Calibrating Calibration: Towards Conceptual Clarity and Agreement in Calculation

Meghan M. Parkinson and Daniel L. Dinsmore
University of Maryland

Author Note
Correspondence concerning this article should be addressed to Meghan M. Parkinson,
Department of Human Development, 3304 Benjamin Building, University of Maryland, College
Park, MD 20742; email: mparkin2@umd.edu
Abstract

Calibration is a common term in a number of literatures but is often operationalized and calculated in very different ways. A comprehensive review was undertaken to examine how calibration is defined, used, and calculated within psychological research over the last 10 years. An EBSCO database search was conducted and 40 studies were identified for a comprehensive data table. Analysis of these data revealed inconsistencies in the conceptualization, operationalization, and calculation of calibration. Further, calibration has rarely been investigated in a domain-specific manner. Implications for more thorough and domain-specific investigation of calibration within educational psychology research are discussed.

*Keywords:* calibration, self-regulation, metacognition, academic domains
Calibrating calibration: Towards conceptual clarity and agreement in calculation

Flavell’s (1979) landmark paper on metacognition sparked an area of inquiry within cognitive and educational psychology that remains of utmost interest today. Current investigations into aspects of self-regulation, which typically incorporate metacognition (Dinsmore, Alexander, & Loughlin, 2008), emphasize the importance of *calibration* in the form of metacognitive experiences outlined by Flavell. In the educational psychology literature, calibration is a common term used to generally describe the correlation between the degree of confidence the learner has about their performance and their actual performance (Fischoff, Slavic, & Lichtenstein, 1977; Glenberg, Sanocki, Epstein, & Morris, 1987). This general definition can also be found in other areas of psychological research, such as perception, memory and cognition, and decision-making (De Bruin, Rikers, & Schmidt, 2005; Wagman, Shockley, Riley, & Turvey, 2001; Weber & Brewer, 2003).

Calibration is particularly important in the self-regulatory cycle because confidence in (or judgments of) students’ own learning may prompt students to review or reread in situations where they have not competently completed a task, such as reading for a class assignment (Dunlosky & Lipko, 2007). Further, self-regulated learners are characterized as being aware of what they do and do not know, suggesting that their self-assessments of learning are well-calibrated with their actual performance (Zimmerman, 1990). Alternatively, as learners improve their calibration to a specific task, they may also improve self-regulation for that task (Stone, 2000). Another role for calibration has been conceived by Bromme, Pieschl, & Stahl (2010), who have described how calibration fits within the COPES Model (Winne & Hadwin, 1998) of self-regulation. They contend that epistemic beliefs are the standards against which learners
frame and evaluate what is to be learned. In this way learners must continually calibrate their learning to different tasks within different domains.

In order to conceptualize and measure calibration, it is necessary to distinguish between the various types of judgments learners make before, during, and after a task, as these are the basis of most calibration calculations. Schraw (2009) categorized eight types of judgments used in calibration research based on when the judgment was made in relation to the learning task. Prospective judgments, made before starting the task, include judgments of learning (JOL), ease of learning judgments (EOL), and feelings of knowing (FOK). Students’ JOLs are their decisions about the likelihood of being able to demonstrate knowledge of recently encountered information on an assessment (Dunlosky, Serra, Matvey, & Rawson, 2005). EOLs are predictions about how easy students believe learning will be for a given task (Koriat, 2008). Students’ FOKs indicate familiarity with something that often cannot be fully recalled (Nelson & Narens, 1980).

Concurrent judgments, made while learning, include online confidence judgments, ease of solution judgments, and online performance accuracy judgments. All three of these judgments capture how certain students are about their responses to a particular task. Ease of solution judgments may indicate both an evaluation of the task as well as the quality of response given (Kelemen, Frost, & Weaver, 2000). Ease of solution judgments and judgments of performance accuracy can be retrospective judgments as well if given after a task has been completed.

Calculating calibration using these judgments of performance on a learning task is desirable because it indicates to what extent students successfully monitor their learning and performance (Thiede, Anderson, & Therriault, 2003). For example, undergraduate performance...
on an exam was found to be related to the accuracy of their judgments for the correctness of their responses (Nietfeld, Cao, & Osborne, 2005). Furthermore, being overconfident in the use of a selected strategy creates a misleading sense of the effectiveness of that strategy that tends to negatively impact students’ learning (Hacker, Dunlosky, & Graesser, 1998). Overconfidence has indeed been a robust finding throughout the calibration literature (Glenberg & Epstein, 1985), but there is some evidence that training can improve calibration (Glenberg, Sanocki, Epstein, & Morris, 1987).

Given the importance of investigating calibration for learning, a recent review has brought attention to methodological issues related to widely used measures and calculation of calibration (Hacker, Bol, & Keener, 2008). The first of several critical gaps mentioned was the near exclusivity of undergraduates in calibration research. This poses an obvious problem for generalizability, but also indicates that we know little about calibration in children and adolescents. In addition to their critique of the limited age range of study samples, the authors called for longitudinal studies to uncover the developmental aspects of calibration that are assumed from its grounding in metacognition. Additionally, the lack of evidence for the role of learner beliefs in calibration and the influences of social factors in the classroom was also highlighted.

Domain-specificity was not explicitly suggested by Hacker, et al., however it should be a purposeful consideration in the study of calibration given the nature of metacognition and more specifically self-regulation (Alexander, Dinsmore, Parkinson, & Winters, in press). Although studies of self-regulation in general often overlook domain-specificity, or choose a domain based on the convenience of the sample, self-regulation should significantly differ based on individual
differences within and across domains (Alexander, et al., in press). What is not known is whether or not domain has been under consideration in studies of calibration.

Many empirical studies have examined calibration, factors that influence it, and its impact on self-regulation (Glenberg & Epstein, 1985; Parkinson, 2009; Thiede, Anderson, & Therriault, 2003). However, findings for the calibration of students’ judgments to their performance have been inconsistent (Van Overschelde & Nelson, 2006). In our view there are three possible reasons for these inconsistencies. First, without a clear conceptual definition, operationalization and calculation of calibration are difficult to consider. At a broader level of analysis Dinsmore, Alexander, & Loughlin (2008) found that conceptualizations of the terms metacognition, self-regulation, and self-regulated learning often were not explicitly defined. It is our view that this lack of explicit definitions also occurs with calibration, a component of metacognition, self-regulation, and self-regulated learning.

Second, inconsistent findings for calibration may simply be an artifact of the different types and methods used to collecting confidence judgments (such as judgments of learning; JOLs) as well as the resulting calculation of calibration utilized by researchers. For example, when learners are asked to provide their confidence judgment in regard to a task might make a difference. In fact, some evidence indicates that the time at which participants are asked to provide confidence judgments for word learning affects the accuracy of those judgments (Parkinson, 2009). Additionally, the choices of response formats for both the confidence judgments (e.g., categorical versus continuous scales) and performance (e.g., multiple choice versus open-ended responses) may place limits on the types of calculations (e.g., calibration curves) and statistical methods (e.g., ANOVA) that can and should be utilized.
Third, individual differences of the learner (e.g., knowledge, strategies, interest) in regards to the experimental task may influence their confidence ratings in different ways. Alexander’s (1997) Model of Domain Learning provides a developmental framework for expertise in academic domains. It is likely that one’s profile of knowledge, interest, and strategies differ from domain to domain, thus it is equally likely that their confidence judgments would differ from domain to domain as well. Additionally the task may have an effect on individuals’ confidence ratings. For example, there is evidence that participants’ confidence ratings differ depending on the genre of text (i.e., expository versus persuasive) read (Dinsmore, Loughlin, & Parkinson, 2009).

Given these identified concerns with the construct of calibration as currently used in psychological research, we completed a systematic review of the literature. This review provides a much-needed starting point for calibrating the conceptualization and investigation of calibration in future research. We sought to document accepted conceptual definitions, measurements of confidence judgments, and statistical methods of calculating calibration scores. In the following sections we describe the sources and methodology of the current review, and then discuss our findings and their implications for research and education.

The purpose of this paper is to investigate, through systematic review, the ways in which the term calibration has been conceptualized, operationalized, and calculated and expand upon previous reviews of calibration such as those of Hacker and colleagues. This review began as an endeavor to use calibration in our own research, but an initial examination of the literature yielded confusion about the construct and how it is measured and calculated. Thus we embarked on a systematic review of calibration in order to determine what is agreed upon, what needs clarifying, and future directions for including calibration in educational research.
Method

Data sources

In order to consider the conceptualization, use, and calculation of calibration in the contemporary psychology literature, we created a pool of studies focusing on calibration. This pool was developed by searching the ERIC and PsycINFO databases through EBSCO and using the keywords *calibrat* or *accuracy* or *monitoring*. Due to the large number of studies initially identified by these search parameters, we limited the search to peer-reviewed empirical works published in the past ten years (2000-2009). We then narrowed the search by subject *metacognition* and then *judgment*. Additionally, we conducted a hand search of certain key journals not in the PsychINFO database. The resulting pool comprised 40 studies. A complete table of the studies can be found at: http://education.umd.edu/EDHD/faculty2/alexander/arl/publications/calibration.html. It is important to note that we did not narrow our search exclusively to educational psychology. We considered all sources meeting the above criteria. This was done to ensure careful consideration of any alternative explorations of calibration outside the purview of educational research.

Coding

We analyzed each article according to the following categories: definitions of the calibration; the participants; when calibration was measured; any time delay between the experimental task and the calibration measure; scaling of the calibration measure; the nature of the experimental task; and the calculation(s) of calibration.

We felt that an explicit coding scheme would make our analysis more useful and ensure that any resulting codes were reliable and valid. First, definitions of calibration were coded using an existing coding scheme which coded the definitions as explicit, conceptual, referential,
or by measure (Dinsmore, Alexander, & Loughlin, 2008; Murphy & Alexander, 2000). Second, we coded the developmental level of the participants as elementary students, middle school students, high school students, college undergraduates, or adults. Third we coded when calibration was measured in regards to the experimental task as occurring before or after the task as well as any time delay (coded yes or no). Lastly, we coded if the experimental task was situated in an academic domain (e.g., mathematics). We denoted the calculation as either a named calculation (e.g., Hamman coefficient) or a simple description of its calculation. For the tabling and the coding of all the columns, we engaged in extensive discussion to ensure that the codes were consistent. To further ensure reliability of these codes over 50% of the studies were coded independently by the authors. The initial level of interrater agreement for all these codes was 92.5%, with the remaining codes resolved in conference.

**Results and Discussion**

The findings of this study indicate specific issues which may underlie the inconsistencies in calibration findings discussed by Van Overschelde & Nelson (2006). These issues relate to three main areas where differences emerged in the data table. As addressed in the theoretical framing, two areas where methodologies differ pertain to: a) methods of collecting confidence ratings and calculating the relationship between confidence and performance; and b) the relationship between the learner and the task.

First and foremost, the method of collecting confidence judgments (which are used to calculate calibration) is dependent on the researcher’s conceptualization of what calibration is. Reviewing the coded definitions of calibration revealed that only 52.5% of the studies reviewed explicitly defined calibration. Of the 52.5% of the studies that explicitly defined calibration, there was general agreement that calibration is the relationship between confidence and
performance. However, in the other 47.5% of studies calibration was not explicitly conceptualized, thus, it was difficult to determine whence the measures and calculation of calibration in these studies were derived.

Using the conceptual definition of calibration as the relationship between confidence and performance, we further explored the operationalization of calibration to see how it, too, varied. We again found differences which may explain why studies dealing with calibration may report inconsistent results. One major difference in these studies was how calibration was measured and calculated. First, we turn to the measurement issue of how confidence ratings were collected (Table 1). Participants’ responses were collected in a variety of ways, including continuous scales (e.g., a 100-millimeter line from not confident to very confident; e.g., Nietfeld & Schraw, 2002), Likert-type scales (e.g., a frowny face for not confident, a neutral face for somewhat confident, and a smiley face for confident; e.g., Lockl & Schneider, 2002), categories representing the percent level of confidence (e.g., 50% for guessing and 100% for absolutely certain; e.g., Allwood, Granhag, & Johansson, 2000), and finally dichotomous or count data (e.g., certain or uncertain; e.g., Nietfeld, Enders, & Schraw, 2006). A majority of the studies (52.5%) used some type of categorical scale (i.e., Likert-type or percentile categories) to collect confidence ratings. These different response scales elicit different patterns of confidence ratings from participants. For example, a continuous scale may yield a much different distribution of scores than a Likert-type scale, particularly one in which there are few categories. We know of no study that has systematically examined these differences.

Further, the types of scales in which confidence ratings are collected (e.g., categorical or Likert-type) can further limit the analyses or calculation of calibration that can be performed. Table 2 lists the different types of analyses run or calculations used for calibration. In many
cases, we found that the scale used was often mismatched to the analyses or calculation performed. For example, categorical data was used for resulting analyses or calculations that require continuous data or at least the assumption of normality (e.g., ANOVA). Additionally the wide range of calculations themselves may be responsible for the reported inconsistent patterns of calibrations between individuals. The most common way to calculate calibration was to use correlation coefficients (such as the Hamman or Gamma coefficients), but these studies only accounted for 35% of the overall studies. In fact, 20% of the studies used the confidence judgments as independent variables in an ANOVA or regression model. It would be interesting to systematically study if the calculations (or analyses) themselves would lead to different interpretations of the results in these calibration studies.

Confounding these operational issues, the type of task and choice of participants may also be causing the inconsistent findings in the calibration research. First, 80% of the studies were not situated in an academic domain such as mathematics, science, reading, or history. Most of the articles we tabled were examining calibration in basic cognitive tasks (generally memory tasks). This is problematic, because as the Model of Domain Learning (MDL) would suggest, a participant’s position along the continuum of expertise (i.e., acclimation, competence, and expertise) as measured by their knowledge, interest, and strategic processing, may influence their confidence. We have seen this influence in a number of our own studies (e.g., Dinsmore, Loughlin, & Parkinson, 2009).

Further, even though only general cognitive tasks have typically been examined, it would be helpful to understand how calibration differs developmentally. Here, not surprisingly, we see a large percentage of the studies (67.5%) using convenience samples of undergraduates. Although undergraduates are an important group to study, we did not find convincing rationales
for why this group was the focus in each of these studies. Without systematic investigation as to the role of individual difference (such as knowledge, interest, and strategic processing) at all developmental levels we are left to wonder what role they have in the often inconsistent findings in the calibration literature.

**Conclusion**

Although the definition of calibration is relatively agreed upon in the literature, its conceptualization remains less clear. Because calibration is typically defined as a simple calculation between confidence ratings and performance, and is often not grounded in any particular framework that would explain or predict an individual’s tendency to be more or less calibrated to certain tasks. Thus we would recommend grounding calibration within a framework (such as metacognition) in order to add some explanatory or predictive power beyond the typical finding that better-calibrated individuals perform more competently.

Second, the way in which calibration is utilized in analysis, and the judgment measures that the calibration calculations are based upon require greater attention. From this systematic review, it was clear there was often a mismatch between the type of measure used to collect judgments and the type of analysis chosen to analyze calibration. It is difficult to draw any conclusions about the nature of calibration, or those factors that influence it, without both valid measurement and analysis. Although several papers and reviews have brought attention to the fact that absolute and relative calculations of accuracy are not correlated (Maki, Shields, Wheeler, & Zacchilli, 2005; Schraw, 2009), there are no studies that we are aware of that examine differences in calibration between continuous and categorical measures.

Further as Hacker et al. (2008) previously noted, participants typically recruited for studies regarding calibration are undergraduates. This type of sample cannot be said to be
typical of all adults, nor does it capture developmental differences that may be present in younger children and adolescents. However, an admitted strength to using undergraduate participants is the opportunity to study self-regulation within a class setting. Despite this advantage, few studies from our reviewed sample studied calibration within a particular domain. This is especially problematic as self-regulatory processes are theoretically framed as domain-specific in nature (Alexander, Dinsmore, Parkinson, & Winters, in press).

Taken together, the vast differences in the conceptualization, operationalization, and task/participant match may all be contributing factors to the inconsistent findings echoed by Van Overschelde & Nelson (2006). Without a systematic examination of their effects (or, more likely, the interactions of these effects), the applicability of the calibration literature for future studies is questionable. One promising area of systematic investigation is that by Thiede, Dunlosky, Griffin, and Wiley (2005) in which they studied the effects of delay between the performance and the confidence ratings. Further targeted efforts towards clarifying understanding and usage of calibration are of utmost importance for both researchers and educators. Well-calibrated individuals are presumably aware of what they do and do not know, thus they are able to identify areas of learning or study in need of improvement. This kind of awareness is critical for effective study strategies and self-guided learning from texts.
References


Table 1

*Raw and Percentage of Calibration Studies by Confidence Scale Type*

<table>
<thead>
<tr>
<th>Confidence scale type</th>
<th>Raw</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert-type</td>
<td>10</td>
<td>25.0</td>
</tr>
<tr>
<td>Continuous</td>
<td>11</td>
<td>27.5</td>
</tr>
<tr>
<td>Categorical (% interval)</td>
<td>11</td>
<td>27.5</td>
</tr>
<tr>
<td>Di- or trichotomous</td>
<td>2</td>
<td>5.0</td>
</tr>
<tr>
<td>Count data</td>
<td>1</td>
<td>2.5</td>
</tr>
</tbody>
</table>
Table 2

*Raw and Percentage of Calibration Studies by Calculation Type*

<table>
<thead>
<tr>
<th>Calculation type</th>
<th>Raw</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>7</td>
<td>17.5</td>
</tr>
<tr>
<td>Calibration curves</td>
<td>3</td>
<td>7.5</td>
</tr>
<tr>
<td>Difference scores</td>
<td>7</td>
<td>17.5</td>
</tr>
<tr>
<td>Correlations</td>
<td>14</td>
<td>35.0</td>
</tr>
<tr>
<td>ANOVA/Regression</td>
<td>8</td>
<td>20.0</td>
</tr>
</tbody>
</table>