Learning and Instruction 47 (2017) 33–42

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

The measurement of calibration in real contexts

Teomara Rutherford

Department of Teacher Education and Learning Sciences, College of Education, North Carolina State University, Raleigh, NC 27695, United States

ARTICLE INFO

Article history:
Received 22 April 2016
Received in revised form
12 October 2016
Accepted 17 October 2016
Available online 28 October 2016

Keywords:
Metacognitive monitoring
Accuracy
Calibration
Data
Self-regulated learning

ABSTRACT

Accurate judgment of performance, or calibration, is an important element of self-regulated learning (SRL) and itself has been an area of growing study. The current study contributes to work on calibration by presenting practical and predictive results of varying calibration measures from authentic educational data: elementary-aged students' interactions with a year-long digital mathematics curriculum. Comparison of predictive validity of measures show only small differences in explained variance in models predicting posttest performance while controlling for pretest. A combined model including Sensitivity and Specificity outperforms other single measures, confirming results in Schraw, Kuch, & Gutierrez (2013); however, results show that student patterns of calibration within these data differ from those assumed in simulation studies and these differences have implications for the calculability of popular calibration measures.

1. Introduction

The ability to accurately judge one's performance is a foundational aspect of self-regulated learning (see Winne & Hadwin, 1998; Zimmerman, 2008). This accuracy is sometimes termed *calibration*, and across numerous contexts has been found to have its own relation with academic achievement (Stone, 2000; e.g., Bol, Riggs, Hacker, Dickerson, & Nunnery, 2010; Ots, 2013; Koku & Qureshi, 2004). Although calibration is noted as an important skill for learning (Alexander, 2013), and is a popular area of research, unanswered questions remain about the nature of calibration. The current study contributes to extant discussions regarding the best way to measure and calculate calibration (e.g., Masson & Rotello, 2009; Nietfeld, Enders, & Schraw, 2006; Schraw, 2009) by presenting a comparison of the practical and predictive results of varying calibration calculations for data obtained from student interactions with a year-long digital mathematics curriculum. In addition, it contributes to the empirical research regarding calibration by presenting results on the predictive power of calibration measures in an authentic elementary mathematics setting.

Calibration generally refers to the agreement between perception of task performance and actual performance (Nietfeld et al., 2006; Stone, 2000) and can be operationalized in a variety of ways. In particular, the choice of how to calculate measures of calibration affects conclusions drawn. Researchers can focus on absolute calibration (e.g., Bol & Hacker, 2001; Huff & Nietfield, 2009) or can investigate the direction of the calibration (e.g., Chen, 2002; Mengelkamp & Bannert, 2010). By way of illustration, in Fig. 1, two students, Sarah and Jenny, have the same level of calibration (looking item-by-item at the agreement between confidence and correctness), but Sarah displays an overconfident bias whereas Jenny is not consistently biased in either direction. Even among measures only looking at agreement, calculations may differ in how they treat these agreements. It is differences between these calculations of calibration upon which this study focuses, asking (1) Which measures of calibration can accommodate real-world data of accuracy and confidence judgments? and (2) Among these measures, which display the greatest predictive validity?

1.1. Calibration's role in self-regulated learning

Although there are a number of theories of self-regulated learning, all generally involve some process in which students set goals, monitor their progress toward these goals, and adjust their performance accordingly (e.g., Pintrich, 2000; Zimmerman, 1989; Winne, 1995). Student ability to accurately assess performance is important at each stage of the process. In the planning or forethought phase, student self-efficacy informs goal-setting (Bandura, 1986). Although slightly over-positive self-efficacy may be most adaptive for setting attainable goals, an over-inflated sense of self-efficacy may result setting too lofty a goal, resulting in failure, accompanied by discouragement and disengagement (Bandura,

E-mail address: teyarutherford@gmail.com.

http://dx.doi.org/10.1016/j.learninstruc.2016.10.006
0959-4752/

Sarah

#	Acc	Conf	Match
1	Y	↑	☑
2	N	↑	☒
3	N	↑	☒
4	N	↓	☑
5	N	↓	☑
	20%	60%	60%

Jenny

#	Acc	Conf	Match
1	Y	↑	☑
2	N	↑	☒
3	Y	↑	☑
4	Y	↓	☒
5	N	↓	☑
	60%	60%	60%

Fig. 1. Illustration of calculation of item-by-item as compared to more macro levels of calibration.

1986; Winne, 2004). As students work toward their goals, they adjust their strategies and resource allocation as they monitor their success (Nelson, 1996; Pintrich, 2004; Winne, 2001). Those students who determine that they are not performing at an appropriate level will attempt to rectify the situation by exercising control (Winne, 1995). It is this determination where the current study is focused: measures of calibration provide us with indications of student ability to accurately assess their performance.

1.2. Comparisons of calibration measures

In selecting measures of calibration, prior research has noted the importance of aligning the purpose of the study with the selected measure (Boekaerts & Rozendaal, 2010; Nietfeld et al., 2006; Schraw, 2009). Various measures may be complementary in that they can provide information on absolute accuracy, bias, or the ability to distinguish between correct and incorrect items (see Boekaerts & Rozendaal, 2010; Schraw, 2009). Choice of measures also can be driven by underlying assumptions about the monitoring process, for example, whether monitoring of potential correct and incorrect answers happens through a single process or separately through distinct processes (Schraw, Kuch, & Gutierrez, 2013); see discussion 1.2.4, below).

Practical considerations beyond the match with research question may also guide the choice of measure. For example, it has been suggested that for young children, measures with fewer choices reduce the cognitive load and allow for more accurate calibration scores (see e.g, Lyons & Ghetti, 2011). The Jenny and Sarah example illustrates a simplified dichotomous measure wherein students indicate whether they feel confident or not confident for each answer given.

1.2.1. The calculation of calibration indices and the practical issue of missing quadrants

The use of such a dichotomous measure in relating accuracy to judgments of confidence results in a 2 × 2 contingency table with cells depicted in Fig. 2. Looking to our examples: of the five quiz questions, Sarah would have one question in cell A, two each in cells B and D, and none in cell C. Jenny would have two questions in cell A and one each in the other three cells. Numerous indices have been created for the calculation of agreement based on the contents of these cells (see Feuerman & Miller, 2008; Schraw, 2009; Schraw et al., 2013). Table 1 presents a number of common indices expressed as functions of cells A through D and largely draws on descriptions of these formulas presented in Schraw et al. (2013) work. Some have emerged as more popular that others: Gamma

A. Confident & Correct	B. Confident & Incorrect
C. Not Confident & Correct	D. Not Confident & Incorrect

Fig. 2. 2 × 2 contingency table expressing the relations between accuracy and confidence.

(e.g., Mengelkamp & Bannert, 2010; Thiede, Anderson, & Therriault, 2003), d' or discrimination (e.g., Boekaerts & Rozendaal, 2010; Macmillan & Creelman, 1996), and G Index (e.g., Schraw, 1995; Tobias & Everson, 1998) have been particularly popular within metacognition and self-regulated learning research. Sensitivity and specificity have been more popular in medical research, where they represent successful detection of the presence or absence of a condition, respectively (e.g., Warnick, Bracken, & Kasl, 2008). Schraw et al. (2013) divided these ten common indices into interpretive families based on the dimensions purportedly captured by each measure—families are specified in Table 1. These and other measures have other empirical justifications (e.g., Gamma may be most useful in determining consistency of judgments whereas G Index may be most useful in measuring changes in calibration, see Nietfeld et al., 2006), and there are also practical ramifications of the selection of one measure over another. Due to the nature of the

Table 1
Common measures of calibration from 2 × 2 contingency tables.

Index	Formula
Sensitivity[a]	A/(A + C)
Specificity[a]	D/(B + D)
Simple Matching[b]	(A + D)/(A + B + C + D)
G Index or Hamann coefficient[b]	(A + D) − (B + C)/(A + B + C + D)
Odds Ratio[c]	AD/BC
Goodman-Kruskal Gamma[c]	(AD − BD)/(AD + BC)
Kappa[c]	2*(AD − BC)/[(A + B)(B + D) + (A + C)(C + D)]
Phi[c]	(AD − BC)/[(A + B)(B + D)(A + C)(C + D)]$^{1/2}$
Sokal Reverse[d]	[1 − [(A + D)/(A + B + C + D)]]$^{1/2}$
Discrimination (d')[e]	z(A/(A + C)) − z(B/(B + D))

Note. Formulas as represented in Schraw et al. (2013). Superscripts indicate the category of measurement as defined by Schraw et al. (2013): (a) Diagnostic efficiency, (b) Agreement, (c) Association, (d) Binary distance, and (e) Discrimination.