Physical education students' calibration accuracy and academic achievement: A longitudinal study

Athanasios Kolovelonis and Marios Goudas
University of Thessaly, Greece

Two studies were conducted to examine associations between undergraduate physical education students' calibration accuracy, assessed both at the local and global level, and their short and long-term achievement in two academic courses (developmental and sport psychology). Participants were 68 Greek senior and 112 first-year physical education students who completed knowledge tests and provided their judgments regarding their performance in these tests. Students' degree grade and time of graduation were also obtained in study two. Results showed that calibration accuracy could explain a small amount of variance of students' academic achievement measured both with short-term (i.e., knowledge tests in the middle and in the end of the semester) and long-term indicators (e.g., degree grade). Calibration at the local level was generally a better predictor of students' short and long-term academic achievement while results regarding differences in the accuracy of the judgments provided at the local and the global level were mixed. These finding are discussed with reference to the theoretical and practical implications for undergraduate students' academic achievement.

Introduction

In educational contexts, students are frequently involved in making judgments for their learning and performance in tasks, tests, or exams. These judgments either made before (i.e., predictions) or after (i.e., postdictions) performance are considered metacognitive in nature, are associated with self-regulated learning, and have attracted researchers' interest in various fields. The present study focused on the accuracy of undergraduate physical education students' judgments and the relations between their accuracy and short and long-term academic achievement.

The term calibration has been widely used to capture the discrepancy between judged and actual performance (Schraw, 2009). In a typical calibration research, students are asked to judge their performance in a test either at a local (i.e., item-by-item) or at a global level (i.e., a cumulative judgement for all items). Then, they perform the test to compare the judged with the actual performance. In the case the judged performance is close to the actual performance, students are considered well calibrated. In the case that the judged performance is higher compared to the actual performance, students are considered overestimators, and if the judged performance is lower than actual performance students are underestimators (Schraw, 2009).

Calibration is considered an important construct in educational contexts due to its associations with motivation and self-regulated learning (Pieschl, 2009). Students' capacity to monitor their own learning and performance is essential for self-regulated learning (Zimmerman, 2000), in the sense that calibration accuracy can inform effective circles of self-regulation (Chen & Rossi, 2013). In fact, reflecting the discrepancy between judged and actual performance, calibration can be considered an indicator of the accuracy of students' awareness regarding what they know and do not know (Gasser & Tan, 2005) affecting their decisions during learning and performance. For example, well-calibrated students set challenging goals and are ready to adjust learning strategies or to focus in aspects of performance that need more practice. In contrast, miscalibrated students who believe that their capabilities are lower than they actually are may avoid challenging tasks and set lower learning goals, thus limiting their potential for mastering new skills (Schunk & Pajares, 2004). Moreover, miscalibrated students who erroneously judged that they have reached high levels of performance may be reluctant to try hard to further develop their skills or may set unrealistic and unachievable goals. Thus, calibration is considered a critical element for enhancing performance and self-regulated learning (Chen & Rossi, 2013).

Calibration has been widely examined in various educational settings. Research focusing on students' capacity to estimate accurately their learning and performance has shown that students are usually inaccurate, with a tendency to overestimate their learning or performance in academic settings (e.g., Chen, 2003; Singer & Alexander, 2017). Evidence of overconfidence in sport performance was also found among recreational basketball players (McGraw, Mellers & Ritov, 2004), golfers (Fogarty & Else, 2005), and elementary students in physical education (Kolovelonis & Goudas, 2012; Kolovelonis, Goudas & Dermitzaki, 2012; Kolovelonis, Goudas, Dermitzaki & Kitsantas, 2013).

Other research focusing on factors related to students' calibration has shown that miscalibration was higher in more difficult tasks (e.g., Chen & Zimmerman, 2007). Also, experienced students were better calibrated (Dinsmore & Parkinson, 2013; Kolovelonis, 2019b), probably because they can discriminate more effectively what they know in specific topics (Bol, Hacker, O'Shea & Allen, 2005). Students' calibration of sport performance was positively associated with self-efficacy and task goal orientation (Kolovelonis & Goudas, 2018), while the characteristics of the tasks, such as the shooting position (Kolovelonis & Goudas, 2019), and students' predictions regarding their peers' performance (Kolovelonis & Dimitriou, 2018) have also been involved in calibration accuracy.

Calibration and academic achievement

An intriguing issue regarding calibration is its association with academic achievement. It has been theorised that accurate metacognitive monitoring is a key element for self-regulated learning and increased performance (Chen & Rossi, 2013; Zimmerman, 2000). Indeed, research findings in elementary and secondary school settings have shown that high performers are usually more accurate in predicting or postdicting their performance (Hacker, Bol & Bahbahani, 2008; Hacker, Bol, Horgan & Rakow, 2000; Kolovelonis & Goudas, 2019; Ots, 2013). Similarly, research evidence has suggested that metacognitive awareness is associated with academic achievement among undergraduate students (Bol & Hacker, 2001; Bol et al., 2005; Young & Fry, 2008). Nietfeld, Cao and Osborne (2005) reported that calibration accuracy remained stable across tests throughout the semester, students were more accurate in their global predictions than their local predictions, and student performance on the tests was related to local calibration accuracy.

Furthermore, students who were involved in interventions to increase calibration accuracy increased their performance too (Bol, Hacker, Walck & Nunnery, 2012; Nietfeld, Cao & Osborne, 2006). A recent meta-analysis suggested that learning strategy instruction can benefit metacognitive monitoring (Gutierrez de Blume, 2021). In an experimental study involving the knowledge of key-term definitions (Dunlosky & Rawson, 2012) greater accuracy was related to greater retention (two days later), while informing university students about the consequences of making overconfident judgments had positive effects on their calibration accuracy (Roelle, Schmidt, Buchau & Berthold, 2017). Valdez (2013) reported that students' absolute accuracy was significantly correlated with concurrent and final exam performance in an undergraduate course of language acquisition. Moreover, the use of appropriate instructional strategies increased students' calibration and improved their performance (Osterhage, Usher, Douin & Bailey, 2019). Follmer, Patchan and Spitznogle (2022) reported that the improvements in students' study time calibration scores followed by a respective intervention was associated with course performance and reported goal setting skills among college students. All this evidence suggested that calibration may be associated with students' academic achievement. However, further empirical evidence regarding calibration and students' short and especially long-term academic achievement is warranted (Bol & Hacker, 2012).

Calibration at the local and global level

Calibration can be measured either at the local (i.e., students provide their judgments item-by-item) or at the global level (i.e., students provide a cumulative judgment for all items) (Schraw, 2009). Both these types of judgments are considered useful measures of online monitoring (Griffin, Wiley & Salas, 2013), are important for self-regulated learning, and have been widely used in calibration research (Chen & Rossi, 2013). Accuracy of judgments at both the local and global level is critical for learning because it informs about learning status in general and also about more specific content.

Grabe and Holfeld (2014) found that both measures were significant and unique predictors of future performance in an introductory college course conducted online. In contrast, local compared to global prediction accuracy was more strongly associated with test performance (Nietfeld et al., 2005) and domain-specific self-concept was more strongly related to global than local judgments and bias (Händel, de Bruin & Dresel, 2020). Moreover, research into metacomprehension judgments has shown that there was little or no relation between absolute (i.e., judgments for overall performance) and relative accuracy (i.e., discrimination of performance across items) (Maki, Shields, Wheeler & Zacchilli, 2005). These mixed results highlight the need for further investigation regarding the relation of the local and global measures of calibration with students' academic achievement.

The present study

Two studies were conducted focusing on the associations between undergraduate physical education students' calibration accuracy and their short and long-term academic achievement. In particular, the main research question for these studies was whether calibration accuracy could predict students' short and long-term academic achievement. Comparisons between local and global calibration accuracy were also conducted in both studies.

Evidence regarding the relations between measures of metacognitive monitoring and future achievement is limited (Schraw, 2009). It has been theorised that current monitoring skill can be predictive of future learning and performance (Pintrich, 2000). High performers are usually more accurate (e.g., Bol & Hacker, 2001; Bol et al., 2005; Hacker et al., 2008; Hacker et al., 2000) while greater accuracy was related with greater retention in short time (two days later) (Dunlosky & Rawson, 2012) and concurrent and final exam performance (Valdez, 2013). However, further evidence regarding the link between monitoring accuracy and achievement, especially involving undergraduate students, is warranted to inform calibration research (Bol & Hacker, 2012). This evidence should include not only short-term but also long-term measures of academic achievement and performance.

Greater calibration accuracy may help undergraduate students to manage their time and effort more effectively, avoiding either premature termination or prolonged duration of study (Hacker et al., 2000). Thus, their success in exams may increase and their general academic achievement including their degree grade and time for graduation may be improved. Graduating is a significant milestone for university students and the factors associated with it may vary. However, research evidence regarding students' graduation and time to degree, or the factors that may lead students to drop out their studies, is generally limited (Yue & Fu, 2017). Some evidence suggests that academic performance is one of the most important factors (Yue & Fu, 2017). For example, research among medical students suggested that struggling academically may be strongly associated with dropout while no specific pattern of demographic variables was particularly important in relation to dropout (O'Neill, Wallstedt, Eika & Hartvigsen, 2011).

Furthermore, although judgments in calibration research can be provided either at the local or at the global level (Schraw, 2009), little is known regarding the differences between these two types of metacognitive judgments within individuals, and whether one or both of these indexes can predict students' academic achievement. Some previous research has provided mixed results (Grabe & Holfeld, 2014; Händel et al., 2020; Maki et al., 2005; Nietfeld et al., 2005). Thus, this study involved comparisons of students' calibration accuracy at the local and global level and examined the predictive value of these types of judgments regarding students' academic achievement.

Moreover, calibration research in real-life classroom contexts, especially over an extended period, is generally limited (Bol & Hacker, 2012). Much of previous calibration research has focused in studying the relationship between judgments and performance using objective assessments of recently studied material (Nietfeld et al., 2006). Thus, following previous research examining calibration in applied settings (Hacker et al., 2008; Hadwin & Webster, 2013; Nietfeld et al., 2005) both studies reported here were conducted in real life learning environments, involving authentic learning materials and evaluation processes that are meaningful for students, thereby increasing the ecological validity of the results.

The aim of this study was to examine if undergraduate students' calibration accuracy could predict their short-term (middle and end of the semester) achievement in two academic courses (developmental and sport psychology), and their long-term academic achievement (degree grade and time of the graduation). Differences between local and global judgments of performance and their relations with students' academic achievement were also explored. It was hypothesised that students' calibration accuracy would predict their short and long-term academic achievement. No specific hypotheses regarding potential differences between local and the global judgments were stated, due to previous mixed results.

Method: Study One

Settings and procedures

The first study was conducted in the context of a developmental psychology course which is delivered as an elective course in the spring semester of fourth year of studies in the local Department of Physical Education and Sport Science. The structure of this Department' program is based on compulsory core courses and elective courses. The studies are completed after 8 semesters and a total number of 240 ECTS is required for graduating. There is not currently an upper time limit, after completing the fourth year of studies, for students to graduate. There are three exam periods in each academic year, in January (for courses delivered in winter semester), in June (for courses delivered in spring semester) and in September (for courses delivered both in winter and in spring semester). Students who have completed the four years of regular study can participate in every exam period for a course, regardless of the semester that the course had been delivered.

The course in developmental psychology included ten 90-minute lectures delivered weekly regarding introduction to developmental psychology, Piaget's theory of cognitive development, cognitive functions, metacognition, intelligence, emotional intelligence, self-esteem, play as a developmental process, and the role of family in children's development. The evaluation included three written tests (after the third, sixth, and tenth lecture) counting 30% each, and one written assignment (10% of the total grade). Attendance was compulsory although students could miss up to 30% of the lectures.

Ethical approval for the study was granted by the University Ethics Review Committee. Students were informed regarding the study and those who agreed to participate in the experiment completed a consent form and responded the additional questions regarding calibration. No credits were provided to students for their participation. No student refused to participate. All knowledge tests used in this study were part of students' official evaluation process for this course.

Participants

Participants were 68 Greek senior physical education students (38 males; mean age = 22.64 years, SD = 3.14) from a physical education and sport science department that elected the developmental psychology course in the spring semester.

Measures

Knowledge tests 1, 2 and 3

Students responded to three knowledge tests each containing 20 multiple choice questions, the first after the third lecture, the second after the sixth lecture (in the middle of the semester) and the third after the tenth lecture (at the end of the semester). The number of correct answers in each test recoded into a 100-point scale was recorded as students' score in the knowledge tests for the developmental psychology course.

Judgments of learning

During the first knowledge test, students provided their judgments of learning both at the local and global level. In particular, after each question of the test students were asked to report how confident they were that they had provided the correct answer, by responding to the question: "How confident are you that you answered this question correctly?" Students responded on a scale ranging from 0 (not at all sure) to 100 (absolutely sure) gradually increasing by 10 points with additional marks for every 5 points. The average of students' confidence scores in the 20 questions was their score at the local judgment in the knowledge test (range: 0 - 100) (Schraw, 2009). After responding to all the questions and providing their local confidence judgments, students were also asked to provide a global judgment regarding the accumulative number of questions they had answered correctly, by completing the following statement: "I think I have answered correctly … out of 20 questions".

Calibration accuracy

The calibration index of absolute accuracy at the local and at the global level was calculated. In particular, the absolute difference between the local or the global judgment and the actual score in the first knowledge test was calculated, respectively. This calibration index of absolute accuracy reflects the magnitude of calibration error with values closer to zero indicating higher calibration accuracy. The advantage of this index is that transforms the negative values of the discrepancy between judged and actual performance into positive ones, making the interpretation of the results easier (Schraw, 2009).

Statistical analyses

Multiple regression analyses were conducted with students' calibration accuracy at the local and global level as independent variables. The outcome variables were students' scores in the knowledge tests of developmental psychology in the middle and at the end of the semester, respectively. Paired samples t-tests were used for comparing students' calibration accuracy at the local and global level. Independent t-tests were used for examining potential gender differences in calibration accuracy. Effect sizes of Cohen's *d* were calculated (Cohen, 1988).

Results

Preliminary analyses

Means, standard deviations, and correlations between all variables are presented in Table 1. Students overestimated their performance in the first knowledge test. In particular, on average, students' judged performance in the first knowledge tests was 32% higher at the local level and 20% higher at the global level compared to their actual performance. Students' scores in the three knowledge tests were positively and highly correlated. Similarly, students' judgments and calibration accuracy at the local and global level were positively and highly correlated.

No gender differences were found in students' score in the knowledge test 1, $t(66) = -0.54$, $p = .591$, test 2, $t(66) = -1.39$, $p = .171$, and test 3, $t(66) = -1.22$, $p = .227$. Moreover, no gender differences were found at the local judgments, $t(66) = 0.35$, $p = .729$, at the global judgments, $t(66) = 0.80$, $p = .425$, at the local calibration accuracy, $t(66) = 0.45$, $p = .652$, and at the global calibration accuracy, $t(66) = 0.12$, $p = .907$.

Local versus global accuracy

No difference between local and global calibration accuracy was found, $t(67) = 1.38$, $p = .173$. However, students provided higher judgments of performance at the local level compared to the global level, $t(67) = 3.56$, $p < .001$, $d = 0.37$.

Regression analyses

Multiple linear regression analysis showed that students' scores at the local and global calibration accuracy could explain significantly a small amount of the variance of their scores in the knowledge test 2, $F(2, 67) = 3.47$, $p = .037$, $R^2 = .10$. However, betas were nonsignificant for both predictors (i.e., local and global accuracy).

Similarly, multiple linear regression analysis showed that students' scores at the local and global calibration accuracy could explain significantly a small amount of the variance of their scores in the knowledge test 3, $F(2, 67) = 4.49$, $p = .015$, $R^2 = .12$. However, betas were nonsignificant for both predictors (i.e., local and global accuracy).

Table 1: Means, standard deviations, and correlations for all variables in Study One

	M	SD	Correlations					
			1	2	3	4	5	6
1. Test 1 score	53.53	22.08	-					
2. Test 2 score	43.60	19.87	.62**	-				
3. Test 3 score	51.84	16.07	.59**	.68**	-			
4. Local judgment	70.43	17.89	.60**	.48**	.45**	-		
5. Global judgment	64.34	15.21	.49**	.43**	.37**	.65**	-	
6. Accuracy local	19.91	15.00	-.62**	-.26*	-.32**	.14	-.09	-
7. Accuracy global	17.72	13.67	-.55**	-.29*	-.30**	-.18	.08	.59**

Note: Test scores have been recoded in a 100-point scale
*p< .05, **p< .01

Method: Study Two

Settings and procedures

The second study was conducted in the context of a sport psychology course delivered to all first-year students in the winter semester of the first year of studies in the local Department of Physical Education and Sport Science. The settings and the procedures of the study were similar to those described in study one. The sport psychology course included twelve 90-minute lectures delivered weekly regarding an introduction to sport psychology, motivation, goal orientation, attribution theories, goal setting, group dynamics, communication, leadership, moral development, and aggressiveness in sport settings. The evaluation process included a final written exam counting 80% and a knowledge test in the middle of the semester consisting of 20 multiple choice questions counting 20% of the grade. Attendance at the lectures for this course was not compulsory.

Participants

Participants were 112 Greek first-year physical education students (57 males), 18 to 20 years old in the beginning of the study (*M*age = 18.77, *SD* = 1.23) who attended the sport psychology course in the winter semester.

Measures

Calibration accuracy
Students' local and global calibration accuracy indexes were calculated following the process described in study one. For measuring actual performance, a knowledge test consisted of 20 multiple choice-questions regarding the content of the first six lectures of the sport psychology course was used. Local and global judgments were measured as described in study one.

Grades in the final exam
A knowledge test, consisted of 20 multiple choice questions, was used in the final exam. The number of correct answers recoded into a 100-point scale was each student's score in the final exam.

Grade of degree
Students' grade of their degree was also obtained after their graduation four years or later. The grade of the degree was calculated as the average of the grades students received in all courses required for acquiring their degree.

Graduation time
Students' time of graduation, further of the compulsory four years period (eight semesters), was also calculated. For students who graduated within four years (at the end of the eighth semester) the score zero was set and for every period of exams a student failed to graduate one more point was added in the measurement scale of his or her graduation time. A higher score on this index indicated a greater delay in graduation time.

Statistical analyses

Statistical analyses were similar to study one. In the three regression analyses conducted, the outcome variables were students' score in final exams, their degree grades and their graduation time.

Results

Preliminary analyses

Means, standard deviations, and correlations between all variables are presented in Table 2. Students overestimated their performance in the knowledge test. In particular, on average, students' judged performance in the knowledge test was 57% higher at the local level and 31% higher at the global level compared to their actual performance. Students' judgments and calibration accuracy at the local and global level were positively correlated. Moreover, students' scores in the final exams test were negatively correlated with calibration accuracy at the local and the global level, while students' grade of their degree were negatively correlated only with calibration at the local level.

Females compared to males performed higher in the knowledge test, $t(110) = -2.47$, $p = .015$, $d = 0.47$. No gender differences were found in students' score in the final exams test, $t(110) = -0.92$, $p = .362$, in the grade of their degree, $t(100) = -1.16$, $p = .249$, and in the time for graduation, $t(110) = 1.91$, $p = .058$. Moreover, no gender differences were found in the local judgments, $t(110) = -0.01$, $p = .997$, in the global judgments, $t(110) = -0.21$, $p = .834$, in the local calibration accuracy, $t(110) = 1.80$, $p = .075$, and in the global calibration accuracy, $t(110) = 1.66$, $p = .101$.

Table 2: Means, standard deviations, and correlations for all variables in Study Two

	M	SD	Correlations						
			1	2	3	4	5	6	7
1. Knowledge test score	48.17	16.38	-						
2. Final exams test score	50.72	18.12	.60**	-					
3. Degree's grade	7.54	0.57	.40**	.37**	-				
4. Time for graduation	1.87	1.78	-.13	-.13	-.45**	-			
5. Local judgment	75.74	14.05	.23*	.20*	-.04	.02	-		
6. Global judgment	63.17	12.48	.26**	.27**	.08	.01	.69**	-	
7. Accuracy local	29.21	16.29	-.71**	-.38**	-.29**	.08	.40**	.18	-
8. Accuracy global	19.82	12.14	-.56**	-.21**	-.12	.12	.15	.29**	.69**

Note: Test scores have been recoded in a 100-point scale
**p*< .05, ***p*< .01

Local versus global accuracy

Students provided lower judgments of performance at the global level compared to the local level, $t(111) = 12.57$, $p < .001$, $d = 0.95$. Moreover, students were more accurate at the global level compared to the local level, $t(111) = 8.42$, $p < .001$, $d = 0.65$.

Regression analyses

Multiple linear regression analysis showed that students' scores at the local and global calibration accuracy measured in the middle of the semester could explain significantly an amount of the variance of their scores in the final exam test, $F(2, 107) = 8.91$, $p < .001$, $R^2 = .15$. The analysis showed that students' scores at the local accuracy significantly predicted their scores in the final exams test, beta = -.44, $p < .001$, but their scores at the global accuracy did not.

Similarly, multiple linear regression analysis showed that students' scores at the local and global calibration accuracy measured in the middle of the semester could explain significantly an amount of the variance of their grade of degree, $F(2, 101) = 4.98$, $p = .009$, $R^2 = .09$. The analysis showed that students' scores at the local accuracy significantly predicted their scores in their grade of degree, beta = -.38, $p = .005$, but their scores at the global accuracy did not.

Multiple linear regression analysis showed that students' scores at the local and global calibration accuracy measured in the middle of the semester could not explain significantly an amount of the variance of their graduation time, $F(2, 111) = 0.85$, $p = .428$.

Discussion

Two studies were conducted involving two academic courses (i.e., sport psychology and developmental psychology) to examine the associations between undergraduate physical education students' calibration and their academic achievement. Students' calibration at the local and the global level were also explored. The results showed that students generally overestimated their performance. Calibration accuracy could explain a small amount of variance of students' academic achievement measured with both short-term indicators (i.e., knowledge tests in the middle and at the end of the semester) and long-term indicators (e.g., grade of degree). Results regarding the differences between calibration accuracy at the local and global level were generally mixed. All these findings are discussed next with reference to the previous finding and the theoretical and practical implications for undergraduate students' academic achievement. Limitations and suggestions for future research are also discussed.

Calibration accuracy and academic achievement

The present studies showed that calibration accuracy accounted for a small amount of variance of students' academic achievement including their performance in the exams in two academic courses (i.e., developmental psychology and sport psychology) within a semester and their degree grade at the end of their studies four or more years later. Moreover, the correlation between calibration accuracy and future achievement was negative indicating that greater accuracy was related with higher achievement. These results are consisted with previous findings showing that high performers are usually more accurate (Bol & Hacker, 2001; Bol et al., 2005; Hacker et al., 2008; Hacker et al., 2000) and calibration accuracy was related with greater retention (two days later) in a study of key-term definitions (Dunlosky & Rawson, 2012) and concurrent and final exam performance within a semester in an undergraduate course on language acquisition (Valdez, 2013). Most importantly, the present results expanded previous finding involving a longitudinal design and showing that calibration accuracy was associated not only with short-term (i.e., within a semester) but also with long-term (i.e., grade of degree four years later) academic achievement.

The results of this study are also consistent with theoretical views suggesting that accurate metacognitive monitoring is a key element in self-regulated learning (Chen & Rossi, 2013; Zimmerman, 2000) and predictive of future learning and performance (Pintrich, 2000). Well calibrated students may manage their time and effort more effectively, improving their successfulness in exams and their general academic achievement including the grade of their degree. It should be noted however that the associations between calibration accuracy and academic achievement were relatively low. Moreover, these associations varied across the level of calibration accuracy (i.e., local and global level), findings that are discussed in the next section. Furthermore, no associations between calibration accuracy and time of graduation were found in the second study. Graduating is a significant milestone for university students and has been associated with academic performance (Yue & Fu, 2017). For example, levels of dropout were higher among medical students struggling academically (O'Neill et al., 2011). However, several other factors including sociological and economic factors may be associated with time to degree (Letkiewicz, Lim, Heckman, Bartholomae, Fox & Montalto, 2014). Future research may further explore the associations between students' calibration accuracy and time of graduation controlling the effects of such factors.

Calibration at the local and the global level

A rather unexplored area in calibration research is the potential differences between calibration accuracy at the local and the global level (Bol & Hacker, 2012). Although both of these types of metacognitive judgments have been widely used in previous research, very little research has directly compared them in a single study and the results of that studies were generally mixed (e.g., Grabe & Holfeld, 2014; Händel et al., 2020; Maki et al., 2005; Nietfeld et al., 2005). In the present studies, students provided lower judgments of performance at the global compared to the local level. This is consistent with theoretical and empirical evidence suggesting that mean post-test estimates usually are smaller than mean confidence ratings for a test (provided item-by-item) (Gigerenzer, Hoffrage & Kleinbölting, 1991; Stankov & Crawford, 1996).

Moreover, evidence that mean post-test estimates (i.e., calibration at the global level) display better calibration accuracy compared to judgments provided at the local level was partially confirmed. In particular, in study two, students were more accurate when providing judgments at the global level compared to the local level while in the study one no differences between these two calibration indexes were found. On the other hand, regression analyses in study two suggested that calibration at the local level was generally a better predictor of students' short- and long-term academic achievement (i.e., final exams test and grade of degree) compared to calibration at the global level. These results are similar to those of Nietfeld et al. (2005) showing that global judgments were more accurate than local judgments, but student performance was related only to local accuracy.

This evidence has suggested that students may be more capable of judging accurately their general level of performance (i.e., at global level). However, providing judgments at the local level is a more sensitive measure of metacognitive monitoring, requiring from students to judge what is actually known in every single item (Efklides, 2014), resulting thus in closer associations with performance. Indeed, it has been supported (Gigerenzer et al., 1991) that item-by-item confidence judgments (i.e., at the local level) and post-test percentage correct judgments (i.e., at global level) are based on different cognitive processes and thus different cues may be involved in making judgments in each case. Undoubtedly, both local and global calibration accuracy can be useful measures of online monitoring (Griffin et al., 2013), providing critical information for learning status in general and more specific content, thus contributing to self-regulated learning (Chen & Rossi, 2013).

Practical implications

In both studies students overestimated both at the local and the global level their performance in the knowledge tests. These results are consistent with previous findings in academic (e.g., Chen, 2003), sport (e.g., Fogarty & Else, 2005; McGraw et al., 2004) and physical education (e.g., Kolovelonis & Goudas, 2012, 2018, 2019; Kolovelonis et al., 2012; Kolovelonis et al., 2013) settings. All this evidence indicates a prevalence of overconfidence in metacognitive judgments of learning and performance and highlights the need for implementing interventions for enhancing students' calibration accuracy.

The present studies, following previous respective research (Hacker et al., 2008; Hadwin & Webster, 2013; Nietfeld et al., 2005, 2006), were conducted in real life learning environments involving learning materials from academic courses delivered to students as a part of their program of study. This has increased the ecological validity and the meaningfulness of the practical implication of the results. From an applied perspective, this suggests that efforts to improve students' calibration should be integral in the process of promoting their academic learning and performance. Studying in university involves students' initiatives and thus self-regulated learning may play a critical role in students' academic achievement (Kitsantas & Zimmerman, 2009). From this perspective, calibration accuracy is considered a critical element in self-regulated learning (Chen & Rossi, 2013). Thus, training students to develop their self-regulatory skills and to enhance their calibration accuracy may help them to improve their academic achievement. For example, an intervention based on the four-level training model of self-regulated learning development (Goudas, Kolovelonis & Dermitzaki, 2013; Zimmerman, 2000) improved students' calibration accuracy in physical education (Kolovelonis, Goudas & Samara, 2022). Such interventions may also provide practice opportunities for calibrating performance (Bol et al., 2012), self-reflection (Zimmerman, Moylan, Hudesman, White & Flugman, 2011), and strategy instruction combined with extrinsic incentives (Gutierrez & Schraw, 2015).

Limitations and future research

Limitations regarding the present studies should be acknowledged. In both studies, students' calibration accuracy was measured through a single knowledge test, one in an elective course during the fourth year of studies and one in a compulsory course during the first year of studies. Thus, the results of this study should be generalised beyond these learning environments with caution. Future research should examine students' calibration accuracy involving multiple knowledge tests from representative academic courses. This may be more critical for research examining the predictive power of students' calibration accuracy in relation to their long-term academic achievement (i.e., degree's grade or time of graduation). Such research may also include qualitative or mixed methods designs involving students in estimations for their performance in various types of knowledge tests (e.g., open ended questions).

Moreover, considering recent evidence showing that students were more accurate in estimating their performance in a sport task compared to sport-related knowledge tests (Kolovelonis, 2019a), calibration research involving physical education students may also include calibration accuracy regarding learning and performance of sport tasks. Longitudinal designs should also involve more calibration measures within the four years of studies (e.g., every year) to capture potential changes in students' calibration accuracy throughout their studies. Moreover, research has shown that calibration accuracy is associated with factors such as students' self-efficacy and task orientation (e.g., Kolovelonis & Goudas, 2018). However, these factors may also be associated with academic achievement. Thus, future research should explore potential associations and interactions between all these factors. The nature of students' calibration accuracy at the local and the global level should be further explored with special emphasis given to examining whether these types of calibration are associated with different learning or performance outcomes.

References

Bol, L. & Hacker, D. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education,* 69(2)*,* 133-151. https://doi.org/10.1080/00220970109600653

Bol, L. & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology*, 3, 229. https://doi.org/10.3389/fpsyg.2012.00229

Bol, L., Hacker, D. J., O'Shea, P. & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education,* 73(4)*,* 269-290. https://doi.org/10.3200/JEXE.73.4.269-290

Bol, L., Hacker, D. J., Walck, C. C. & Nunnery, J. A. (2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology*, 37(4), 280-287. https://doi.org/10.1016/j.cedpsych.2012.02.004

Chen, P. P. (2003). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences*, 14(1), 79-92. https://doi.org/10.1016/j.lindif.2003.08.003

Chen, P. P. & Rossi, P. D. (2013). Utilizing calibration accuracy information with adolescents to improve academic learning and performance. In H. Bembenutty, T. Cleary & A. Kitsantas (Eds.), *Applications of self-regulated learning across diverse disciplines: A tribute to Barry J. Zimmerman* (pp.263-297). Greenwich, CT: Information Age. https://www.infoagepub.com/products/Applications-of-Self-Regulated-Learning-across-Diverse-Disciplines

Chen, P. & Zimmerman, B. (2007). A cross-national comparison study on the accuracy of self-efficacy beliefs of middle-school mathematics students. *The Journal of Experimental Education*, 75(3), 221-244. https://doi.org/10.3200/JEXE.75.3.221-244

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum. https://doi.org/10.4324/9780203771587

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. https://doi.org/10.1016/j.learninstruc.2012.06.001

Dunlosky, J. & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, 22(4), 271-280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Efklides, A. (2014). How does metacognition contribute to the regulation of learning? An integrative approach. *Psihologijske Teme [Psychological Topics]*, 23(1), 1-30. https://psycnet.apa.org/record/2014-25618-001

Fogarty, G. J. & Else, D. (2005). Performance calibration in sport: Implications for self-confidence and metacognitive biases. *International Journal of Sport and Exercise Psychology,* 3(1), 41-57. https://doi.org/10.1080/1612197X.2005.9671757

Follmer, D. J., Patchan, M. & Spitznogle, R. (2022). Supporting college learners' study time calibration: Relations to course achievement and self-regulated learning skills. *Journal of College Reading and Learning,* 52(2), 75-96. https://doi.org/10.1080/10790195.2022.2033646

Gasser, M. & Tan, R. (2005). Performance estimates and confidence calibration for a perceptual-motor task. *North American Journal of Psychology*, 7(3)*,* 457-468. link.gale.com/apps/doc/A159922654/AONE?u=anon~54861d27&sid=googleScholar&xid=513cff3c

Gigerenzer, G., Hoffrage, U. & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98(4)*,* 506-528. https://doi.org/10.1037/0033-295x.98.4.506

Goudas, M., Kolovelonis, A. & Dermitzaki, I. (2013). Implementation of self-regulation interventions in physical education and sports contexts. In H. Bembenutty, T. Cleary & A. Kitsantas (Eds.), *Applications of self-regulated learning across diverse disciplines: A tribute to Barry J. Zimmerman* (pp. 383-415). Greenwich, CT: Information Age. https://www.infoagepub.com/products/Applications-of-Self-Regulated-Learning-across-Diverse-Disciplines

Grabe, M. & Holfeld, B. (2014). Estimating the degree of failed understanding: A possible role for online technology. *Journal of Computer Assisted Learning,* 30(2)*,* 173-186. https://doi.org/10.1111/jcal.12038

Griffin, T. D., Wiley, J. & Salas, C. R. (2013). Supporting effective self-regulated learning: The critical role of monitoring. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 19-34). New York: Springer. https://doi.org/10.1007/978-1-4419-5546-3_2

Gutierrez de Blume, A. P. (2021). Calibrating calibration: A meta-analysis of learning strategy instruction interventions to improve metacognitive monitoring accuracy. *Journal of Educational Psychology.* Online first. https://doi.org/10.1037/edu0000674

Gutierrez, A. P. & Schraw, G. (2015). Effects of strategy training and incentives on students' performance, confidence, and calibration. *The Journal of Experimental Education,* 83(3), 386-404. https://doi.org/10.1080/00220973.2014.907230

Hacker, D. J., Bol, L. & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning,* 3, 101-121. https://doi.org/10.1007/s11409-008-9021-5

Hacker, D. J., Bol, L., Horgan, D. D. & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, 92(1), 160-170. https://doi.org/10.1037/0022-0663.92.1.160

Hadwin, A. F. & Webster, E. A. (2013). Calibration in goal setting: Examining the nature of judgments of confidence. *Learning and Instruction*, 24, 37-47. https://doi.org/10.1016/j.learninstruc.2012.10.001

Händel, M., de Bruin, A. B. H. & Dresel, M. (2020). Individual differences in local and global metacognitive judgments. *Metacognition and Learning*, 15, 51-75. https://doi.org/10.1007/s11409-020-09220-0

Kitsantas, A. & Zimmerman, B. (2009). College students' homework and academic achievement: The mediating role of self-regulatory beliefs. *Metacognition and Learning*, 4, 97-110. https://doi.org/10.1007/s11409-008-9028-y

Kolovelonis, A. (2019a). Greek physical education students' calibration accuracy in sport and knowledge tasks – a comparison. *International Sports Studies,* 41(1), 16-28. https://doi.org/10.30819/iss.41-1.03

Kolovelonis, A. (2019b). Relating students' participation in sport out of school and performance calibration in physical education. *Issues in Educational Research,* 29(3)*,* 774-789. http://www.iier.org.au/iier29/kolovelonis.pdf

Kolovelonis, A. & Dimitriou, E. (2018). Exploring performance calibration in relation to better or worse than average effect in physical education. *Europe's Journal of Psychology,* 14(3), 665-679. https://doi.org/10.5964/ejop.v14i3.1599

Kolovelonis, A. & Goudas, M. (2012). Students' recording accuracy in the reciprocal and the self-check teaching styles in physical education. *Educational Research and Evaluation,* 18(8), 733-747. https://doi.org/10.1080/13803611.2012.724938

Kolovelonis, A. & Goudas, M. (2018). The relation of physical self-perceptions of competence, goal orientation, and optimism with students' performance calibration in physical education. *Learning and Individual Differences,* 61, 77-86. https://doi.org/10.1016/j.lindif.2017.11.013

Kolovelonis, A. & Goudas, M. (2019). Does performance calibration generalize across sport tasks? A multiexperiment study in physical education. *Journal of Sport and Exercise Psychology,* 41(6), 333-344. https://doi.org/10.1123/jsep.2018-0255

Kolovelonis, A., Goudas, M. & Dermitzaki, I. (2012). Students' performance calibration in a basketball dibbling task in elementary physical education. *International Electronic Journal of Elementary Education,* 4(3)*,* 507-517. https://www.iejee.com/index.php/IEJEE/article/view/193

Kolovelonis, A., Goudas, M., Dermitzaki, I. & Kitsantas, A. (2013). Self-regulated learning and performance calibration among elementary physical education students. *European Journal of Psychology of Education,* 28*,* 685-701. https://doi.org/10.1007/s10212-012-0135-4

Kolovelonis, A., Goudas, M. & Samara, E. (2022). The effects of a self-regulated learning teaching unit on students' performance calibration, goal attainment, and attributions in physical education. *The Journal of Experimental Education,* 90(1)*,* 112-129. https://doi.org/10.1080/00220973.2020.1724852

Letkiewicz, J., Lim, H., Heckman, S., Bartholomae, S., Fox, J. J. & Montalto, C. P. (2014). The path to graduation: Factors predicting on-time graduation rates. *Journal of College Student Retention: Research, Theory & Practice*, 16(3), 351-371. https://doi.org/10.2190/CS.16.3.c

McGraw, A. P., Mellers, B. A. & Ritov, I. (2004). The affective costs of overconfidence. *Journal of Behavioral Decision Making,* 17(4), 281-295. https://doi.org/10.1002/bdm.472

Maki, R. H., Shields, M., Wheeler, A. E. & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology,* 97*,* 723-731. https://doi.org/10.1037/0022-0663.97.4.723

Nietfeld, J. L., Cao, L. & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education,* 74(1), 7-28. https://www.jstor.org/stable/20157410

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning,* 1*,* article 159. https://doi.org/10.1007/s10409-006-9595-6

Ots, A. (2013). Third graders' performance predictions: calibration deflections and academic success. *European Journal of Psychology of Education*, 28, 223-237. https://doi.org/10.1007/s10212-012-0111-z

O'Neill, L. D., Wallstedt, B., Eika, B. & Hartvigsen, J. (2011). Factors associated with dropout in medical education: A literature review. *Medical Education,* 45(5)*,* 440-454. https://doi.org/10.1111/j.1365-2923.2010.03898.x

Osterhage, J. L., Usher, E. L., Douin, T. A. & Bailey, W. M. (2019). Opportunities for self-evaluation increase student calibration in an introductory biology course. *CBE-Life Sciences Education*, 18(2), ar16, 1-10. https://doi.org/10.1187/cbe.18-10-0202

Pieschl, S. (2009). Metacognitive calibration - an extended conceptualization and potential applications. *Metacognition and Learning,* 4, 3-31. https://doi.org/10.1007/s11409-008-9030-4

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Self-regulation: Theory, research, and applications* (pp. 451-502). San Diego, CA: Academic Press. https://doi.org/10.1016/B978-012109890-2/50043-3

Roelle, J., Schmidt, E. M., Buchau, A. & Berthold, K. (2017). Effects of informing learners about the dangers of making overconfident judgments of learning. *Journal of Educational Psychology,* 109(1)*,* 99-117. https://doi.org/10.1037/edu0000132

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning,* 4, 33-45. https://doi.org/10.1007/s11409-008-9031-3

Schunk, D. H. & Pajares, F. (2004). Self-efficacy in education revisited: Empirical and applied evidence. In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited, Vol. 4: Research on sociocultural influences on motivation and learning* (pp. 115-138). Greenwich, CT: Information Age. https://www.infoagepub.com/products/Big-Theories-Revisited

Singer, L. M. & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education*, 85(1), 155-172. https://doi.org/10.1080/00220973.2016.1143794

Stankov, L. & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, 21(6)*,* 971-986. https://doi.org/10.1016/S0191-8869(96)00130-4

Valdez, A. J. (2013). Student metacognitive monitoring: Predicting test achievement from judgment accuracy. *International Journal of Higher Education,* 2(2)*,* 141-146, https://doi.org/10.5430/ijhe.v2n2p141

Young, A. & Fry, J. D. (2008). Metacognitive awareness and academic achievement in college students. *Journal of the Scholarship of Teaching and Learning*, 8(2), 1-10. https://files.eric.ed.gov/fulltext/EJ854832.pdf

Yue, H. & Fu, X. (2017). Rethinking graduation and time to degree: A fresh perspective. *Research in Higher Education*, 58, 184-213. https://doi.org/10.1007/s11162-016-9420-4

Zimmerman, B. (2000). Attaining self-regulation: A social-cognitive perspective. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13-39). San Diego, CA: Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

Zimmerman, B., Moylan, A., Hudesman, J., White, N. & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, 53(1), 108-127. http://mathedseminar.pbworks.com/w/file/fetch/94722140/enhancing_self_reflection.pdf

Athanasios Kolovelonis *PhD* (corresponding author) is member of the teaching staff and researcher in the Department of Physical Education & Sport Sciences, University of Thessaly. He has published three books, four book chapters and more than 40 papers. His research focuses on life skills, motivation, self-regulated learning and cognitive development through physical activity.
Email: akolov@pe.uth.gr

Marios Goudas *PhD* is Professor of Psychology of Physical Education in the Department of Physical Education & Sport Sciences, University of Thessaly. He has published seven books and more than 120 papers and book chapters. His research focuses on the development and assessment of life-skills and self-regulatory skills programs in physical education and sport.
Email: mgoudas@uth.gr

Please cite as: Kolovelonis, A. & Goudas, M. (2022). Physical education students' calibration accuracy and academic achievement: A longitudinal study. *Issues in Educational Research*, 32(2), 575-592. http://www.iier.org.au/iier32/kolovelonis.pdf