ISSN 1303-0485 • **eISSN** 2148-7561
DOI 10.12738/estp.2016.1.0033

Educational Sciences: Theory & Practice • 2015 December • 15(6) • 1575-1584

Received | April 9, 2015
Accepted | August 5, 2015
OnlineFirst | October 2, 2015

The Confidence–Accuracy Relationship in Diagnostic Assessment: The Case of the Potential Difference in Parallel Electric Circuits

Murat Saglam[a]
Ege University

Abstract
This study explored the relationship between accuracy of and confidence in performance of 114 prospective primary school teachers in answering diagnostic questions on potential difference in parallel electric circuits. The participants were required to indicate their confidence in their answers for each question. Bias and calibration indices were calculated for each prospective teacher. A resolution index was calculated for each of the 43 participants who had some variance in both their performance scores and confidence ratings. In this study, the participants were more likely to know when they were giving the correct answer than when they were giving an incorrect answer. The resolution of confidence was positively related to the calibration of confidence. The findings indicate that having good resolution of confidence may be a prerequisite to, but not sufficient for, being well calibrated. In general, the prospective teachers in this study were poorly calibrated in their understanding of the potential difference in parallel circuits, the main reason for this being overconfidence. We also compared the participants for gender differences and noted that female students were more under-confident and less overconfident than their male counterparts in their understanding of the potential difference in parallel circuits. Implications of the findings for teaching and future research are discussed.

Keywords: Accuracy • Calibration of confidence • Resolution of confidence • Diagnostic assessment • Electric circuits

a **Correspondence**
Assoc. Prof. Murat Saglam (PhD), Department of Primary Science Education, Faculty of Education, Ege University, Izmir 35100 Turkey
Research areas: Children's understanding of science; Metacognition; Problem-based learning; Teacher education
Email: murat.saglam@hotmail.com

Research in science education has revealed that students have conceptions that differ from the accepted scientific ones in many science subjects (Driver, Guesne, & Tiberghien, 1985; Driver, Squires, Rushworth, & Wood-Robinson, 1994; Duit, 2009). Although several research methods have been employed to elicit these alternative ways of reasoning (White & Gunstone, 1992), written diagnostic questions have been the primary method for data collection from a large sample in a relatively short time. A diagnostic question analyzes an individual's performance to uncover his or her strengths or weaknesses in the subject tested. In addition to the information provided by students' answers to such questions, the knowledge of their confidence in their answers may prove to be useful. Confidence judgements provide insights into "how well a person evaluates and monitors his or her performance" (Stankov & Crawford, 1996, p. 971). Students with unwarranted high confidence in their knowledge in a given domain are considered "overconfident," whereas those with unwarranted low confidence are "under-confident." Overconfidence in a given area of knowledge indicates that (*a*) "such knowledge is frequently invoked," (*b*) "such knowledge is rarely questioned or checked against other criteria, so that wrong conclusions are not detected," and (*c*) "such knowledge may be maintained, even in the face of explicit counterevidence, so that wrong conclusions persist" (Reif & Allen, 1992, p. 28). Conversely, under-confidence in a given area indicates that "such knowledge may not be believed, even if correct, and may be rejected in favor of incorrect conclusions" (Reif & Allen, 1992). Research on over/under-confidence has revealed that many people are overconfident in their knowledge (Lichtenstein, Fischhoff, & Phillips, 1982).

In the research literature, overconfidence and under-confidence are related to the concept of calibration investigating the relationship between confidence in and accuracy of performance. Alexander (2013) defined calibration as "the distance between perceived and demonstrated levels of understanding, capability, competence, or preparedness" (p. 1). She indicated that there is an increasing interest in the topic of calibration among educational researchers and argued that one reason for this is that constructs and processes, such as conceptual change, metacognition, and self-regulation, are closely linked to learners' ability to accurately assess their own capacities. Furthermore, Alexander argued that learners' subsequent efforts and strategic behaviors will be affected by their calibration.

Another important concept in the research literature on the confidence-accuracy relationship is resolution (Mengelkamp & Bannert, 2012). Sharp, Cutler, and Penrod (1988) defined resolution as students' ability "to discriminate correct from incorrect judgments by differentially assigning confidence judgments to accurate and inaccurate judgments" (p. 272). Students who can appropriately discriminate between what they know and what they do not (i.e., the students with good resolution) may use their study time more effectively than the students with poor resolution of confidence.

Data on calibration and resolution are considered valuable because the findings from the research on the confidence-accuracy relationship imply that students' confidence in their science knowledge is a part of their understanding of natural phenomena.

Confidence Judgements in Science Education Research

Several studies in science education research have employed the Certainty of Responses Index (CRI) proposed by Hasan, Bagayoko, and Kelley (1999) to distinguish a lack of knowledge from alternative conceptions. In their study, the confidence levels of the CRI were "totally guessed answer," "almost a guess," "not sure," "sure," "almost certain," and "certain." According to Hasan et al., for a student (or for a group of students) and a given question, a correct or incorrect answer with a low CRI suggests a lack of knowledge; a correct answer with a high CRI indicates knowledge of correct concepts; and an incorrect answer with a high CRI indicates the presence of an alternative conception. They used a mechanics diagnostic test with 36 five-option multiple-choice questions and required 106 university students to provide a CRI along with each answer. The authors concluded that the method could be easily employed to differentiate alternative conceptions from a lack of knowledge. Then, the method was used to distinguish a lack of knowledge from alternative conceptions in mechanics (Planinic, Boone, Krsnik, & Beilfuss, 2006; Potgieter, Malatje, Gaigher, & Venter, 2010); simple direct-current circuits (Peşman & Eryılmaz, 2010; Planinic et al., 2006); radioactivity (Colclough, Lock, & Soares, 2011); diffusion and osmosis (Odom & Barrow, 2007); and environmental education (Arslan, Çiğdemoğlu, & Moseley, 2012). The review of these studies suggests that the Hasan et al.'s method can be employed to distinguish a lack of knowledge from alternative conceptions. However, these studies did not provide data on calibration and resolution.

In psychology, several indices have been developed to study the confidence-accuracy relationship. Recently, science education researchers have begun to employ these indices to generate data on calibration and resolution of confidence. For example, Caleon and Subramaniam (2010a) designed a 14-item three-tier diagnostic test to explore secondary school students' understanding of the nature and propagation of mechanical waves. They asked students how confident they were in their answers to two-tier questions. The confidence ratings were 1 for "just guessing," 2 for "very unconfident," 3 for "unconfident," 4 for "confident," 5 for "very confident," and 6 for "absolutely confident." For each question, they calculated the mean confidence accuracy quotient (CAQ) to examine students' discrimination between what they knew and what they did not (i.e., resolution of confidence). The following CAQ formula was used in the study where CFC stands for the mean confidence of students when they gave a correct answer, CFW for the mean confidence of students when they gave a wrong answer, and SD for standard deviation:

$$CAQ = \frac{CFC - CFW}{\text{SD of Confidence of the students for the question}}$$

Caleon and Subramaniam (2010a) found that five out of the fourteen questions in the test had low CAQs (less than .3), and the CAQ for all items across the entire sample was .08. They reported that, predominantly, the students poorly discriminated between what they knew and what they did not. Other studies employing the CAQ reported similar results (Caleon & Subramaniam, 2010b; McClary & Bretz, 2012; Sreenivasulu & Subramaniam, 2013). In their study, Caleon and Subramaniam (2010a) also calculated the confidence bias (CB) for each question (CB = confidence rating recorded in a scale of 0 to 1—proportion of students who gave correct responses). They stated that the CB suggests overconfidence when positive, under-confidence when negative, and perfect calibration when equal to zero. The authors found that all questions had positive confidence bias values. They concluded that, in general, the students were overconfident. In another study employing the CB, Sreenivasulu and Subramaniam found that, overall, the students' confidence matched the accuracy of their responses. The mean CB value, when averaged across the test for all items, was .05. In the studies reviewed, calibration and resolution of confidence were examined for each question. Alternatively, the confidence-accuracy relationship can be investigated by examining calibration and resolution for each student (Schraw, 2009). Such an approach allows a researcher to probe individual differences, such as gender, in data on calibration and resolution.

This study attempts to explore how prospective primary school teachers' accuracy of performance on some diagnostic questions relates to their confidence in performance. In particular, it examines over/under-confidence, the calibration of confidence, and the resolution of confidence. The research questions that guided the study are as follows:

(1) How does the accuracy of prospective primary school teachers' answers to some diagnostic questions on the potential difference in direct-current (DC) parallel circuits relate to their declared confidence in their answers?

(2) Is there any gender difference in the confidence-accuracy relationship in prospective primary school teachers' understanding of the potential difference in direct-current (DC) parallel circuits?

There are several reasons for choosing the potential difference (voltage) in direct-current (DC) parallel circuits in physics as the topic of the study. First, research on students' understanding of DC circuits has found that, when faced with potential difference questions in parallel circuits, many students fail to use the scientific idea that if the components of electric circuits, such as batteries, resistors, and lamps, are connected in parallel, they have the same potential difference across their ends (Cohen, Eylon, & Ganiel, 1983; McDermott & Shaffer, 1992; Shipstone, 1984; Shipstone et al., 1988). This makes the topic suitable for the investigation of possible interactions between students' correct or incorrect answers and their confidence in them. Second reason for choosing the topic was the availability of related diagnostic questions. The questions were developed by science education researchers in the UK (Millar, 2003), and each included a four-point confidence rating scale. Finally, the topic was a part of a general physics course the researcher has taught in a state university in Turkey.

Method

Participants

This preliminary study explores how prospective primary school teachers' accuracy of performance on

some diagnostic questions in science is related to their confidence in performance. As the researcher does not intend to generalize to a wider population on the basis of this small sample, a convenience sampling method was used. The sample comprised 114 second-year prospective primary school teachers (43 male and 71 female) enrolled in a one-term general physics course in a state university in Turkey. During the course, the author instructed the students in the basic ideas of general physics and solved conceptual/quantitative questions related to these ideas. The topic of potential difference and current in parallel and series electric circuits lasted four weeks with the students listening to a one two-hour lecture per week.

Diagnostic Questions and Administration

The study included three diagnostic questions about students' understanding of potential difference in parallel circuits. Each diagnostic question included two or three sub-questions (see Appendix). The questions were taken from a large bank of diagnostic questions developed by the EPSE Research Network in the UK (Millar, 2003). Written permission was obtained to use the questions in the study. After each question, students were required to indicate their level of confidence in their answers. The confidence judgement question was "How confident are you that your answers to this question are correct? Tick ONE box." The confidence levels were "just guessing," "not confident," "fairly confident," and "very confident." As several sub-questions were testing the same idea, i.e., potential difference in parallel circuits, in the same context, an overall confidence score was used. The questions were translated by the author into Turkish, and the accuracy of the translations was checked by another science education researcher, and the questions were modified in light of the comments received. Both researchers had a satisfactory level of English-language proficiency and were familiar with the terms relating to simple electric circuits. The prospective teachers answered the potential difference and confidence judgement questions as a part of their final exam.

Methods of Analysis

To acquire one point from a potential difference question, students had to answer all sub-questions correctly; otherwise, they were given zero points. Thus, for each question, a student's performance score was either 0 or 1. The reason for scoring the potential difference questions as "all or none" was to allocate 1 point for each correct response pattern. To investigate possible interactions between students' correct or incorrect answers and their confidence in them, ratings were scored as follows: "just guessing" (0), "not confident" (1), "fairly confident" (2), and "very confident" (3). These confidence-rating scores were then scaled to a 0–1 range to match the performance scores on the questions. Therefore, the scaled confidence ratings used were 0, .33, .67, and 1.

To evaluate each student's overconfidence or under-confidence on each question, each performance score was subtracted from the corresponding confidence rating. The new scores obtained in this way are called "bias scores" (Pallier et al., 2002). The bias score formula used in this study is as follows:

Bias Score: $c_i - p_i$

where c_i stands for a confidence rating, and p_i for a performance score. In this study, a student who had a bias score of less than –.33 was considered under-confident as he or she felt less than "fairly confident" after correctly answering a question. A student who had a bias score of more than .33 was considered overconfident as he or she felt more than "not confident" after incorrectly answering a question.

Participants' *overall* overconfidence or under-confidence in a given task can be evaluated through the bias index (Schraw, 2009; Yates, 1990). The bias index for a participant is the sum of the participant's bias scores divided by the number of questions. The bias index formula used in this study is shown below:

$$\text{Bias Index: } \frac{1}{N}\sum_{i=1}^{N}(c_i - p_i)$$

where N stands for the number of questions, c_i for a confidence rating, and p_i for a performance score. Similar to the bias scores, in this study, a student who had a bias index of less than –.33 was considered under-confident, and a student who had a bias index of more than 0.33 was considered overconfident.

Participants' *overall* calibration in a given task can be evaluated through the calibration index, also termed "absolute accuracy" (Schraw, 2009, p. 35). The calibration index formula used in this study is as follows:

$$\text{Calibration Index: } \frac{1}{N}\sum_{i=1}^{N}(c_i - p_i)^2$$

where N stands for the number of questions, c_i for a confidence rating, and p_i for a performance score. Similar to the bias index calculation, the calibration

index for a participant is the sum of the participant's calibration scores for each question divided by the number of questions. The calibration index ranges from zero (perfect calibration) to 1 (no calibration). In this study, a student who had a calibration index of less than or equal to .11 was considered "well calibrated," as the square of the cut-off indicates over/under-confidence (that is, .33 and –.33, respectively) was .11. Conversely, a student who had a calibration index of more than .11 was considered "poorly calibrated." These cut-off points were used because the calibration index was calculated using the term $(c_i - p_i)$ in the formula of the bias index.

Participants' resolution of confidence in a given task can be evaluated using Pearson's *r*, also termed "relative accuracy" (Schraw, 2009, p. 36), to measure the strength of correlation between a participant's performance score and confidence ratings for the questions in the set. In this study, some participants answered all three potential difference questions correctly while others answered all of them incorrectly. Similarly, some prospective teachers provided the same confidence rating for all three questions. Thus, it was not possible to calculate the correlation between performance scores and confidence ratings for these participants. Consequently, Pearson's *r* was calculated for each of the 43 participants who had some variance in both their performance scores and confidence ratings. In this study, students who had a Pearson's $r \geq .5$ were classified as "participants with good resolution." These prospective teachers could appropriately discriminate between what they knew and what they did not. Conversely, participants who had a Pearson's $r < .5$ were classified as "participants with poor resolution." This criterion was adopted to ensure that the common variance between a participant's performance score and confidence ratings for the questions in the set was at least 25%.

Construct Validity and Reliability of the Research Instrument

The three potential difference questions and the three confidence judgement questions in the study were factor analyzed using Principle Axis Factoring with Varimax (orthogonal) rotation to establish the construct validity of the research instrument. The analysis yielded two unique factors explaining a total of 43.6% of the variance accounted by all the factors. Factor 1 was labeled "participants' understanding of potential difference in parallel electric circuits due to the high loadings by the questions about potential difference in parallel circuits." This factor explained 28.9% of the variance accounted by all the factors. Factor 2 was labeled "the self-confidence (or self-monitoring) trait" (Kleitman & Stankov, 2001; Pallier et al., 2002; Stankov, 2000; Stankov & Crawford, 1997) due to the high loadings by the confidence judgement questions. The variance explained by this factor was 14.7%. The eigenvalue for each factor was greater than 1. The six questions loaded above .50 on their respective factors, and they did not load on any other factor. Only factor loadings >.30 were considered relevant. Bartlett's test of equal variance across groups and the Kaiser-Maiyer-Olkin (Leech, Barrett, & Morgan, 2005) measure of adequacy both indicated that the variables were adequately related for factor analysis. Furthermore, the same factors were obtained employing a Principle Axis Factoring with a Promax (non-orthogonal) rotation. The factor analysis, which was conducted on a small set of questions, indicates that there are two factors at play here are as follows: (*1*) knowledge of the content being tested and (*2*) confidence. For research purposes, "scores with modest reliability (coefficients in the range of .50 to .60) may be acceptable" (Ary, Jacobs, Sorensen, & Razavieh, 2010, p. 249). The internal consistency reliability coefficient was .707 (KR-20) for Factor 1 and .638 (Cronbach's Alpha) for Factor 2. This suggests that the research instrument used in this study can be considered adequate for exploring the relationship between the current study group's accuracy of and confidence in performance in answering diagnostic questions on potential difference in parallel electric circuits.

Limitations of the Study

Evaluation of the study results reveals certain limitations. First, the sample used in this study was a convenience sample. Therefore, it was not possible to use a statistical test, such as a *t*-test, to generalize the results obtained to a wider population. Second, the students answered only three diagnostic questions about potential difference in parallel circuits since the final exam included questions on other physics topics. Third, although the students were informed that honesty in answering the confidence judgement questions would be rewarded, some students may have exaggerated their confidence levels to portray a positive image to the instructor. Fourth, due to the upcoming mid-term break, no interview/qualitative data were collected to probe students' overconfidence and under-confidence. However, this study does generate some preliminary findings that can later be tested in a more comprehensive study.

Results and Discussion

Of the 342 performance scores, 181 (53%) were correct, indicating that the overall difficulty of the three potential difference questions was at an appropriate level for the participants (neither too easy nor too difficult). The mean confidence rating was .66, suggesting that the students were fairly confident in their answers. The rest of this section analyses the confidence-accuracy relationship in two parts: (*1*) prospective teachers' calibration and bias and (*2*) their resolution of confidence (i.e., the extent to which confidence judgements discriminate correct from incorrect answers).

Students' Calibration and Bias

Calibration of confidence is defined as "the distance between perceived and demonstrated levels of understanding, capability, competence, or preparedness" (Alexander, 2013, p. 1). In this study, prospective teachers who had a calibration index of less than or equal to .11 were considered "well calibrated," and those who had a calibration index of more than .11 were considered "poorly calibrated." Of the 114 participants, 33.3% were well calibrated. Of the 43 male participants, 20.9% were well calibrated while the corresponding figure was 40.8% for the female participants (see Table 1). In summary, both sexes were poorly calibrated in their understanding of the potential difference in parallel circuits and that the female students' calibration was better than that of their male counterparts.

Table 1
Percentages of Well Calibrated, Overconfident, and Under-Confident Participants on the Three Questions

	Well Calibrated Participants (%)	Overconfident Participants (%)	Under-Confident Participants (%)
All participants	33.3	24.6	7.0
Male	20.9	34.9	2.3
Female	40.8	18.3	9.9

The percentages of over/under-confident participants in Table 1 indicate that overconfidence made the largest contribution to the poor calibration of the students. In this study, the prospective teachers were overconfident when they felt "fairly confident" or "very confident" after incorrectly answering a question. The analyses of the participants' responses to the diagnostic questions indicated that many of their incorrect answers were based on some alternative conceptions or inappropriate use of some formulae. For example, of the 44 participants who incorrectly answered Question 1, 35 equally divided the potential difference of the battery between the resistors. Thirty-eight of the 43 participants who incorrectly answered Question 3 also equally divided the potential difference of the battery between the resistors. Of the 74 participants who incorrectly answered Question 2, 25 divided the potential difference of the battery between the resistors proportionally to their resistance. These participants did not differentiate between the potential difference and current in parallel circuits. Of the 44 participants who incorrectly answered Question 1, six indicated that the potential difference across each resistor was 16 volts. It appears that these participants first correctly calculated the total resistance, which was R/2 ($1/R_T = 1/R + 1/R$). The potential difference across the battery was 8 volts. Therefore, the main current in the circuit was 16/R volts ($I = V/R$). Perhaps, they inappropriately multiplied the main current with the resistance of each resistor ($V = IxR$) and found that the potential difference across each resistor was 16 volts. This inappropriate use of Ohm's Law ($V = IxR$) was also found in the 11 incorrect responses to Question 2.

Participants' calibration of confidence can be improved by providing specific instruction on the alternative conceptions or inappropriate use of formulae in a science topic. If students have the knowledge of the alternative forms of reasoning in a science topic, they will be more likely to monitor their occurrence when responding, resulting in a better evaluation of the accuracy of their responses. In terms of gender, the male participants were more overconfident whereas the female participants were more under-confident (see Table 1). Interventions to improve calibration may, therefore, need to be differentiated to meet the needs of both sexes.

Students' Resolution of Confidence

In this study, students' resolution (i.e., the extent to which confidence judgements discriminate correct from incorrect answers) was evaluated through Pearson's *r*. Students who had a Pearson's $r \geq .5$ were classified as "participants with good resolution." These participants could appropriately discriminate between what they knew and what they did not on the topic of the potential difference in parallel circuits. Conversely, participants who had a Pearson's $r < .5$ were classified as "participants with poor resolution." As explained earlier, this can only be calculated for participants who had some variance in both their performance scores and confidence ratings. Of the 43 participants who met this condition, 60.5% had good resolution. Of the 19 male participants who met this condition, 63.2%

had good resolution. The corresponding figure was 58.3% for the female participants. Irrespective of their sexes, most of the 43 participants could appropriately discriminate between what they knew and what they did not. For the 17 participants who had poor resolution, Pearson's *r* was less than .200. This indicates that the common variance between their performance scores and confidence ratings was less than 4%. The 26 participants with good resolution of confidence may use their study time more effectively than the 17 participants with poor resolution of confidence.

Table 2
Distribution of Correct and Incorrect Answers between Confidence Levels for the 43 Participants with Good or Poor Resolution

	Very Confident	Fairly Confident	Not Confident	Just Guessing
Correct	27	26	10	7
Incorrect	9	21	22	7

Note: The numbers in the table are the number of responses.

Table 2 shows the distribution of the correct and incorrect answers of the 43 participants between the confidence levels. Of the 70 correct answers, 75.7% were either in the "very confident" or in the "fairly confident" category. Of the 59 incorrect answers, 49.2% were either in the "not confident" or in the "just guessing" category. The corresponding figures for the whole sample were 77.9% and 39.8%, respectively. This indicates that the participants were more likely to know when they were giving the correct answer than when they were giving an incorrect answer. This suggests that, compared with the participants who knew when they were giving an incorrect answer those who knew when they were giving the correct answer made the larger positive contribution to the resolution observed in this study.

Table 3
Cross-Tabulation of Calibration and Resolution Data for the 43 Participants with Good or Poor Resolution

	Well Calibrated Participants	Poorly Calibrated Participants
Students with good resolution	12	14
Students with poor resolution		17

Note: The numbers in the table are the number of students.

Table 3 shows the distribution of the 43 participants between the categories of calibration and resolution. Twenty-nine of the 43 participants (a percentage of 67.4) had either good calibration and good resolution or poor calibration and poor resolution. In this study, the resolution of confidence was positively related to the calibration of confidence. This suggests that the concepts of calibration and resolution deal with two different, but related, aspects of the confidence-accuracy relationship. In Table 3, all the participants with poor resolution were also poorly calibrated. The participants with good resolution were distributed between the two categories of calibration. This suggests that having good resolution of confidence may be a prerequisite to, but not sufficient for, being well calibrated. The 17 participants who had poor resolution and calibration first need to improve their resolution of confidence. The 14 participants who had good resolution but poor calibration should focus on improving their calibration of confidence. The 12 participants who had good resolution and calibration may be more aware of their understanding of the topic of the potential difference in parallel circuits compared with the other 31 participants.

Conclusion and Implications of the Study

This study explored the relationship between accuracy of performance and confidence in performance among 114 prospective primary school teachers in answering diagnostic questions on potential difference in parallel electric circuits. In general, the participants in this study were poorly calibrated in their understanding of the potential difference in parallel circuits. They need to learn how to adjust their confidence ratings so that they match better to their performance scores. Otherwise, they may keep underestimating or overestimating their knowledge in science, resulting in poor learning. Individual feedback on calibration performance in science classes may help students to become less biased and better calibrated on diagnostic assessments (Stone & Opel, 2000). In this study, overconfidence made the largest contribution to the participants' poor calibration. This meant that many participants were "fairly confident" or "very confident" in their incorrect answers. The analyses of the participants' responses to the diagnostic questions indicated that many of their incorrect answers were based on some alternative conceptions or inappropriate use of some formulae. Therefore, one way to improve students' calibration of confidence may be to provide specific instruction on the alternative conceptions or inappropriate uses of formulae in a science topic. This study found that female participants, compared with male participants, were more under-confident and less overconfident in their understanding of the potential difference in parallel circuits. Therefore, interventions to

improve calibration may need to be differentiated to meet the needs of both sexes.

In this study, the prospective teachers were more likely to know when they were giving the correct answer than when they were giving an incorrect answer. Future research may consider the development of interventions designed to help students identify deficiencies in their understanding of a science topic. This study found that the resolution of confidence was positively related to the calibration of confidence. The science education researchers interested in the confidence-accuracy relationship may want to explore both calibration and resolution in their studies, as these concepts deal with two different, but related, aspects of the confidence-accuracy relationship. This study found that having good resolution of confidence may be a prerequisite to, but not sufficient for, being well calibrated. Future research may assess the reliability of this finding.

References

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning & Instruction, 24*, 1–3.

Arslan, H. Ö., Çiğdemoğlu, C., & Moseley, C. (2012). A three-tier diagnostic test to assess pre-service teachers' misconceptions about global warming, greenhouse effect, ozone layer depletion, and acid rain. *International Journal of Science Education, 34*(11), 1667–1686.

Ary, D., Jacobs, J. C., Sorensen, C., & Razavieh, A. (2010). *Introduction to research in education* (8th ed.). Canada: Wadsworth, Cengage Learning.

Caleon, I. S., & Subramaniam, R. (2010a). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, *32*(7), 939–961.

Caleon, I. S., & Subramaniam, R. (2010b). Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education*, *40*(3), 313–337.

Cohen, R., Eylon, B., & Ganiel, U. (1983). Potential difference and current in simple electric circuits: A study of students' concepts. *American Journal of Physics, 51*(5), 407–412.

Colclough, N. D., Lock, R., & Soares, A. (2011). Pre-service teachers' subject knowledge of and attitudes about radioactivity and ionising radiation. *International Journal of Science Education, 33*(3), 423–446.

Driver, R., Guesne, E., & Tiberghien, A. (1985). *Children's ideas in science*. Milton Keynes: Open University Press.

Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). *Making sense of secondary science*. London: Routledge.

Duit, R. (2009). *Students' and teachers' conceptions and science education*. Kiel: Institude for Science Education (IPN) (Distributed electronically).

Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education, 34*(5), 294–299.

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15*(3), 321–341.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation* (2nd ed.) Mahwah, NJ: Lawrence Erlbaum Associates.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic., & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge: Cambridge University Press.

McClary, L. M., & Bretz, S. L. (2012). Development and assessment of a diagnostic tool to identify organic chemistry students' alternative conceptions related to acid strength. *International Journal of Science Education, 34(5),* 2317–2341.

McDermott, L. C., & Shaffer, P. S. (1992). Research as a guide for curriculum development: An example from introductory electricity. Part I: Investigation of student understanding. *American Journal of Physics, 60*(11), 994–1003.

Mengelkamp, C., & Bannert, M. (2012). Confidence judgements in learning. In N. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 756–759). New York, NY: Springer.

Millar, R. (2003). *Evidence-based Practice in Science Education (EPSE) Project 1. Sample Diagnostic Questions. Electric circuits Set 4: Potential difference*. York: University of York Science Education Group. Retrieved July 31, 2013 from www.york.ac.uk/media/educationalstudies/documents/research/epse/Potentialdifference.pdf

Odom, A. L., & Barrow, L. H. (2007). High school biology students' knowledge and certainty about diffusion and osmosis concepts. *School Science and Mathematics, 107*(3), 94–101.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology 129*(3), 257–299.

Peşman, H., & Eryılmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *The Journal of Educational Research, 103*(3), 208–222.

Planinic, M., Boone, W. J., Krsnik, R., & Beilfuss, M. L. (2006). Exploring alternative conceptions from Newtonian dynamics and simple DC circuits: Links between item difficulty and item confidence. *Journal of Research in Science Teaching*, *43*(2), 150–171.

Potgieter, M., Malatje, E., Gaigher, E., & Venter, E. (2010). Confidence versus performance as an indicator of the presence of alternative conceptions and inadequate problem-solving skills in mechanics. *International Journal of Science Education, 32*(11), 1407–1429.

Reif, F., & Allen, S. (1992). Cognition for interpreting scientific concepts: A study of acceleration. *Cognition & Instruction, 9*(1), 1–44.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition & Learning, 4*(1), 33–45.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes, 42*, 271–283.

Shipstone, D. M. (1984). A study of children's understanding of electricity in simple DC circuits. *European Journal of Science Education, 6*(2) 185–198.

Shipstone, D. M., Rhöneck, C. V., Jung, W., Kärrqvist, C., Dupin, J.-J., Johsua, S., & Licht, P. (1988). A study of students' understanding of electricity in five European countries. *International Journal of Science Education, 10*(3), 303–316.

Sreenivasulu, B., & Subramaniam, R. (2013). University students' understanding of chemical thermodynamics. *International Journal of Science Education, 35*(4), 601–635.

Stankov, L. (2000). Complexity, metacognition and fluid intelligence. *Intelligence, 28*(2), 121–143.

Stankov, L., & Crawford, J. D. (1996). Confidence judgements in studies of individual differences. *Personality and Individual Differences, 21*(6), 971–986.

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence, 25*(2), 93–109.

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes, 83*(2), 282–309.

White, R., & Gunstone, R. (1992). *Probing understanding*. London: The Falmer Press.

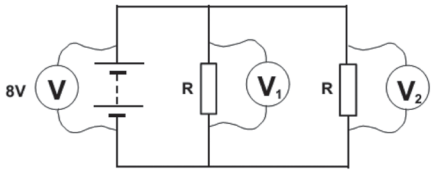
Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall.

Appendix

The three diagnostic questions used in the study. After answering each of the three questions, prospective teachers were required to indicate their confidence in their answers.

Question 1

The two resistors in this circuit are identical. The voltmeter connected across the battery reads 8V.



(a) What is the reading on voltmeter V_1? _____ volts

(b) What is the reading on voltmeter V_2? _____ volts

How confident are you that your answers to this question are correct? Tick ONE box (✓)

Very confident ☐

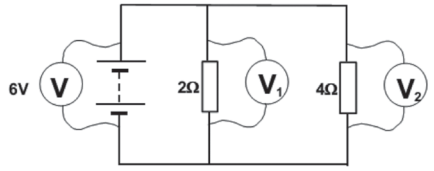
Fairly confident ☐

Not confident ☐

Just guessing ☐

Question 2

In this circuit, the voltmeter across the battery reads 6V.



(a) What is the reading on voltmeter V_1? _____ volts

(b) What is the reading on voltmeter V_2? _____ volts

How confident are you that your answers to this question are correct? Tick ONE box (✓)

Very confident ☐

Fairly confident ☐

Not confident ☐

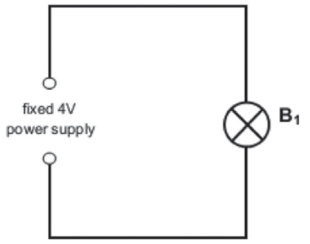
Just guessing ☐

Question 3

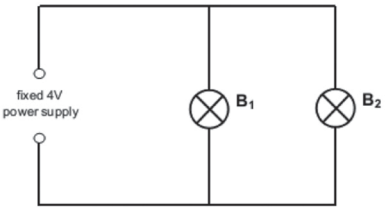
The power supply in this circuit has a fixed output voltage of 4V.

(a) A bulb B_1 is connected to the power supply.

What is the voltage across bulb B_1? _____ volts



(b) A second identical bulb B_2 is then connected, to make this circuit.



What is the voltage now across bulb B_1? _____ volts
What is the voltage across bulb B_2? _____ volts

How confident are you that your answers to this question are correct? Tick ONE box (✓)

Very confident ☐
Fairly confident ☐
Not confident ☐
Just guessing ☐