International Journal of Science and Mathematics Education (2019) 17:1519–1540
https://doi.org/10.1007/s10763-018-09941-4

Sixth Grade Students' Performance, Misconceptions, and Confidence When Judging the Reasonableness of Computational Results

Der-Ching Yang[1] · Iwan Andi J. Sianturi[1]

Received: 21 May 2018 / Accepted: 6 November 2018 / Published online: 3 December 2018

Abstract

Judging the reasonableness of computational results is pivotal for students to understand mathematical concepts. This domain is the most sensitive to the presence of misconceptions in mathematics. Confidence ratings can serve as a measure of the strength of students' conceptual understanding. This study investigated the performance, misconceptions, and confidence ratings of 942 Hong Kong sixth grade students when they were asked to judge the reasonableness of computational results. The results showed that the students performed unsatisfactorily at judging the reasonableness, with an average score of 3.45 (out of 8). Slightly more than half of the students (53.72%) selected the correct computational results, but more than 60% of those students could not judge the reasonableness of the computational results (49.71% had misconceptions and 11.24% simply guessed the answers). In addition, only 20.82% and 18.23% of the students could apply number-sense- and rule-based methods to judge the reasonableness, respectively. Moreover, only 5.73% of the students showed high performance with a high confidence rating, 3.18% exhibited low performance with a low confidence rating, and 35.46% of them showed low performance with a high confidence rating. Furthermore, this study discusses students' misconceptions, the implications of the study, and suggestions for future research.

Keywords Confidence · Hong Kong · Reasonableness · Misconceptions · Performance

✉ Iwan Andi J. Sianturi
iwansianturi75@gmail.com

Der-Ching Yang
dcyang@mail.ncyu.edu.tw

[1] Graduate Institute of Mathematics and Science Education, National Chiayi University, 85,Wen Lung, Ming-Hsiung, 621 Chiayi, Taiwan

Introduction

In the current century, students' understanding of mathematical concepts cannot be assessed only by the accuracy of the answer they provide. For the students to develop effective conceptual understanding in mathematics, they should be able to judge the reasonableness of computational results (Bragg & Herbert, 2017; Yang, 2017). This ability is regarded as "the glue that holds everything together, the lodestar that guides learning" (Kilpatrick, Swafford, & Findell, 2001, p. 129). Through the judgment of the reasonableness of computational results, the student can navigate through various facts, procedures, concepts, and solution methods, thereby determining that they fit together in a way that clarifies the concept (Kilpatrick et al., 2001). Therefore, the ability to "judge the reasonableness of computational results" is internationally emphasized (National Council of Teachers of Mathematics [NCTM], 2000, p. 32), and its importance has been highlighted in studies worldwide (e.g. Alajmi & Reys, 2007; McIntosh, Reys, Reys, Bana, & Farrel, 1997; Reys & Noda, 1994; Yang, 2017). In particular, the capability to judge the reasonableness of computational results is considered a pivotal characteristic of number sense (Alajmi & Reys, 2007; McIntosh et al., 1997; NCTM, 2000; Verschaffel, Greer, & De Corte, 2007; Yang, Li, & Lin, 2008). Interestingly, number sense is the most important mathematical domain in twenty-first-century K-12 education (Devlin, 2017; NCTM, 2000, 2017).

In this study, recognition of a reasonable explanation of a computational result refers to the likelihood of the obtained answer (provided) being compatible and satisfactory explanation for a computational result (Alajmi & Reys, 2010; Yang, 2017). The ability to judge the reasonableness has not been widely highlighted or prioritized in the teaching and learning of mathematics (Alajmi & Reys, 2007; Reys & Noda, 1994). In mathematics, judging the reasonableness is highly sensitive to misconceptions (Van Dooren, De Bock, Depaepe, Janssens, & Verschaffel, 2003). A misconception can persist over a long period and become entrenched (Eryilmaz, 2002; McNeil & Alibali, 2005). Therefore, misconceptions should be diagnosed earlier to develop compatible teaching strategies and instructions for mathematics to help students develop better conceptual knowledge. Previous studies have demonstrated a positive linear relationship between students' confidence level and mathematics performance (Liu & Meng, 2010; Stankov, Morony, & Lee, 2014). In addition, the strength of students' conceptual understanding can be examined on the basis of their confidence ratings (Caleon & Subramaniam, 2010; Cetin-Dindar & Geban, 2011; Pesman & Eryilmaz, 2010). Based on these points, the study aims to investigate and analyze students' performance, misconceptions, and confidence level when judging the reasonableness of a computational result. For this study, the following research questions were considered:

1. How do sixth grade students in Hong Kong perform at judging the reasonableness of a computational result?
2. What misconceptions do sixth grade students in Hong Kong encounter when judging the reasonableness of a computational result?
3. How are the performance and confidence levels of sixth grade students in Hong Kong in judging the reasonableness of a computational result distributed?

To the best of our knowledge and from the review of the literature, this study is the first to investigate and analyze the performance, misconceptions, and confidence levels of Hong Kong sixth grade students in judging the reasonableness of a computational result. Hong Kong Chinese students have consistently been top performers in international mathematics assessment results, such as the Trends in International Mathematics and Science Study (TIMSS) 2015 (Mullis, Martin, Foy, & Hooper, 2016) and the Programme for International Student Assessment (PISA) 2015 (Organization for Economic Co-operation & Development [OECD], 2016). However, these studies did not consider the presence of misconceptions and the students' confidence levels when solving mathematical problems, especially when judging the reasonableness of computational results (part of number domain). The findings of the current study can shed light on the screening results of Hong Kong students' performance, misconceptions, and confidence ratings in solving mathematical problems and in judging the reasonableness of the computational results. In addition, the results are expected to contribute to the literature and have implications for teachers, curriculum designers, and education stakeholders in Hong Kong and overseas.

Theoretical Background

Judging the Reasonableness of a Computational Result

Clarke, Clarke, and Sullivan (2012) reported that many teachers emphasize explaining mathematics-related knowledge, but neglect judging reasonableness in teaching and learning mathematics. Nevertheless, previous studies have suggested that students use one of the following criteria to judge the reasonableness of a computational problem (Gagne, 1983; McIntosh & Sparrow, 2004; Reys, 1985): (1) number relationships and the effect of operations and (2) the practicality of the answer. In addition, recognition of the reasonableness of a computational result depends on two models: characterization of mathematical proficiency (Kilpatrick et al., 2001) and the framework for number sense (McIntosh, Reys, & Reys, 1992). Moreover, Alajmi and Reys (2010) pointed out that identifying and recognizing reasonableness require an integration of capabilities and qualities relating to conceptual understanding (number sense and working with operations), strategic competence (work flexibility with numbers), adaptive reasoning (observation and explanation of relationships), procedural fluency (operation with whole numbers, fractions, and decimals), and productive disposition (making connections to real-life situations). Alajmi and Reys (2010) argued that students can proficiently identify and recognize the reasonableness of results if they possess the ability to integrate the abovementioned capabilities and qualities.

Students who are able to judge the reasonableness of a computational result can mentally apply a strategy to solve mathematical problems without using paper and pencil (McIntosh et al., 1992) as well as justify the reasonableness of a computational result (Alajmi & Reys, 2010; Yang, 2017). For example, consider the following computational problem:

(1) Less than 1000
(2) Equal to 1000
(3) Greater than 1000

In solving this question, students were able to state that 249 is about 250, 250 multiplied by 4 is 1000, and 249 is less than 250, so 249 × 4 is less than 1000. This explanation showed that students can judge the reasonableness of a computational result by estimation based on the information provided in the question (McIntosh et al., 1992; Reys & Yang, 1998; Yang & Lin, 2015). Moreover, Yang (2017) highlighted that, in judging the reasonableness of a computational result, students can apply different methods (e.g. number-sense (NS)-based method, rule-based method, and guessing); that study also reported that students had substantial misconceptions when justifying the reasonableness.

Students' Misconception and Confidence Levels

[1]Misconceptions are symptomatic of a faulty line of thinking that causes systematic errors (Green, Piel, & Flowers, 2008; Riccomini, 2005) and indicate that students' ideas are incompatible with currently accepted knowledge and the given related problem (Clement, Brown, & Zietsman, 1989). A misconception can persist over a long period and become entrenched (Eryilmaz, 2002; McNeil & Alibali, 2005). For instance, majority of students and teachers and more than one-tenth of the students majoring in mathematics at the university level have misconceptions when solving computational problems (Merenluoto & Lehtinen, 2002).

Previous studies have shown that a misconception exists when students solve a problem incorrectly with a high level of confidence (certainty) that the answer is correct (Caleon & Subramaniam, 2010; Kaltakci Gurel, Eryilmaz, & McDermott, 2015). In addition, Hiebert (1992) found that many of students' misconceptions arose because they relied on memorizing a procedure with inadequate understanding of the associated concepts as a foundation. However, other studies highlighted that as students' self-confidence level increases, their achievement level increases (Liu & Meng, 2010; Stankov et al., 2014). Shen (2002) investigated the correlation between students' confidence levels and mathematics achievement in TIMSS 1999 for 38 countries; it was observed that students who had high confidence in mathematics were more successful in almost all countries.

The previous studies have identified students' confidence ratings as a measure for the degree of certainty of students' responses. These ratings serve as a predictor regarding the strength of students' conceptual understanding (Caleon & Subramaniam, 2010; Cetin-Dindar & Geban, 2011; Pesman & Eryilmaz, 2010). The following observations have been presented regarding confidence ratings: (1) a correct or wrong answer with a low confidence level indicates lack of knowledge (Caleon & Subramaniam, 2010; Clement et al., 1989; Stankov & Crawford, 1997); (2) a correct answer with a high confidence level indicates thorough understanding of the related

[1] For more details regarding misconceptions on computational results and number sense, see our previous study (Yang, 2017; Yang & Lin, 2015; Yang & Wu, 2010). Judging the reasonableness is a component of number sense.

concept (Caleon & Subramaniam, 2010); (3) a wrong answer with a high confidence level indicates the presence of a misconception (Caleon & Subramaniam, 2010); and (4) a low confidence rating with high performance (high scores) can probably be attributed to students' personality style (Yang, 2017) or lack of knowledge leading to the need for guessing (Pesman & Eryilmaz, 2010).

Methods

Study Participants

A total of 942 sixth grade students in Hong Kong willingly participated in the test conducted for this study. The sample included 39 classes from 14 elementary schools located in different districts of Hong Kong. The participants came from various socioeconomic backgrounds.

Study Instrument

We designed an instrument to investigate students' performance, misconceptions, and confidence ratings on number sense (Yang, 2017; Yang & Lin, 2015; Yang & Wu, 2010). The instrument involves five components of number sense based on earlier studies. One of the number sense components concentrates on judging the reasonableness of a computational result (Berch, 2005; Markovits & Sowder, 1994; McIntosh et al., 1997; Verschaffel et al., 2007; Yang, 2017; Yang & Lin, 2015; Yang & Wu, 2010). The instrument involves a two-tier test (Yang & Lin, 2015), where the first-tier examines content knowledge and the second examines the reasons supporting the responses provided for the first-tier. In addition, we adapted a third-tier for confidence ratings by using a five-point Likert scale (very confident [5], confident [4], neutral [3], uncertain [2], and very uncertain [1]). The instrument involves eight mathematics-related questions. An example of the instrument can be retrieved from data depository provided at https://osf.io/7n63a/.

For the eight questions, the Cronbach's α value was .823, the difficulty indices were between .421 and .737, and the discrimination powers were between .082 and .448. In this study, the questions and reasons were designed and developed on the basis of related studies (e.g., Berch, 2005; Cheung & Yang, 2018; Markovits & Sowder, 1994; McIntosh et al., 1997; NCTM, 2000; Reys & Yang, 1998; Sowder, 1992; Yang, 2017; Yang et al., 2008; Yang & Lin, 2015; Yang & Wu, 2010). In addition, the reasons for choosing an answer were designed and created by analyzing data collected from thousands of students in the past two decades through paper-and-pencil methods and interviews (e.g., Cheung & Yang, 2018; Reys & Yang, 1998; Yang, 2017; Yang et al., 2008; Yang & Lin, 2015; Yang & Wu, 2010). We developed the second tier of selecting reasons by analyzing students' responses in the data collection as well as their most frequent misconceptions (Yang & Lin, 2015). Consequently, we provided only the reasons relevant and compatible for each answer in the first-tier. The students were required to select the compatible options in the reason tier instead of posting their own reasons, thus allowing students to focus on fewer but more relevant and compatible reasons, which would result in more meaningful results.

To determine whether students could fully understand each question and whether the questions used in the test were appropriate for sixth grade students, 20 students were initially interviewed for content clarity. Furthermore, we revised several questions to rectify their ambiguous contents and make the eight questions sufficiently clear for sixth grade students. To ensure that the questions and answers are representative and not beyond the curriculum scope for sixth grade students, two mathematics researchers and three elementary school teachers were invited to review the questions. They unanimously agreed that the questions were effectively designed, reflected judging the reasonableness of a computational result, and were not beyond the curriculum scope for sixth grade students.

Procedure, Data Collection, and Analysis

The test was administered online and the students used computers to respond to the test questions. The selected Hong Kong elementary schools were equipped with computer labs, and the sixth grade students of those schools were familiar with computer operation. First, all the students were provided with test instructions. To answer each of the eight questions, students were required to choose one of the answer options, one of the reason options for the selected answer, and the confidence ratings within 40 s, 60 s, and 20 s, respectively.

The answers and reasons were subsequently analyzed using SPSS 21.0. For the first test (answers), the answers were classified as correct or incorrect, and the corresponding percentages were calculated. For the second test (reasons), the method based on which the students selected the reasons was classified as NS based, rule based, misconception, and guessing. In addition, a chi-squared test was also used to examine the differences between the methods. Moreover, we analyzed the students' misconceptions and the possible causes of the misconceptions. Furthermore, this study reported the distribution of the students' performance and confidence ratings (certainty) when judging the reasonableness.

A scoring rule (Table 1) was designed based on earlier studies concerning students' performance and confidence ratings (Caleon & Subramaniam, 2010; Cetin-Dindar & Geban, 2011; Pesman & Eryilmaz, 2010; Yang et al., 2008; Yang & Lin, 2015; Yang & Wu, 2010). For treatment of data and analysis, students' responses to the first- and second-tier test (indicating performance) were divided into four groups: (1) high performance: score ≥ 6 (out of 8); (2) high-medium performance: $4.8 \leq$ score < 6; (3) medium-low performance: $3 <$ score < 4.8; (4) low performance: score ≤ 3. For the third-tier test (confidence ratings), the students' responses were divided into three groups: (1) high confidence rating: confidence rate index (CRI) ≥ 3.3 (out of 5); (2) medium confidence: $2.9 <$ CRI < 3.3; and (3) low confidence: CRI ≤ 2.9.

Results

Students' Performance in Judging the Reasonableness of a Computational Result

Table 2 displays the students' performance as percentages of correct answers (first test) and response type (second test) for each of the eight questions. The results were used to examine students' capability of judging the reasonableness of a computational result, and the results of the chi-squared test were used to examine the differences between the methods.

The results revealed that the average score of students was 3.45 (out of 8) for the eight questions, indicating that the students performed poorly at judging the reasonableness of

Table 1 Scoring rules applied to the students' responses (data)

Students' performance at judging the reasonableness of a computational result						
1st stage	Answer options	Correct answer				Wrong answer
		4 points				0
2nd stage	Reason options	NS-based	Rule-based	Misconception	Guessing	
		4	2	1	0	0
	Score given	8	6	5	4	0
3rd stage	CRI in the test (Scored only in a correct answer)					
	CRI	Very confident	Confident	Neutral	Unconfident	Very unconfident
	Score given	5	4	3	2	1

CRI (Confidence Rate Index)

Table 2 Students' performance in judging the reasonableness of a computational result

Question	First-tier test	Second-tier test (%)				Chi square (Significance testing for all methods)
	Correct %	NS-based (N)	Rule-based (R)	Misconception (M)	Guessing (G)	Significance ($\alpha = .05$)
Question 1	*73.78*	*27.81*	19.00	49.57	3.62	M > N, M > R, M > G, N > R, N > G, R > G
Question 2	42.14	13.27	24.20	49.57	12.96	M > N, M > R, M > G, R > N, R > G
Question 3	68.47	21.87	16.88	56.26	4.99	M > N, M > R, M > G, N > R, N > G, R > G
Question 4	46.39	23.99	11.15	47.45	*17.41*	M > N, M > R, M > G, N > R, N > G
Question 5	42.57	18.79	0.00	*66.89*	14.32	M > N, M > R, M > G, N > R, N > G, R > G
Question 6	51.59	13.06	*33.33*	39.82	13.79	M > N, M > R, M > G, R > N, R > G
Question 7	58.39	20.59	33.23	36.09	10.09	M > N, M > R, M > G, R > N, R > G, N > G
Question 8	46.39	27.18	8.07	52.01	12.74	M > N, M > R, M > G, N > R, N > G
Total	53.72	20.82	18.23	49.71	11.24	M > N, M > R, M > G, N > R, N > G, R > G

The italics indicate the highest percentage of responses related to correct responses and the methods to answer the questions (number sense-based, rule-based, misconception, and guessing)

NS-based (N): Number Sense-based, R: Rule-based, M: Misconception, and G: Guessing

M > N indicates a significant difference between the misconception and NS-based method when judging the reasonableness of the computational results computational results. In addition, Table 1 shows that slightly more than half of the students (53.72%) were able to judge the reasonableness of the computational results correctly. More specifically, the percentages of students who answered the eight questions correctly ranged from 42.14 to 73.78%. Moreover, 20.82% and 18.23% of them could apply the NS- and rule-based methods to judge the reasonableness, respectively. This indicates that slightly more than 40% of students understood the questions and were able to justify the answers in meaningful ways. On the other hand, less than 18% of students guessed their answers. This might partly be because of the students' personality style or lack of knowledge (Yang, 2017). Furthermore, the data revealed that a considerable number of students had misconceptions when judging the reasonableness, with an average of 49.71% for the eight questions (the highest proportion).

Regarding the methods adopted by the students in judging reasonableness, the chi-squared test indicated significant differences between (1) misconception and each of the other three methods (NS-based, rule-based, and guessing) for all the questions as well as for the overall test; (2) NS-based and rule-based methods for all the questions as well as for the overall test; (3) NS-based method and guessing for questions 1, 3, 4, 5, 7, and 8 as well as for the overall test; and (4) rule-based method and guessing for questions 1, 2, 3, 5, 6, and 7 as well as for the overall test.

Students' Responses and Misconceptions When Judging Reasonableness

This section briefly presents the proportion and details of students' responses to each question. Table 3 displays the results of the responses to question 1 in the first- and second-tier tests.

> Question 1: Without using paper and pencil, determine which of the following is the most reasonable answer of 249 × 4?

Table 3 Distribution of students' responses to question 1 in the two-tier tests

Answer	Percentage	Reasons	Method
Less than 1000* (73.78%)	26.54%	249 is about 200, 200 × 4 = 800, so 249 × 4 is less than 1000.	M
	18.47%	$\frac{\begin{array}{r}249\\ \times 4\end{array}}{996}$, so 249 × 4 is less than 1000	R
	27.81%#	249 is about 250, 250 multiplied by 4 is 1000, and 249 is less than 250, so 249 × 4 is less than 1000.	N
	0.96%	I am guessing.	G
Equal to 1000 (3.82%)	3.18%	249 is about 250, 250 × 4 = 1000, so 249 × 4 is equal to 1000.	M
	0.64%	I am guessing.	G
Greater than 1000 (21.13%)	19.85%	249 is about 300, 300 × 4 = 1200, so 249 × 4 is greater than 1000.	M
	1.28%	I am guessing.	G
Cannot tell (1.27%)	0.53%	A correct judgment cannot be made without calculating the answers.	R
	0.74%	I am guessing.	G

The difficulty index and discrimination power of this item were .737 and .448, respectively

*Correct answer; # NS-based method

The results revealed that 49.57% of students had several misconceptions on this question. The misconceptions presented in Table 3 indicate that these students did not understand the concept of estimation and hence did not apply it to arrive at a compatible argument for the correct computational result (McIntosh et al., 1992; Reys & Yang, 1998; Yang & Lin, 2015). Furthermore, these students did not show strategic competence (work flexibility with numbers) to judge the reasonableness and were unable to apply their conceptual understanding appropriately (e.g. number sense and working with operations). Some students guessed the answers (3.62%) probably because they did not understand how to manipulate the relationship between numbers and operations and how to use a benchmark or estimation to determine the answer (Yang, 2017).

Table 4 reports the results of the responses to question 2 in the two-tier tests.

> Question 2: Without using paper-and-pencil, find which of the following is the closest to 1? (1) 0.9 + 1; (2) 0.9 × 1; (3) 0.9 + 0.9; (4) 0.9 × 0.9.

The purpose of question 2 was to determine whether the students could judge the reasonableness of the given estimations without using paper and a pencil. The results indicated that 56.26% of the students had misconceptions on this question, which were commonly related to estimation and computation errors. These students evidently could not understand the concept of operations with numbers. Some students were unable to

Table 4 Distribution of students' responses to question 2 in the two-tier tests

Answer	Percentage	Reasons	Method
0.9 + 1 (19.84%)	15.39%	0.9 + 1 = 1.9, this is quite close to 1.	M
	2.65%	The addend is 1.	M
	1.80%	I am guessing.	G
0.9 × 1* (68.47%)	27.92%	The multiplier is 1.	M
	21.87%#	0.9 + 1 and 0.9 + 0.9 are approximately 2, while 0.9 is less than 1, 0.9 × 0.9 < 0.9 × 1, so the product of 0.9 × 1 is the closest to 1.	N
	16.88%	0.9 + 1 = 1.9, 0.9 + 0.9 = 1.8, 0.9 × 0.9 = 0.81, so the product of 0.9 × 1 is quite close to 1.	R
	1.80%	I am guessing.	G
0.9 + 0.9 (7.64%)	5.20%	The sum of 0.9 + 0.9 is larger than 0.9, which is quite close to 1.	M
	1.38%	The sum of 0.9 + 0.9 is 0.99, which is quite close to 1.	M
	1.06%	I am guessing.	G
0.9 × 0.9 (4.04%)	1.70%	The product of multiplication is larger than the multiplicand and the multiplier, so the product of 0.9 × 0.9 is the most likely to be closest to 1.	M
	0.96%	0.9 × 0.9 = 0.99, which is quite close to 1.	M
	1.06%	0.9 is quite close to 1, and the product of two numbers which are quite close to 1 is quite close to 1, so 0.9 × 0.9 is quite close to 1.	M
	0.32%	I am guessing.	G

The difficulty index and discrimination power of this item were .684 and .275, respectively

find the correct answer of multiplication and addition because they may have lacked the ability to consider operations in general when solving problems. Overall, the misconceptions indicated that the students had no capabilities and qualities pertaining to strategic competence (work flexibility with numbers), procedural fluency (operation with whole numbers, fractions, and decimals), and adaptive reasoning (observation and explanation of relationships).

Table 5 illustrates the results of the students' responses to question 3 in the two-tier tests.

> Question 3: The Taipei 101 skyscraper has 101 floors. Which of the following has a height that is approximately equal to the height of the skyscraper?

This question aimed to determine the students' understanding of numbers and operations, estimation, and measurement. The students were asked to make the most appropriate estimation of the height of Taipei 101 skyscraper in another representation. This question generated the highest proportion (66.89%) of students who had misconceptions. From Table 6, most students (19.96%) overestimated the height of the

Table 5 Distribution of students' responses to question 3 in the first two-tier tests

Answer	Percentage	Reasons	Method
4000 m-high mountain (23.57%)	19.96%	A 101-story building is very tall, so 4000 m will most likely be close to the height of the building.	M
	1.27%	A 101 skyscraper is very tall, just like a mountain.	M
	2.34%	I am guessing.	G
A 100 m-high Ferris wheel (18.90%)	10.62%	A 101-story building makes a height just like the Ferris wheel, so it will most likely be close to the height of the Ferris wheel.	M
	3.82%	One floor is approximately 1 m in height, so a 101-floored building will make a height of about 100 m.	M
	1.70%	All of the other heights provided are impossible because they are too large.	M
	2.76%	I am guessing.	G
500 m-high waterfall* (42.57%)	16.99%	500 m is the middle value of all the heights provided, which makes it more probable.	M
	18.79%#	One floor is approximately 5 m in height, so a 101-floored building will make a height of about 500 m.	N
	6.79%	I am guessing.	G
The total height of 101 people stacking together (14.97%)	11.04%	The height of one floor is approximately the height of a person.	M
	1.49%	The height of the skyscraper is about the height of 101 people stacked together.	M
	2.44%	I am guessing.	G

The difficulty index and discrimination power of this item were .427 and .342, respectively

skyscraper without any adaptive reasoning - the skyscraper is very tall just like a mountain (4000 m). This misconception showed that these students were unable to perform productive disposition (i.e. make connections to the real-world situation) (Alajmi & Reys, 2010). The others could not estimate the height of one floor. In addition, some students tried to estimate by comparing the given answers, but with no essential connection to the given problem.

Table 6 displays the results of the responses to question 4 in the two-tier tests.

> Question 4: A coconut palm in a schoolyard is about three floors tall. Which of the following is approximately the same height as the coconut palm?

This question was used to determine students' understanding on numbers and operations and measurement. The data revealed that 47.45% of the students had several misconceptions on this question because they had no complete understanding of numbers, quantity, and the height of the three floors under normal circumstances. The students failed to define the height of the floor in real-life situations. The other misconceptions were related to estimation and computation errors. Students also failed to compare arguments (reasons), such as (1) it is too short for the 3 m coconut palm and too tall for the 9 m or 30 m coconut palm and (2) it is too short for the 3 m or 6 m coconut palm and too tall for the 30 m coconut palm.

Table 7 reports the results of the students' responses to question 5 in the two-tier tests.

> Question 5: Which of the following students' descriptions is the most feasible?
>
> Alice: One piece of tissue paper is approximately 0.1 kg in weight.
> Tom: One piece of eraser is approximately 1 kg in weight.
> Frank: My satchel is approximately 500 g in weight.
> Ruby: The math textbook is approximately 300 g in weight.

This question was used to determine students' understanding of numbers and operations, estimations, and measurement. The students were asked to decide whether the given problem is correct and match with the corresponding reasons. The results revealed that 39.82% of students had several misconceptions on this question. They tended to choose Alice's description because they thought that one piece of tissue paper is very light, thus they chose the lightest estimation (0.1 kg). Overall, the related misconceptions showed that students could not estimate the correct description when it is given in an unusual context; they assumed the answer on the basis of an incompatible comparison. However, Bonotto (2005) found that capability in estimation allows students to compare their own work with prior predictions, and the methods used in estimation fostered a connection between solutions and the reasonableness.

Table 8 reports the results of the responses to question 6 in the two-tier tests.

> Question 6: Please estimate how large the area of the floor of a general classroom probably is.

Table 6 Distribution of students' responses to question 4 in the first two-tier tests

Answer	Percentage	Reasons	Method
3 m (14.55%)	9.66%	One floor is approximately 1 m in height, so three floors will make a height of about 3 m.	M
	0.85%	Because of "3" floors, they will make a height of about 3 m.	M
	0.64%	Because 1 m is very tall, and the coconut palm is not, the answer will be 3 m.	M
	3.40%	I am guessing.	G
9 m* (46.40%)	23.99%#	One floor is approximately twice as tall as my height, so one floor is approximately 3 m in height and three floors will make a height of about 9 m.	N
	11.15%	The teacher taught that one floor is approximately 3 m in height, so three floors will make a height of about 9 m.	R
	4.78%	It's too short for the three-meter or six-meter coconut palm and too tall for the 30-m coconut palm.	M
	6.48%	I am guessing.	G
15 m (23.14%)	14.33%	One floor is approximately 5 m in height, so three floors will make a height of about 15 m.	M
	4.99%	It's too short for the three-meter coconut palm and too tall for the nine-meter or 30-m coconut palm.	M
	3.82%	I am guessing.	G
30 m (15.92%)	2.12%	The coconut palm must be very tall; Other answers are too short for the tall of it.	M
	10.08%	One floor is very tall and it's approximately 10 m in height, so three floors will make a height of about 30 m.	M
	3.72%	I am guessing.	G

The difficulty index and discrimination power of this item were .463 and .368, respectively

This question was aimed to examine whether students could use their prior knowledge to determine the area of a general classroom floor. The results revealed that 36.09% of the students had several misconceptions. According to the students' responses, most students (15.18%) argued that the given numbers (except 9 m^2; the chosen answer) were too large. Lack of integration of mathematics and real-life situations might have influenced this issue. In this case, they were not aware of the possible length and width for the chosen area (9 m^2), because this is quite uncommon for classroom size. Some students could not apply their content knowledge to a real-life situation (i.e. they were unable to perform productive disposition).

Table 7 Distribution of students' responses to question 5 in the first two-tier tests

Answer	Percentage	Reasons	Method
Ruby* (51.59%)	24.63%	The teacher taught that before. The math textbook is approximately 300 g in weight.	R
	8.70%	300 g = 0.3 kg and the math textbook is light, so it will be the most feasible answer.	R
	13.06%#	One piece of tissue paper is lighter than 0.1 kg; one piece of eraser is lighter than 1 kg and the satchel is heavier than 500 g, so Ruby's description is correct.	N
	5.20%	I am guessing.	G
Frank (18.04%)	7.54%	My satchel is heavy, so 500 g will be the most feasible answer.	M
	3.61%	500 g = 0.5 kg, so it will be the most feasible answer.	M
	4.88%	One piece of tissue paper is lighter than 0.1 kg; one piece of eraser is lighter than 1 kg and the math textbook is lighter than 300 g, so Frank's description is correct.	M
	2.02%	I am guessing.	G
Tom (12.74%)	5.84%	One piece of eraser is very light, so 1 kg will be the most feasible answer.	M
	3.40%	One piece of tissue paper is lighter than 0.1 kg; the satchel is heavier than 500 g and the math textbook is lighter than 300 g, so Tom's description is correct.	M
	3.50%	I am guessing.	G
Alice (17.63%)	9.45%	One piece of tissue paper is very light, so 0.1 kg will be the most feasible answer.	M
	5.10%	One piece of eraser is lighter than 1 kg; the satchel is heavier than 500 g and the math textbook is lighter than 300 g, so Alice's description is correct.	M
	3.08%	I am guessing.	G

The difficulty index and discrimination power of this item were .516 and .082, respectively

Table 9 reports the results of the responses to question 7 in the two-tier tests.

> Question 7: Which of the following students' descriptions is the most feasible?
>
> John: An elevator can carry about 100 adults.
> Amy: The capacity of a classroom is about 30 m^3.
> Lily: A human drinks about 50 L every day.
> Emma: The julep we drink when we get cold is about 5 cL at a time.

This question was aimed to check students' understanding of numbers and operations, estimation, and measurement. The results showed that 36.09% of the students had several misconceptions on this question. Most of these students (11.04%) chose

Table 8 Distribution of students' responses to question 6 in the first two-tier tests

Answer	Percentage	Reasons	Method
9 square meters (21.87%)	15.18%	The other answers are too large, so only 9 square meters is feasible.	M
	4.14%	The long and wide of the classroom are almost 3 m, 3 × 3 = 9, so 9 square meters is closed.	M
	0.85%	The wide of the general classroom is probably 2 m, and the long is probably 4.5 m, so the area is probably 9 square meters.	M
	1.70%	I am guessing.	G
90 square meters* (46.38%)	27.18%#	The wide of the general classroom is probably 9 m, and the long is probably 10 m, so the area is probably 90 square meters.	N
	1.49%	Because the teacher has taught that the area of floor of the general classroom probably is 90 square meters.	R
	3.70%	The wide of the general classroom is probably 3 m, and the long is probably 30 m, so the area is probably 90 square meters.	M
	6.58%	9 square meters is too small, 900 and 9000 square meters is too large, so 90 square meters is feasible.	R
	7.43%	I am guessing.	G
900 square meters (28.99%)	11.15%	Because the area is almost seen to be 900 square meters.	M
	8.07%	The wide of the general classroom is probably 15 m, and the long is probably 60 m, so the area is probably 900 square meters.	M
	6.69%	9 and 90 square meters are too small, and 9000 square meters is too large, so 900 square meters is feasible.	M
	3.08%	I am guessing.	G
9000 square meter (2.76%)	1.06%	Because the area is almost seen to be 9000 square meters.	M
	0.32%	The classroom is so large that the answer is probably 9000 square meters.	M
	0.85%	The wide of the general classroom is probably 90 m, and the long is probably 100 m, so the area is probably 9000 square meters.	M
	0.53%	I am guessing.	G

The difficulty index and discrimination power of this item were .469 and .378, respectively

Amy's description. According to their responses, they failed to estimate the capacity of a classroom (30 m^3 is too small). The students' misconceptions reported in Table 10 indicated that they did not apply their mathematical knowledge, but only argued on the basis of their logical thinking. In addition, some students applied their relevant knowledge to solve the problem, but failed to compare the given problems to judge the most reasonable answer. Moreover, some students were able to perform a mathematical computation, yet overestimated the quantity of the object (i.e. they lacked real-life experience). For instance, "one bottle is about 10 L and I drink five bottles everyday" – this is an overestimation.

Table 10 reports the results of the responses to question 8 in the two-tier tests.

> Question 8: A pack of candies has 12 candies. Sandy and her five classmates divide two packs of candies among themselves. How many packs does everyone have?

Table 9 Distribution of students' responses to question 7 in the first two-tier tests

Answer	Percentage	Reasons	Method
John (11.57%)	9.98%	The space in an elevator is very large, so it can carry about 100 adults.	M
	0.21%	An elevator can carry 1 ton, which means it can carry many people, so there is no problem to carry 20 adults.	M
	0.32%	30 m^3 is too small, 50 L is too much and 5 cc is too less, so John's statement is the most feasible.	M
	1.06%	I am guessing.	G
Amy (18.79%)	11.04%	A classroom is very large, so it is approximately 300 m^3	M
	5.41%	A classroom is approximately 3 m in wide, 10 m in long and 1 m in height, so the capacity is about 30 m^3	M
	2.34%	I am guessing.	G
Lily (11.25%)	5.73%	We drink much water every day, so 50 L will be the most feasible answer.	M
	2.02%	20 adults are too many, 30 30 m^3 is too small and 5 cc is too less, so Lily's statement is the most feasible.	M
	1.38%	One bottle is about 10 L and I drink 5 bottles every day.	M
	2.12%	I am guessing.	G
Emma* (58.38%)	33.23%	The nurse asked we to drink one scale at a time and one scale is about 5 cL	R
	20.59%#	An elevator can carry about 10 adults, The solidity of a classroom is about 300 m^3, and a human drinks about 2 L every day, so Emma's statement is the most feasible.	N
	4.56%	I am guessing.	G

The difficulty index and discrimination power of this item were .584 and .147, respectively

The results revealed that 56.26% of the students had several misconceptions on this question because they failed to notice and understand the properties (the important features of the problem) and rushed to determine the reasoning. Although they could

Table 10 Distribution of students' responses to question 8 in the two-tier tests

Answer	Percentage	Reasons	Method
$\frac{12}{5}$ packs (14.22%)	11.78%	12 candies bisect to 5 people, so everyone gets $\frac{12}{5}$ candies.	M
	2.44%	I am guessing.	G
4 packs (28.13%)	24.52%	$12 \times 2 \div 6 = 4$, so there are 4 packs.	M
	1.91%	The number of pack must be an integer, so it cannot be a fraction.	M
	1.70%	I am guessing.	G
$\frac{2}{5}$ packs (15.50%)	11.36%	Two packs of 12 candies bisect to 5 people, so everyone gets $\frac{2}{5}$ candies.	M
	4.14%	I am guessing.	G
$\frac{2}{6}$ packs* (42.14%)	24.20%	$12 \times 2 = 24$, $24 \div 6 = 4$ and $\frac{4}{12} = \frac{2}{6}$.	R
	13.27%#	Two packs bisect to 6 people, so everyone get $\frac{2}{6}$ packs.	N
	4.67%	I am guessing.	G

The difficulty index and discrimination power of this item were .421 and .347, respectively

Table 11 Distributions of students across their performance and confidence levels

Categories	High confidence	Medium confidence	Low confidence	Total
High performance	54 (5.73%)	1 (0.11%)	1 (0.11%)	56 (5.94%)
High-medium performance	116 (12.31%)	6 (0.64%)	6 (0.64%)	128 (13.59%)
Medium-low performance	295 (31.32%)	38 (4.03%)	23 (2.44%)	356 (37.79%)
Low performance	334 (35.46%)	38 (4.03%)	30 (3.18%)	402 (42.68%)
Total	799 (84.82%)	83 (8.81%)	60 (6.37%)	942 (100%)

(1) high performance: score ≥ 6 (the highest score was 8); high-medium performance: score between 4.8 and 6; medium-low performance: score between 3 and 4.8; low performance: score ≤ 3. (2) High confidence rating: confidence rate index (CRI) ≥ 3.3 (the highest score was 5); medium confidence: CRI between 2.9 and 3.3; low confidence: CRI ≤ 2.9

perform multiplication and division, they could not apply the concepts to solve the related problems.

Distribution of Students Across Their Performance and Confidence Levels

Table 11 displays the distribution of students across their performance and confidence levels. The average CRI of the students' responses was 4.17 (the highest was 5). This indicates that students were apparently overconfident when judging the reasonableness of the computational results, regardless of whether their responses were correct. The data revealed that 54 students (5.73%) exhibited high performance with high confidence, indicating that these students had a profound understanding and capability to judge the reasonableness of a computational result. In addition, 30 students (3.18%) exhibited low performance with low confidence, which implies that these students may have lacked related knowledge to judge the reasonableness of the computational results. Moreover, 334 students (35.46%) exhibited low performance with high confidence, indicating that these students had severe misconceptions on the related concepts pertaining to computational problems. Furthermore, 38 students (4.03%) showed low performance and medium confidence, which indicates that the students had moderate misconceptions on concepts related to computational problems.

Discussion and Conclusion

Our first research question focused on Hong Kong sixth grade students' performance in judging the reasonableness of computational results. The results indicated that the students showed an unsatisfactory performance in mathematical computational problems. According to their responses, most of them typically could not judge the reasonableness of the computational results, with an average of 60.95% for the eight questions (49.71% had misconceptions and 11.24% guessed the answers). This study also revealed that students could apply different methods to arrive at their responses (20.82% and 18.23% for NS- and rule-based methods, respectively). In addition, the average of the correct percentage was approximately 53.72% for the eight questions, which is in marked contrast with their performance on international mathematics assessments, such as PISA and TIMSS; particularly, in the results of number domain, Hong Kong elementary students performed extremely well in mathematics (Mullis et al., 2016; OECD, 2016).

Due to the absence of studies concerning Hong Kong students' capabilities at judging the reasonableness, the results were compared with those of studies in different countries. They were consistent with those of Yang (2005) and Yang et al. (2008), which have reported that Taiwanese elementary students consistently showed a low performance in judging the reasonableness of computational results, although they performed satisfactorily in both national and international mathematics tests (e.g. PISA and TIMSS). Those studies claimed that the results might be partially influenced by less of exposure of students to judging the reasonableness and the design of Taiwanese mathematics textbooks, which emphasize rule-based problem-solving. Moreover, Alajmi and Reys (2010) found that Kuwaiti eight-grade students performed poorly in recognizing reasonableness. Therefore, many students worldwide apparently encounter difficulties in judging the reasonableness of computational results, which might partially be because the importance and practicality of judging reasonableness are neglected in teaching and learning mathematics (Clarke et al., 2012). In addition, studies have argued that the ability to judge the reasonableness of computational results might be affected by the current philosophy of mathematics teaching that greatly relies on written algorithms and procedures to solve mathematical problems (Menon, 2004; Yang et al., 2008).

Our second research question focused on students' misconceptions when judging the reasonableness of computational results. According to the students' responses, misconceptions were the highest proportion of responses among all types (including NS-based response, rule-based response, and guessing); misconceptions had the highest percentage on the overall test and each of the eight questions. This finding confirms that the domain of judging the reasonableness in mathematics is the most sensitive to the presence of misconceptions (Van Dooren et al., 2003). The misconceptions would hinder students' mathematics learning ability, conceptual understanding, and the advanced mathematics development (Batanero & Sanchez, 2005; Steinle & Stacey, 2003). Therefore, a new treatment and strategy should be designed to prevent students from lagging behind their counterparts in other countries or from continuing to have misconceptions in upper levels of education. However, researchers argued that misconceptions are highly resistant to change and problematic for gaining advanced scientific knowledge (Caleon & Subramaniam, 2010; Cetin-Dindar & Geban, 2011; Pesman & Eryilmaz, 2010). Nonetheless, misconceptions might arise naturally in the mathematics learning process and are partially the consequence of inappropriate teaching style or explanation of related conceptual knowledge (Steinle & Stacey, 2003; Swan, 2001).

The results also clarified that the students who were able to judge the reasonableness of the computational results possessed the capabilities and qualities regarding conceptual understanding (number sense and working with operations), strategic competence (work flexibility with numbers), adaptive reasoning (observation and explanation of relationships), procedural fluency (operation with whole numbers, fractions, and decimals), and productive disposition (making connections to the real world) (Alajmi & Reys, 2010; Kilpatrick et al., 2001; McIntosh & Sparrow, 2004). This is evidenced by the facts that the students used different methods to judge the reasonableness, such as the NS- and rule-based methods. These methods essentially reflect the related capabilities and qualities (Alajmi & Reys, 2007; McIntosh et al., 1997; Reys & Noda, 1994; Menon, 2004; Verschaffel et al., 2007). However, the majority of the students evidently lacked related capabilities and qualities, leading to their misconceptions. For instance, most students overestimated the height of the skyscraper – "the skyscraper is very tall just like a mountain (4000 m)." This response indicated that the students were unable to perform productive disposition and adaptive reasoning. The others argued that (1) the product of multiplication is larger than the multiplicand and the multiplier, so the product of 0.9 × 0.9 is the most likely to be closest to 1, (2) 0.9 is quite close to 1, and the product of two numbers which are quite close to 1 is quite close to 1, so 0.9 × 0.9 is quite close to 1, (3) $0.9 \times 0.9 = 0.99$, which is quite close to 1, (4) the sum of $0.9 + 0.9$ is 0.99, which is quite close to 1. These misconceptions show that the students do not have the capabilities and qualities relating to strategic competence (work flexibility with numbers), adaptive reasoning (observation and explanation of relationships), and procedural fluency (operation with whole numbers, fractions, and decimals). The results demonstrated that students had considerable difficulty with even the simplest fraction ("Two packs of 12 candies are divided among 5 people, so everyone gets $\frac{2}{5}$ candies") and decimals ($0.9 + 0.9$ is 0.99). The students harbored significant misconceptions about fractions.

In addition, students' misconceptions were substantially related to conceptual changes, which are usually well established based on their daily experiences (Vosniadou & Verschaffel, 2004). For instance, "One bottle is about 10L and I drink 5 bottles every day, so a human drinks about 50L every day is the most feasible among the given descriptions." However, the quantity of 10 L in a drinking bottle is contradictory to the daily life situation. This response shows that students could not perform adaptive reasoning (i.e. observation and explanation of relationships) and productive disposition (i.e. making connections to the real-world situation) (Alajmi & Reys, 2010; Kilpatrick et al., 2001; McIntosh & Sparrow, 2004). Such misconceptions are affected by the conflicts between the new information and prior knowledge, which is termed as conceptual change (Vosniadou & Verschaffel, 2004). Conceptual change is often observed in the number domain in mathematics; particularly, in the transition of learning concepts from natural numbers to rational numbers (Merenluoto & Lehtinen, 2002; Vamvakoussi & Vosniadou, 2004), negative numbers (Vlassis, 2004), fractions (Stafylidou & Vosniadou, 2004), and decimal numbers (Stacey & Steinle, 1998). Moreover, misconceptions exist because of students' reliance on memorizing a procedure with inadequate understanding of the associated concepts and the problem's properties (Hiebert, 1992). For instance, when students were asked to find how many candies everyone gets if Sandy and her four classmates divide two packs of candies among themselves, with a pack containing 12 candies, 11.78% (110 students) argued that 12 candies are divided among 5 people, so everyone gets $\frac{12}{5}$ candies. This shows that student typically assumed that 12 should be the numerator and 5 should be the denominator. They neglected that Sandy and her four classmates divided two packs of candies.

Our third research question focused on the distribution of students on the basis of their performance and confidence rates. By using CRI, we identified the degrees of students' misconceptions. For instance, a higher CRI with low performance indicates that students have a strong misconception and are more confident about the accuracy of the misconception. This observation is consistent with the finding of earlier studies that it may be more difficult to revise students' misconceptions (Caleon & Subramaniam, 2010; Cetin-Dindar & Geban, 2011; Eryilmaz, 2002; McNeil & Alibali, 2005; Pesman & Eryilmaz, 2010). In addition, a low CRI with high performance can be discussed from two perspectives: (1) students' personality style (Yang, 2017) and (2) the inclination for guessing the answers (Pesman & Eryilmaz, 2010; Yang, 2017). Merenluoto and Lehtinen (2004) argued that students may be slightly overconfident about the results of mathematical problems because they found the problem familiar from their mathematics textbooks; they have low certainty in more difficult problems because such problems need spontaneous justification that depends on social evidence. They argued that an overestimation of certainty (overconfidence) with a higher performance indicates an illusion of understanding. Nevertheless, Fischbein (1987) indicated that number domains in mathematics have a high intuitive acceptance that encourages students to be self-evident, self-justifiable, or self-explanatory, which results in overconfidence.

In conclusion, this study revealed that the majority of Hong Kong sixth grade students performed poorly at judging the reasonableness, partly due to the existence of misconceptions and overconfidence (which can be attributed to the students' personality style or lack of knowledge). However, some students could judge the reasonableness in more meaningful ways, such as thorough NS- and rule-based methods. Misconceptions arose because the students partially did not understand the concepts that they learned in the school or lack related knowledge, and they had less exposure to explanation and mathematical problems concerning the reasonableness of computational results. Some students applied numerical operations incorrectly and misunderstood the meaning of the operation, thus resulting in misconceptions. In addition, the students might have had difficulties in understanding the problems, which might have been exacerbated by the poor ability to judge the reasonableness of the computational results.

Implications and Future Researches

The unsatisfactory performance of the sixth grade students in judging the reasonableness and small proportion of students who use NS-based method to judge the reasonableness bear a meaning in the development of teaching strategy and learning instruction that significantly address students' exposure to recognizing the reasonableness and applying the NS-based method in solving computational problems. For instance, teachers should attempt to improve students' relevant concepts and abilities (Yang, 2017). In addition, regarding mathematics, previous studies have shown that textbooks affect students' opportunities to learn and their performance in mathematics (Fan, 2013; Tornroos, 2005; Xin, 2007; Yang, 2017) In this case, the educational stakeholders could select or develop textbooks that emphasize more on judging and recognizing reasonableness. In addition, it is pivotal that further studies could examine and analyze mathematics textbooks concerning reasonableness, especially textbooks used in Hong Kong. Moreover, this research justifies the necessity to emphasize recognizing the reasonableness of results in solving mathematical problems when developing the next release of the curriculum or textbooks for students. The curriculum must provide teachers with the strategies and styles for solving problems; not only what to teach but also how to teach in order to provide deep understanding as well as procedural fluency.

Most students had several misconceptions when judging the reasonableness and a considerable number of students were in the range of low performance and high confidence ratings, indicating that they had severe misconceptions on related concepts regarding the computational problems. Therefore, teachers should be informed about prevalent misconceptions in schools and be able to rectify the misconceptions as well as develop further teaching contents and strategy to overcome the emergence of such misconceptions. In addition, in-depth analyses of the misconceptions in judging reasonableness in different mathematical contexts and how teachers perform when judging the reasonableness of a computational result should be investigated. Moreover, the development council of education in Hong Kong should establish an additional course to assist teachers in understanding the students' possible misconceptions and enhancing their teaching skills. This might aid school mathematics teachers in learning how to prevent the emergence of misconceptions (Yang, 2017). Finally, this study only investigated Hong Kong sixth graders' performance, misconceptions, and confidence ratings when judging the reasonableness of computational results. Therefore, the reader should be generalized to different contexts with caution.

Acknowledgments This article was a part of a research project supported by the Ministry of Science and Technology, Taiwan, with grant no. MOST 105-2511-S-415-003-MY3. Any opinions expressed here are those of the authors and do not necessarily reflect the views of the National Science Council, Taiwan (Republic of China). The authors thank Dr. Ka Luen Cheung (The Education University of Hong Kong) and Cheng-Yi Chuang, M.Ed. (National Chiayi University, Taiwan) for their support in handling the data for the study.

References

Alajmi, A., & Reys, R. (2007). Reasonable and reasonableness of answers: Kuwaiti middle school teachers. *Educational Studies in Mathematics, 65*(1), 77–94. https://doi.org/10.1007/s10649-006-9042-4.

Alajmi, A., & Reys, R. (2010). Examining eighth grade Kuwaiti students' recognition and interpretation of reasonable answer. *International Journal of Science and Mathematics Education, 8*(1), 117–139. https://doi.org/10.1007/s10763-009-9165-z.

Batanero, C., & Sanchez, E. (2005). What is the nature of high school students' conception and misconceptions about probability? In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 241–266). New York, NY: Springer.

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities, 38*(4), 333–339. https://doi.org/10.1177/00222194050380040901.

Bonotto, C. (2005). How informal out-of-school mathematics can help students make sense of formal in-school mathematics: The case of multiplying by decimal numbers. *Mathematical Thinking & Learning: An International Journal, 7*(4), 313–344. https://doi.org/10.1207/s15327833mtl07043.

Bragg, L. A., & Herbert, S. (2017). A "true" story about mathematical reasoning made easy. *Australian Primary Mathematics Classroom, 22*(4), 3–6.

Caleon, I., & Subramaniam, R. (2010). 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. https://doi.org/10.1080/09500690902890130.

Cetin-Dindar, A., & Geban, O. (2011). Development of a three-tier test to assess high school students' understanding of acids and bases. *Procedia-Social and Behavioral Sciences, 15*(2011), 600–604. https://doi.org/10.1016/j.sbspro.2011.03.147.

Cheung, K. L., & Yang, D. C. (2018). Examining the differences of Hong Kong and Taiwan students' performance on the number sense three-tier test. *Eurasia Journal of Mathematics, Science and Technology Education, 14*(7), 3329–3345. https://doi.org/10.29333/ejmste/91682.

Clarke, D. M., Clarke, D. J., & Sullivan, P. (2012). Reasoning in the Australian curriculum: Understanding its meaning and using the relevant language. *Australian Primary Mathematics Classroom, 17*(3), 28–32.

Clement, J., Brown, D., & Zietsman, A. (1989). Not all preconceptions are misconceptions: Finding anchoring conceptions for grounding instruction on students' intuitions. *International Journal of Science Education, 11*(5), 554–565.

Devlin, K. (2017). Number sense: The most important mathematical concept in 21st Century K-12 Education. *HUFFPOST.* https://www.huffingtonpost.com/entry/number-sense-the-most-important-mathematical-conceptus58695887e4b068764965c2e0. Accessed 11 April 2018.

Eryilmaz, A. (2002). Effects of conceptual assignments and conceptual change discussions on students' misconceptions and achievement regarding force and motion. *Journal of Research in Science Teaching, 39*(10), 1001–1015. https://doi.org/10.1002/tea.10054.

Fan, L. (2013). Textbook research as scientific research: Towards a common ground on issues and methods of research on mathematics textbooks. *ZDM Mathematics Education, 45*(5), 765–777. https://doi.org/10.1007/s11858-013-0530-6.

Fischbein, E. (1987). *Intuition in science and mathematics: An educational approach*. Dordrecht: Reidel.

Gagne, R. M. (1983). Some issues in the psychology of mathematics instruction. *Journal for Research in Mathematics Education, 14*, 7–18.

Green, M., Piel, J., & Flowers, C. (2008). *Reversing education majors'arithmetic misconceptions with short-term instruction using manipulatives*. Charlotte, NC: Heldref Publications.

Hiebert, J. (1992). Mathematical, cognitive, and instructional analyzes of decimal fractions. In G. Leinhardt, R. Putman, & R. Hattrup (Eds.), *Analysis of arithmetic for mathematics teaching* (pp. 283-322). Hillsdale, NJ: LEA.

Kaltakci Gurel, D., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science and Technology Education, 11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Liu, S., & Meng, L. (2010). Re-examining factor structure of the attitudinal items from TIMSS 2003 in cross-cultural study of mathematics self-concept. *Educational Psychology: An International Journal of Experimental Educational Psychology, 30*, 699–712. https://doi.org/10.1080/01443410.2010.501102.

Markovits, Z., & Sowder, J. T. (1994). Developing number sense: An intervention study in grade 7. *Journal for Research in Mathematics Education, 25*(1), 4–29.

McIntosh, A., Reys, B. J., & Reys, R. E. (1992). A proposed framework for examining basic number sense. *For the Learning of Mathematics, 12*, 2–8.

McIntosh, A., Reys, B. J., Reys, R. E., Bana, J., & Farrel, B. (1997). *Number sense in school mathematics: Student performance in four countries*. Perth, Australia: Edith Cowan University.

McIntosh, A., & Sparrow, L. (2004). *Beyond written computation*. Perth, Australia: Mathematics, Science & Technology Education Centre (MASTEC).

McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development, 76*(4), 883–899. https://doi.org/10.1111/j.1467-8624.2005.00884.x.

Menon, R. (2004). Elementary school children's number sense. *International Journal for Mathematics Teaching and Learning*. Retrieved 12 Aug 2016 from http://www.cimt.plymouth.ac.uk/journal/default.htm.

Merenluoto, K., & Lehtinen, E. (2002). Conceptual change in mathematics: Understanding the real numbers. In M. Limo'n & L. Mason (Eds.), *Reconsidering conceptual change. Issues in theory and practice* (pp. 233–258). Dordrecht, The Netherlands: Kluwer Academic publishers.

Merenluoto, K., & Lehtinen, E. (2004). Number concept and conceptual change: Towards a systemic model of the processes of change. *Learning and Instruction, 14*(5), 519–534. https://doi.org/10.1016/j.learninstruc.2004.06.016.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics.* Retrieved 11 April 2018 from Boston College, TIMSS & PIRLS international study center website: http://timssandpirls.bc.edu/timss2015/international-results/. Accessed 11 April 2018.

National Council of Teachers of Mathematics [NCTM]. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics [NCTM]. (2017). *Instructional programs from prekindergarten through grade 12*. Retrieved 18 December 2017 from http://www.nctm.org/Standards-and-Positions/Principles-and-Standards/Algebra/.

Organization for Economic Co-operation & Development [OECD]. (2016). *PISA 2015 Results in Focus*. Paris, France: OECD.

Pesman, H., & Eryilmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *The Journal of Educational Research, 103*(3), 208–222. https://doi.org/10.1080/00220670903383002.

Reys, R. (1985). Estimation. *Arithmetic Teacher, 32*, 37–41.

Reys, R. E., & Noda, N. (1994). *Computational alternative for the 21th century: Cross cultural perspectives from Japan and the United States*. Reston, VA: NCTM.

Reys, R. E., & Yang, D. C. (1998). Relationship between computational performance and number sense among sixth- and eighth-grade students in Taiwan. *Journal for Research in Mathematics Education, 29*(2), 225–237.

Riccomini, P. J. (2005). Identification and remediation of systematic error patterns in subtraction. *Learning Disability Quarterly, 28*(3), 233–242. https://doi.org/10.2307/1593661.

Shen, C. (2002). Revisiting the relationship between students' achievement and their selfperceptions: A cross-national analysis based on the TIMSS 1999 data. *Assessment in Education, 9*(2), 161–184. https://doi.org/10.1080/0969594022000001913.

Sowder, J. (1992). Estimation and number sense. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Teachers of Mathematics* (pp. 371–389). New York: Macmillan.

Stacey, K., & Steinle, V. (1998). Refining the classification of students' interpretations of decimal notation. *Hiroshima Journal of Mathematics Education, 6*, 49–59.

Stafylidou, S., & Vosniadou, S. (2004). The development of students' understanding of the numerical value of fractions. *Learning and Instruction, 14*(5), 503–518. https://doi.org/10.1016/j.learninstruc.2004.06.015.

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence, 25*(2), 93–109. https://doi.org/10.1016/S0160-2896(97)90047-7.

Stankov, L., Morony, S., & Lee, Y. P. (2014). Confidence: The best non-cognitive predictor of academic achievement? *Educational Psychology, 34*, 9–28. https://doi.org/10.1080/01443410.2013.814194.

Steinle, V., & Stacey, K. (2003). Grade-related trends in the prevalence and persistence of decimal misconceptions. In N. A. Pateman, B. Dougherty, & J. T. Zilliox (Eds.), *Proceedings of the 2003 joint meeting of PME and PMENA* (pp. 259–266). Honolulu, HI: CRDG, College of Education, University of Hawaii.

Swan, M. (2001). Dealing with misconceptions in mathematics. In P. Gates (Ed.), *Issues in mathematics teaching* (pp. 147–165). London, England: Routledge Falmer.

Tornroos, J. (2005). Mathematics textbooks, opportunity to learn and student achievement. *Studies in Educational Evaluation, 31*(4), 315–327. https://doi.org/10.1016/j.stueduc.2005.11.005.

Vamvakoussi, X., & Vosniadou, S. (2004). Understanding the structure of the set of rational numbers: A conceptual change approach. *Learning and Instruction, 14*, 453–467.

Van Dooren, W., De Bock, D., Depaepe, F., Janssens, D., & Verschaffel, L. (2003). The illusion of linearity: Expanding the evidence towards probabilistic reasoning. *Educational Studies in Mathematics, 53*(2), 113–138. https://doi.org/10.1023/A:1025516816886 .

Verschaffel, L., Greer, B., & De Corte, E. (2007). Whole number concepts and operations. In F. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 557–628). Charlotte, NC: Information Age Publishing.

Vlassis, J. (2004). Making sense of the minus sign or becoming flexible in "negativity". *Learning and Instruction, 14*(5), 469–484. https://doi.org/10.1016/j.learninstruc.2004.06.012 .

Vosniadou, S., & Verschaffel, L. (2004). Extending the conceptual change approach to mathematics learning and teaching. *Learning and Instruction, 14*(5), 445–451. https://doi.org/10.1016/j.learninstruc.2004.06.014.

Xin, Y. P. (2007). Word problem solving tasks in textbooks and their relation to students' performance. *Journal of Educational Research, 100*(6), 347–359. https://doi.org/10.3200/JOER.100.6.347-360.

Yang, D. C. (2005). Number sense strategies used by 6th-grade students in Taiwan. *Educational Studies, 31*(3), 317–333.

Yang, D. C. (2017). Performance of fourth graders when judging the reasonableness of a computational result. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-017-9862-y.

Yang, D. C., & Wu, W. R. (2010). The study of number sense realistic activities integrated into third-grade math classes in Taiwan. *The Journal of Educational Research, 103*(6), 379–392. https://doi.org/10.1080/00220670903383010.

Yang, D. C., & Lin, Y. C. (2015). Assessing 10- to 11-year-old children's performance and misconceptions in number sense using a four-tier diagnostic test. *Educational Research, 57*(4), 368–388.

Yang, D. C., Li, M. N., & Lin, C. I. (2008). A study of the performance of 5th graders in number sense and its relationship to achievement in mathematics. *International Journal of Science and Mathematics Education, 6*(4), 789–807.