Metacognitive calibration: a methodological expansion and empirical application

International Society of
the Learning Sciences

Metacognitive calibration: a methodological expansion and empirical application

Samuel Tobler & Manu Kapur
samuel.tobler@gess.ethz.ch, manukapur@ethz.ch
Professorship for Learning Sciences and Higher Education, ETH Zurich, Switzerland

Abstract: The ability to judge performance accurately is essential for successful learning. However, statistics or measures to do so are frequently limited to binary judgments and not scalable. Moreover, they primarily assess only one dimension of the metacognitive calibration accuracy. In this methodological paper, we develop and discuss a new set of statistics to determine the calibration accuracy and the direction of miscalibration. Together, they indicate the extent of confidence accuracy and whether learners are overconfident or underconfident in their judgments. These statistics are scalable to non-binary judgment data. We then illustrated them in an empirical study with 34 doctoral students' performance judgment data which were assessed when answering domain-specific conceptual questions. Results from traditional measures were calculated, serving as a reference for the new measures' reliability. In addition, we developed an *R*-package implementing and visualizing the latter. The theoretical and practical implications are discussed.

Introduction

The concept of *metacognitive calibration* describes the idea of a learner's ability to correctly judge their task performance (Keren, 1991). In other words, learners that are able to judge their correct performance as correct and incorrect performance as wrong display a high metacognitive calibration. Since the seminal literature review by Lin and Zabrucky (1998), the importance of metacognitive calibration became a highly acknowledged concept in the field of educational psychology and the learning sciences. Metacognitive miscalibration through overconfidence regarding the own performance appeared to negatively influence learning by reducing cognitive processing efforts (Lin & Zabrucky, 1998). Inadequate underconfidence, on the other hand, might negatively impact self-confidence, which is essential for successful learning as through the affected intrinsic motivation (i.e., through perceived competence). Similarly, it might influence potential ability-grounded failure attribution (Bandura, 1986; Dweck, 1975; Ryan & Deci, 2000). As such, empirical studies provided evidence for the predictive power of anxiety and reduced self-confidence (i.e., underconfidence) on lower test performance (Barrows et al., 2013). Furthermore, failure-driven problem-solving (e.g., Productive Failure; Kapur, 2014), in contrast to success-driven problem-solving, was shown to partly increase students' metacognitive calibration, potentially due to greater opportunities for self-evaluation (Sinha & Kapur, 2021a). In other words, the failure to successfully solve a problem might be beneficial to increase calibration accuracy, next to all other alleged positive effects of initial struggle as effective preparation for future learning (Sinha & Kapur, 2021b). In conclusion, high accuracy in the judgment of performance appears to constitute often a beneficial prerequisite for knowledge acquisition in any domain. The inability to accurately assess the own metacognitive calibration might hinder learning.

Whereas one major field of research is concerned with the analysis of metacognitive calibration in different situations, as presented above, another research branch aims to explore various possibilities of enhancing learners' calibration accuracy to ultimately facilitate learning. For instance, Xia and colleagues (2019) could show that students' reflections on their own performance contributed to a more accurate metacognitive calibration. In contrast, the calibration assessment after repeated judgments of learning revealed enhanced underconfidence, indicating that students who are continually asked to judge their performance might become less confident about their responses over time (Koriat et al., 2002). Yet, providing feedback on the actual performance might play an important role in an individual's performance judgment improvement, as empirical evidence suggested (Callender et al., 2016). Similarly, delayed conceptual summarizing (Thiede & Anderson, 2003) or strategy training for assessing calibration (Nietfeld & Schraw, 2002) supported students' calibration accuracy. Hence, next to the suggested significance of metacognitive calibration for learning, there is substantial empirical evidence on how to promote calibration accuracy.

However, to profit from theoretical and empirical work that investigated various ways of improving metacognitive calibration accuracy to positively affect learning, it is irrevocable to have a well-substantiated statistic to estimate this accuracy.

Theoretical background

In a recent explorative calibration accuracy comparison study, Schraw and colleagues (2014) evaluated the most commonly used statistics (e.g., *d'*, *gamma*, *G*-index). They found that the appropriate metrics of choice depend on the research question. One measure alone was shown to only be rarely sufficient to establish an estimation of calibration accuracy (Schraw et al., 2014). A particular reason for this conclusion is, however, inherent to the different measures themselves. Whereas *d'* measures the standardized difference between correctly judging the correct answer and wrongly judging a wrong answer as correct, this measure alone mainly indicates whether learners are rather over- or underconfident, hence the direction of the miscalibration. Additionally, the standardization of this measure does not allow for a direct comparison of calculated indices from different studies, as the results will depend on the measured variance in the specific sample population. The *G*-index, in contrast, looks at the proportion of correct judgments to incorrect judgments. Thereby, this index determines the accuracy of individuals but does not make any statement regarding the direction of the miscalibration. In other words, a low accuracy result does not indicate whether participants are rather underconfident or overconfident, despite the essential difference between these two concepts regarding learning. Lastly, *gamma* follows a similar approach to the *G*-index by subtracting the product of the wrong judgments from the product of the correct judgments, over the sum of both products. Whereas the weighting of over- and underconfidence is different from the *G*-index, also in this case, it remains impossible to determine the direction of the miscalibration. Overall, and in agreement with the suggestion from Schraw and colleagues, it appears that no measure drastically outperforms the others but that they are rather covering different aspects of the concept of calibration accuracy.

A further limitation of some of these statistics for performance judgment assessment is their restriction to dichotomous data sets. Calibration estimations can then only be determined if the judgments are assessed in a yes-or-no format (i.e., "Are you sure about your response?" with answer options "yes" and "no"). However, in practice, an individual's judgment of their own answers might not always be so straightforward, and more fine-grained answering options could yield more accurate approximations of the calibrations. One way to increase the sensitivity of calibration analyses is by asking a similar question but assessing responses on a 4-point Likert scale (yes / rather yes / rather no / no). Having four items to choose from still forces the participant to decide but allows them to indicate uncertainties. Nonetheless, the currently available statistics for examining calibration accuracy do often fail to come up for the need for greater sensitivity. Thus, the generally used statistics might answer specific research questions very well but not coherently report accuracy and the miscalibration direction. Also, they are often limited to binary data sets, thus restricting their application in empirical studies.

In this paper, we propose a novel set of statistical measures for assessing calibration accuracy and the direction of a miscalibration. Moreover, we demonstrate the applicability of these statistics for binary data as well as data assessed in 4-point Likert scales, thereby taking into account the non-binary nature of actual performance judgments. Thus, we aim to advance the methodological standards of determining and interpreting metacognitive calibration. Lastly, we present and apply an *R*-software package to easily calculate and plot the calibration accuracy and miscalibration direction on empirical data from a study with 34 doctoral students to illustrate the suggested methodological advancements and compare the results with conventional measures for reliability.

Methodological expansion of calibration accuracy and miscalibration

As previous research showed, scholars mostly focused on binary confidence judgment data (Schraw et al., 2014). Thereby, students' judgments are evaluated based on whether they correspond to the actual performance. These different combinations of performance and judgment can be visualized in a 2×2 matrix (Table 1, left). However, if working with non-binary data, this matrix must be expanded, for example, to a 2×4 matrix in the case of a 4-point Likert scale-based performance judgment (Table 1, right). Consequently, new statistics are needed.

A starting point for establishing new statistics comes from defining a robust measure for a binary system, which then can be scaled up, for instance, to a 4-point scale system. Of primary importance are thereby measures for overconfidence (1–sensitivity; Feuerman & Miller, 2008) and underconfidence (1–specificity; Feuerman & Miller, 2008). Simply put, the overconfidence ratings indicate the frequency with which a learner wrongly judges their answer as correct when they are wrong, thus being overconfident in their abilities. Likewise, underconfidence describes the frequency of correct answers that are wrongly judged wrong, thus indicating a learner's lack of recognizing their abilities (Table 2). The accuracy of one's metacognitive calibration depends on these two measures. High underconfidence and high overconfidence, as well as a combination of these two, must be reflected in such a value. Additionally, the frequency of their actual occurrence must be considered as well. If this is neglected, a participant with all judgments and performance correct beside one obtains the same calibration accuracy score as a participant with all judgments wrong. Thus, the values for under- and overconfidence must be considered in relation to their actual frequency. The resulting formula for the calibration accuracy is shown in Table 2.

Table 1

Performance-evaluation matrices for dichotomous and 4-point-based confidence judgments

	Performance			*Performance*	
Confidence Judgment	Correct	Incorrect	*Confidence Judgment*	Correct	Incorrect
Yes	a (true positive)	b (false positive)	Yes	a	b
No	c (false negative)	d (true negative)	Rather Yes	c	d
			Rather No	e	f
			No	g	h

Note. Letters *a* to *d* (left) and *a* to *h* (right) indicate the variables needed for the formulas used for the calculations displayed in Table 2. *Performance* specifies whether a specific problem was solved correctly or not. *Confidence judgment* indicates students' judgments of their own answers' correctness, either collected in dichotomous format (yes / no) or on a 4-point Likert scale (yes / rather yes / rather no / no).

Like the limitations of the statistics *gamma* and *G*-index, this new calibration accuracy measure does not make any statement regarding the direction of the miscalibration. However, it is possible to apply the same theoretical and mathematical reasoning to determine whether any calibration inaccuracy is due to over- or underconfidence. By relatively subtracting the false positive (*b*) from the false negatives (*c*), one obtains a similar statistic to *d'* that allows investigating the miscalibration direction (Table 2). Additionally, this miscalibration value is not based on standardization, as is the case for *d'*, but yields relative and comparable miscalibration estimates. Combining these newly established statistics (i.e., calibration accuracy and miscalibration), we can assess the accuracy of a performance judgment and the cause of any inaccuracy.

Table 2

Calibration calculation formulas for dichotomous confidence judgments

Statistic	Formula	Explanation of the formula
Overconfidence *O*	$O = b/(b+d)$	This value explains how often a student wrongly answers a question while wrongly believing to have answered it correctly.
Underconfidence *U*	$U = c/(a+c)$	This value explains how often a student correctly answers a question while wrongly believing to have answered it wrongly.
Calibration accuracy *C*	$C = 1 - \left(\frac{1}{a+b+c+d}\right) \cdot \left(\frac{b^2}{b+d} + \frac{c^2}{c+d}\right)$	The metacognitive calibration accuracy value is based on the relative occurrence of a students' over- and underconfidence judgment. 1 indicates perfect calibration, and 0 indicates full miscalibration.
Miscalibration *M*	$M = \frac{1}{a+b+c+d} \cdot (b-c)$	The miscalibration value explains the cause of any metacognitive calibration inaccuracy. *M* = 1 indicates full overconfidence, and *M* = –1 indicates full underconfidence.

Note. The variables *a* to *d* correspond to the participant-specific count values established as described in Table 1. If the nominator of any fraction is equal to 0, this specific fraction must be set to 0. For instance, if a student never judges an incorrect answer as correct (*b*), the overconfidence value must be set to 0. In this case, the calibration and miscalibration values remain dependent only on the underconfidence statistic.

Generally, there are two major advantages of using these statistics instead of a combination of the commonly used ones. First, the obtained measures are mathematically based on the same underlying construct and are, thus, directly comparable within and across studies. And second, both are directly scalable to any dimension of interest. Being able to estimate calibration for not only binary response judgments but those of a higher level might deepen the understanding of learners' actual calibrations. As such, we derived the formulas for performance judgments of a 4-point Likert scale (Table 3). The exact calculations to obtain the formulas in Tables 2 and 3 can be found in the supplementary materials on OSF (see methods).

Table 3
Calibration calculation formulas for 4-point-based confidence judgments

Statistic	Formula	Explanation of the formula
Overconfidence *O*	$O = \frac{xb+yd}{x(b+h)+y(d+f)}$	This value explains the weighted ratio of how often a student wrongly overestimates their performance in case the given answer is incorrect.
Underconfidence *U*	$U = \frac{xg+ye}{x(a+g)+y(c+e)}$	This value explains the weighted ratio of how often a student wrongly underestimates their performance in case the given answer is correct.
Calibration accuracy *C*	$C = 1 - \frac{1}{\text{tot}'} \cdot \left(\frac{(xb+yd)^2}{x(b+h)+y(d+f)} + \frac{(xg+ye)^2}{x(a+g)+y(c+e)}\right)$ with $\frac{1}{\text{tot}'} = \frac{1}{x(a+b+g+h)+y(c+d+e+f)}$	The calibration accuracy describes the inversed relative sum of the weighted overconfidence and underconfidence values. *C* = 1 indicates perfect calibration, and *C* = 0 indicates full miscalibration.
Miscalibration *M*	$M = \frac{1}{\text{tot}'} \cdot (x(b-g) + y(d-e))$	The miscalibration score describes the relative difference between the weighted incorrect answers (overconfident – underconfident). *M* = 1 indicates full overconfidence, and *M* = –1 shows full underconfidence.

Note. The variables *a* to *h* correspond to the participant-specific count values established as described in Table 1. If the numerator of any fraction is equal to 0, this specific fraction must be set to 0. For instance, if a student never judges an incorrect answer as correct (*b*) or as rather correct (*d*), the overconfidence value must be set to 0, and the calibration and miscalibration values remain dependent only from the underconfidence statistic. The factors *x* and *y* indicate the relative weighting of the individual values from the performance-evaluation matrix to come up for the different judgment certainty levels (*x* for "yes" and "no"; *y* for "rather yes" and "rather no"). The generalized formulas are described in the supplementary materials on OSF (see methods).

Weighting confidence ratings

Suppose working with a non-dichotomous performance-judgment matrix, as in those cases in which the response confidence was assessed with 4-point Likert scales. The weighting of the individual confidence judgments then gains importance. Not differently weighting the answers would reduce them again to a binary measure. Thus, introducing the weighting factors *x* and *y* could overcome the shortcoming of presently available measures that categorize judgments in a binary manner (Table 3).

Founding the weighting ratio ($w = x/y$) in theoretical elaborations, we propose one specific solution for this problem: attributing a three times higher weight to those ratings with greater confidence ($w = 3$). The rationale for doing so is motivated by the literature on and common practice of interpreting Likert-based data as interval data despite its ordinal nature (Wu & Leung, 2017). To understand why ordinal-scaled Likert data can be treated as interval scales in some instances, Boone's and Boone's (2012) distinguishment between Likert-type and Likert scale data comes into play. Whereas the first describes situations in which single items are compared, the latter is based on multiple items that describe together one characteristic. Thus, when having multiple items that constitute one composite characteristic, there is evidence in favor of analyzing them on an interval scale (e.g., Boone & Boone, 2012; Sullivan & Artino Jr, 2013).

Having concluded that interpreting Likert scales on interval data might be appropriate in specific situations, we need to assign values to the individual judgment options. On a 4-point Likert scale from "no" to "yes," the interval is set around the value 0 ("neither yes nor no"), whereby 0 is not a selectable option, aiming to enforce students' decisions. Looking at the two intermediate values ("rather no" and "rather yes"), it appears that they are mathematically twice as much represented on any interval scale than the border values ("yes" and "no"). In other words, if a student selects the option "rather yes," this answer implies that the student's decision of performance judgment was either in the interval of (0; "rather yes"] or ["rather yes"; "yes"). In contrast, a student's answer of "yes" suggests that the student's decision was only in the interval of ("rather yes"; "yes"]. Expanding this train of thought, we find the numerical decision interval of (0, 1.5) for the answer "rather yes" (at the interval value of 1) and [1.5,2] for the answer "yes" (at the interval value of 2). The absolute decision interval of the answer "rather yes" is thus three times as large as the interval of the answer "yes." Consequently, we can only mathematically account for this double representation if the weighting is set to 3 (see endnotes 1 & 2).

Empirical application

Methods

All data sets and annotated *R*-scripts used for the present analysis are openly available in an OSF online repository (https://osf.io/6pdjt/). The various functions of the novel *R*-package for the metacognitive calibration analysis can be directly installed as *R*-package (https://github.com/samueltobler/mcc).

Participants

The participants of the application study were 34 doctoral students in natural sciences at a highly-ranked European university. The participants were, in average, 27.1 years old (*SD* = 2.3), whereby 41% indicated to be female, 59% male, and 0% non-binary. Participation in the study was voluntary, and three vouchers from a local grocery store were raffled among all participants. The university's ethics commission approved all studies before their conductance.

Materials

The test materials consisted of nine multiple-choice questions that covered a fundamental concept of the participants' study field and were published as part of a validated concept inventory (α = 0.69; 95% CI: 0.54-0.84) (Tobler et al., 2023). The students' self-reported performance judgment was assessed for each question by asking them, "How confident are you with your response?" on a 4-point Likert scale with the descriptors *very unconfident, rather unconfident*, *rather confident*, *very confident*.

Metacognitive calibration *R*-package

The for this purpose developed *R*-package directly calculates metacognitive calibration accuracy *C* values and the miscalibration *M* estimations for data sets with performance results and performance judgments on a 4-point Likert scale. The calculations are based on the proposed formulas in Table 2. Moreover, conventional measures, including *d'*, *gamma*, and *G*-index are functionally integrated to directly compare the different statistics. Eventually, the package allows plotting the results for more informative analyses of the data set. The extensively annotated *R*-package can be directly installed in the *R* software environment from GitHub (see link above).

Procedure and analysis

The participants were recruited through university-internal mailing lists and asked to complete the online test alone and without further resources. There was no time limit for taking the test. However, we excluded participants who showed statistical duration outliers (*n* = 3) and those who were faster in finishing the test than it would take to read the individual questions (*n* = 4). The final sample size consisted of 27 participants (Age: *M* = 27.2, *SD* = 2.3; 30% female, 70% male, 0% non-binary).

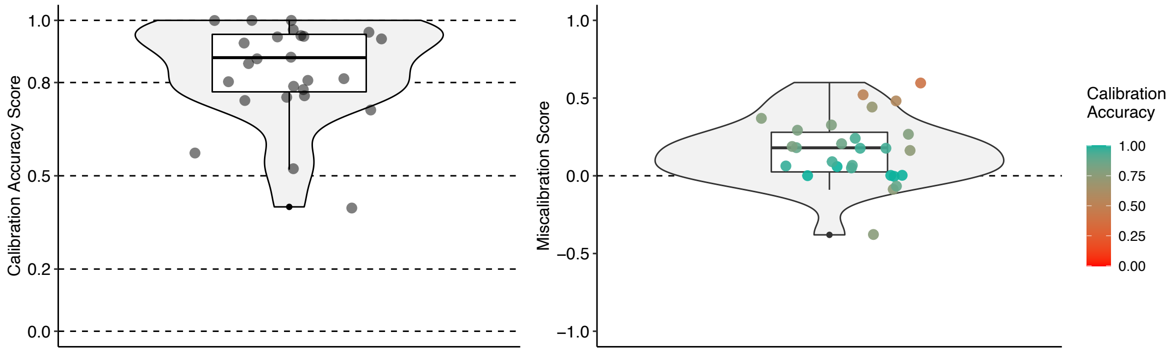
The test results were descriptively analyzed. The metacognitive calibration accuracy and miscalibration values were analyzed and plotted using the hereby introduced metacognitive calibration *R*-package. Additionally, we compared the results from the newly proposed calibration accuracy statistics with the results that would have been obtained by applying the commonly used calibration accuracy measures (i.e., *d'*, *gamma*, and *G*-index). Moreover, we investigated the correlation between calibration scores and actual performance. All analyses were conducted in the *R* software environment (R version 4.2.1; R Core Team, 2022). A list of all *R*-packages used for the analysis can be found in the supplementary materials on OSF.

Results

The normalized performance score analysis indicated that the participant understood the tested concept relatively well (*M* = 0.72, *SD* = 0.21, *min* = 0.11, *max* = 1.00). Furthermore, the results from the empirical application of the novel statistics revealed that most doctoral students answered the performance judgments with relatively high accuracy (*M* = 0.84, *SD* = 0.16; Figure 1, left). The miscalibration scores further indicated that most of the students, if demonstrating some calibration inaccuracy, were somewhat overconfident regarding their performance (*M* = 0.16, *SD* = 0.22). This finding is in line with the color-coded miscalibration scores demonstrating that those students with lower calibration accuracy rather were over- or underconfident but not both (Figure 1, right).

Figure 1

Metacognitive calibration and miscalibration values



Note. The dots in both sub-figures indicate the individual participants. The miscalibration scores (Figure 1, right) are color-coded according to the individual calibration accuracy. Green indicates perfect accuracy; red indicates complete inaccuracy.

Looking at the Pearson's correlation values determined for each comparison of newly proposed and priorly discussed measures, the results showed significant correlations between the new statistics and the *G*-index, but fewer between the new ones and *d'* or *gamma* (Table 4). Additionally, the correlation of *gamma* or *d'* accuracy values with the 4-point calibration accuracy and miscalibration scores were weaker compared to the binary values of the new measures. These results indicate that the 4-point-based calculations contain more information that the other measures cannot capture, explaining more variance and, thus, revealing more precise estimates. Lastly, we found no significant correlation between performance and calibration accuracy *C* (*r*(25) = 0.26, *p* < .18). Instead, the calibration results obtained by using the *G*-index statistics on the artificially binarized data set revealed a significant correlation with performance (*r*(25) = 0.57, *p* < .01). No significant correlations were found between performance score and *d'* (*r*(25) = -0.01, *p* = .97) or *gamma* (*r*(25) = -0.37, *p* = .06). However, a performance score-dependent visual breakdown of calibration accuracy and miscalibration descriptively indicates greater variability in accuracy and miscalibration with lower performance (Figure S1 in the supplementary materials).

Table 4

Statistic comparisons with empirical data

Measure	*M*	*SD*	1	2	3	4	5	6
1. Calibration accuracy *C* (4-point)	0.84	0.16	–					
2. Miscalibration *M* (4-point)	0.16	0.21	-0.64^{***}	–				
3. Calibration accuracy *C* (binary)	0.84	0.13	0.86^{***}	-0.62^{***}	–			
4. Miscalibration *M* (binary)	0.21	0.17	-0.56^{**}	0.89^{***}	-0.73^{***}	–		
5. *d'*	0.00	1.46	0.41^{*}	–0.19	0.43^{*}	-0.30^{*}	–	
6. *Gamma*	0.48	0.33	0.34	–0.29	0.46^{*}	–0.34	0.18	–
7. *G-index*	0.14	0.75	0.72^{***}	-0.42^{*}	0.89^{***}	-0.65^{***}	0.47^{*}	0.29

Note. The performance judgment ratings have been transformed to binary values to calculate the various statistics (*d'*, *gamma*, *G-index*). The newly proposed statistics have been evaluated with both the binary-transformed and the original 4-point Likert scale data. Statistically significant correlations are marked with an asterisk sign (* *p* < .05; ** *p* < .01; *** *p* < .001). *N* = 27.

General discussion and conclusion

The significance of being metacognitively well calibrated and, thus, able to accurately judge the performance has been repeatedly shown to positively affect learning (Lin & Zabrucky, 1998). Moreover, prior work has documented and compared various statistical approaches to accurately estimate a learner’s metacognitive calibration (Schraw et al., 2014). By comparing different statistics, they investigated which of them might show the best suitable measure for accuracy. Regrettably, they did not find a one-size-fits-all statistic to measure the latter but concluded that the appropriate measure must be chosen based on the research question. Furthermore, some of these measures can only be applied to dichotomous data but are not directly scalable to judgment data assessed on higher order Likert scales, for instance.

We developed a methodology that compensates for these two shortcomings. Our novel approach is grounded in two complementary statistics that are based on the relative occurrence of false positive and false negative performance judgments. These two statistics, the calibration accuracy and the direction of any miscalibration, yield a direct estimate of an individual’s metacognitive calibration. Furthermore, they explain any deviation from a perfect calibration in terms of underconfidence or overconfidence. Additionally, these two statistics are directly scalable from binary judgment inputs to 4-point Likert scale ratings, and generalizable beyond that as well. Thus, it appears that the newly developed methodology to estimate metacognitive calibration accuracy might overcome the major limitations of commonly used statistics. Also, it presents easily applicable measures that might be valuable for researchers in and outside the field of the learning sciences when working with calibration measurements. Nonetheless, triangulating the results with other metrics such as *d’* or *G*-index, recommended by Benjamin and Diaz (2008) or Schraw (1995), might provide supplementary validity of the calculated accuracy estimates.

Moreover, we empirically tested the new measures to demonstrate their performance. Statistical comparisons with the established measures revealed significant correlations in most cases, indicating that the new statistics assess a similar construct to conventional measures but explain more variance. Lastly, we developed a freely available and directly implementable *R*-package to apply the proposed formulas as well as more conventional ones to calculate calibration accuracy and visualize the results for facilitated interpretation.

Limitations and future directions

One limitation of the current approach is that the ordinal nature of Likert-scale data is ignored and interpreted as interval data. However, treating the data as ordinal data would not allow determining a calibration score, which emphasizes the different extent of judgment certainty. Instead, it would lead again to a dichotomous data set. Furthermore, assessing calibration accuracy over multiple items was shown to approximately resemble an interval scale (Boone & Boone, 2012). Nonetheless, the technical advancements as present in online conducted studies (Evans & Mathur, 2005) would allow collecting performance judgment data directly on true interval scales. Yet, the herein presented statistics could easily be expanded for higher degree matrices.

Future work could focus on applying these statistics in classrooms where most participants are novice learners and not experts in the field. Whereas students at the end of their educational career (i.e., post-graduate students) might have learned well over the years to accurately judge their own capabilities, learners at lower educational levels might struggle more to do so. Thus, using these measures to continuously investigate the students’ metacognitive calibration accuracy and the impact of success or failure on it might reveal more detailed insights regarding the development of calibration accuracy throughout their education. Similarly, future studies could emphasize assessing calibration accuracy with the proposed measures when testing new educational interventions to investigate their impact on this trait.

Endnotes

(1) Intervals are described according to the general notation standards for mathematical intervals. Round brackets indicate that all values until but without the start- or endpoint are included, and square brackets indicate that all values until and with the start- or endpoint are included. Mathematically expressed, this means $(0,1] = \{x \mid 0 < x \leq 1\}$.

(2) A weighting of $w = 2$ might appear more logical at first glance. To explain why this might be less exact, let’s reconsider the interval of [-2; 2] on a 4-point Likert scale. This interval would then result in the values of [-2; -1; 1; 2] for [“no”; “rather no”; “rather yes”; “yes”]. Like before, we have three 0.5 interval steps for the answer “rather yes” (0, 1.5), but only one for the answer “yes” [1.5, 2]. Suppose we now set the weight of “yes” answers to be double as high as that for “rather yes” answers. In that case, the relative weight of the different answers with respect to their abundance on the interval scale corresponds to $1 \cdot 1.5$ for the “rather yes” answer and $2 \cdot 0.5$ for the “yes” answer. Thus, the “rather yes” answer option still weighs more than the “yes” option $((1 \cdot 1.5)/(2 \cdot 0.5) = 1.5 \neq 1)$. Only if we set the weight to $w = 3$, we get an equal ratio (i.e., 1) for the weighting of “yes” and “rather yes” answers.

References

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. In *Social foundations of thought and action: A social cognitive theory.* (pp. xiii, 617–xiii, 617). Prentice-Hall, Inc.

Barrows, J., Dunn, S., & A. Lloyd, C. (2013). Anxiety, Self-Efficacy, and College Exam Grades. *Universal Journal of Educational Research*, *1*(3), 204–208. https://doi.org/10.13189/ujer.2013.010310

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. *Handbook of Memory and Metamemory*, 73–94.

Boone, H. N. Jr., & Boone, D. A. (2012). Analysing Likert Data. *Journal of Extension*, *50*(2).

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, *11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6

Dweck, C. S. (1975). The role of expectations and attributions in the alleviation of learned helplessness. *Journal of Personality and Social Psychology*, *31*(4), 674–685. https://doi.org/10.1037/h0077149

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research*, *15*(2), 195–219. https://doi.org/10.1108/10662240510590360

Feuerman, M., & Miller, A. R. (2008). Relationships between statistical measures of agreement: Sensitivity, specificity and kappa. *Journal of Evaluation in Clinical Practice*, *14*(5), 930–933. https://doi.org/10.1111/j.1365-2753.2008.00984.x

Kapur, M. (2014). Productive Failure in Learning Math. *Cognitive Science*, *38*(5), 1008–1022. https://doi.org/10.1111/cogs.12107

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273. https://doi.org/10.1016/0001-6918(91)90036-Y

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*(2), 147–162. https://doi.org/10.1037/0096-3445.131.2.147

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of Comprehension: Research and Implications for Education and Instruction. *Contemporary Educational Psychology*, *23*(4), 345–391. https://doi.org/10.1006/ceps.1998.0972

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *Journal of Educational Research*, *95*(3), 131–142. https://doi.org/10.1080/00220670209596583

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology*, *9*(4), 321–332. https://doi.org/10.1002/acp.2350090405

Schraw, G., Kuch, F., Gutierrez, A. P., & Richmond, A. S. (2014). Exploring a three-level model of calibration accuracy. *Journal of Educational Psychology*, *106*(4), 1192–1202. https://doi.org/10.1037/a0036653

Sinha, T., & Kapur, M. (2021a). Robust effects of the efficacy of explicit failure-driven scaffolding in problem-solving prior to instruction: A replication and extension. *Learning and Instruction*, *75*, 101488. https://doi.org/10.1016/j.learninstruc.2021.101488

Sinha, T., & Kapur, M. (2021b). When Problem Solving Followed by Instruction Works: Evidence for Productive Failure. *Review of Educational Research*, *91*(5), 761–798. https://doi.org/10.3102/00346543211019105

Sullivan, G. M., & Artino Jr, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542.

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, *28*(2), 129–160. https://doi.org/10.1016/S0361-476X(02)00011-5

Tobler, S., Köhler, K., Sinha, T., Hafen, E., & Kapur, M. (2023). Understanding Randomness on a Molecular Level: A Diagnostic Tool. *CBE—Life Sciences Education*, *22*(2). https://doi.org/10.1187/cbe.22-05-0097

Wu, H., & Leung, S.-O. (2017). Can Likert Scales be Treated as Interval Scales? A Simulation Study. *Journal of Social Service Research*, *43*(4), 527–532. https://doi.org/10.1080/01488376.2017.1329775

Xia, Y., Lee, H. Y., & Borge, M. (2019). Exploring students' self-assessment on collaborative process, calibration, and metacognition in an online discussion environment. *13th International Conference on Computer Supported Collaborative Learning (CSCL) 2019*, 945–946.