Learning and Instruction 24 (2013) 62–66

Contents lists available at SciVerse ScienceDirect

Learning and Instruction

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

# Calibration and confidence: Where to next?

John Hattie

*University of Melbourne, Australia*

ARTICLE INFO

*Article history:*
Received 24 May 2012
Accepted 31 May 2012

*Keywords:*
Calibration
Self-regulation
Confidence
Accuracy

ABSTRACT

One of the key feedback questions is "where to next?" and this article provides some directions as to where to next for research based on a review of the five articles in this special issue. The directions relate to the critical importance of calibration, the multidimensionality of calibration, the relation of calibration to self-regulation strategies, whether calibration is specific to the task or more general within the student, how to measure calibration, how much confidence should be given to partial knowledge when calibrating, the role of overconfidence and knowing "when one does not know", and how to improve the accuracy of judgments.



## 1. Introduction

A major purpose of a special issue is to generate a solid basis for a topic, to generate excitement for further research on the topic, but most of all to bring a new perspective to a worthwhile problem. Calibration is often sidelined with questions raised as to why teachers should care so much about how students calibrate their learning. The five articles discussed in this commentary turn this way of thinking upside down. These articles show how students' calibrations of their own confidence and accuracy can be important enablers or barriers to learning; they show how critical it can be for teachers to attend to these calibrations, especially for struggling learners, and they provide many arguments in favor of the *backward design* approach to teaching. In terms of this backward design, teachers should:

- explain to students exactly what they will be learning before beginning the teaching process;
- provide clear success criteria for them;
- adjudge what students already know and believe relative to these goals;
- ensure teaching is directed to reducing the gap between what students believe they know and understand, and what teachers want them to know and understand;
- ensure feedback is provided and received to reduce these gaps (Sadler, 1989).

The basic messages are: to attend to what students feel confident about, the accuracy they have of their prior learning, and their beliefs about the effectiveness of the learning methods they typically use—when they enter the classroom.

One of the strengths of these articles is to 'bring back' into the equation the (somewhat novel) notion of student calibration for effective teaching and learning. To *bring back* is important, as these notions of the power of students' beliefs and confidence have been a key part in many of the learning theories from Piaget (1962), Ausubel (1968), and Bandura (1986). So often the accuracy of these levels of confidence and beliefs is not sufficiently taken into account, with claims made about how students learn (best), how to help students construct knowledge, and how to ensure a greater number of students know more. A major message in these articles is that by ignoring students' beliefs concerning their confidence and accuracy, we are ignoring a major precursor to their learning.

From these articles, I can see at least nine major messages that should aim to spur further research on calibration. These messages relate to the importance of calibration, the polymorphous nature of calibration, the importance of self-regulation, whether the task is specific or more general, the measurement issues, the way students consider the role of partial knowledge, the place of knowing when one does not know, and the improvement of the accuracy of judgments.

## 2. The importance of calibration

In my review of more than 900 meta-analyses on the factors that most affect student learning, student expectations and self-reported grades came out on top (Hattie, 2009, 2012). This means that if a teacher were to say to a group of students "it is now *test time*, however, before you complete the test, I would like you to estimate the score or grade you think you would get," the students would be very adept at such a task; they can calibrate their expected

*E-mail address:* jhattie@unimelb.edu.au.

0959-4752/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.learninstruc.2012.05.009

performance across a series of items quite accurately. Kuncel, Crede, and Thomas (2005) completed a meta-analysis of 37 independent samples of more than 60,000 students concerning the accuracy of self-reported grades compared to recorded grades. They found college students were more accurate at calibrating their expected performance than high school students ($r = .84$ vs. .70), although the proportion of accurately reported grades was higher for high school GPA (82%) than for college GPA (54%). The accuracy of expected performance was higher in mathematics and social science ($r = .85$) and lowest in art and music ($r = .67$); it was similar when predicting rank in-class (.76) with the raw score (.77); it did not differ for males (.79) and females (.82); and was higher for white (.80) rather than non-white students (.66). Most importantly, the accuracy declined markedly for students of lower ability (see also Mabe & West, 1982). On the one hand, this is good news as it supports the claims in the current articles that students, especially brighter students, are very good at judging their own performance. However, the news is not so good for students of lower ability who are not only lower achievers, but who also have poorer calibrations of their performance. As Dinsmore and Parkinson (2013) argue, not only do lower achieving students have less accuracy and less skills at learning, they also tend to be less likely to know they are inaccurate or understand how close they are to the desired learning.

There are some important differences between the levels of calibration examined in these articles in this special issue and in the meta-analyses reviewed in Hattie (2009). The former are more concerned with calibration at a specific or item level, whereas the latter involves calibration at a slightly higher level (i.e., "How well will I perform in this test?" as opposed to "How well will I perform on each item?"). Further, these articles are more concerned with the discrepancy between prior or current judgments of performance and a desired standard, whereas the aforementioned review sets the standard as *performance on a test*. This suggests that learners may be more effective calibrators of a generalized rather than specific performance. It also brings into sharp focus the importance of learners having clear understandings about what *success* looks like in order to effectively calibrate their performance.

## 3. Calibration is not a unidimensional notion

These articles show that our ability to accurately judge our performance, confidence, strategies, or problem-solving is not only learned (and thus can be taught), but can also be an enabler or barrier when engaging in more challenging learning tasks. Similarly, the accuracy of judgments by teachers about what students can or cannot do is essential to engaging students in learning and successful teaching. Calibration of the *gap* between a student's current desired performance should inform the provision of feedback aimed at closing this gap. When we have closed the gap, we have success in learning. As calibration relates to many parts of the learning system, it is clearly not a unidimensional but a polymorphous concept with various forms and appearances.

Students often receive too much information about their current performance (usually expressed in terms of whether or not it is sufficient). However, the more critical calibration relates to the accuracy of judgments *between* "where they are now" and "where they are meant to be" – that is, the difference between current to target status. Effective calibration can be used to inform students of what success in the domain of learning might look like, and how far away they are from success. In many classroom situations, learners have little or no knowledge of the standards or success criteria of lessons. Frequently, they devolve to generalized beliefs about what is needed to accomplish the standard (e.g., looking busy until the bell rings, neatness, or length of answers). Success can come in many forms and appearances: measures of knowledge, correctness of an answer, effectiveness of a strategy, or attainment of a specific learning goal. The multifaceted nature of calibration makes it difficult to define, and consequently, to measure.

## 4. The importance of calibration to self-regulation

It seems a tautology to claim that learners' calibration of their confidence and accuracy are key parts of metacognition and self-regulation. There can be little, if any, self-regulation unless students have some knowledge or beliefs about their current and desired learning state. This involves not only a sense of accuracy about both states, but also a sense of confidence that the student has the required learning or study strategies to reduce the gap – without this confidence, the student could sit back and wait for the teachers to reduce the gap. Teachers may exacerbate this by ignoring the status of the learner's current understanding, reducing the challenge of the goal, changing the goal from understanding to mere engagement in tasks, or over-talking and under-listening to the student. The question missing from these current studies is the degree to which students can accurately gauge their progress. This represents an opening for others to build on the studies presented in this special issue.

There have been many studies showing the importance of confidence or self-efficacy in learning, but these articles are more concerned with the accuracy of this confidence. If our confidence is high and our accuracy is low then there is a major problem. These articles show that struggling learners are more likely to have high confidence in their prior knowledge (even if it is inaccurate) and this can become a major barrier to their learning and receptiveness for learning. When new information comes before these students it can easily be rejected as it does not fit with prior (and often inaccurate) knowledge, and consequently, confidence in learning such new material can decrease. This can lead to the student being less likely to retry, to re-strategize, to seek help, or to persist in learning: a vicious cycle of learning.

## 5. Is calibration specific to the task or more general to the student?

Dinsmore and Parkinson (2013) were interested in students' explanations for their confidence ratings. They use Bandura's (1986) model of reciprocal determinism to show the personal, behavioral, and environmental factors that influence the forming of such ratings for students. They showed that students were able to take multiple factors into account when making their confidence judgments, including text characteristics, item characteristics, prior knowledge, and guessing (in that order), although 76% used only one and not multiple sources. This echoes the findings of Brunswick researchers that although people typically claim to use multiple cues, most will use a very small subset when making judgments, although this small subset can differ among people (Cooksey, 1996).

## 6. How to measure calibrations

The core of the notion of calibration is a discrepancy between our judgment and the accuracy of the situation. Discrepancy scores, also known as change or gain scores, have a long and often, notorious history, which many of the arguments in these papers rehearse in the context of calibration. It has long been known that discrepancy scores tend to have low reliability under certain circumstances, leading to Lord (1956) noting that "differences between scores tend to be much more unreliable than the scores themselves" (p. 429). In one of the most cited articles in psychology, Cronbach and Furby (1970) concluded their classic analysis of change scores with the statement that, "it appears that investigators

who ask questions regarding gain scores would ordinarily be better advised to frame their question in other ways" (p. 80). The measurement of discrepancy has come a long way, and there are now many worthwhile methods such as growth models, latent change models, and value added models. However, these methods are not the ones referred to in these articles. Future studies of calibration could use these more recent and reliable methods, rather than depending on many of the outdated discrepancy measures outlined in these articles.

The measurement of calibration may be made unnecessarily difficult by dichotomizing the confidence and accuracy responses. Schraw, Kuch, and Gutierrez (2013) considered calibration as a dichotomous relation (correct, incorrect; high and low) between task performance and judgment about that performance. Subsequently, they sought the optimal measurement procedure for these 2*2 matrices. This search has a long history (Warrens, 2008), but perhaps could best be avoided by simply increasing the number of categories. When there are about five categories, then the advantages of a more continuous measurement take effect.

Schraw et al. selected ten measures and used Monte Carlo simulations in which they varied different parameters, particularly the degree of correspondence between prediction and actual success. The correlations between these measures were close to perfect in nearly all cases and thus it is somewhat surprising that two factors could be fitted to such a saturated set of correlations. The loadings on the first factor show that there is remarkable redundancy in these measures. Simple math *c*, *g* index, odds, gamma, kappa, phi, Sokal and d′ all perfectly load on this factor and have close to perfect correlation between them. There is a similar pattern for the accuracy monitoring data. The researchers named one factor "specificity" and the other "sensitivity," which is unusual as these were the only two measures that share less than perfect covariance. Their claim that these two factors are uncorrelated is not convincing as they purposefully constrained them to be so (by choosing varimax rotation). The results of this study suggest that any of these measures may be sufficient to measure the accuracy of calibration, if it is conceptualized in terms of binary measures (e.g., correct–incorrect).

## 7. How much confidence should be given to partial knowledge?

In the 1970–1980's there was much work on confidence or the differential weighting of multiple-choice items with the objective of allowing for partial information in the distracters, rather than all information residing in the correct response. Wang and Stanley (1970), however, showed that any extra information gained from asking students to declare probabilities or weightings on items to reflect their confidence in choosing the correct answer, led to minimal changes in the estimates of reliability, and was not worth the extra cognitive and administrative workload for participants. Shuford and Brown (1975), probably the most well-known promoters of confidence-weighting of the era, argued that it was "patently desirable to broaden the responses that students are permitted to make to multiple-choice questions" (p. 142) because this allowed students to provide more information, taught students to weigh their strength of conviction, and verified that students were not thereby *chafing* under the limitations of the conventional one-choice, multiple choice test. With the advent of item response models, far more sophisticated methods became available for estimating the properties of items, and the optimal weighting models.

Dinsmore and Parkinson (2013) reinvented some of these earlier methods. They constructed items such that the correct option was weighted 4, the option that was not correct but within the same topic as the correct response was weighted 2, the in-domain incorrect response was weighted 1, and the incorrect, popular lore option received zero weighting. These led to random scores, as the estimate of reliability for their items was about alpha = .30 (Wainer & Thissen, 1996). They also used the more interesting method of magnitude scaling. While Dinsmore et al., note its rarity, I have used it in a variety of studies (e.g., Hattie & Fletcher, 2006; see also Cooksey, 1996). Not only do participants seem to enjoy using this method, there are additional measures of reliability such as the $R^2$ for each student (Hattie & Fletcher, 2006). These studies show that measures derived from magnitude scaling have excellent statistical properties including, ratio scaling, high reliabilities for the measure of discrepancy, and known anchoring properties.

## 8. The role of overconfidence and knowing when one does not know

Over the past ten years I have overseen the building and implementation of a national assessment reporting system for elementary and high schools (Hattie, Brown, & Keegan, 2005). In one of the research projects we asked teachers to estimate the difficulty of items that had known difficulty properties. In general, elementary teachers underestimated the average difficulty of the items by about a year and high school teachers overestimated by about a year. In another part of the application, students were invited to set targets for their learning, it soon became evident that if their calibrations were not defensible (most often they were too high) it became very difficult to recalibrate. We then only allowed the teachers to set targets, but used a non-linear regression based on previous scores to estimate a growth line to strongly hint where the calibration should be placed. Smith (2009) showed that when teachers set targets using these methods, the subsequent learning of their students successfully met or exceeded these targets, compared to teachers who did not or refused to accept targets.

A theme that comes through in these articles is that the problem seems to be more about overconfidence rather than under-confidence. As Hadwin and Webster (2013) notes, when learners are overconfident, they may fail to recognize when they should be actively regulating or experimenting with strategies to increase the likelihood of achieving their goals. If learners are under-confident, they may overinvest in cognitive and effective resources to monitor and regulate goals that they would successfully achieve anyway.

Following this finding of a higher prevalence of over- rather than under-confidence is the question of whether students know, when they do not know. It may be that students know *when* they know (i.e., calibrating success), but are less able to judge how *well* they do not know (calibrating failure). For example, van Loon, Bruin, van Gog, and van Merriënboer (2013) noted that when students are asked for judgments about which items they get incorrect (so they can study further on these concepts), they too often choose the wrong items. There is a tendency for overconfidence ("I chose the option I thought correct, and thus right now, it is correct in my opinion"). As van Loon et al. noted, poor accuracy can result when learners base their judgments on cues that are not valid indicators of test performance, such as: the ease of processing when studying the information; the perceived familiarity with the topic prior to study; and the amount of information that comes to mind at the recall test, instead of the quality of this information.

## 9. How to improve the accuracy of judgments

These articles hint at five major directions for teachers to improve the accuracy of judgments. First, van Loon et al. (in this

issue) demonstrated that the way to reverse the inaccuracy of judgments is to focus learners' attention on valid cues when judging learning, particularly when these are delayed, rather than immediate judgments. There may need to be some time between reflecting on the performance of a task and the accuracy of this performance. This relates to the comments above about the importance of calibrating at a general, rather than a specific level and this attends to patterns across a series of items and concepts rather than concentrating on specific items.

Second, one powerful way to enhance judgments is to provide learners with worked examples. Worked examples not only illustrate what success looks like, they also avoid students believing that if they find a solution, they thereby have the correct answer. Moreover, the presence of a worked example reduces the cognitive load for students in such a way that they can then concentrate on the processes that lead to the correct answer and not just on providing an answer. Crissman (2006) used 62 studies in a meta-analysis on the effects of worked examples on achievement and found an overall effect-size of .57, which is a reasonably high effect-size. There is more work needed, as Van Loon et al. has shown, even though calibration improved to a certain extent, when students were asked to compare their response to a standard, they still showed a high level of overconfidence.

Third, attend closely to students' prior knowledge. Teachers should have a clear understanding of the beliefs and knowledge that students bring to the lesson. We know that learners' prior knowledge has a large influence on what they understand, and this is particularly influential when the prior knowledge is inaccurate. Such inaccuracy clearly affects the monitoring of accuracy and confidence, and not surprisingly, when confidence in judgment (of prior wrongs) is high, then new learning is all the more difficult. As Van Loon et al. show, students rely more heavily on their prior knowledge than on studied information when answering recall test questions. Further, these researchers noted that learning outcomes might actually be better when learners are not able to activate prior knowledge compared to when they activate inaccurate prior knowledge. They showed that learners were highly overconfident when they had inaccurate prior knowledge (37%), but were far less overconfident (9%) when they were not able to activate any prior knowledge. Further, learners often decided to restudy the concepts for which they were not able to activate any prior knowledge. The simple conclusion would be that it is better to *know nothing*. However, the more appropriate conclusion would be that it is critical to know about and deal with inaccurate prior knowledge before learning can lead to reasonable progress. Again, lower achieving students are the most hampered by not knowing that what they know may be inaccurate, or only partially correct; preferring like all learners to believe that what they know is probably accurate.

Hadwin and Webster (2013) appear to contradict these claims by showing that confidence judgments are more calibrated with perceptions of current goal attainment than with perceptions of past goal attainment. That is, the current task at hand appears to exert a stronger influence on judgments of confidence than past experience. Their argument, not inconsistent with van Loon et al. (2013), states that students begin tasks with relatively stable notions of their confidence, but students who overestimate their abilities tend to be more variable in their judgments of goal attainment over time, but not in their judgments of confidence. The problem is that learners too rarely have external evaluations of their progress, and I would add, too rarely have notions of what success on the task should look like. Hence learners rely on their own goals and their own judgments of how well those goals are being met. Too often they may aim for performance goals such as completing the task on time and to the required length although accuracy is not necessarily part of these performance goals.

Fourth, these articles continually return to the power of diagnosis and feedback. Such clinical analyses on behalf of teachers requires them to understand students' prior knowledge, what the students believe to be accurate, how confident the students are as learners, and the degree to which they ensure students know what is being learnt and what success looks like in the lesson(s). Feedback to ensure confidence and accuracy in learning is necessary therefore, not only to learning, but to enhance self-regulation by the students (Hattie & Gan, 2011).

Fifth, Hadwin and Webster (2013) concluded that teachers might need to examine the quality of goals that learners set (i.e., are they good and reasonable goals, given the purpose of the lessons?). I would add that it is important to also examine whether the goals include concepts of success (e.g., do students know what an A, B, C, or such look like and where they are aiming?). It may be necessary to be more explicit about the nature of the differences between an A and B, B and C, and so on; about where students are in their progress toward the goals; and about whether learners' evaluations of their own goal attainment are calibrated with actual evidence of goal attainment within and across study episodes.

All these claims beg more *listening* to learners during the teaching process. Listening to what prior knowledge they bring; listening to their understanding of the goals of learning; listening to where they are moving, from priors to goals; and, listening to their conceptions of confidence and accuracy in answering these questions. Hence, more self-talk, more self-consequences, and more self-calibration are the keys for teachers to understand how to structure lessons, structure tasks, structure scaffolding, and thus enhance learning and the student's confidence in their accuracy of learning. This means not only being explicit to students about learning intentions (what we are learning, rather than what we are doing) but also being explicit about the goals of the lessons (what does success look like). Most classrooms, however, are dominated by teacher talk, tests which reinforce a student's current ranking in their class, and a lack of feedback and teaching about skills and strategies (Kane and Staiger, 2012). In such conditions, it is not that necessary for students to calibrate their current or long-term progress through their learning. May be the most fruitful next step in calibration research is to ask how and when teachers are successful in calibrating the impact they are having on students learning (see Connolly, Klenowski, & Wyatt-Smith, 2011; Wyatt-Smith, Klenowski, & Gunn 2010).

## References


Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart & Winston.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs NJ: Prentice Hall.

Connolly, S. R., Klenowski, V., & Wyatt-Smith, C. (2011). Moderation and consistency of teacher judgement: teachers' views. *British Educational Research Journal, 30*(1), 1–22. http://dx.doi.org/10.1080/01411926.2011.569006.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. San Diego, CA: Academic Press.

Crissman, J. K. (2006). *The design and utilization of effective worked examples: A meta-analysis*. Unpublished doctoral dissertation, The University of Nebraska–Lincoln.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change" – or should we? *Psychological Bulletin, 74*, 68–80. http://dx.doi.org/10.1037/h0029382.

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14.

Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: examining the nature of judgments of confidence. *Learning and Instruction, 24*, 37–47.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. Oxford, UK: Routledge.

Hattie, J. A. C. (2012). *Visible learning for teachers. Maximizing impact on achievement*. Oxford, UK: Routledge.

Hattie, J. A. C., Brown, G. T., & Keegan, P. (2005). A national teacher-managed, curriculum-based assessment system: Assessment Tools for Teaching & Learning (asTTle). *International Journal of Learning, 10*, 770–778.

Hattie, J. A., & Fletcher, R. B. (2006). Self-esteem=success/pretensions: assessing pretensions/importance in self-esteem. InMarsh, H. W., Craven, R. G., & McInerney, D. M. (Eds.), (2006). *International advances in self research, Vol. II*. Charlotte, NC: Information Age Publishers.

Hattie, J. A. C., & Gan, M. (2011). Instruction based on feedback. In R. H. Mayer, & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York: Routledge.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill and Melinda Gates Foundation. http://metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf.

Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: a meta-analysis and review of the literature. *Review of Educational Research, 75*(1), 63–82. http://dx.doi.org/10.3102/00346543075001063.

van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merrienboer, J. J. G. (2013). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction, 24*, 15–25.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*, 321–437. http://dx.doi.org/10.1177/001316445601600401.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: a review and meta analysis. *Journal of Applied Psychology, 67*(3), 280–296. http://dx.doi.org/10.1037/0021-9010.67.3.280.

Piaget, J. (1962). *The language and thought of the child*. London: Routledge & Kegan Paul.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144. http://dx.doi.org/10.1007/BF00117714.

Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: calibrating ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57.

Shuford, E. H., & Brown, T. A. (1975). Elicitation of personal probabilities and their assessment. *Instructional Science, 4*, 137–188. http://dx.doi.org/10.1007/BF00051729.

Smith, S. L. (2009). *Academic target setting: Formative use of achievement data*. Unpublished doctoral dissertation, University of Auckland, New Zealand.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22–29. http://dx.doi.org/10.1111/j.1745-3992.1996.tb00803.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: a review of methods and empirical studies. *Review of Educational Research, 40*(5), 663–705. http://dx.doi.org/10.3102/00346543040005663.

Warrens, M. J. (2008). *Similarity coefficients for binary data*. Oisterwijk, Netherlands: Proefschriftmaken, Retrieved from. https://openaccess.leidenuniv.nl/bitstream/handle/1887/12987/Full?sequence=2.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy and Practice, 17*(1), 59–75. http://dx.doi.org/10.1080/09695940903565610.