

# Effectiveness and consistency of WIT benchmarking

Nathan Sanders

Associate Professor, Teaching Stream  
Department of Linguistics

work done in collaboration with  
Lisa Sullivan (UofT PhD 2022, now at Carleton) and  
Erin Vearncombe (UTM Institute for the Study of University Pedagogy)

November 6th, 2025  
Faculty of Arts & Science  
Teaching & Learning Community of Practice

# Roadmap of the talk

- 1** Benchmarking
- 2** Context of this study
- 3** Many-facet Rasch modelling
- 4** Using MFRM in LIN200
- 5** Wrap-up

# Benchmarking

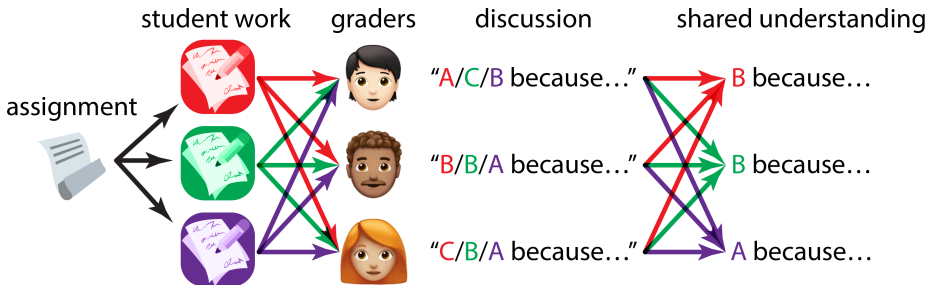
**Variability in grading** is a widespread and inequitable problem throughout education (see Brookhart et al. 2016 for an overview of more than a century of research).

- ▶ highly salient with multiple graders in the same course, as is typical in large courses at UofT like LIN200 (the course of focus in this presentation)
- ▶ even individual instructors may vary significantly in their own grading practices (Eells 1930)

**Benchmarking** is a consensus-based collaborative process designed to mitigate grading variability by helping align grading practices across graders.

- ▶ also known in the pedagogical literature as *calibration*, *moderation*, *norming*, or *tuning*
- ▶ ultimate goal is increase the likelihood of students with similar *ability* receiving similar *marks*, regardless of who grades their work (an outcome they are certainly entitled to; Sadler 2013)

During benchmarking, graders meet together to grade the same set of sample student work for a particular assignment with a rubric, working to **construct a shared understanding** of how to interpret and apply the rubric, by explaining and comparing their justifications for their grading decisions.



Lots of evidence from a variety of fields and educational contexts show that **benchmarking is effective at reducing grader variability** (Langer et al. 2003, Lawton and Braz 2011, O'Connell et al. 2016, Karnas-Heines 2021, Middleton et al. 2024, *inter alia*).

- ▶ particularly important in our context: LIN200 typically has many relatively inexperienced graders who may be unfamiliar with the expectations and standards of the department, the university, or even the country
- ▶ even experienced graders can benefit, since openly discussing different perspectives can highlight subtle aspects of the assignment (and thus, appropriate marks and feedback) that might be missed by individuals

Benchmarking helps **clarify instructor expectations** and the objectives of the assignment and the rubric.

- ▶ avoids multiple TAs independently asking the instructor for clarification later
- ▶ helps minimize students requesting reconsideration of their marks
- ▶ a few hours of benchmarking can save many hours of work and frustration later!

Benchmarking can help **reduce mental labour**.

- ▶ grading often has duplicated mental labour, with graders independently thinking through the same various issues
- ▶ benchmarking distributes some of this mental labour across all graders at the same time in one session, which improves grading efficiency
- ▶ more efficient grading benefits students (student work returned more quickly)
- ▶ more efficient grading benefits TAs (more time to spend on their own studies and research)

Benchmarking is also a structured opportunity for more **general pedagogical conversations** to happen, which can benefit the TAs' professional growth.

- ▶ graders may share useful pedagogical strategies and tips for grading, bookkeeping, time management, etc. that can be applied more broadly
- ▶ these conversations can and do happen spontaneously without benchmarking, but benchmarking provides a space for these conversations to arise naturally and be relevant to immediate concerns

Explicit protocol for benchmarking is crucial for its success (see Maki 2011, Holmes and Oakleaf 2013, Crisp 2017, and Schoepp et al. 2018 for steps to consider in benchmarking).

## Before our sessions:

- ▶ Lead WIT TA (LWTA) selects and anonymizes 4–6 student responses to the assignment to be benchmarked (sometimes called *anchor papers* in the pedagogical literature), demonstrating a range of student performance
  - ▶ 1+ that fairly clearly meets the expectations of an overall excellent response
  - ▶ 1+ that fairly clearly fails to meet the expectations an overall adequate response
  - ▶ 2+ intermediate responses (ideally, with different strengths and weaknesses)

## During our two-hour sessions:

- ▶ instructor introduces the assignment, rubric, and any other grading guidelines
- ▶ LWTA, instructor, and all TAs grade the same first sample student response, independently writing down scores for each component of the rubric
- ▶ when all are finished, the group collectively discusses the overall successes and challenges of the student response and how these relate to specific components of the rubric

- ▶ variation in grading practices are discussed as they arise
- ▶ instructor clarifies expectations as needed
- ▶ group consensus (hopefully!) achieved
- ▶ process is repeated for each sample student response
- ▶ throughout, everyone is given the opportunity to ask questions and seek or give advice
- ▶ LWTA monitors group discussion and tries to balance who is contributing by encouraging quieter TAs to speak up

## Context of this study

The **Writing-Integrated Teaching (WIT) program** in Arts & Science provides ongoing, practical, and discipline-specific training to all TAs in designated WIT courses.

- ▶ designed to help TAs develop their skills in teaching and assessing undergraduate writing
- ▶ implementation varies across departments in A&S
- ▶ in Linguistics, WIT training in large introductory courses like LIN200 typically includes benchmarking (among other training)

Our study concerns implementation of WIT in **LIN200: Introduction to Language**, a large introductory general-interest course intended for non-majors and requiring no previous background in linguistics, typically offered once per semester, with approximately 8–12 TAs.

- ▶ most of the 250–300 students are upper-year students from other majors trying to complete distributional requirements, so they are unlikely to pursue linguistics further
- ▶ disciplinary diversity of the students means there is much variation in their academic backgrounds and writing experience
- ▶ course's explicit connection to language also tends to attract students from diverse, multilingual backgrounds
- ▶ some homework assignments are short essay-style writing assignments, approximately 2–3 pages each

The **writing assignment in LIN200** for this study requires revision of a pre-existing sample partial essay.

- ▶ students revise the four paragraphs of the essay, each with different writing issues, corresponding to the four components of the general writing rubric used in the course
  - ▶ erroneous factual content
  - ▶ incorrect logical argumentation
  - ▶ problematic overall structure and cohesion
  - ▶ problematic mechanics and tone
- ▶ students also write a suitable concluding paragraph of their own

This assignment was used across two iterations of the course.

- ▶ **Winter 2020:** assignment was **not benchmarked** because a different first writing assignment was benchmarked
- ▶ **Fall 2020:** assignment was **benchmarkd** because the first writing assignment from Winter 2020 was not used

It wasn't the original intention, but this is a reasonable accidental set-up for testing the effects of benchmarking, with a control (Winter 2020) versus treatment (Fall 2020).

There are some caveats...

- ▶ different students
- ▶ mostly different TAs (though one TA was common)
- ▶ different modality due to pandemic-era education (mostly in-person with disruption in Winter, entirely online in Fall)
- ▶ slightly different instructors (I taught it solo in Winter and then co-taught it with Suzi Lima in Fall)

How can we meaningfully compare grading in the two iterations to determine whether benchmarking was an effective pedagogical intervention?

**Wait, but...** what does *effective* even mean anyway?!

# Many-facet Rasch modelling

Suppose  $p_{ij}$  is the probability that a student  $i$  gets the correct answer on some assessment item  $j$ . Then the **odds of the student succeeding are**  $\frac{p_{ij}}{1-p_{ij}}$ . Rasch (1960) used **logistic regression to model the logarithm of these odds (log-odds)** as the difference of the student's underlying unknown **ability**  $A_i$  and the item's underlying unknown **difficulty**  $D_j$ .

$$\underbrace{\ln \left( \frac{p_{ij}}{1-p_{ij}} \right)}_{\substack{\text{log-odds} \\ \text{measured in logits} \\ \text{(logistic units)}}} = \overset{\text{ability}}{\underset{|}{A_i}} - \underset{\substack{| \\ \text{item difficulty}}}{D_j}$$

Basic properties of a simple Rasch model:

- ▶ if student ability matches item difficulty (i.e.  $A_i = D_j$ ), then student is just as likely to answer the item correctly as incorrectly (i.e.  $p_{ij} = 0.5$ )
- ▶ with higher ability and/or easier item ( $A_i > D_j$ ), success rate goes up ( $p_{ij} > 0.5$ )
- ▶ with lower ability and/or more difficult item ( $A_i < D_j$ ), success rate goes down ( $p_{ij} < 0.5$ )

Note that while Rasch modelling is framed here in terms of students and grading, it is a **general model** with many other uses beyond educational assessment: agriculture (Moral and Rebollo 2012), health (Apon et al. 2021), psychology (Teye-Kwadjo and de Bruin 2022), etc.

Some key assumptions and limitations:

- ▶ **linearity:** ability and difficulty share the same linear scale (not required to be known, but assumed to underlie the system; common assumption across statistical modelling)
- ▶ **unidimensionality:** ability and items are all related to the same singular skill (which can be broadly construed as consisting of many inter-related subskills; Andrich 2002)
- ▶ **binarity:** items are dichotomous (i.e. either right or wrong, with no partial credit, as on typical multiple choice questions)

Binarity is a serious limitation, because it precludes Rasch analysis of many typical grading scenarios that involve polytomous ordinal grading (partial credit, percentages, rankings, etc.).

Andrich's (1978) solution to the binarity problem was to extend the model to include an additional term for the underlying unknown **difficulty  $G_k$  in progressing to a particular grade level  $k$** , given success at the next lowest grade level  $k - 1$ . Thus, if a student is able to get a C+,  $G_{B-}$  would represent how hard it is for them to get a B-. Andrich also changed the odds from Rasch's original overall success/failure ratio to a ratio of the probabilities of student  $i$  achieving grade level  $k$  versus  $k - 1$  on item  $j$ .

$$\ln \left( \frac{p_{ijk}}{p_{ij(k-1)}} \right) = A_i - D_j - G_k$$

odds  
ability      grade level difficulty  
log-odds      item difficulty

Masters (1982) refined Andrich's solution by relativizing grade level difficulty to individual items, allowing each item to have distinct scoring structure rather than a single universal scale, so  $G_{jk}$  represents the difficulty of progressing to grade level  $k$  on item  $j$ , given success at grade level  $k - 1$  on the same item.

$$\underbrace{\ln \left( \overbrace{\frac{p_{ijk}}{p_{ij(k-1)}}}^{\text{odds}} \right)}_{\text{log-odds}} = \overbrace{A_i}^{\text{ability}} - \underbrace{D_j}_{\text{item difficulty}} - \overbrace{G_{jk}}^{\substack{\text{grade level} \\ \text{difficulty} \\ \text{by item}}}$$

**Many-facet Rasch modelling (MFRM):** Linacre (1989) had the insight that the difficulty a student faces from being graded by a particular person could also be added as an additional term (“facet”) in the model (indeed, *any* factor that could potentially affect student performance could be included as a facet: time of day for timed assessment, tutorial section, POST, gender, etc.). This facet represents the underlying unknown **severity  $S_m$  of grader  $m$**  (with the odds adjusted appropriately to include  $m$ ).

$$\underbrace{\ln \left( \frac{p_{ijkm}}{p_{ij(k-1)m}} \right)}_{\text{log-odds}} = \overset{\text{ability}}{A_i} - \underset{\text{item difficulty}}{D_j} - \overset{\text{grade level difficulty by item}}{G_{jk}} - \underset{\text{grader severity}}{S_m}$$

The model outputs many different statistics about the various facets used, as well as the model itself. For this presentation, the two MFRM statistics of interest relate to the individual graders in the grader facet:

- ▶ **severity**: an estimate of how inherently severe/lenient a grader is (i.e. their unknown value for  $S$ ), based on the marks they assigned in comparison to estimates of student ability across all items and graders
- ▶ **fit**: a measure of how erratic/indiscriminate their actual grading is (i.e. how much the variability in their grading deviates from the variability we expect to see due to natural randomness)

Recall that grader severity  $S$  is subtracted from student ability when calculating the odds of student success.

- ▶ **positive severity score** ( $S > 0$ ): grader tends to assign **lower marks** than expected; that is, they **grade harshly**
- ▶ **negative severity score** ( $S < 0$ ): grader tends to assign **higher marks** than expected; that is, they **grade leniently**
- ▶ **zero severity score** ( $S \approx 0$ ): grader tends to assign **expected marks**; that is, they **grade fairly** (desired outcome)

The fit statistic is a positive number that can be used to see how much the actual observed marks fit the idealized predictions made by the model, with tolerance for variation due to natural randomness.

- ▶ **fit much greater than 1** (i.e.  $\text{fit} \gg 1$ , called a “**misfit**”): observed marks **vary much more** than expected; that is, the grader **grades too erratically** (e.g. giving unpredictably different marks to students with similar ability)
- ▶ **fit much less than 1** ( $\text{fit} \ll 1$ , “**overfit**”): observed marks **vary much less** than expected; that is, the grader **grades too indiscriminately** (e.g. giving similar marks to students with different ability); special case of fit of zero means the grader gives the same mark to everyone, regardless of ability
- ▶ **fit close to 1** ( $\text{fit} \approx 1$ ): observed marks **vary as expected** (desired outcome)

Severity and fit can be used to help diagnose different patterns of grading, making it easier to determine the best response to grading issues (targeted training, reviewing the rubric, mark calibration, etc.).

The effectiveness of some pedagogical intervention related to grading can also be measured by comparing severity and fit from before and after the intervention, which I present here for LIN200.

# Using MFRM in LIN200

Initial concern: are the student populations between the two iterations of LIN200 comparable?

Overall course marks had **similar means** (with slight increase) but **different variances** (with significant decrease).

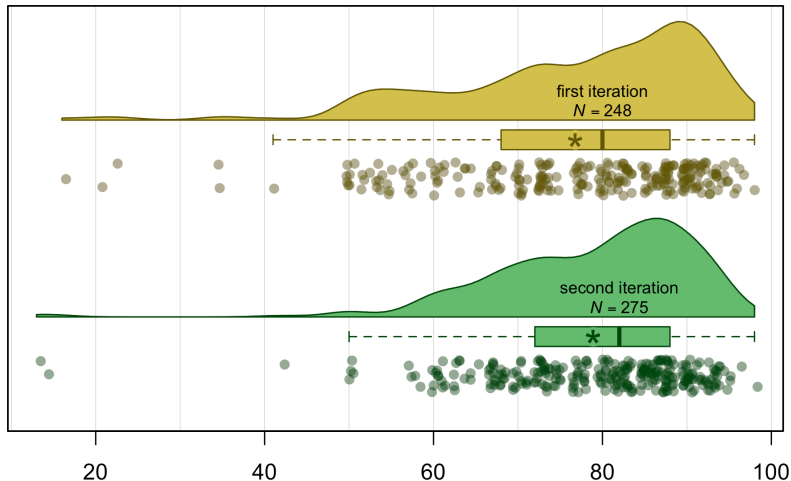
course marks	mean	variance
first iteration	76.8	219.6
second iteration	78.9	140.9
	$p = 0.07$	$p < 0.001$

This is likely due to the effects of the late emergency switch to online learning in Winter 2020 and resulting policy changes concerning late CR/NCR declaration, since students who might otherwise have dropped in Winter 2020 stayed in the course.

This would have an impact on overall course marks, but would not have been a factor for the writing assignment near the beginning of the semester.

Ignoring the lowest marks (below 55) eliminates the variance difference ( $p = 0.081$ ). So the bulk of the students, especially at the time of the writing assignment, are probably comparable.

## Overall course marks

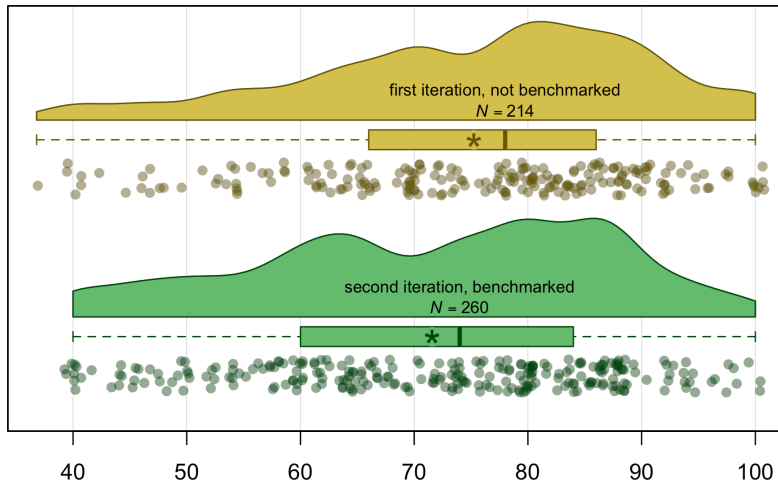


Interestingly, the marks for the writing assignment are quite different, much more than the overall course marks, and the differences go in the opposite direction from the overall course marks.

- ▶ **decreased mean in second iteration**, opposite direction of difference from the course marks!
- ▶ **slightly increased variance in second iteration**, again, opposite direction of difference from the course marks!

assignment marks	mean	variance
first iteration	75.26	213.36
second iteration	71.62	230.23
	$p = 0.009$	$p = 0.573$

## Marks for the writing assignment



So, it seems reasonably valued to compare the two iterations of the course based on overall population pattern.

Writing assignment marks had a decreased mean and increased variance, which is the opposite from how course marks changed (comparable but slightly higher mean, decreased variance).

Since this assignment also differed with the introduction of benchmarking in Fall 2020, that seems like a plausible cause of the changes in the assignment marks.

**Key question:** Are the changes in assignment marks reflected as expected changes in grader severity and/or grader fit according to MFRM (based on what was intended by benchmarking)?

Instructors intended for benchmarking to result in **lower marks** on this writing assignment to counterbalance high marks on easier non-writing homework assignments, so **mean severity should increase**, which is did. **Group success!**

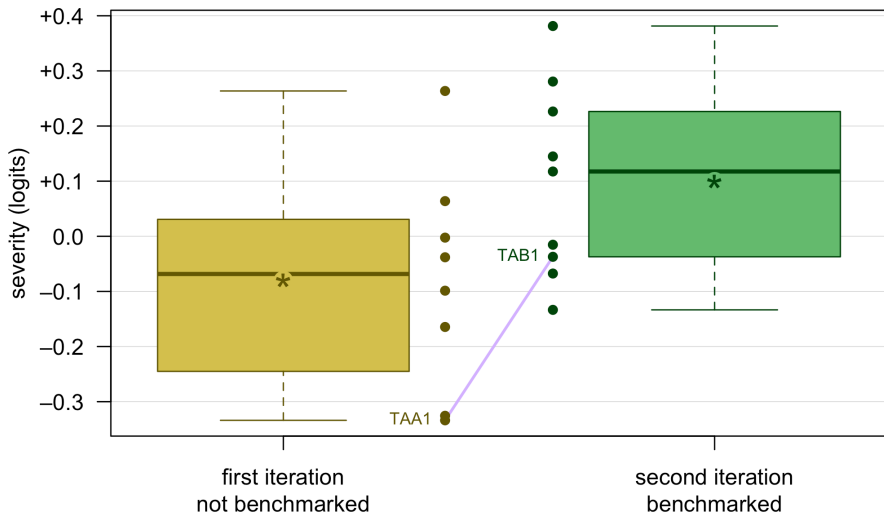
Benchmarking is also intended to **increase consistency** between graders, so **variance in severity should decrease**, which is did (marginally). **Trending towards group success!**

grader severity	mean	variance
first iteration	-0.0795	0.0305
second iteration	0.0997	0.0242
	$p = 0.035$	$p = 0.358$

In addition to group success, the one TA who was common to both iterations of the course went from being one of the most lenient graders ( $S = -0.335$ ) to having a severity close to zero ( $S = -0.037$ ).

**Individual success!**

## Grader severity



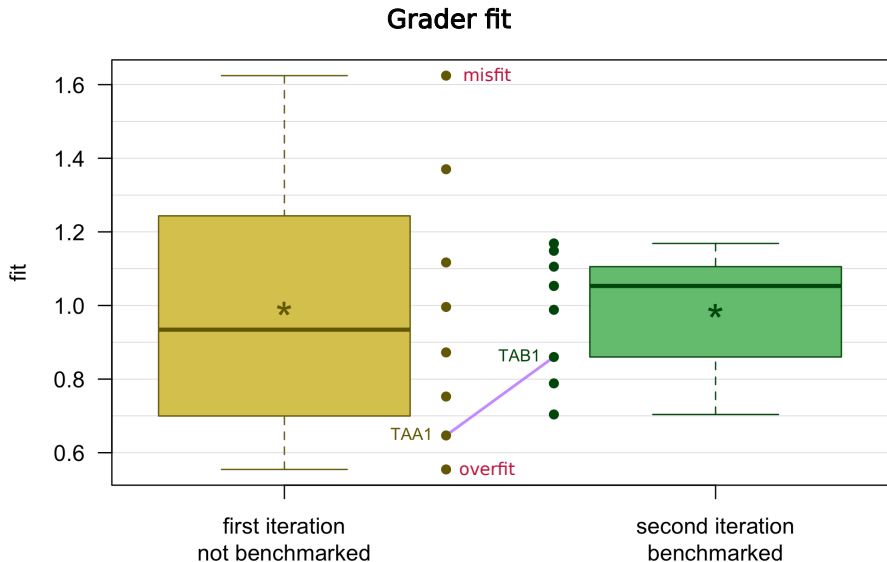
Benchmarking should **reduce the number of misfits and overfits**, so that fewer TAs are grading outside of expectations.

The first iteration has one misfit (fit = 1.624) and one overfit (fit = 0.555), while the second iteration had no misfits or overfits (all fits were between 0.788 and 1.169). **Group success!**

Again, due to expected consistency, benchmarking should also result in a **decrease in the variance in fit**, which it did. **Group success!**

grader fit	variance
first iteration	0.1029
second iteration	0.0215
	$p = 0.020$

And again, the common TA changed in the desired direction, from a fit of 0.647 to 0.860 (closer to the ideal fit of 1). **Individual success!**



**Wrap-up**

# Wrap-up

Inherent differences in grading practices between the two different groups of TAs could be the cause of differences in grader severity and fit. It might have nothing at all to do with benchmarking.

However, grader severity and fit after benchmarking were not different in arbitrary ways: they were different in exactly the ways that match instructor expectations for the results of benchmarking, for both the group and the common TA.

# Wrap-up

This suggests that **benchmarking is an effective pedagogical intervention for helping align TAs' grading practices with instructor expectations**, at least in terms of increased consistency (for both severity and fit), lower marks (due to higher severity), and better discernment (as measured by elimination of misfits and overfits).

But even if we had seen no effect on grading practices with benchmarking, there are still many other benefits as noted earlier, such as increasing efficiency of grading and promoting pedagogical discussions among the TAs.

# Wrap-up

Finally, I want to highlight MFRM as a powerful tool that can give us different kinds of information about the factors that can affect student performance (even beyond grader severity and fit).

It is useful to know whether graders are grading too harshly or too leniently (MFRM severity), as well as whether they are grading too erratically or without enough discernment (MFRM fit).

These are distinct properties, and using MFRM could help an instructor better identify the most appropriate interventions when grading issues arise, at any point in the course.

**Thanks!**

# References I

- Andrich, David. 1978. A rating formulation for ordered response categories. *Psychometrika* 43(4): 561–573. DOI: 10.1007/BF02293814.
- Andrich, David. 2002. Implications and applications of modern test theory in the context of outcomes based education. *Studies in Educational Evaluation* 28(2): 103–121. DOI: 10.1016/S0191-491X(02)00015-9.
- Apon, Inge, Nikki van Leeuwen, Alexander C. Allori, Carolyn R. Rogers-Vizena, Maarten J. Koudstaal, Eppo B. Wolvius, Stefan J. Cano, Anne F. Klassen, and Sarah L. Versnel. 2021. Rasch analysis of patient- and parent-reported outcome measures in the International Consortium for Health Outcomes Measurement Standard Set for cleft lip and palate. *Value Health* 24(3): 404–412. DOI: 10.1016/j.jval.2020.10.019.
- Brookhart, Susan M., Thomas R. Guskey, Alex J. Bowers, James H. McMillan, Jeffrey K. Smith, Lisa F. Smith, Michael T. Stevens, and Megan E. Welsh. 2016. A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research* 86(4): 803–848. DOI: 10.3102/0034654316672069.

# References II

- Crisp, Erin A. 2017. Calibration: Are you seeing what I'm seeing. *Intersection: A Journal at the Intersection of Assessment and Learning* Winter 2017: 7–13.
- Eells, Walter Crosby. 1930. Reliability of repeated grading of essay type examinations. *Journal of Educational Psychology* 21(1): 48–52. DOI: 10.1037/h0071103.
- Holmes, Claire and Megan Oakleaf. 2013. The official (and unofficial) rules for norming rubrics successfully. *Journal of Academic Librarianship* 39(6): 599–602. DOI: 10.1016/j.acalib.2013.09.001.
- Karnas-Heines, Colleen. 2021. Making assessment actionable through assessor training: A tool for building trust through moderation and calibration. *Intersection: A Journal at the Intersection of Assessment and Learning* 2(3). DOI: 10.61669/001c.24570.
- Langer, Georgea M., Amy B. Colton, and Loretta S. Goff. 2003. *Collaborative analysis of student work: Improving teaching and learning*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Lawton, Bessie Lee and Mary Braz. 2011. A grade-norming exercise to increase consistency and perceived consistency in grading among public speaking instructors. *Basic Communication Course Annual* 23: Article 7 (29–60).

# References III

- Linacre, John Michael. 1989. Many-facet Rasch measurement. Doctoral dissertation, University of Chicago, Chicago.
- Maki, Peggy L. 2011. *Assessing for learning: Building a sustainable commitment across the institution*. New York: Routledge, 2nd ed. DOI: 10.4324/9781003443056.
- Masters, Geoff N. 1982. A Rasch model for partial credit scoring. *Psychometrika* 47(2): 149–174. DOI: 10.1007/BF02296272.
- Middleton, Rebekkah, Kelly Lewer, Carolyn Antoniou, Helen Pratt, Suzanne Bowdler, Carley Jans, and Kaye Rolls. 2024. Understanding the processes, practices and influences of calibration on feedback literacy in higher education marking: A qualitative study. *Nurse Education Today* 135: Article 106106. DOI: 10.1016/j.nedt.2024.106106.
- Moral, Francisco J. and José M. Terrón Rebollo. 2012. Analysis of soil fertility and its anomalies using an objective model. *Journal of Plant Nutrition and Soil Science* 175(6): 912–919. DOI: 10.1002/jpln.201100361.

# References IV

- O'Connell, Brendan, Paul De Lange, Mark Freeman, Phil Hancock, Anne Abraham, Bryan Howieson, and Kim Watty. 2016. Does calibration reduce variability in the assessment of accounting learning outcomes. *Assessment & Evaluation in Higher Education* 41(3): 331–349. DOI: 10.1080/02602938.2015.1008398.
- Rasch, Georg. 1960. *Studies in mathematical psychology I: Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.
- Sadler, D. Royce. 2013. Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy & Practice* 20(1): 5–19. DOI: 10.1080/0969594X.2012.714742.
- Schoepp, Kevin, Maurice Danaher, and Ashley Ater Kranov. 2018. An effective rubric norming process. *Practical Assessment, Research & Evaluation* 23: Article 11. DOI: 10.7275/z3gm-fp34.
- Teye-Kwadjo, Enoch and Gideon P. de Bruin. 2022. Rasch analysis of the Proactive Personality Scale. *Psychological Reports* 125(5): 2788–2806. DOI: 10.1177/00332941211028110.