Research on calibration remains a popular line of inquiry. Calibration is the degree of fit between a person’s judgment of performance and his or her actual performance. Given the continued interest in this topic, the questions posed in this article are fruitful directions to pursue to help address gaps in calibration research. In this article, we have identified six research directions that if productively pursued, could greatly expand our knowledge of calibration. The six research directions are: (a) what are the effects of varying the anchoring mechanisms from which calibration judgments are made, (b) how does calibration accuracy differ as a function of incentives and task authenticity, (c) how do students self-report the basis of their calibration judgments, (d) how do group interactions and social comparisons affect calibration accuracy, (e) what is the relation between absolute and relative accuracy, and (f) to what extent does calibration accuracy predict achievement? To help point the way to where we go from here in calibration research, we provide these research questions, propose research methods designed to address them, and identify prior, related studies that have shown promise in leading the way to fill these gaps in the literature.

**Keywords:** calibration, metacognition, self-regulated learning, social cognition, research methods
areas of forethought for several reasons, two of which we address here. Estimates of performance have been shown to be biased toward some initial anchor, and learners often do not adequately adjust from these anchors. Therefore, knowing the psychological bases for these anchors and how learners can debias their judgments is not a trivial matter (research question a). In addition, because studies on the effects of incentives on motivation to achieve greater accuracy have been mixed, greater attention needs to be focused on how motivation can be manipulated (research question b).

During the performance phase, the learners gain feedback concerning self-control and self-observation, processes that are essential for continued attention to and action on a task. Learners self-explanations about how and why they use their self-observations to self-control and whether those self-explanations are mediated or moderated by other factors such as attributional style (research question c) or social influences (research question d) are critical areas of investigation. Maintaining attention and action on a task also requires that ongoing performance is being judged accurately. Performance can be judged at a global level (e.g., How prepared is a learner for an upcoming test?) and at local levels (e.g., Is the answer to this question correct?). Knowing whether there is a relation between these global and local levels of judgment can provide insights into the psychological mechanisms upon which they are made (research question e).

Finally, during the self-reflection phase, the learner makes self-judgments and self-reactions on their performance. This feedback on whether actual achievement, the end product of self-regulation, is high or low, satisfactory or unsatisfactory, or simply just good enough, then exerts an influence on whether further action will be taken in the forethought phase. Therefore, knowing whether there is a payoff to self-regulation is instrumental to continued self-regulation (research question f).

PROPOSED DIRECTIONS
Our attention now turns more specifically to our proposed research agenda on how knowledge of calibration can be promoted in further lines of inquiry. Table 1 presents the six research areas and questions already addressed as well as designs and variables aligned with these questions. In addition, we identify prior, related studies that have shown promise in filling these gaps in the literature.

WHAT ARE THE EFFECTS OF VARYING THE ANCHORING MECHANISMS FROM WHICH CALIBRATION JUDGMENTS ARE MADE?
In their seminal article, Judgment under Uncertainty: Heuristics and Biases, Tversky and Kahneman (1974) proposed that people make estimates of their performance by starting from some initial value and make adjustments that are biased toward that initial value. Their claim was that these adjustments are often insufficient so that the estimates continue to appear biased. The anchoring-and-adjustment effect has been widely recognized in the decision-making literature (e.g., Mussweiler et al., 2000; Epley and Gilovich, 2004, 2005), and it has been used to explain underconfidence in calibration research (e.g., Scheck et al., 2004).

Researchers of anchoring-and-adjustment effects make a distinction about who sets the initial value from which subsequent adjustments are made. In some cases, the initial value is set by another person (e.g., a salesperson setting the price for a new car) or by oneself (e.g., when guessing how long it takes Mars to orbit the sun, people often select an anchor on Earth’s orbit and then adjust from that value; Epley and Gilovich, 2004). These self-generated anchors are the ones that we believe could potentially influence people’s calibration judgments.

Scheck and Nelson (2005) used anchoring-and-adjustment as an explanation for the underconfidence-with-practice (UWP) effect proposed by Koriat et al. (2002). The UWP effect is a robust finding in which people initially show overconfident calibration when making judgments of learning (JOLs) but subsequently become underconfident after their second study trial. Scheck and Nelson (2005) hypothesized that people form a psychological anchor for their JOLs somewhere between 30 and 50% of correct recall and adjust their JOLs either upwards or downwards based on whether performance is above or below this band. They found that when performance was above 50% after the second study trial, participants adjusted both immediate and delayed JOLs downward in relation to recall, thereby appearing to be underconfident. When performance was below 30% after the second study trial, participants adjusted their delayed JOLs upwards in relation to recall, thereby appearing to be overconfident; and when performance was at 30%, their immediate JOLs were near perfect. If accuracy of calibration is a necessary condition for self-regulated learning, then a potent research question is how can higher achievers overcome their underconfident JOLs and lower achievers overcome their overconfident JOLs without adversely impacting students who are between the two groups?

Research into this question can be informed by research conducted by Epley and Gilovich (2005) in which insufficient...
Table 1 | Proposed questions, design, and exemplar studies in calibration research.

<table>
<thead>
<tr>
<th>Research questions</th>
<th>Design</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the effects of varying the anchoring mechanisms from which calibration judgments are made?</td>
<td>True experimental</td>
<td>Treatment: manipulating initial judgments</td>
</tr>
<tr>
<td>How does calibration accuracy differ as a function of incentives and task authenticity?</td>
<td>Correlational</td>
<td>Measures: extent and type of adjustments for judgments</td>
</tr>
<tr>
<td>How do students self-report the basis for their calibration judgments?</td>
<td>Qualitative</td>
<td>Measures: interviews and think-aloud protocols</td>
</tr>
<tr>
<td>How do group interactions and social comparisons affect calibration accuracy?</td>
<td>Experimental, factorial</td>
<td>Measures: calibration accuracy, group interactions</td>
</tr>
<tr>
<td>What is the relationship between absolute and relative accuracy?</td>
<td>Correlational</td>
<td>Measures: absolute and relative on items, topics/concepts, and overall performance</td>
</tr>
<tr>
<td>To what extent does calibration accuracy predict achievement?</td>
<td>Correlational</td>
<td>Measures: calibration accuracy and achievement</td>
</tr>
</tbody>
</table>

adjustments were compensated by providing financial incentives that motivated participants to make additional adjustments. Also, forewarning participants that initial adjustments are inadequate helped them to engage in additional adjustments, resulting in greater accuracy. Finally, Zhao and Linderholm (2011) found that providing information about peer performance on a task can serve as an anchor for metacomprehension judgments and can be used to debias judgments. We propose initiating research on this topic by posing the following question, “What are the effects of varying the anchoring mechanism from which calibration judgments are made?” Experimental work following any one of these lines of research could be pursued to demonstrate viable ways to encourage adjustment away from initial judgments and toward greater calibration accuracy.

HOW DOES CALIBRATION ACCURACY DIFFER AS A FUNCTION OF INCENTIVES AND TASK AUTHENTICITY?

Granted, motivation is a broad construct, but as in other areas of education research, calibration cannot be thoroughly understood without reference to motivational variables. To narrow the focus of this line of inquiry, we have selected incentives and task authenticity as starting points. These represent reasonable starting points because they represent salient constructs in the forethought phase of self-regulated learning and can be readily generalized to classroom contexts. In our conceptualization, incentives would be some type of course credit, and task authenticity would be learning course content. Focusing on learning course content stands in stark contrast to typically used tasks in calibration research, such as learning paired associates in a remote tribal language to control for prior learning.

Studies examining the impact of incentives on calibration are rare. In the preceding section, we referenced Epley and Gilovich’s (2005) study in which financial incentives were employed to motivate participants to adjust their metacognitive judgments. In another more ecologically valid study, we manipulated incentives and reflection in a fully crossed quasi-experiment conducted in college course (Hacker et al., 2008a). Students in an extrinsic reward condition were told they would receive one to four additional points on each of three exams, depending on their calibration accuracy, with more points given for greater accuracy. We found that incentives significantly improved calibration accuracy but only among lower-achieving students, which may have been the result of greater motivation on their part to be accurate to earn the additional points. Schraw et al. (1993) also provided evidence for the effectiveness of incentives for promoting calibration accuracy. In their procedure, students also received extra credit either for improving performance on a test or increasing their calibration accuracy. Though performance improved in both incentive conditions, the findings revealed that incentivizing accurate calibration was more effective than incentivizing improved performance.

Thus, the research question we propose is how does calibration accuracy differ as a function of incentive or task authenticity? Because it would be difficult to isolate these variables while controlling for all other influences, comparative studies might be more appropriate initially. In future research, the influence of both incentives and task authenticity could be investigated in the same study. For example, in a within-subjects design, calibration accuracy might be compared for students completing more authentic versus contrived tasks in conditions where incentives are or are not present.

HOW DO STUDENTS SELF-REPORT THE BASIS FOR THEIR CALIBRATION JUDGMENTS?

Earlier we posed a research question related to how individuals might anchor and then adjust their calibration. What we have not adequately addressed is how students self-report the basis for calibration judgments. Gathering self-report data from students regarding their calibration judgments is rare but not unprecedented. In a mixed methods study, Hacker et al. (2008a) asked college students to identify factors that influenced the accuracy of their predictions and postdictions. Attributional style was used as the theoretical framework to organize the data. The most frequent explanations focused on internal, student-centered constructs. Students were most likely to attribute discrepancies between their scores and calibration judgments to how much or how well they studied or how well they felt they knew the material. Another frequently reported factor was test-taking ability and prior performance on tests as well as expectations for test content and difficulty. Similar results were reported in a study.
with middle school students (Bol et al., 2010). In this study, immediately after making their predictions, students were asked why they predicted that score. The most frequent categories of responses centered on time and effort spent studying, global perceptions of their own abilities, and past performance. After making their postdictions, students again were asked to explain why these were accurate or inaccurate. Explanations focused on knowing the number answered correctly, their expectations of test difficulty, the effort exerted in studying, and their global sense of self-confidence.

Bandura’s model of reciprocal determinism has also been used to categorize responses (Dinsmore and Parkinson, 2012). Participants were asked to explain how they arrived at or what was considered when making confidence judgments. Instances of the a priori categories observed in student responses included prior knowledge, characteristics of the text and the items, and guessing. Students often cited a combination of personal and task characteristics. The combination of how students explain their calibration judgments and how judgments are measured may be important for understanding how judgments contribute to performance.

Although the results from a few studies have helped us understand how students describe the basis for their calibration judgments, more research is warranted. As employed in the studies just described, a qualitative approach grounded in a theoretical framework would be most appropriate and revealing. Qualitative data could be collected via open-ended responses to surveys and think-aloud protocols in which real-time data would be collected as students are considering, making, and explaining their calibration judgments.

Threats to validity in self-reported data cannot be avoided. Consequently, in the qualitative tradition, we might rely on triangulation strategies to support the credibility or transferability of findings. Researchers (Azevedo et al., 2010; Winne, 2010) recommend combining self-report data about self-regulated learning with trace evidence that reflects students’ cognitive operations (e.g., highlighting text). Data collected in computer-based learning environments should facilitate these types of studies. For example, if students attribute their lack of understanding of a topic to lack of study time, actual time spent studying the topic could be calculated. We might also follow their navigation patterns to determine whether they return to content judged to be in need of further study.

**HOW DO GROUP INTERACTIONS AND SOCIAL COMPARISONS AFFECT CALIBRATION ACCURACY?**

Experimental manipulations centered on group interactions and social comparisons have already shown promise in improving calibration accuracy. In our recent factorial experiment with high school biology students (Bol et al., 2012), half of the students practiced calibration in groups while the other half practiced calibration individually. The second treatment variable was whether students used guidelines to gage their judgments of how well they mastered the content. We found both group settings and guidelines to be effective in promoting calibration accuracy and achievement. Other studies have demonstrated that the combination of group learning contexts and guiding questions promoted metacognitive skills and achievement (Kramarski and Mevarech, 2003; Kramarski and Dudai, 2009).

Group work logically elicits implicit or explicit social comparisons. Carvalho and Yuzawa (2001) manipulated social comparisons by presenting some participants with information concerning the mean percentage of correctly answered questions that a fictitious group of fellow students presumably scored. This information was presented prior to participants making their own metacognitive judgments on their performance. The results indicated that social comparisons did impact the magnitude of metacognitive judgments with greater magnitude in judgments associated with higher performance. Other results suggested that participants with little confidence in their judgments may be particularly susceptible to social influences.

Given that group settings and social comparisons can influence metacognitive judgments, the next logical step would be to manipulate both of these variables in a factorial experiment. The question posed is how do group interactions and social comparisons affect calibration accuracy? Students would be asked to calibrate in group or individual settings and would do so with or without social comparisons. Because earlier studies suggest that guidelines promote accuracy in metacognitive judgments (Kramarski and Dudai, 2009; Bol et al., 2012), all four groups would receive guidelines. The social comparisons could be presented to half of the students as part of the guidelines employed by students, and could take the form of mean accuracy scores achieved by low, middle, and high achievers. However, rather than using fictitious scores as in the Carvalho and Yuzawa (2001) research, actual scores could be presented, which may help students differentiate between more and less reasonable calibration judgments and how they may be tied to achievement levels. This may be especially beneficial for lower-achieving students who tend to overestimate their performance (e.g., Bol et al., 2010). Noting the group interactions among students assigned to this condition may further illuminate how social interactions or comparisons may affect calibration judgments.

**WHAT IS THE RELATIONSHIP BETWEEN ABSOLUTE AND RELATIVE ACCURACY?**

Calibration (aka absolute accuracy) provides estimates of overall memory retrieval (Lichtenstein et al., 1982; Keren, 1991; Nietsfeld et al., 2006), and relative accuracy (aka discrimination) provides estimates of whether a person’s judgments can predict the likelihood of correct performance of one item relative to another (Nelson, 1996). For instance, a student can judge that overall he or she will get 85% of the items on a test correct and in fact get 85% correct (i.e., perfect calibration accuracy), but upon closer examination, the student may have given high confidence judgments to items answered incorrectly and low confidence to items answered correctly, in which case, relative accuracy may be close to chance. Both types of accuracy are important for students to self-regulate their learning; however, whether there are shared psychological processes contributing to both is an important area for investigation (Maki et al., 2005).

Accuracy of both global and item-level calibration judgments plays an important role in current and future study efforts because poor judgments can lead to either premature or protracted termination of study of general and specific content. Current research of metacomprehension judgments has shown that there is little or no relation between absolute and relative accuracy (Maki et al., 2005).
Although widespread acceptance has been given to the theoretical underpinnings of calibration research, there has been much less attention to the actual retention of knowledge and the empirical evidence of the importance of accurate monitoring and its relative accuracy. Correlational and experimental studies have shown that monitoring can positively impact decisions about what to study (e.g., Metcalfe and Finn, 2008; Hines et al., 2009), but whether that studying leads to gains in achievement is a question in need of further support (Dunlosky and Rawson, 2011). Some calibration studies have shown a positive relation between calibration accuracy and achievement level (e.g., Hacker et al., 2000b; Bol and Hacker, 2001). In addition, Nietfeld et al. (2006) and Bol et al. (2012) demonstrated that students who participated in interventions to increase calibration accuracy realized higher gains in achievement than students who did not participate in them. However, in both these studies, treatment and classroom assignment were conflated, leaving open the question of whether internal validity was potentially compromised.

Dunlosky and Rawson (2011) experimentally manipulated judgment accuracy by asking participants to study key-term definitions, one group used an idea-unit standard in which the participants were shown their responses with the idea units contained in the correct answer, and another group used their responses but without access to the correct answer. After being shown their responses, participants made a self-score judgment about whether their answer was correct. The test-judge-study trials continued until a definition was judged as correct three times. Two days later, all participants were administered a retention test. Findings indicated that greater accuracy was related to greater retention. Moreover, they found that participants who were overconfident in their judgments prematurely terminated study, and as a consequence their retention suffered.

Similar experimental manipulations of accuracy need to be conducted to more firmly establish the link between calibration accuracy and achievement. Empirical findings provide multiple ways to manipulate accuracy: study guidelines used in group settings (Bol et al., 2012), self-assessment with feedback (Nietfeld et al., 2006), and feedback with idea-unit standards (Dunlosky and Rawson, 2011). Employing these accuracy manipulations and measuring subsequent retention could provide valuable support for the importance of accurate monitoring and provide information about whether specific manipulations of accuracy lead to greater retention. In addition, the kinds of learning could be manipulated. The tasks used in an experiment could be varied from simple paired associates to multiple-choice tests to text comprehension. Firmly establishing the link between monitoring accuracy and achievement is a critical goal for calibration research.

SUMMARY AND CONCLUSION

We do not assume that these are the only fruitful research directions to pursue in order to more thoroughly understand calibration and its role in promoting self-regulated learning. However, they represent a good start. The research agenda outlined is based on the social cognitive model of self-regulation developed by Zimmerman and his colleagues (Schunk and Zimmerman, 1997; Zimmerman, 2008; Zimmerman and Moylan, 2009). In alignment with the forethought phase, we propose investigating the effects of anchoring, incentives, and task authenticity on calibration judgments to reflect psychological processes linked to self-efficacy and motivation. Research questions focused on social influences, self-explanations, and the basis for metacognitive judgments are represented in the performance phase of the model. More specifically, we propose addressing student explanations for calibration judgments, the impact of group interactions and social comparison on calibration accuracy, and the relationship between absolute and relative accuracy. In the self-reflection phase, learners judge and react to their performance or achievement. The final question reflects the extent to which calibration accuracy predicts achievement. Feedback on performance and self-reflection influences subsequent, cyclical phases of self-regulated learning.

Similarly, the methods we propose are not exhaustive and reflect examples of how these questions may be pursued. The methods we propose range from qualitative approaches investigating how students explain their calibration judgments, comparing motivational factors linked to task characteristics, correlating absolute and relative accuracy, predicting achievement based on calibration accuracy, and manipulating anchoring and adjustment effects as well as social interactions in controlled experiments.

Calibration research will be further advanced when we identify patterns of findings guided by sound theoretical models and based on precise descriptions of terms, measures, contexts, tasks, and populations. As we have argued previously (Hacker et al., 2008b), calibration has been measured in different ways but largely studied in more controlled contexts using college students. Granted, we must consider the trade-offs between internal and external validity as we move into more naturalistic settings, such as classrooms and employ more authentic tasks. Various research methods with varying levels of control will better inform our questions overall.
REFERENCES


Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 May 2012; accepted: 19 June 2012; published online: 09 July 2012.

This article was submitted to Frontiers in Educational Psychology, a specialty of Frontiers in Psychology.

Copyright © 2012 Bol and Hacker. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.