported that correlations among global constructs, such as comprehensibility and accentedness, tended to be higher for longer scales (e.g., 100- and 1,000-point scales) than for shorter scales (e.g., 5-, 7-, and 9-point scales). Therefore, empirical pursuit of a better rating scale has great potential to minimize the inconsistent research findings, potentially boosting replicability among primary studies.

Against this backdrop, the present study aims to provide further empirical evidence regarding potential limitations of long rating scales—defined here as scales containing nine or more score categories—used in L2 pronunciation research. The study first reviews previous studies in view of a validation framework, identifying the aspects of long rating scales for which more validity evidence is warranted. Then, the study reports on the reanalysis of the 9-point scale in Saito et al. (2020) and the 100-point scale in Huensch and

[1]Some scales contain 0 as the minimum score, and thus some 100-point and 1,000-point scales in fact contain 101 or 1,001 score categories. In the present study, such scales are referred to based on their maximum score (e.g., 100-point scale, 1,000-point scale) to maintain consistency with previous studies. For example, 100-point scales can refer to scales that contain 100 score categories and those that contain 101 score categories.

[2]Score categories refer to response options that are associated with scores to be assigned to examinees (see also Linacre, 2002; Toland & Usher, 2016; Tsai et al., 2025). The present study uses this term to distinguish response options on a rating scale and scores that are assigned to speakers.

Nagle (2023) as these two studies present dense, well-controlled rating data. Using these two datasets as a point of illustration, the present study further proposes post hoc score category collapsing as a countermeasure against suboptimal scale functioning in research contexts. Overall, in response to Saito and Plonsky's (2019) call, the study contributes to the empirical refinement of rating scales in L2 pronunciation research while also aiming to enhance researchers' measurement practices more broadly.

Literature review

Assessment of comprehensibility in L2 pronunciation research

L2 pronunciation research has measured L2 learners' comprehensibility using rating scales with varying granularity, such as 9-point, 100-point, and 1,000-point scales (Chau & Huensch, 2025; Kostromitina et al., 2025). According to Kostromitina et al.'s (2025) methodological review of perception-based L2 pronunciation studies, among the rating scales whose reliability was examined via Cronbach's alpha, the 9-point scale was the most frequently used ($n = 64$, 48%), followed by 1,000-point ($n = 34$, 26%), 7-point ($n = 20$, 15%), and 5-point scales ($n = 11$, 8%). A major reason for the widespread adoption of the 9-point scale can be the seminal paper by Munro and Derwing (1995), where they asked 18 L1 speakers of English to rate the comprehensibility of 10 L1 Mandarin speakers and two L2 speakers of English with a 9-point scale. In their study, a score of 1 was labeled as *extremely easy to understand*, while a score of 9 was labeled as *impossible to understand.* The rest of the score categories (i.e., 2 through 8) were visually presented but came with no performance descriptors. Adapting this 9-point scale in different ways, many studies defined a score of 9 as a representation of comprehensible speech and a score of 1 as incomprehensible speech. For example, Saito et al. (2020) labeled a score of 1 as *very difficult to understand* and a score of 9 as *very easy to understand*, and Isaacs and Trofimovich (2012) labeled a score of 1 as *hard to understand* and a score of 9 as *easy to understand*. This endpoint labeling has been applied to 100-point and 1,000-point scales. Huensch and Nagle (2023) labeled a score of 0 on their 100-point scale as *very difficult to understand* and a score of 100 as *very easy to understand.* Saito et al. (2017) labeled a score of 0 on their 1,000-point scale as *difficult to understand* and a score of 1,000 as *easy to understand.* Unlike the 9-point scale, where all possible scores from 1 through 9 are visually presented to raters, 100-point and 1,000-point scales often omit numerical markers, including those for the endpoints. Accordingly, raters were often asked to adjust the location of a slider bar to express their intuitive evaluation of comprehensibility, instead of choosing a discrete, prelabeled score (e.g., Crowther et al., 2015; Huensch & Nagle, 2021, 2023; Saito et al., 2017). These unlabeled scales require raters to construct their own benchmarks and constantly think about what it means to be a score of 5 out of 9 and a score of 50 out of 100, for example, while simultaneously being influenced by their teaching experience and accent familiarity (Saito et al., 2019; Tsunemoto et al., 2023; Winke et al., 2013), among other individual difference factors.

Validity evidence for these rating scales has been pursued primarily in two ways. First, many researchers examine interrater reliability among raters (e.g., Ali, 2023; Bergeron & Trofimovich, 2017; Crowther et al., 2015; Galante & Thomson, 2017; Saito et al., 2017; Suzuki & Kormos, 2020; Tergujeff, 2021), attempting to validate generalization inference that scores assigned by a rater are generalizable to other raters (Knoch & Chapelle, 2018). Although their methodological review focused on assessment instruments intended for accentedness, comprehensibility, and intelligibility, Kostromitina et al. (2025) found that approximately 48% of the reviewed instruments in L2 pronunciation

research reported Cronbach's alpha and 32% reported intraclass correlation coefficients. Those Cronbach's alpha values were largely sufficient, ranging from .85 (5-point scale) to .94 (9-point scale). Similarly, intraclass correlation coefficients also ranged from .91 (1,000-point scale) to .97 (100-point scale). This evidence supports the claim that scores obtained from one rater are generalizable to other raters.

Secondly, another line of research examined the relationship between comprehensibility ratings and linguistic correlates in L2 speakers' performance (Bergeron & Trofimovich, 2017; Isaacs & Trofimovich, 2012; Saito et al., 2016, 2017; Suzuki & Kormos, 2020; Tsunemoto & Trofimovich, 2024), validating the explanation inference that scores accurately represent the construct being measured (Knoch & Chapelle, 2018). Isaacs and Trofimovich's (2012) pioneering study asked 60 L1 speakers of English to assess the comprehensibility of picture description performances elicited from 40 L1 French learners of English. They also analyzed L2 learners' performance in terms of 19 linguistic features, reporting that comprehensibility ratings were strongly related to both phonological and lexico-grammatical features, such as unique words and word stress errors. This line of research provides evidence that scores reflect differences in the language use observed in speech samples.

However, L2 pronunciation research has provided limited evidence regarding the functioning of a scale. Commonly used reliability indices, such as Cronbach's alpha and intraclass correlation, do not provide direct evidence because those indices are no more than indicators of rater agreement (Isbell, 2018). When Cronbach's alpha is used to assess raters' agreement, different raters are considered to be fixed test items; thus, Cronbach's alpha indicates the extent to which different raters measure the same construct (i.e., unidimensionality). Regarding intraclass correlation, pronunciation researchers often use two-way consistency, average-measure intraclass correlation known as ICC (2, k) (e.g., Ali, 2023; Huensch & Nagle, 2023; Isbell, 2018; Nagle & Rehman, 2021). This type of intraclass correlation assumes that raters are randomly drawn from a pool of raters and is appropriate when raters' averaged scores are of interest (Shrout & Fleiss, 1979). ICC(2, k) partitions variance attributable to examinees and variance attributable to raters, better reflecting raters' consistency than Cronbach's alpha (Isbell, 2018). While commonly used among L2 pronunciation researchers, both Cronbach's alpha and intraclass correlation deal with obtained ratings that are decontextualized from the scale design. Suppose that one rater assigns scores of 1, 3, and 9 and another rater assigns scores of 2, 4, and 6 for the same three examinees. Cronbach's alpha is .86, and ICC(2, k) is .90. However, these ratings can be obtained either with a 9-point scale or with a 100-point scale. When a 9-point scale is used, raters appear to make use of diverse score categories (i.e., functional scale), but they use a narrow range of score categories on a 100-point scale (i.e., dysfunctional scale). As this example illustrates, while Cronbach's alpha and intraclass correlation inform us about rater agreement, the functionality of a rating scale remains unknown unless ratings are interpreted in relation to the scale design.

Examining scale functioning pertains to evaluation inference that "Observations are evaluated using procedures that provide observed scores with intended characteristics" (Knoch & Chapelle, 2018, p. 483). Validating evaluation inference requires evidence that raters use the given rating scale as intended and that the rating scale functions as intended. For example, when raters are asked to use a 9-point scale, a score of 8 needs to represent more comprehensible speech than that represented by scores up to 7. Isaacs and Thomson's (2013) pioneering study tested this assumption by asking 40 novice and experienced raters to use 5-point and 9-point scales to assess comprehensibility of narrative speeches performed by 38 L2 speakers of English. They found that the 9-point scale did not reliably distinguish L2 speakers who were assigned different scores,

especially due to raters' confusion caused by the large number of score categories. Consequently, L2 speakers who were assigned middle score categories were particularly muddled, and speakers who were assigned a score of 2 were also likely to receive a score of 5 on the 9-point scale. Similarly, Isbell (2018) asked 10 L1 speakers of Korean to use a 9-point scale to assess picture description and read-aloud performances elicited from 36 L2 learners of Korean, finding that the middle scores, such as 4 and 6, were essentially interchangeable with adjacent score categories. More recently, Kermad (2024) compared 30 untrained raters and 30 trained raters in their use of a 5-point scale while rating prompted speeches performed by 20 L2 learners of English. The study found that untrained raters did not use score categories appropriately, observing that the scores of 3 and 4 were essentially interchangeable. Trained raters, in contrast, used the rating scale as intended, assigning different scores to speakers who displayed speaking performance that differed in comprehensibility. These findings collectively suggest that long rating scales may be overly segmented and thereby impose greater cognitive load on raters, especially due to the presence of potentially redundant mid-points.

Despite potential confusion among raters caused by long rating scales, such scales continue to be used in L2 pronunciation research (Kostromitina et al., 2025). Indeed, some studies did not explicitly justify their scale length (e.g., Galante & Thomson, 2017; Isaacs & Trofimovich, 2012; Saito et al., 2019), and others justified their scale length on the grounds that previous studies had used the same length (e.g., Ali, 2023; Bergeron & Trofimovich, 2017; Crowther et al., 2015; Suzuki & Kormos, 2020; Thorpe et al., 2025; Uchihara et al., 2023). A few studies provided more concrete justifications, stating that long rating scales were expected to provide raters with sufficient response options (Tergujeff, 2021) and to enable scores to be treated as continuous variables (Huensch & Nagle, 2021), for example. It is noteworthy that Nagle and Rehman (2021) is one of the few studies that explicitly drew on the rebuttal evidence from Isaacs and Thomson (2013) and opted for a 7-point scale as an eclectic solution to address the drawbacks of 5-point and 9-point scales. While potentially reflecting the needs of their own research context on one hand, researchers' arbitrary choice of scale on the other hand makes it difficult to draw solid conclusions due to variations in scale design.

As Isbell (2018) rightly pointed out, perhaps long rating scales have been tacitly justified because many studies primarily aimed to examine laypeople's intuitive, impressionistic assessment of comprehensibility. Yet, this persistent methodological convention may also be ascribed to a lack of communication between second language acquisition (SLA) researchers and language testing researchers (Isaacs & Thomson, 2013). In fact, language testing researchers emphasize that rating scales need to undergo a sequence of validation processes that go far beyond calculating interrater reliability (Knoch & Chapelle, 2018). In addition, language testing researchers often use Rasch models that allow closer examination of rater behavior and scale functioning (McNamara et al., 2019), but such an approach has rarely been used in L2 pronunciation research (Kostromitina et al., 2025). Although a few studies employed Rasch models, these studies mainly focused on speaker ability and listener severity without fully examining scale functioning (Nagle & Rehman, 2021; Shintani et al., 2019). Clearly, more validity evidence is needed to justify researchers' use of long rating scales.

Ordered thresholds and threshold distances as psychometric qualities of a rating scale

Two critical assumptions for a functional rating scale are "scale steps are adequate to distinguish among the levels that appear in the scale" and "raters are able to identify

differences in performances across score levels" (Knoch & Chapelle, 2018, p. 483). These assumptions have been commonly examined via a Rasch measurement approach (Linacre, 2002), which allows the investigation of two key psychometric properties—ordered thresholds and threshold distances. First, the Rasch-Andrich threshold, or "the location corresponding to the equal probability of observing adjacent categories *k-1* and *k*" (Linacre, 2002, p. 88), needs to increase for higher score categories (Linacre, 1999). Thresholds, also referred to as step calibrations, are estimated by polytomous Rasch models, such as the partial credit model (Masters, 1982) and the rating scale model (Andrich, 1978). They indicate the relative probability of receiving a score versus a score one point lower. Negative threshold measures suggest that the higher score category is more likely to be observed than the lower score category, whereas positive threshold measures indicate that the lower score category is more likely to be observed. When threshold measures for higher score categories exceed those for lower score categories across any pairs of two adjacent score categories on a rating scale, such evidence supports the claim that different scores meaningfully reflect differences in L2 speakers' ability level. Conversely, when threshold measures for lower score categories exceed those for higher score categories, the rating scale is said to exhibit threshold disordering. Threshold disordering indicates that a score category captures an overly limited range of the latent trait or represents a concept that is not well defined in respondents' minds (Linacre, 2002). In practice, threshold disordering often occurs when a scale has too many categories relative to the ability distribution, or when raters struggle to link certain score categories to specific performance levels.

Another essential condition for a functional rating scale is that obtaining a certain score is sufficiently more difficult than obtaining the score one point lower. When two adjacent categories are equally likely to be chosen for the same examinee, it indicates that the same examinee's ability can be represented by two score categories that are potentially interchangeable. Threshold distances can be calculated by the difference in logits between the Rasch-Andrich threshold of a score and that of the score one point lower. According to Linacre (2002), the optimal threshold distance falls within 1.4 to 5.0 logits. Ordered thresholds and threshold distances can also be visually examined in probability curves, which depict the changing likelihood of each score category across examinees' ability levels. Score categories are said to function properly when each score category displays a distinct peak in its probability curve.

There are two good reasons to assume threshold disordering and narrow threshold distances in the long rating scales used for speech comprehensibility. First, because it is usually difficult for a study to recruit learners with a wide range of proficiency, the sampled learners may display relatively homogeneous comprehensibility that cannot be meaningfully categorized into nine or more score categories. Indeed, Isaacs and Thomson (2013) reported that raters perceived redundancy of some score categories of their 9-point scale (see also Isbell, 2018; Kermad, 2024). Such redundancy potentially forces raters to choose a score category that does not precisely match the ability level constructed in the raters' minds (Linacre, 2002). Secondly, for 100-point and 1,000-point scales, raters are asked to assign a score without knowing the exact score they are assigning because numerical values are often hidden on the scale (Bergeron & Trofimovich, 2017; Crowther et al., 2015; Huensch & Nagle, 2021; Nagle et al., 2022; Saito et al., 2017; Trofimovich et al., 2020). For example, in Crowther et al.'s (2015) 1,000-point scale, no numerical values were presented except for brief descriptions at the endpoints (i.e., 1 = *hard to understand*, 1,000 = *easy to understand*). As such, the relative difficulty of score categories may rest upon not only raters' intuitive assessment of comprehensibility but also their operation of the slide bar. For example, raters may

accidentally place the slide bar at a score of 60 for a speech performance that they think is better than another speech performance they assigned a score of 61 as these two score categories are placed extremely close on the continuum. Indeed, though they used a 1,000-point scale for audio-based measures, such as segmental errors and word stress errors, Saito et al. (2016) reminded raters that even a slight adjustment of the slider could cause a substantial change in the rating, implying the critical role played by raters' delicate operation of the slide bar. When threshold disordering and narrow threshold distances are observed, interpreting scores becomes difficult as higher scores do not necessarily represent higher ability levels.

Score category collapsing as a potential solution

One way to address disordered thresholds and narrow threshold distances is score category collapsing, whereby two or more score categories on a rating scale are merged into a single score category (Linacre, 2002). For example, when Rasch models suggest that scores of 2 and 3 are disordered on a 9-point scale (i.e., 123456789), those two score categories can be merged into a single score category, resulting in a new 8-point scale (i.e., 122345678). When scores of 2 and 3 and scores of 7 and 8 are disordered, both pairs can be merged, creating a new 7-point scale (i.e., 122345677). Score category collapsing allows collapsed score categories to represent more distinct ability levels by aggregating minimally different ability levels represented by original score categories. As score categories are collapsed after data collection, L2 pronunciation researchers can still use conventional long rating scales when they design their study, assuring the compatibility of their study with previous studies. Furthermore, while score category collapsing usually requires one to redefine performance descriptors for newly created score categories, comprehensibility rating scales usually come only with the endpoint labeling (e.g., 1 = *hard to understand*, 9 = *easy to understand*) and omit descriptors for other score categories. Accordingly, researchers can still interpret obtained scores as listeners' intuitively perceived comprehensibility of heard speech samples.

Score category collapsing is not a completely new idea in L2 pronunciation research. Isaacs and Trofimovich (2013) mentioned the possibility of combining the score categories of 3, 4, 5, and 6 in a 9-point scale for comprehensibility, and Isbell (2018) mentioned the possibility of combining the score category of 7 with adjacent score categories in a 9-point scale for accentedness. However, very few studies have empirically tested its impact on measurement quality, and to my knowledge, Kermad and Bogorevich (2022) is the only study. They asked 56 raters, including L1 and L2 speakers of English, to assess speech performances elicited by four tasks—read-aloud, spontaneous speech, elicitation, picture description—which were performed by 15 L2 learners of English. Raters used a 9-point scale (i.e., 123456789) to assess comprehensibility and accentedness. The study collapsed score categories based on statistical information derived from the many-facet Rasch measurement analysis. Specifically, they combined the score categories of 3 and 4 and the score categories of 6 and 7, creating a 7-point scale (i.e., 123345567). They also combined the score categories of 1 and 2, 3 and 4, 6 and 7, and 8 and 9, creating a 5-point scale (i.e., 112234455). The results suggested that the score categories in the 5-point scale represented different ability levels more clearly than those in the 7-point and 9-point scales. Meanwhile, the midpoint of the 5-point scale still did not function well, being muddled with adjacent score categories. While Kermad and Bogorevich's (2022) study demonstrated score category collapsing as a potential solution, they tested only two collapsing patterns, leaving other collapsing

patterns to be further explored. In addition, other popular rating scales, such as 100-point and 1,000-point scales, have yet to be examined, though these substantially long rating scales are arguably more likely to confuse raters than a 9-point scale. Expanding Kermad and Bogorevich's (2022) pioneering study, the present study aims to illustrate the impact of score category collapsing on psychometric quality, using 9-point and 100-point scales as illustrative cases.

Research questions

The literature review suggests that many L2 pronunciation studies have conventionally used long rating scales, despite rebuttal evidence such as that presented by Isaacs and Trofimovich's (2013) pioneering study. The present study thus aims to raise researchers' awareness of the potential drawbacks of long rating scales by examining key psychometric properties—ordered thresholds and threshold distances. I chose 9-point and 100-point scales because Saito et al. (2020) and Huensch and Nagle (2023) provide well-controlled, rich datasets collected with these two scales (see Methods for more details). I have no intention to exclusively problematize these two scales nor to generalize the present findings to other scales in L2 pronunciation research. Rather, my purpose is to provide empirical evidence that demonstrates the potential limitations of long scales by using 9-point and 100-point scales as illustrative cases. The present study also aims to showcase post hoc score category collapsing as a potential countermeasure against suboptimal scale functioning. With these overarching goals, the present study addresses the following research questions:

RQ1: To what extent do 9-point and 100-point scales function effectively when used to assess comprehensibility in terms of scale functioning indices (i.e., threshold ordering, threshold distances, and probability curves)?

RQ2: To what extent does score category collapsing improve the functioning of 9-point and 100-point scales in terms of scale functioning indices (i.e., threshold ordering, threshold distances, and probability curves)?

RQ3: To what extent are collapsed scales compatible with the original 9-point and 100-point scales?

Method

Datasets in Saito et al. (2020) and Huensch and Nagle (2023)

The present study reanalyzed the dataset in Saito et al. (2020) and the dataset in Huensch and Nagle (2023). Saito et al. (2020) used a 9-point scale, and Huensch and Nagle (2023) used a 100-point scale. These studies were chosen for three major reasons. First, both datasets contain relatively large samples of L2 speakers and first language (L1) listeners. Saito et al. (2020) recruited 110 L2 English speakers and 10 L1 English raters, while Huensch and Nagle (2023) recruited 42 L2 Spanish speakers and 80 L1 Spanish raters. These large samples provide rich rating data, which is essential for Rasch analysis. Second, both 9-point and 100-point scales are widely used among L2 pronunciation studies (Chau & Huensch, 2025; Kostromitina et al., 2025), and thus the present findings were expected to be relevant to many previous studies. Third, Saito et al. (2020) and Huensch and Nagle (2023) differed in several important respects, including their targeted L2 (i.e., English vs. Spanish), speaking task (i.e., picture description vs. prompted

speech),[3] and data collection method (i.e., in-person vs. Amazon Mechanical Turk). This methodological variation allowed the present study to demonstrate the effectiveness of score category collapsing for different research contexts.

It should be noted that the present methodological showcase would not have been possible without these researchers' pioneering commitment to open science. Also, I have no intention to challenge any research claims made in these studies; rather, my purpose is to propose a methodological approach that allows us to examine long rating scales from new perspectives and to spark constructive discussions among L2 pronunciation researchers.

Analysis of Saito et al.'s (2020) dataset derived from 9-point scale

In Saito et al. (2020), 110 L2 learners of English performed a picture description task, and their comprehensibility was rated by 10 L1 speakers of English based in the UK. L2 speakers represented diverse L1 backgrounds (e.g., Italian, German, and Bengali), ranging in age from 20 to 59. Some had little to no experience of practicing English in classroom contexts, whereas others reported up to 23 years of study. In the picture description task, L2 speakers were given five seconds to prepare and then asked to describe seven pictures while using three key words pertaining to each picture. The first four pictures were used for L2 speakers to practice, and their performance on the last three pictures was used as the final data. The first 10 seconds trimmed from each of the three comprised one single audio file, which was subsequently rated by 10 L1 English speakers. The rating scale contained nine score categories from 1 (*very difficult to understand*) to 9 (*very easy to understand*) (i.e., 123456789). Raters first familiarized themselves with the picture description prompt and practiced rating with sample audio files. Afterward, they were asked to rate 110 audio files, listening to each file once. This rating design yielded a total of 1,100 ratings (i.e., 110 L2 speakers times 10 L1 raters).

To answer the first research question (i.e., threshold ordering, threshold distance, probability curves of the original 9-point scale), the present study conducted the many-facet Rasch measurement (MFRM) analysis, specifying speaker and rater as two facets. The speaker facet was centered (i.e., average speaker ability was set at 0.0 logit), and the rater facet was not centered. This configuration of facets is essential to find unique estimates and ensure identifiability, and it is customary that the one focal facet is noncentered while the other facets are centered (see Linacre, 2025, pp. 161–162 for more details). For the second research question (i.e., the impact of score category collapsing on threshold ordering, threshold distances, probability curves of the original 9-point scale), the present study explored the four collapsing schemes in Figure 1 (i.e., 111222333, 112234455, 112223344, 112233344). The schemes of 111222333 and 112234455 were explored because they retain the mid-point option in the original 9-point scale (Tsai et al., 2025). The scales of 112223344 and 112233344 were tested because previous studies reported muddled score categories around the mid-point on a 9-point scale (Isaacs & Trofimovich, 2013; Isbell, 2018). Aggregating three score categories around the mid-point, especially scores of 3, 4, and 5 and scores of 5, 6, and 7, was thus expected to optimize the

[3]A reviewer rightly pointed out the possibility of incorporating task as another facet in the MFRM analysis and thereby taking into account the difficulty of the task. While this is potentially methodologically defensible on one hand, the inclusion of another facet was expected to blur the focus of my study, which is to examine the functioning of a scale. The primary purpose of analyzing two tasks separately but reporting them together in the same paper was to demonstrate the impact of score category collapsing in different research contexts.

	Original score categories								
Scale	1	2	3	4	5	6	7	8	9
3-point	1			2			3		
5-point	1		2		3	4		5	
4-point (112223344)	1		2			3		4	
4-point (112233344)	1		2			3		4	

Figure 1. Collapsing schemes tested for the 9-point scale in Saito et al. (2020).

functioning of score categories. To answer the third research question (i.e., the compatibility between the original 9-point scale and collapsed scales), the present study compared the original 9-point scale and collapsed scales across different aspects, such as speaker statistics and rater statistics, to provide direct evidence regarding their compatibility.

Analysis of Huensch and Nagle's (2023) dataset derived from 100-point scale

Huensch and Nagle (2023) recruited 42 L2 learners of Spanish who were enrolled in Spanish courses at two universities in the United States. These L2 learners were all L1 speakers of English. Their speech performance was elicited by a prompted task, where learners were asked to speak for about one minute on a given topic after planning. From each learner, two utterances were extracted and submitted to the rating session. Unlike Saito et al. (2020), Huensch and Nagle (2023) conducted a rating session via Amazon Mechanical Turk, through which they recruited 80 L1 speakers of Spanish based in Spain, Venezuela, Mexico, Colombia, and Argentina. These raters first underwent practice rating items and then assessed speech samples. They were allowed to listen to each audio file one time and given 45 seconds to transcribe the utterance and assess its comprehensibility and accentedness. The rating scale for comprehensibility contained 101 score categories, ranging from 0 to 100 in increments of one point. The left endpoint was labeled as *very difficult to understand*, and the right endpoint was labeled as *very easy to understand.* The rating scale did not present numerical values to raters, but their placements of the slider were converted to scores from 0 to 100.

To answer the first research question (i.e., threshold ordering, threshold distance, probability curves of the original 100-point scale), the present study conducted the MFRM analysis that specified speaker (centered) and rater (noncentered) as two facets. To answer the second research question (i.e., the impact of score category collapsing on threshold ordering, threshold distance, probability curves of the original 100-point scale), the present study employed Rasch-based score category collapsing (Kermad, 2024; Linacre, 2002) because, unlike a 9-point scale, it was impractical to explore many possible collapsing patterns for the 100-point scale.[4] The present study screened score

[4]The results of score category collapsing cannot be reasonably compared between Huensch and Nagle's (2023) 100-point scale and Saito et al.'s (2020) 9-point scale because these two scales were collapsed in different ways. I attempted the Rasch-based collapsing for the 9-point scale, finding that the collapsing schemes I report in this paper facilitated the scale functioning more effectively. Also, my purpose was not to make side-by-side comparisons between Saito et al. (2020) and Huensch and Nagle (2023); rather, my purpose was to explore best collapsing patterns separately for 9-point and 100-point scales as these scales were designed differently.

	Original score categories										
Scale	0	1-10	11-20	21-25	26-40	41-60	61-70	71-80	81-85	86-90	91-100
10-point	0	1	2	3	4	5	6	7	8	9	10
5-point	0	1	2			3	4		5		
4-point	0	1	2			3			4		

Figure 2. Collapsing schemes for the 100-point scale in Huensch and Nagle (2023).

categories based on three criteria: (1) whether Rasch-Andrich threshold measures monotonically increased as score categories increased, (2) whether threshold distances between adjacent score categories were positive and within the recommended range (i.e., from 1.4 to 5.0 logits), and (3) whether each score category displayed a clear peak in its probability curve (Linacre, 2002). Score categories that failed to meet these criteria were combined, and the MFRM analysis was rerun iteratively until adequate scale functioning was achieved. Through these iterations, the present study tested 10-point, 5-point, and 4-point collapsed scales as presented in Figure 2.

To answer the third research question (i.e., the compatibility between the original 100-point scale and the collapsed scale), the present study compared the original 100-point scale and a collapsed scale across several dimensions (e.g., variable map, speaker statistics, rater statistics). It should be noted that Huensch and Nagle (2023) extracted two audio files from each speaker's performance, but the present study randomly sampled one of the two files from each speaker and submitted it to the analysis to ensure independent observations in the final dataset.[5] As such, the final dataset contained a total of 3,360 ratings (i.e., 42 speakers times 80 raters). The MFRM analysis was performed in the FACETS program 4.3.3. (Linacre, 2025), and all the output files are available in the OSF platform (https://osf.io/8jh3u/overview?view_only=01e99a1d7f09445c9e07a657f5a26a0c).

Results of Saito et al.'s (2020) 9-point scale

Functioning of the original 9-point scale (RQ1)

Table 1 summarizes the statistics regarding the functioning of the nine score categories in Saito et al. (2020). The number of times a score category being chosen monotonically increases from 1 (19 times, 2%) to 7 (204 times, 19%) with a slight drop at the scores of 8 (196 times, 18%) and 9 (160 times, 15%). Measure indicates the difficulty of test takers obtaining the score category, suggesting that the difficulty linearly increases as raw scores increase from 1 (–.80 logits) to 9 (2.35 logits). Outfit suggests the extent to which the score category conforms to the specified Rasch model, and values below 2.0 are generally considered acceptable (Linacre, 2002). The nine score categories all meet this criterion. Threshold measures—points at which the likelihood of test takers obtaining two adjacent score categories is equal—monotonically increase as raw scores increase from 2 (–1.54 logits) to 9 (2.07 logits). All distances between adjacent Rasch-Andrich

[5]If those two utterances had been elicited by two different tasks, I could have specified the task facet in the MFRM analysis, for example. However, two files are not associated with any variable in Huensch and Nagle (2023). Accordingly, this data cleaning is essential as the MFRM assumes observations are independent of each other (Wright, 1991).

Table 1. Score category statistics of the original 9-point scale

Score	Count	%	Measure	Outfit	Threshold		Threshold distance
					Measure	SE	
1	19	2	–0.80	1.4	NA	NA	NA
2	38	3	–0.59	1.3	–1.54	0.25	NA
3	78	7	–0.41	1.1	–1.30	0.16	0.24
4	113	10	–0.21	1.1	–0.69	0.11	0.61
5	137	12	–0.02	0.9	–0.24	0.10	0.45
6	155	14	0.41	0.9	0.14	0.09	0.38
7	204	19	0.87	0.9	0.37	0.08	0.23
8	196	18	1.50	0.8	1.19	0.09	0.82
9	160	15	2.35	0.9	2.07	0.11	0.88

Note. Count = the number of times the score category being chosen, Measure = difficulty of test takers obtaining the score, outfit = fit to observed data, Threshold measure = test taker ability at which obtaining two adjacent score categories is equally likely, Threshold SE = standard error of threshold measure, Threshold distance = the difference between two adjacent thresholds.

thresholds were positive, suggesting that obtaining higher scores required higher ability levels. However, threshold distances were relatively narrow for score categories from 3 to 7, ranging from .23 to .61 logits. In contrast, threshold distances for 8 and 9 were relatively large, being .82 and .88 logits, though falling short of the recommended lower bound of 1.4 (Linacre, 2002).

This suboptimal scale functioning is also evident in the probability curves (Figure 3), where the probability of receiving each score (y-axis; 0.0 to 1.0) is plotted as a function of speakers' ability (x-axis; –4.0 to 4.0 logits). For example, speakers

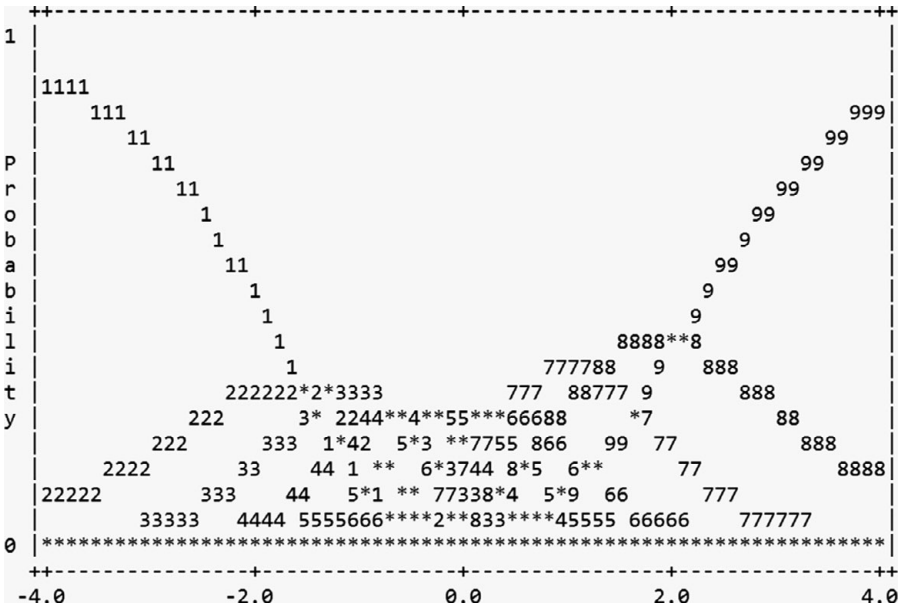

Figure 3. Probability curves of the original 9-point scale visualizing the probabilities of score categories (y-axis) as a function of examinee ability (x-axis).

associated with 1.5 logits on the x-axis likely receive a score of 7. However, many score categories from 2 to 7 fail to display clear peaks. As shown in the curves, speakers around 0 logits are equally likely to receive scores of 3, 4, 5, 6, and 7, which suggests that raters assigned different scores to speakers who did not substantially differ in comprehensibility.

Functioning of rating scales derived from score category collapsing (RQ2)

Table 2 presents Rasch-Andrich thresholds and threshold distances for the four rating scales derived from score category collapsing. Threshold measures monotonically increased from low to high score categories for the 3-point and 4-point scales (i.e., ordered thresholds). Meanwhile, in the 5-point scale, the threshold measure for the score category of 3 (.26 logits) was higher than that for the score category of 4 (–.28 logits), indicating disordering between these two score categories (see the Measure column of 112234455 in Table 2). Threshold distances in the 3-point and 4-point scales all fell within the recommended range from 1.4 to 5.0 (Linacre, 2002), suggesting that each score category reflected a distinct difficulty level. In contrast, the 5-point scale contained one negative distance between the score categories of 3 and 4 (–.54 logits), further confirming their disordering.

Figure 4 presents the probability curves for the four rating scales derived from score category collapsing. Most score categories display distinct peaks, suggesting that those score categories represent distinct ability levels. Meanwhile, the score category of 3 in the 5-point scale remained muddled with adjacent score categories, which suggests that L2 speakers assigned a score of 3 are also likely to receive a score of 2 or 4.

Comparing the original 9-point scale with the collapsed scales (RQ3)

Figure 5 presents the variable maps for the original 9-point scale and the three collapsed ratings scales that met the criteria for thresholds and probability curves. The leftmost column denotes the common logit scale that applies to the speaker, rater, and scale. The second column presents the distribution of speakers' ability, where higher values toward the top of the column represent more able speakers. The third column visualizes the distribution of rater severity, where higher values toward the top of the column represent more severe raters. The rightmost column shows each score category and its coverage of measures.

According to these variable maps, it seems that speakers' abilities became more distinct for the collapsed rating scales. For example, speakers are clustered closely

Table 2. Score category statistics of the four collapsed scales

	111222333		112234455		112223344		112233344	
Score	Measure	Distance	Measure	Distance	Measure	Distance	Measure	Distance
1	NA	NA	NA	NA	NA	NA	NA	NA
2	–1.46	NA	–1.95	NA	–2.76	NA	–2.06	NA
3	1.46	2.92	0.26	2.21	0.40	3.16	–0.59	1.47
4			–0.28	–0.54	2.36	1.96	2.65	3.24
5			1.98	2.26				

Note. Measure = test taker's ability at which obtaining two adjacent score categories is equally likely, Distance = the difference between two adjacent thresholds.

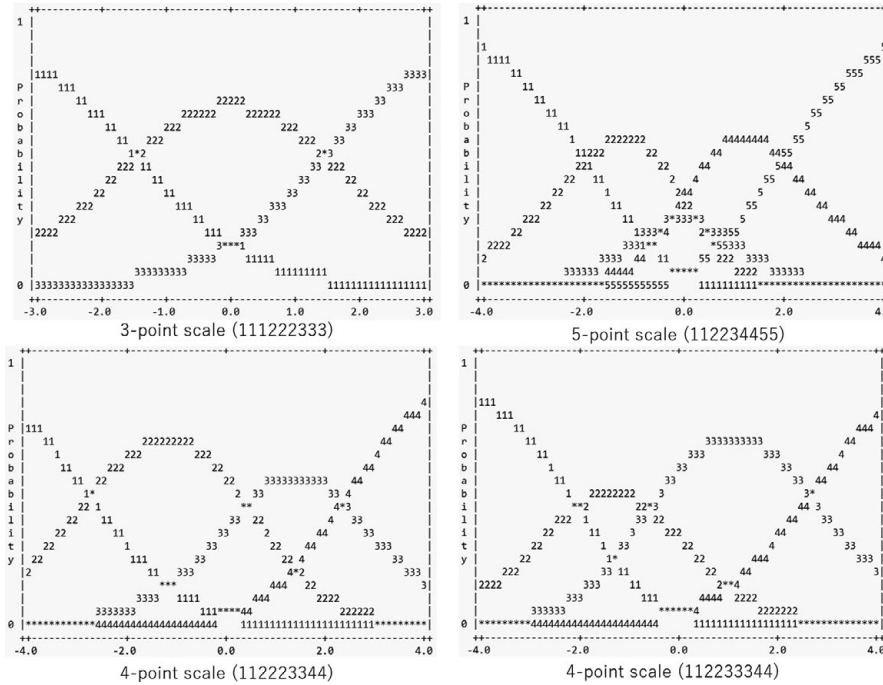

Figure 4. Probability curves for the 3-point (111222333, top left), 5-point (112234455, top right), 4-point (112223344, bottom left), and 4-point scales (112233344, bottom right).

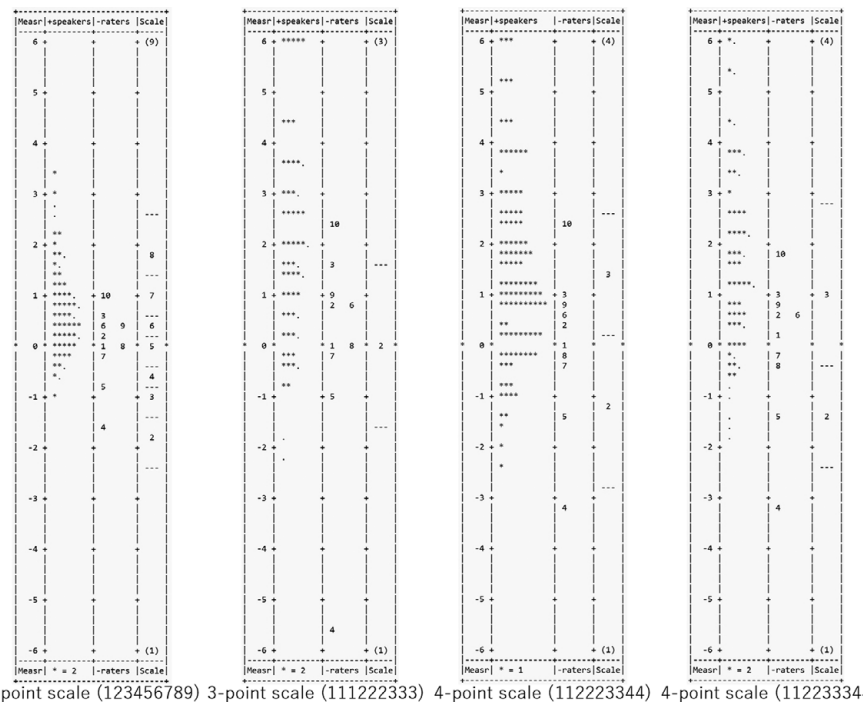

Figure 5. Variable maps for the original 9-point scale and the collapsed 3-point, 4-point (112223344), and 4-point (112233344) scales.

Table 3. Comparison among the original 9-point, 3-point (111222333), 4-point (112223344), and 4-point (112233344) scales

Criteria	123456789	111222333	112223344	112233344
Cronbach's alpha[6]	.89 [.86, .91]	.84 [.79, .87]	.87 [.84, .90]	.86 [.81, .89]
Two-way consistency, average-measure intraclass correlation	.89 [.85, .92]	.82 [.77, .87]	.86 [.82, .90]	.84 [.79, .88]
Variance explained by Rasch	63.59%	52.82%	60.80%	54.75%
Standardized residuals over |2.0|	3.2%	3.8%	4.2%	4.5%
Stara for person ability	3.13	2.97	3.79	3.65
Separation reliability for person	.91	.80	.87	.86
Increasing thresholds	Yes	Yes	Yes	Yes
Threshold distances over 1.4	No	Yes	Yes	Yes
Peaks in probability curves	No	Yes	Yes	Yes
Raters with infit over 1.5	1	0	0	0
Strata for rater severity	8.35	7.98	11.76	10.75
Separation reliability for rater	.97	.97	.99	.98

together on the 9-point scale—particularly those with logits between –0.5 and 2.0—whereas the 3-point scale seems to cluster speakers associated with logits of 6.0, 4.5, 3.5, 3.0, and so on. The ranking of rater severity does not seem to differ across the scales. Rater 10 was distinctively severe, while Raters 4 and 5 were lenient. The rest of the raters largely fell within a range from –1.0 to 1.0 logits. As the scale collapsed, each score category covered a wider range of speakers' ability levels. On the 9-point scale, the score categories from 3 to 6 each covered approximately 0.5 logits at most, whereas the score categories of 7, 8, and 9 each represented a wider range of ability levels. In contrast, the 4-point scale of 112233344 (the rightmost figure) allowed each score category to cover at least 2.0 logits of the speaker's ability. In short, echoing the observations from their probability curves, these variable maps indicate that the collapsed scales pool minimally different ability levels into one score category and thereby link each score category with a distinctly different ability level.

Table 3 compares the original 9-point scale and the three collapsed rating scales that met the criteria for thresholds and probability curves. It seems that these scales are not substantially different in terms of Cronbach's alpha and intraclass correlation, as their 95% confidence intervals largely overlap, mostly falling within the range from .80 to .90. As for the model fit, the original 9-point scale and the collapsed 4-point scale of 112223344 explained approximately 60% of the variance, whereas approximately 50% of the variance was explained by the 3-point scale and the other 4-point scale of 112233344. Rasch residuals over |2.0| were minimal across the scales. These explained variances and minimal residuals suggest that the Rasch model adequately fits the observed data across the four scales.

Regarding speaker ability, all the scales were associated with speaker strata of approximately 3, indicating there were three statistically distinguishable ability levels. Separation reliability for person, which can be interpreted as Cronbach's alpha in classical test theory (Linacre, n.d.), was sufficient, ranging from .80 to .91. As for rater severity, the original 9-point scale and the collapsed 3-point scale found eight rater severity levels, but the 4-point scales found 11 rater severity levels. Separation reliability for rater was high for all four scales, being above .95. These statistics suggest that while the four scales reliably identified three distinct ability levels among speakers, the 4-point scales distinguished rater severity more clearly than the 9-point and 3-point scales.

[6]Cronbach's alpha and intraclass correlation can be redundant in a Rasch context. My intention to include these reliability indices is to highlight the fact that these indices were not negatively impacted by score category collapsing.

Person measures or estimated speakers' ability derived from the collapsed scales were highly correlated with those from the original scale. Pearson's product-moment correlations were .97 (95% CI[.95, .98]) for the 3-point scale, .99 (95% CI[.98, .99]) for the 4-point scale of 112223344, and .99 (95% CI[.98, .99]) for the 4-point scale of 112233344. These results suggest that ability estimates derived from the collapsed scales were highly compatible with those derived from the original scale. In short, score category collapsing helped score categories to represent more distinct ability levels than those originally represented by the nine score categories, without compromising the psychometric properties of the original 9-point scale.

Results of Huensch and Nagle's (2023) 100-point scale

Functioning of the original 100-point scale (RQ1)

Out of 99 possible threshold distances, 46 distances were positive (i.e., ordered thresholds), and 53 distances were negative (i.e., disordered thresholds). Table 4 presents the 10 score categories that were associated with the highest threshold distances on the 100-point scale[7]. Interestingly, nine of the 10 highest threshold distances were found at score categories that denote one point above multiples of 5 (i.e., 11, 21, 26, 41, 61, 71, 81, 86, 91). This may suggest that raters distinguished scores that are up to exact multiples of 5 (e.g., 20, 60, 70) and scores that denote one point above multiples of 5.

Score category collapsing for rectifying threshold disordering and distances (RQ2)

Given that approximately half of the score categories were disordered, score categories were collapsed based on threshold distances under the assumption that large threshold distances suggest the anchoring function of those score categories. Specifically, scores in the ranges 1–10, 11–20, 21–25, 26–40, 41–60, 61–70, 71–80, 81–85, 86–90, and 91–100 were each collapsed into a single score category. This collapsing scheme resulted in a 10-point scale that contains 11 score categories from 0 to 10, with 0 retained as a separate category (see Figure 2). Table 5 summarizes the score category statistics from the MFRM analysis. The threshold measures suggest several disordered categories, such as scores of 3 (.46 logits) and 4 (–1.41 logits) and scores of 8 (1.41 logits) and 9 (.84 logits). Those score categories were also associated with negative threshold distances, including a score of 4 (–1.87 logits) and a score of 9 (–.57 logits).

To further refine this 10-point scale, another rating scale was created. Specifically, Rasch-Andrich threshold measures and threshold distances were screened in two ways. First, threshold measures and threshold distances were compared for two adjacent score categories from the lowest score category—whether measures and distances associated with higher score categories were higher. Secondly, threshold measures and threshold distances were compared for two adjacent score categories from the highest score category—whether measures and distances associated with lower score categories were lower.

These two screening methods resulted in the slightly different, yet consistent, identification of disordered categories. Neither screening method flagged scores of

[7]For the sake of readability, I present only the 10 score categories that were associated with highest threshold distances in this table. Threshold measures and distances for the entire set of score categories can be found in the supplementary file in the OSF platform (https://osf.io/8jh3u/overview?view_only=01e99a1d7f09445c9e07a657f5a26a0c).

Table 4. Score category statistics of the 10 score categories associated with the highest threshold distances on the 100-point scale

					Threshold		
Score	Count	%	Measure	Outfit	Measure	SE	Threshold distance
11	17	1	–.06	1.0	.96	.07	1.51
13	14	0	–.06	.8	.60	.07	1.12
21	12	0	–.04	1.2	1.26	.06	2.60
26	20	1	–.03	.8	.56	.06	1.47
41	26	1	.00	.8	.57	.05	1.23
61	20	1	.03	.8	1.13	.05	1.61
71	34	1	.05	.4	.68	.05	1.53
81	36	1	.07	.9	.71	.05	1.39
86	36	1	.09	.6	.78	.05	1.31
91	37	1	.12	.5	.80	.06	1.29

Note. Count = the number of times the score category being chosen, Measure = difficulty of test takers obtaining the score, outfit = fit to observed data, Threshold measure = test taker's ability at which obtaining two adjacent score categories is equally likely, Threshold SE = standard error of threshold measure, Threshold distance = the difference between two adjacent thresholds.

Table 5. Score category statistics of the 10-point scale

					Threshold		
Score	Count	%	Measure	Outfit	Measure	SE	Threshold distance
0	54	2	–.74	1.5	NA	NA	NA
1	235	7	–.62	1.2	–2.23	.14	NA
2	230	7	–.36	1.1	–.51	.07	1.72
3	107	3	–.31	.8	.46	.06	0.97
4	399	12	–.03	.9	–1.41	.06	–1.87
5	609	18	.19	1.0	–.30	.05	1.11
6	375	11	.46	.8	.82	.05	1.12
7	408	12	.68	.9	.45	.05	–0.37
8	210	6	.87	.8	1.41	.05	0.96
9	237	7	1.05	1.0	.84	.06	–0.57
10	496	15	1.33	1.1	.46	.06	–0.38

Note. Count = the number of times a score category being chosen, Measure = difficulty of test takers obtaining the score, outfit = fit to observed data, Threshold measure = test taker ability in logit at which obtaining two adjacent score categories is equally likely, Threshold SE = standard error of threshold measure, Threshold distance = the difference between two adjacent thresholds.

1 and 5, which were thus maintained. Scores from 2 to 4 were found to be disordered and were collapsed into a single category. The screening flagged scores from 6 to 10. While collapsing these five score categories into one score category was possible, their associated measures seemed notably different. Specifically, scores of 6 and 7 were associated with notably lower measures (.82 and .45 logits, respectively) than the score of 8, which was associated with a threshold measure of 1.41 logits. While scores of 9 and 10 were similar to scores of 6 and 7 in threshold measures (.84 and .46 logits, respectively), it was thought that the score of 8 denoted an ability level that was somewhat different from ability levels represented by scores of 6 and 7. Accordingly, scores of 6 and 7 were collapsed into one score category, and scores of 8, 9, and 10 were collapsed into another score category. This collapsing transformed the 10-point scale of 012345678910 to a 5-point scale of 01222344555 (see Figure 2).

Table 6. Score category statistics of the 5-point scale

Score	Count	%	Measure	Outfit	Threshold		Threshold distance
					Measure	SE	
0	54	2	–.93	1.6	NA	NA	NA
1	235	7	–.62	1.1	–2.41	.15	NA
2	736	22	.19	1.0	–1.33	.07	1.08
3	609	18	.87	.9	.75	.05	2.08
4	783	23	1.63	1.0	1.05	.05	0.30
5	943	28	2.71	1.0	1.95	.05	0.90

The 5-point scale did not show any disordered categories, and higher score categories were associated with higher threshold measures (Table 6). However, threshold distances were relatively narrow between score categories of 3 and 4 (.30 logits), compared with the other threshold distances of 1.08 (scores from 1 to 2), 2.08 (scores from 2 to 3), and .90 (scores from 4 to 5) logits.

Due to their narrow threshold distance between scores of 3 and 4 in this 5-point scale (.30 logits), these two score categories were collapsed into one score category. This collapsing transformed the 5-point scale of 012345 into a 4-point scale of 012334 (see Figure 2), which maintained the score categories appropriately ordered while also displaying adequately large threshold distances across the score categories (Table 7).

Figure 6 presents the probability curves for the original 100-point scale and the three collapsed scales. The probability curves of the 100-point scale show no clear peaks, suggesting that many different scores represented essentially equivalent ability levels and that the same score represented different ability levels. Combining score categories helped the 10-point scale to display a clear peak of the collapsed score category of 1, which represented ability levels around –2.0 logits. Further merging score categories helped the 5-point scale to display a clear peak for the collapsed score category of 2, which distinctly represented ability levels around 0.0 logits. Finally, the 4-point scale boosted the functioning of the collapsed score category of 3, which uniquely represented ability levels around 2.0 logits. Although these collapsing schemes were implemented based on the statistical information derived from the MFRM analysis, each collapsing scheme happened to target a unique range of ability levels from low-ability speakers (10-point scale), middle-ability speakers (5-point scale), to high-ability speakers (4-point scale).

Comparing the original 100-point scale and the collapsed 4-point scale (RQ3)

Figure 7 presents the variable maps of the original 100-point scale and the collapsed 4-point scale. In the original 100-point scale, many speakers and raters are located at 0.0

Table 7. Score category statistics of the 4-point scale

Score	Count	%	Measure	Outfit	Threshold		Threshold distance
					Measure	SE	
0	54	2	–1.00	2.0	NA	NA	NA
1	235	7	–.53	1.2	–2.56	.15	NA
2	736	22	.48	.9	–1.16	.08	1.40
3	1392	41	1.88	1.0	.60	.05	1.76
4	943	28	3.68	.9	3.11	.05	2.51

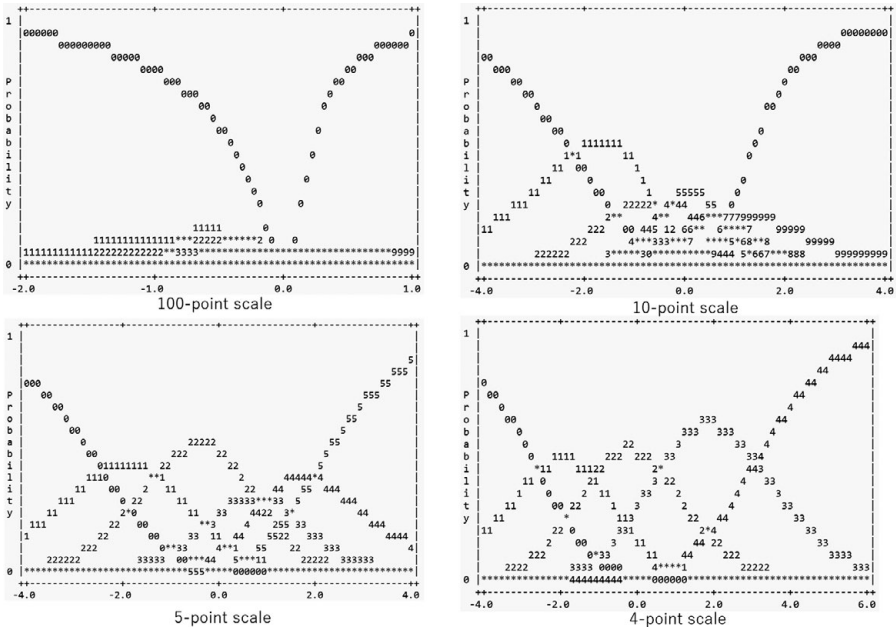

Figure 6. Probability curves of the 100-point (top left), 10-point (top right), 5-point (bottom left), and 4-point (bottom right) scales.[8]

logits, and many score categories cluster in this same region. In contrast, in the 4-point scale, the score categories of 1, 2, 3, and 4 largely corresponded to speaker abilities below –1.0, around 0.0, around 2.0, and above 3.0 logits, respectively.

Table 8 summarizes the comparison between the original 100-point scale and the collapsed 4-point scale. Cronbach's alpha and intraclass correlation were not substantially different between the two scales. As for the model fit, while the original scale explained more variance than the 4-point scale, there were slightly more unexpected responses for the original scale than for the 4-point scale. The 100-point scale reliably distinguished 10 ability levels, whereas the 4-point scale distinguished nine. The original scale detected eight rater severity levels, whereas the 4-point scale found seven severity levels. Person measures derived from the two scales were highly correlated, with Pearson's product-moment correlation of .99 (95% CI[.99, 1.00]). Taken together, these findings suggest that the reduction of scale granularity aggregated minimally different ability levels and revealed meaningfully distinct ability levels among speakers without a substantial loss of the information that was originally obtained with the 100-point scale.

Discussion

The present study aimed to reexamine the functioning of the 9-point and 100-point scales used to assess comprehensibility in Saito et al. (2020) and Huensch and Nagle (2023) and

[8]In the probability curves for the 100-point scale, there are two curves that are denoted as 0. The one on the left side represents the score of 0, but the one on the right side represents the score of 100. Similarly, 0 on the right side of the 10-point scale represents the score of 10.

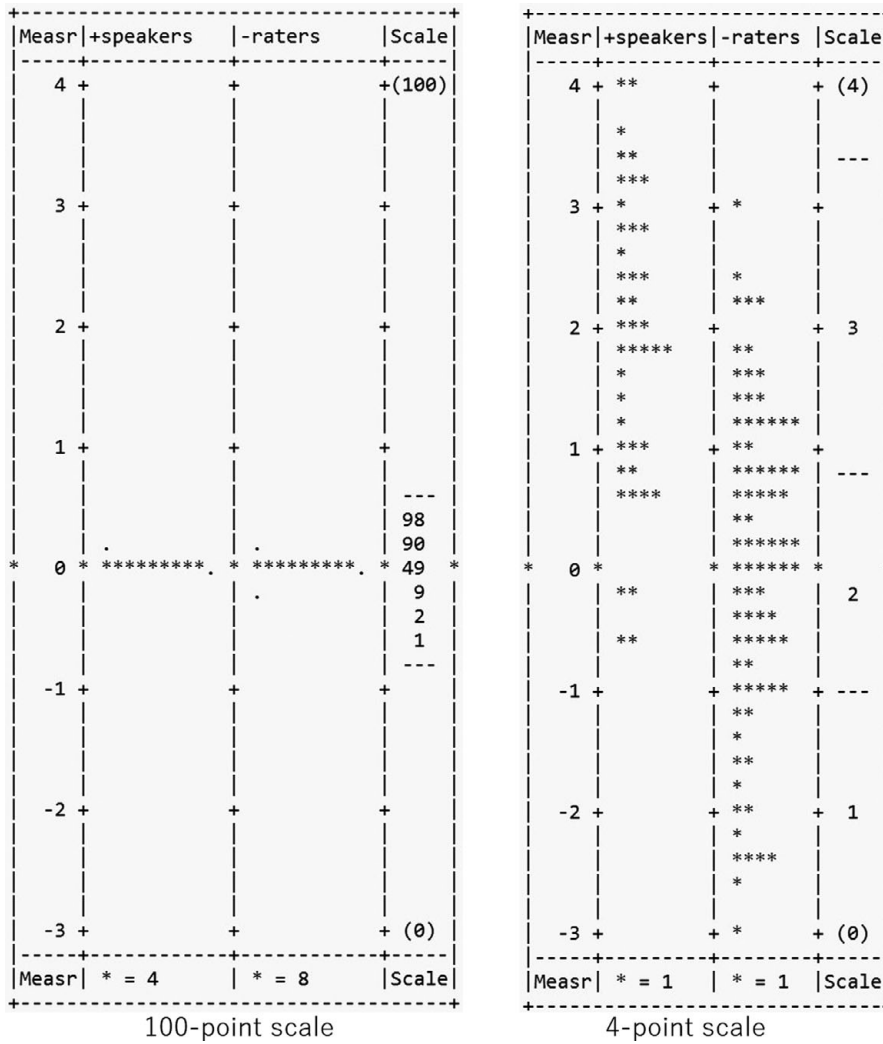

Figure 7. Variable maps of the original 100-point scale (left) and the collapsed 4-point scale (right).

to demonstrate the usefulness of score category collapsing as a potential countermeasure against suboptimal rating scale functioning. For the 9-point scale, while the difficulty of obtaining higher scores monotonically increased, score categories from 2 to 7 were muddled as suggested by their narrow threshold distances and unclear peaks in the probability curves. Regarding the 100-point scale, approximately half of the score categories were disordered, but large threshold distances were observed among score categories that denote one point above multiples of 5. The 9-point scale was effectively collapsed into 3-point and 4-point scales, whereas the 100-point scale was effectively collapsed into a 4-point scale. These results are discussed in relation to previous studies below.

Addressing the first research question (i.e., threshold ordering, threshold distances, probability curves), the study reanalyzed Saito et al.'s (2020) 9-point scale. The results

Table 8. Comparison between the original 100-point scale and the collapsed 4-point scale

Criteria	Original 100-point scale	Collapsed 4-point scale
Cronbach's alpha	.99 [.98, .99]	.98 [.97, .99]
Two-way consistency, average-measure intraclass correlation	.98 [.98, .99]	.98 [.97, .99]
Variance explained by Rasch	63.50%	56.77%
Standardized residuals over |2.0|	3.0%	1.0%
Stara for person ability	10.91	9.21
Separation reliability for person	.98	.98
Increasing thresholds	No	Yes
Threshold distances over 1.4	No	Yes
Peaks in probability curves	No	Yes
Raters with infit over 1.5	13	13
Strata for rater severity	8.21	7.21
Separation reliability	.97	.96

suggested that while Rasch–Andrich thresholds increased monotonically with higher score categories, all threshold distances fell below the recommended minimum of 1.4. Such narrow threshold distances suggest that the difficulty of obtaining a score was not substantially different from the difficulty of obtaining a score higher. Indeed, the probability curves suggested that speakers with 0.0 logits were equally likely to receive scores of 4, 5, 6, or 7. These findings align well with previous studies (Isaacs & Trofimovich, 2013; Isbell, 2018; Kermad & Bogorevich, 2022), which have reported that raters often struggled to associate specific differences in speaking performance with numerical values on the scale, particularly near the midpoint.

Regarding the second research question (i.e., the impact of score category collapsing on the functioning of the 9-point scale), the present study tested four collapsing patterns. As a result, the 3-point scale and the two 4-point scales demonstrated monotonically increasing threshold measures and threshold distances above 1.4. Their probability curves also displayed clear peaks for collapsed score categories. Kermad and Bogorevich (2022) similarly collapsed a 9-point scale (123456789) for comprehensibility into 7-point (123345567) and 5-point (112234455) scales. They found that while this collapsing improved the probability curves to some extent, the midpoint of the 7-point and 5-point scales still failed to display a clear peak, possibly because the original midpoint remained intact in these collapsed scales. Indeed, the present study also showed that the collapsed 5-point scale, where the midpoint of 3 stayed intact, did not show a clear peak for this midpoint. Perhaps, when the ability levels are muddled around the midpoint —as is commonly seen among comprehensibility ratings (Isaacs & Trofimovich, 2013; Isbell, 2018)—the scale functioning of a 9-point scale may be effectively enhanced by combining the midpoint score category with adjacent categories.

For the third research question (i.e., the compatibility between the original 9-point scale and collapsed scales), the original scale and the newly created scales were not substantially different in terms of widely used reliability indices, such as Cronbach's alpha and intraclass correlation. The original scale and the new 3-point scale identified eight levels of rater severity, whereas the two 4-point scales identified 11 rater severity levels. In principle, fewer severity strata are preferable, as scores from different raters can be interpreted in a similar way. From this viewpoint, the original 9-point scale and the 3-point scale may seem advantageous. However, the 9-point scale does not allow each score category to represent a distinct ability level, and the 3-point scale may risk

oversimplifying ability differences in a sample of learners. Given these trade-offs, the two 4-point scales could be an eclectic solution. Although these two scales may accentuate severity levels among raters, the increase from eight severity levels in the 9-point scale to 11 severity levels in these 4-point scales may be inconsequential compared with the interchangeability of score categories on the 9-point scale. It should be noted that in Saito et al. (2020), raters did use a 9-point scale and thus post hoc score category collapsing does not necessarily simulate how raters would behave with shorter scales derived from the 9-point scale. Nevertheless, while I have no intention to prescribe a single optimal scale, these observations highlight how researchers can weigh the trade-offs of scale length when justifying their methodological choices.

To answer the first research question (i.e., threshold ordering, threshold distances, probability curves), the present study also reanalyzed Huensch and Nagle's (2023) 100-point scale, finding that approximately half of the score categories were disordered, likely due to the large number of score categories that potentially caused raters' confusion. Threshold distances tended to be relatively large and positive for score categories that denote one point above multiples of 5 (e.g., 21, 61, 71). This finding is interesting, especially because the numerical values were hidden in the 100-point scale. One possibility is that raters attempted to distinguish different ability levels by placing the slide bar noticeably farther apart. In fact, score categories that denote multiples of 5 were chosen relatively frequently, compared with score categories around them. By collapsing infrequently endorsed score categories, the threshold-based score category collapsing successfully tailored the rating scale to the raters' scale use. Although their study focused on a Likert-scale survey that was intended to measure self-efficacy of mathematics, Toland and Usher (2016) also found that respondents tended to choose multiples of 5 on a 100-point scale, where only 0, 50, and 100 were labeled (e.g., *not at all confident*, *somewhat confident*). While speculative, it is possible that when unsure about the meaning of 100 score categories, respondents might consciously or unconsciously segment the 100-point scale into 10 to 20 small blocks and thereby lessen their cognitive load on their own.

Addressing the second (i.e., the impact of score category collapsing on threshold ordering, threshold distances, probability curves of the original 100-point scale) and third research questions (i.e., the compatibility between the original 100-point scale and collapsed scale), the present study collapsed the 100-point scale based on the MFRM analysis and compared the original 100-point scale and the collapsed 4-point scale. While approximately half of the score categories were originally disordered in the 100-point scale, the finalized 4-point scale had all the score categories appropriately ordered with adequate threshold distances and clear peaks in the probability curves. Although generalizing this finding to other contexts requires great caution, the present finding seems to align with Toland and Usher (2016), who also used the Rasch measurement approach and found a 4-point scale as the best scale derived from a 100-point scale intended to measure mathematics self-efficacy. Notably, in the present study, scores as far apart as 41 and 80 on the original 100-point scale were merged into the same category in the 4-point scale, indicating that comprehensibility originally represented by these scores was similar. Similarly, the score of 11 on the original 100-point scale was relabeled as 2, and the score of 80 was relabeled as 3 on the 4-point scale. That is, this difference of 1 on the collapsed 4-point scale was originally presented as a difference of 69 points on the 100-point scale. These observations critically highlight that interpretations of raw scores can run the risk of overly accentuating differences among L2 speakers. It should also be noted that the original 100-point and the collapsed 4-point scales were similar in terms of many psychometric qualities, such

as speaker and rater statistics. In sum, while these two scales seem to substantially differ at first glance, their psychometric properties are largely comparable; however, the collapsed 4-point scale affords more interpretable and substantively meaningful score categories.

Conclusion

The present study reexamined Saito et al.'s (2020) 9-point scale and Huensch and Nagle's (2023) 100-point scale, focusing on their rating scale functioning. The findings suggested that some score categories were disordered and associated with narrow threshold distances. This suboptimal scale functioning was successfully addressed by score category collapsing, and newly created scales allowed each score category to represent a distinct ability level while largely preserving the psychometric qualities of the original long scale.

However, these findings should be interpreted with three major limitations in mind. First, the study only examined Saito et al.'s (2020) 9-point scale and Huensch and Nagle's (2023) 100-point scale that were used to measure comprehensibility in L2 learners' speaking performance. While some evidence of the suboptimal rating scale functioning was obtained, it does not mean that all 9-point and 100-point scales fail to function well, and generalizations cannot be extended to other 9-point and 100-point scales. Especially, scale functioning is closely related to an array of factors, such as rater characteristics and expertise, performance descriptors, the clarity of the construct being assessed, and linguistic variability in examinees' performance samples to be rated. For example, the present study analyzed only monologic tasks, and thus generalizability to other tasks, such as interactive speech, should be made with great caution. Similarly, raters with different backgrounds, such as accent familiarity and teaching experience, may use the same scale differently. Accordingly, the scale functioning reported in the present study may not be found when other rater samples are recruited. Secondly, the present study as a methodological showcase explored only a small set of collapsing patterns using real datasets, leaving other collapsing strategies untested. Future studies thus could explore other collapsing strategies by, for example, simulating artificial data and testing the impact of different collapsing patterns in different rating situations (Tsai et al., 2025). Third, it should be borne in mind that while the present study suggested that 3-point and 4-point scales may function better than 9-point and 100-point scales, the shorter versions examined in the present study were derived from collapsing score categories based on the data obtained with long rating scales. Therefore, the study provides no direct evidence on how raters would use standalone 3-point or 4-point scales, leaving this as an avenue for future investigation.

Despite those limitations, the present study offers implications for L2 pronunciation researchers. First of all, they may want to consider using shorter rating scales, such as those that contain five score categories, whose psychometric qualities have been empirically supported by previous research (Isaacs & Trofimovich, 2013; Kermad, 2024). Indeed, the present study found that 9-point and 100-point scales could be collapsed into more functional scales that contain four to five score categories, regardless of substantial contextual differences between Saito et al. (2020) and Huensch and Nagle (2023). Second, if future studies want to stick to long rating scales, they may want to provide more robust rater training in order to lessen raters' cognitive burden and facilitate their consistent use of the scale (Kermad, 2024; Tsunemoto et al., 2022). Perhaps, researchers can present speech samples that represent scores of 10, 30, 60, and 90, helping evaluators to link concrete speech performance to scores on a 100-point scale, for example. Third, if providing such a robust rater training session is not ideal

given their research purpose, they can implement post hoc score category collapsing and tailor the scale design to the obtained data (Kermad & Bogorevich, 2022). Post hoc score category collapsing allows researchers to use conventionally used rating scales, ensuring the compatibility between their study and previous studies. Alternatively, researchers can analyze their data from a Rasch measurement approach and use the obtained logit measures, which they can treat as an interval variable, for further statistical analysis (see Isbell & Lee, 2022). While rarely used in L2 pronunciation research, the Rasch measurement approach has great potential to provide evidence regarding scale functioning, unlike Cronbach's alpha and intraclass correlation. Last but not least, if none of these countermeasures is feasible, future studies are encouraged to provide context-specific, evidence-based justifications for their choice of scale length, rather than adopting conventional practices uncritically.

As a final note, while long rating scales may be easy to use for untrained raters and easy to develop for researchers, such convenience comes at the cost of accuracy. The use of long rating scales may not only render it difficult to make sound score interpretations but also contribute to the uncritical reproduction of long rating scales in future studies, if not intended by previous studies. While the present study is indeed exploratory and no more than a proof of concept, it is my humble hope that the present study will guard against SLA researchers' uncritical development and use of long rating scales in future L2 pronunciation research and beyond, contributing to more meaningful and easy-to-interpret research findings.

Data availability statement. The experiment in this article earned Open Materials badge for transparent practices. The materials are available at https://osf.io/8jh3u/overview?view_only=01e99a1d7f09445c9e07a657f5a26a0c.

Acknowledgments. I'm grateful for the reviewers' constructive comments on the previous versions of this manuscript. I used ChatGPT (GPT-5.2) to improve the clarity of my writing but not to generate ideas. The author bears sole responsibility for any errors in this manuscript.

References

Ali, M. M. (2023). The foreign-accentedness, comprehensibility, and intelligibility of L2 Arabic speech. *Language Teaching Research*. https://doi.org/10.1177/13621688231158787

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals*, *50*(3), 547–566. https://doi.org/10.1111/flan.12285

Chau, T., & Huensch, A. (2025). The relationships among L2 fluency, intelligibility, comprehensibility, and accentedness: A meta-analysis. *Studies in Second Language Acquisition*, *47*(1), 282–307. https://doi.org/10.1017/S0272263125000014

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility?. *The Modern Language Journal*, *99*(1), 80–95. https://doi.org/10.1111/modl.12185

Galante, A., & Thomson, R. I. (2017). The effectiveness of drama as an instructional approach for the development of second language oral fluency, comprehensibility, and accentedness. *TESOL Quarterly*, *51*(1), 115–142. https://doi.org/10.1002/tesq.290

Huensch, A., & Nagle, C. (2021). The effect of speaker proficiency on intelligibility, comprehensibility, and accentedness in L2 Spanish: A conceptual replication and extension of Munro and Derwing (1995a). *Language Learning*, *71*(3), 626–668. https://doi.org/10.1111/lang.12451

Huensch, A., & Nagle, C. (2023). Revisiting the moderating effect of speaker proficiency on the relationships among intelligibility, comprehensibility, and accentedness in L2 Spanish. *Studies in Second Language Acquisition*, *45*(2), 571–585. https://doi.org/10.1017/S0272263122000213

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*(3), 475–505. https://doi.org/10.1017/S0272263112000150

Isbell, D. (2018). Assessing pronunciation for research purposes with listener-based numerical scales. In O. Kang & A. Ginther (Eds.), *Assessment of second language pronunciation* (pp. 89–112). Routledge.

Isbell, D. R., & Lee, J. (2022). Self-assessment of comprehensibility and accentedness in second language Korean. *Language Learning*, *72*(3), 806–852. https://doi.org/10.1111/lang.12497

Kermad, A. (2024). Training the "everyday" listener how to rate accented speech. *International Journal of Listening*, *38*(1), 58–78. https://doi.org/10.1080/10904018.2021.1987910

Kermad, A., & Bogorevich, V. (2022). Using statistical transformation methods to explore speech perception scale lengths. *Language Teaching Research Quarterly*, *29*, 65–91. https://doi.org/10.32038/ltrq.2022.29.05

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Kostromitina, M., Sudina, E., & Baghlaf, E. (2025). Study and instrument quality in perception-based L2 pronunciation research: A methodological synthesis. *Studies in Second Language Acquisition*, 1–34. https://doi.org/10.1017/S027226312500018X

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369–377. https://doi.org/10.2307/3588485

Linacre, J. M. (n.d.). Reliability and separation of measures. https://www.winsteps.com/winman/reliability.htm

Linacre, J. M. (1999). Category disordering (disordered categories) vs. threshold disordering (disordered thresholds). *Rasch Measurement Transactions*, *13*(1), 675.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106.

Linacre, J. M. (2025). *A user's guide to FACETS (64-bit) Rasch-model computer programs. Program manual 4.4.4.* https://www.winsteps.com/a/Facets64-Manual.pdf

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice and language assessment*. Oxford University Press.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2015). A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation*, *1*(1), 11–42. https://doi.org/10.1075/jslp.1.1.01mun

Nagle, C. L., & Rehman, I. (2021). Doing L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition*, *43*(4), 916–939. https://doi.org/10.1017/S0272263121000292

Nagle, C. L., Trofimovich, P., O'brien, M., & Kennedy, S. (2022). Comprehensible to whom? Examining rater, speaker, and interlocutor perspectives on comprehensibility in an interactive context. *The Modern Language Journal*, *106*(4), 675–693. https://doi.org/10.1111/modl.12809

Saito, K., Macmillan, K., Mai, T., Suzukida, Y., Sun, H., Magne, V., … & Murakami, A. (2020). Developing, analyzing and sharing multivariate datasets: Individual differences in L2 learning revisited. *Annual Review of Applied Linguistics*, *40*, 9–25. https://doi.org/10.1017/S0267190520000045

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech?: Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. *Studies in Second Language Acquisition*, *41*(5), 1133–1149. https://doi.org/10.1017/S0272263119000226

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439–462. https://doi.org/10.1093/applin/amv047

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217–240. https://doi.org/10.1017/S0142716414000502

Shintani, N., Saito, K., & Koizumi, R. (2019). The relationship between multilingual raters' language background and their perceptions of accentedness and comprehensibility of second language speech. *International Journal of Bilingual Education and Bilingualism*, *22*(7), 849–869. https://doi.org/10.1080/13670050.2017.1320967

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167. https://doi.org/10.1017/S0272263119000421

Tergujeff, E. (2021). Second language comprehensibility and accentedness across oral proficiency levels: A comparison of two L1s. *System*, *100*, 102567. https://doi.org/10.1016/j.system.2021.102567

Thorpe, W. C., Baker-Smemoe, W., Hartshorn, K. J., McMurry, B. L., & Wilcox, M. (2025). The relationship of language anxiety and English learners' accentedness, comprehensibility, and speech rate across three communication tasks. *System*, *133*, 103721. https://doi.org/10.1016/j.system.2025.103721

Toland, M. D., & Usher, E. L. (2016). Assessing mathematics self-efficacy: How many categories do we really need?. *The Journal of Early Adolescence*, *36*(7), 932–960. https://doi.org/10.1177/0272431615588952

Trofimovich, P., Nagle, C. L., O'Brien, M. G., Kennedy, S., Taylor Reid, K., & Strachan, L. (2020). Second language comprehensibility as a dynamic construct. *Journal of Second Language Pronunciation*, *6*(3), 430–457. https://doi.org/10.1075/jslp.20003.tro

Tsai, C. L., Wind, S., & Estrada, S. (2025). Exploring the effects of collapsing rating scale categories in polytomous item response theory analyses: An illustration and simulation study. *Measurement: Interdisciplinary Research and Perspectives*, *23*(1), 66–89. https://doi.org/10.1080/15366367.2023.2288791

Tsunemoto, A., Trofimovich, P., & Kennedy, S. (2023). Pre-service teachers' beliefs about second language pronunciation teaching, their experience, and speech assessments. *Language Teaching Research*, *27*(1), 115–136. https://doi.org/10.1177/1362168820937273

Tsunemoto, A., & Trofimovich, P. (2024). Coherence and comprehensibility in second language speakers' academic speaking performance. *Studies in Second Language Acquisition*, *46*(3), 795–817. https://doi.org/10.1017/S0272263124000305

Tsunemoto, A., Trofimovich, P., Blanchet, J., Bertrand, J., & Kennedy, S. (2022). Effects of benchmarking and peer-assessment on French learners' self-assessments of accentedness, comprehensibility, and fluency. *Foreign Language Annals*, *55*(1), 135–154. https://doi.org/10.1111/flan.12571

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2023). Frequency of exposure influences accentedness and comprehensibility in learners' pronunciation of second language words. *Language Learning*, *73*(1), 84–125. https://doi.org/10.1111/lang.12517

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252. https://doi.org/10.1177/0265532212456968

Wright, B. (1991). Scores, reliabilities and assumptions. *Rasch Measurement Transactions*, *5*(3), 157–158. Retrieved from https://www.rasch.org/rmt/rmt53a.htm

Taichi Yamashita is an Assistant Professor of Applied Linguistics at the Faculty of Foreign Language Studies, Kansai University, Japan. He earned a Ph.D. in Applied Linguistics and Technology from Iowa State University. His research interests include computer-assisted language learning, language assessment, second language writing, and research methods. He has published articles in peer-reviewed journals, such as *Applied Linguistics*, *Language Learning & Technology*, and *Language Testing*.

Cite this article: Yamashita, T. (2026). Post hoc score category collapsing for L2 pronunciation research. *Studies in Second Language Acquisition,* 1–26. https://doi.org/10.1017/S0272263126101582