GLODAL

OPEN ACCESS

*Brief Reports in Language Research*

*Brief Reports in Language Research*, 1, 260101 (2026)
https://doi.org/10.65553/BRLR.260101

# *Rare Categories, Anxious Learners: Using Relative Efficiency Curves to Evaluate Category Collapsing in L2 Likert Scales*

Mark Feng Teng*

*Department of English, The Hang Seng University of Hong Kong, China*

## Abstract

When few respondents choose some categories of a Likert scale, L2 researchers often merge those categories and report an unchanged Cronbach's alpha as evidence that nothing was lost. This methodological note shows why that evidence is uninformative. This article thus offers a simple alternative: relative efficiency (RE) curves, which show where along the trait a merged scale loses precision. In a 10-item anxiety scale completed by 714 Chinese EFL learners, merging the two lowest categories left alpha unchanged (.92–.93) but kept only 39–45% of the information about the most anxious learners; merging the three lowest kept 10–12%. The result replicated in a held-out half of the sample, and a simulation showed the same pattern. I provide an R function and five reporting guidelines for L2 researchers.

*Keywords:* category collapsing; item response theory; relative efficiency; foreign language anxiety

Submission date: June 20, 2026; Revision date: October 16, 2026; Online: November 6, 2026

Copyright © 2026
The authors

(corresponding author: markteng@hsu.edu.hk)

## Introduction

Imagine a teacher-researcher who wants to find the most anxious students in a class using a six-point anxiety questionnaire. Few students choose the most extreme option, so, following common practice, she merges it with the next one before analyzing the data. Cronbach's alpha does not change, and she concludes that nothing was lost. This methodological note shows that the conclusion is wrong, and that the students most affected are exactly the ones she wanted to find.

Likert-type scales are the main instrument of research on language learning anxiety (Sudina, 2021; Teimouri et al., 2019). In most samples, the categories at one end of these scales are rarely used, simply because few learners have extreme trait levels. Sparse categories cause practical problems. Their thresholds are estimated poorly, polychoric correlations become unstable, and a category may be empty in one group of a multi-group analysis (Rhemtulla et al., 2012). These problems arise often in measurement invariance testing, which is increasingly common in L2 research but still applied inconsistently (Sudina et al., 2026). The usual remedy is to merge sparse categories with their neighbours, often following rules of thumb such as a minimum of ten responses per category (Linacre, 2002).

Recent methodological work has started to examine what merging costs. Quan and Wang (2025) showed that collapsing can make an item response model appear to fit better even when it biases item parameters. Liu et al. (2025) found that collapsing in confirmatory factor analysis makes misspecified models more likely to be accepted. In language research, Brysbaert (2026) showed that the number of scale points matters little when responses spread across the whole scale, but that precision drops when responses use only part of it. Teng (under review) shows that high Cronbach's alpha alone can mask serious measurement problems: a five-stage SEM/IRT workflow showed that a mixed composite of emotion regulation and language learning anxiety items produced alpha (.946) as high as either scale, and that careless responding can mimic measurement quality. The above studies therefore demonstrate that alpha is insufficient evidence of measurement quality and that researchers must probe what it conceals. These studies concern model fit, parameter recovery and scale length. None of them asks the question an applied linguist most needs answered: *which learners* lose measurement precision when categories are merged?

That question cannot be answered with alpha. Alpha summarizes precision as one number for the whole sample (McNeish, 2018; Sijtsma, 2009). Rare categories are rare because few learners need them, so the precision they provide is concentrated in a small group and barely affects a sample average. Merging them can leave alpha unchanged while removing much of what the scale tells us about the learners at the extreme. In anxiety research, those learners are usually the ones that screening studies try to identify, interventions try to help and longitudinal studies try to follow.

This methodological note makes three contributions to research methods in applied linguistics. First, it identifies a blind spot: using real L2 data, it shows that alpha cannot detect the cost of category collapsing, because that cost falls on a small group of learners. Second, it introduces a diagnostic, adapting relative efficiency curves (Lord, 1980) into a four-step procedure with ready-to-use R code that shows where on the trait a merged scale loses

precision. Third, it offers guidelines that link the decision to merge categories to the research question, so that researchers interested in learners at the extremes can judge whether merging is acceptable for their purpose.

## Methodological Framework

### *The idea in plain terms*

An item response theory (IRT) model such as the graded response model (GRM; Samejima, 1968) describes how likely a learner at a given trait level θ is to choose each category of each item. From these probabilities it computes the test information function I(θ), which shows how precisely the scale measures learners at every trait level (Embretson & Reise, 2000). Higher information means a smaller standard error; the standard error is approximately $1/\sqrt{I(\theta)}$.

Each boundary between two adjacent categories carries information about learners whose trait level lies near that boundary. Merging the two categories removes the boundary, and so removes that information. Because the boundaries between extreme categories lie at extreme trait levels, merging them costs precision only for learners at those levels.

For a merging scheme *s*, the relative efficiency curve compares the information of the merged scale with that of the original scale at each trait level:
$RE_s(\theta) = I_s(\theta) / I0(\theta)$,
where I0 is the information function of the original scale. The curve is easy to read. An RE of 1 means that merging has no cost for learners at that trait level. An RE of .50 means that the merged scale measures those learners only as precisely as the original scale would with half as many items; to restore the original precision, the merged scale would need about twice as many items. Both models fix the latent mean at 0 and the variance at 1, so the two information functions refer to approximately the same metric.

### *Four steps*

1. Inspect category use. Tabulate how often each category is chosen and list the candidate categories for merging.
2. Fit the GRM to the original coding and to each candidate scheme. Compute $RE_s(\theta)$ over a grid of trait values, for example from −3 to 3.
3. Define the region of interest from the research question. For example, use $\theta < -1.5$ for a study about highly anxious learners, or the whole range for a study of the full distribution. Keep a scheme only if its RE stays above a stated threshold throughout that region. We suggest .80, which equals the precision lost by dropping one item in five.
4. Report the RE curve together with sample-level reliability. Alpha shows average precision; the RE curve shows where precision is lost and for whom.

The supplementary R function re_curve() carries out steps 2 and 3 with the mirt package (Chalmers, 2012).

## Evaluation

### *Data and analysis*

The data came from an online survey of Chinese undergraduates learning English as a foreign language, conducted in April 2020. The same data set was analyzed for a different purpose in Teng (under review). We used a locally developed 10-item language learning anxiety scale rated from 0 (*not at all true of me*) to 5 (*completely true of me*). Higher scores indicate lower anxiety, so the most anxious learners lie at the low end of θ and use the lowest categories.

Of 803 submissions, we excluded later duplicate submissions and respondents who gave the same answer to every item, leaving 714 learners. They were randomly divided into a discovery half ($n$ = 361), which was analyzed first, and a replication half ($n$ = 353), which was analyzed afterwards with the same code. The replication was not preregistered.

We compared five schemes: the original six categories; merging 0 and 1; merging 0 to 2; merging 4 and 5; and merging both 0 and 1 and 4 and 5. In both halves, the GRM fitted better than the generalized partial credit and partial credit models (AIC = 9,528 vs 9,824 and 9,917 in the discovery half; 9,279 vs 9,578 and 9,647 in the replication half). No item showed misfit in the discovery half and one item did in the replication half (S-$X^2$, Benjamini–Hochberg adjusted $p < .01$). Absolute fit of the unidimensional GRM was mixed (C2-based CFI = .938 and .923, SRMSR = .091 and .093, RMSEA = .153 and .179), which is consistent with the two correlated facets (absence of anxiety, items 1–5; engagement, items 6–10) reported for this scale in Teng (under review). We therefore repeated the analysis within each five-item facet as a robustness check. Uncertainty in RE was estimated with 200 bootstrap resamples per half; resamples in which a merged category was empty (3%) were discarded. Data and R codes were available at the following link (https://osf.io/w3qse/overview).

## Empirical results

Categories 0 and 1 together accounted for 13–14% of all responses. Pooling both halves, 5.6% of learners had GRM trait estimates below −1.5. By the ten-responses rule of thumb (Linacre, 2002), category 0 was too sparse in three items in the discovery half and one item in the replication half, so the rule would have recommended merging it.

Table 1 and Figure 1A show the main pattern. Across all schemes, alpha varied by no more than .012, and marginal reliability fell by at most .05. The RE curves, however, showed large losses confined to specific regions of the trait. Merging 0 and 1 kept only 45% (95% bootstrap CI [.32, .64]) and 39% [.29, .56] of the information at θ = −2 in the two halves, but cost nothing from θ = −1 upwards. Merging 0 to 2 was more costly, keeping 12% [.08, .20] and 10% [.07, .16] of the information at θ = −2 and 68–74% at θ = −1. Merging 4 and 5 produced the mirror-image loss at the low-anxiety end, keeping 16% [.13, .21] and 12% [.09, .16] at θ = 2. In every case, the curves from the two halves were nearly identical, and their confidence intervals overlapped.

The pattern did not depend on the misfit of the ten-item model. Within each five-item facet, merging 0 and 1 kept 30–52% of the information at θ = −2, and merging 0 to 2 kept 6–15%. This held for the engagement facet, whose unidimensional model fitted well in the replication

half (CFI = .991, SRMSR = .056).

**Table 1.** *Reliability and relative efficiency of merging schemes in the discovery / replication halves*

| Scheme | Categories | α | Marginal reliability | RE at θ = −2 | RE at θ = −1 | RE at θ = 0 | RE at θ = 2 |
|---|---|---|---|---|---|---|---|
| Original | 6 | .920 / .928 | .940 / .944 | 1 | 1 | 1 | 1 |
| Merge 0+1 | 5 | .923 / .929 | .935 / .940 | .45 / .39 | .98 / .99 | .99 / 1.01 | 1.01 / 1.00 |
| Merge 0+1+2 | 4 | .925 / .930 | .921 / .922 | .12 / .10 | .74 / .68 | 1.06 / 1.04 | .98 / .99 |
| Merge 4+5 | 5 | .913 / .925 | .894 / .899 | .99 / 1.03 | 1.00 / 1.02 | .96 / .99 | .16 / .12 |
| Merge 0+1 and 4+5 | 4 | .917 / .926 | .890 / .894 | .46 / .39 | .97 / 1.01 | .95 / 1.00 | .17 / .12 |

*Note.* Discovery half n = 361; replication half n = 353. RE = relative efficiency, the test information of the merged scale divided by that of the original scale. Lower θ indicates higher anxiety. RE values slightly above 1 reflect re-estimation of discrimination parameters.

## Simulation

To check the pattern against known true values, we generated 200 data sets ($N$ = 700) from the GRM parameters estimated on the pooled sample, with $\theta \sim N(0, 1)$. Each data set was analyzed in its original form and after merging 0 and 1 or 0 to 2.

Alpha hardly moved: it averaged .941, .940 and .936 across the three codings, and its 95% ranges across replications overlapped almost completely (.934–.947, .933–.946 and .928–.943). For learners with true $\theta < -1.5$, however, the error of their trait estimates (root mean square error) rose from 0.26 (95% range 0.19–0.33) to 0.33 (0.25–0.43; +29%) and to 0.45 (0.32–0.58; +75%). In the middle range it barely changed (0.23 to 0.25), and above 1.5 it did not change at all (Figure 1B).

The practical consequence was a poorer ability to find anxious learners. Of the truly most anxious 5%, the original sum score placed 79% (95% range 69–89%) in its own bottom 5%, compared with 76% (63–86%) and 63% (49–77%) after merging. The population RE at θ = −2, computed from the true parameters, was .42 and .12, close to the empirical values.

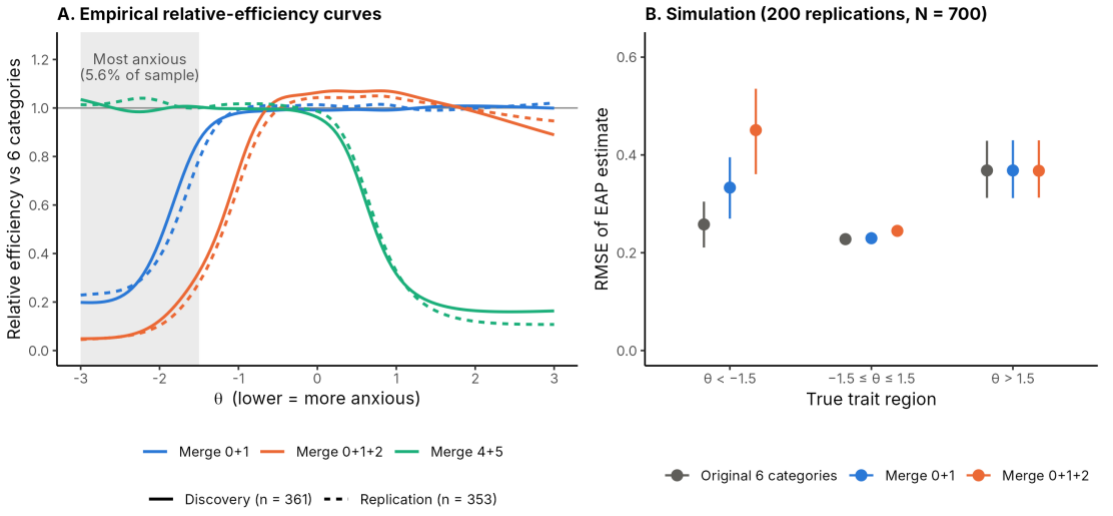



**Figure 1**. *Relative efficiency of merging schemes and simulated estimation error*
*Note.* (A) Relative-efficiency curves of three merging schemes in the discovery (solid) and replication (dashed) halves; the shaded band marks $\theta < -1.5$, the most anxious learners. (B) Simulated root mean square error of trait estimates by true trait region; bars span the 10th–90th percentiles across 200 replications.

## Implications

### *What this methodological note adds to L2 research methods*

The main message is simple: an unchanged alpha is not evidence that merging categories is harmless. Merging takes precision away from the learners who used the merged categories, and in anxiety research these are often the learners of greatest interest. Earlier work showed that collapsing can distort model fit and item parameters (Liu et al., 2025; Quan & Wang, 2025). This note adds the learner-level view: it shows *who* pays for collapsing and gives L2 researchers a tool to check this before they decide. The tool needs no new data, runs in a few lines of R, and produces a single figure that reviewers and readers can interpret.

### *Guidelines for researchers*

1. Do not justify merging with alpha or any other sample-level reliability coefficient. Report the RE curve for every scheme.
2. State the region of interest before merging. A study of the whole distribution can tolerate low-end losses that a study of highly anxious learners cannot.
3. Treat sparseness as a property of the sample, not of the category. A category chosen by 4% of a general sample may be chosen by 40% of a sample of anxious learners. Do not carry merging decisions over to new populations by default.
4. Prefer analytic remedies to merging where possible. IRT models can usually estimate sparse categories directly. If merging is unavoidable for one analysis, such as an empty category in one group of an invariance test, confine it to that analysis and report its cost.
5. When designing new scales, check where the categories are used. If the extreme categories are routinely merged, the remedy may be better-targeted items rather than fewer categories.

## Limitations

The evidence comes from one scale and one population, and the simulation covered a single condition ($N$ = 700, normally distributed trait). RE curves also depend on the chosen IRT model. Absolute fit of the ten-item model was only moderate here; the facet-level results suggest that the conclusions are robust to this, but researchers should check model fit before interpreting the curves. Within these limits, the RE curve gives L2 researchers a transparent and easily computed way to see what category merging costs, and whom it costs.

## References


Brysbaert, M. (2026). Are 7-point Likert scales preferable to 5-point scales in language research? *Behavior Research Methods, 58*(8), Article 218. https://doi.org/10.3758/s13428-026-03101-1

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

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

Liu, J., Bao, Y., DiStefano, C., & Jiang, W. (2025). Collapsing sparse responses in Likert-type scale data: Advantages and disadvantages for model fit in CFA. *Educational and Psychological Measurement*. Advance online publication. https://doi.org/10.1177/00131644251401097

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Quan, Y., & Wang, C. (2025). Collapsing or not? A practical guide to handling sparse responses for polytomous items. *Methodology, 21*(1), 46–73. https://doi.org/10.5964/meth.14303

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series, 1,* Pages i-169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sudina, E. (2021). Study and scale quality in second language survey research, 2009–2019: The case of anxiety and motivation. *Language Learning, 71*(4), 1149–1193. https://doi.org/10.1111/lang.12468

Sudina, E., Quan, Y., & Ozdemir, H. (2026). Measurement and structural invariance testing in L2 research: A methodological synthesis. *Studies in Second Language Acquisition, 48*(3), 873–906. https://doi.org/10.1017/S0272263126101624

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition, 41*(2), 363–387. https://doi.org/10.1017/S0272263118000311

Teng, M. F. (Under review). An integrated SEM–IRT approach to evaluating emotion regulation and language learning anxiety measures in L2 research.

**Author bio**
Mark Feng Teng, Ph.D., is a PT faculty member at English Department, The Hang Seng University of Hong Kong. His research portfolio mainly focuses on computer-assisted L2 vocabulary acquisition, reading and writing. His publications have appeared in international journals, including *Applied Linguistics, TESOL Quarterly, Language Teaching Research, System, Applied Linguistics Review, Computer Assisted Language Learning, Computers & Education, Foreign Language Annals,* and *IRAL,* among others. His recent monographs were published by Cambridge, Routledge, Springer, and Bloomsbury.
E-mail: markteng@hsu.edu.hk
ORCID: 0000-0002-5134-8504