



## Assessing 10- to 11-year-old children's performance and misconceptions in number sense using a four-tier diagnostic test

Der-Ching Yang & Yung-Chi Lin

To cite this article: Der-Ching Yang & Yung-Chi Lin (2015): Assessing 10- to 11-year-old children's performance and misconceptions in number sense using a four-tier diagnostic test, Educational Research, DOI: [10.1080/00131881.2015.1085235](https://doi.org/10.1080/00131881.2015.1085235)

To link to this article: <http://dx.doi.org/10.1080/00131881.2015.1085235>



Published online: 21 Sep 2015.



Submit your article to this journal [↗](#)



Article views: 21



View related articles [↗](#)



View Crossmark data [↗](#)

## Assessing 10- to 11-year-old children's performance and misconceptions in number sense using a four-tier diagnostic test

Der-Ching Yang<sup>a</sup> and Yung-Chi Lin<sup>b\*</sup>

<sup>a</sup>Graduate Institute of Mathematics and Science Education, National Chiayi University, Chiayi, Taiwan; <sup>b</sup>Graduate Institute of Science Education, National Changhua University of Education, Changhua, Taiwan

(Received 17 October 2014; final version received 3 August 2015)

**Background:** Number sense is a key topic in mathematics education, and the identification of children's misconceptions about number is, therefore, important. Information about students' serious misconceptions can be quite significant for teachers, allowing them to change their teaching plans to help children overcome these misconceptions. In science education, interest in children's alternative conceptions has led to the development of three- and four-tier tests that not only assess children's understandings and misconceptions, but also examine children's confidence in their responses. However, there are few such tests related to mathematical content, especially in studies of number sense.

**Purpose:** The purpose of this study was to investigate children's performance and misconceptions with respect to number sense via a four-tier diagnostic test (Answer Tier → Confidence rating for Answer Tier → Reason Tier → Confidence rating for Reason Tier).

**Design and method:** A total of 195 fifth graders (10–11 years old) from Taiwan participated in this study. The four-tier test was web-based and contained 40 items across five components of number sense.

**Findings:** The results show that (1) students' mean confidence rating for the answer tier was significantly higher than for the reason tier; (2) an average of 68% of students tended to have equal confidence ratings in both answer and reason tiers; (3) students who chose correct answers or reasons had higher mean confidence ratings in most items (36 out of 40) than those who did not; and (4) 16 misconceptions were identified and most of them were at a strong level.

**Conclusion:** The four-tier test was able to identify several misconceptions in both the answer and reason tier and provide information about the confidence levels. By using such information, teachers may be better positioned to understand the nature of learners' misconceptions about number sense and therefore support their pupils' progress in mathematics.

**Keywords:** assessment; computer-based; four-tier diagnostic test; number sense; mathematics

### Introduction

There are many studies in mathematics education which focus on examining students' misconceptions and how these misconceptions may hinder students' mathematics learning and conceptual understanding (Batanero and Sanchez 2005; Clement 1982; Hirsch and O'Donnell 2001). Therefore, the identification and investigation of students'

---

\*Corresponding author. Email: [yclin@cc.ncue.edu.tw](mailto:yclin@cc.ncue.edu.tw)

mathematical misconceptions plays a key role in the field of mathematics education. In addition, number sense plays an important role in our daily life (Dehaene 2001). The need to help children develop number sense has been highlighted internationally by many studies and reports (Berch 2005; Dunphy 2007; Jordan, Raminent, and Watkin 2010; Li and Yang 2010; NCTM 2000; Sood and Jitendra 2007; Verschaffel, Greer, and De Corte 2007). To a certain degree, lack of number sense can often lead to learning difficulties in mathematics (Dyson, Jordan, and Glutting 2013; Jordan, Glutting and Ramineni 2010).

An earlier study of Li and Yang (2010) developed a valid and reliable 'two-tier' test to assess students' performance and misconceptions about number sense. The two-tier test can provide both quantitative data (students' number sense performance and misconceptions) and qualitative data (student thinking in terms of the reasons for their answers and possible causes of their misconceptions). However, this two-tier test system (options + reasons) for number sense does not provide information about the confidence levels with which the students selected their answers and reasons. Chinn and Brewer (1993) have suggested that strong misconceptions (misconceptions but with high confidence levels) are firmly established and difficult to change. Therefore, the two-tier test system can be strengthened with the addition of a certainty of response index.

Using the resulting four-tier test can help us reveal students' levels of confidence in their number sense (i.e. how much confidence students have about their responses). This additional information can inform judgments about the students' level of conceptual understanding. For example, a correct answer with a low confidence rating may imply that the student did not truly understand a concept, but might have only a tenuous grasp. In addition, students' confidence ratings may also provide further information about the level of students' misconceptions. For example, a misconception with a more than 80% confidence rating (mean confidence above four points, where 5 is the highest confidence rating) could be categorised as a strong misconception (Caleon and Subramaniam 2010).

In parallel research carried out in science education, Caleon and Subramaniam (2010) showed that students tend to think problems in the answer tier are easier than those in the reason tier. Accordingly, students' confidence ratings in the answer tier are probably higher than those in the reason tier. In addition, Caleon and Subramaniam also found that for some items, students have higher confidence ratings for their incorrect answers. At the same time, studies have shown a relationship between students' high performance and high confidence ratings (Stankov and Crawford 1997; Zakay and Glicksohn 1992). However, there have been few studies about four-tier tests in mathematics; these studies have been limited to examining science concepts. It remains unclear what the results of applying a four-tier test in mathematics education would be or even how to apply effectively a four-tier test in mathematics education.

In our study, we were curious as to whether students who had higher performance on number sense would also have higher confidence levels. Therefore, we augmented the previously developed two-tier test into a four-tier test. The number sense four-tier test (NS4TT) described in this paper can be used to distinguish differences in confidence level between students' answer tiers and reason tiers as well as to detect students' significant misconceptions. This tool can, therefore, help us to examine the relationships between the answer and reason tiers and their respective confidence levels. Thus, the purposes of this study were to use the four-tier test to detect students' number sense, misconceptions and confidence levels. The research questions are as follows:

- (1) How do students perform on the number sense four-tier diagnostic test?
- (2) What is the relationship between students' confidence ratings in the answer and reason tiers?
- (3) What is the difference in confidence ratings between students who give correct and incorrect responses?
- (4) What are students' significant misconceptions with respect to number sense?

## Background

During the past two decades, there have been many two-tier multiple-choice tests that examine students' alternative conceptions in science education (Tan et al. 2002; Treagust 1988; Tsai and Chou 2002). Several studies have addressed the weaknesses of two-tier tests in science education (Tobin and Capie 1981; Treagust 1988). This has led to the development of three-tier or four-tier tests used not only to assess students' understandings and misconceptions, but also to examine students' confidence in their responses (Caleon and Subramaniam 2010; Cetin-Dindar and Geban 2011; Sreenivasulu and Subramaniam 2013). However, there are few two-tier, three-tier or four-tier tests related to mathematical content, especially in studies of number sense.

Number sense is a key topic in elementary and middle grades mathematics teaching and learning (Berch 2005; Dunphy 2007; Dyson, Jordan, and Glutting 2013; Jordan et al. 2010; Li and Yang 2010; McIntosh, Reys, and Reys 1992; NCTM 2000; Sood and Jitendra 2007; Verschaffel, Greer, and De Corte 2007). Number sense refers to one person's comprehensive understanding of numbers, operations, their relationships and the ability to handle daily life situations that include numbers. A student with number sense can 'decompose numbers naturally, use particular numbers like 100 or 1/2 as referents, use the relationships among arithmetic operations to solve problems, understand the base-ten number system, estimate, make sense of numbers, and recognise the relative and absolute magnitude of numbers' (NCTM 2000, 32). The characteristics of number sense ability, as described above, can be summarised into five components (McIntosh, Reys, and Reys 1992; Sowder 1992; Yang 2005; Yang, Li, and Lin 2008):

- (1) Understanding the meanings of numbers and operations: this includes understanding of the basic meaning of numbers, base-ten system, place-value concepts and number patterns – for example, knowing that if one moves the decimal to the left, the original number will be decreased.
- (2) Recognising the number size: for example, a student can use meaningful ways (not based on standard written methods) to compare two fractions.
- (3) Using multiple representations of numbers and operations: this means that one can use multiple ways to represent a number, such as knowing two different ways to represent  $\frac{3}{4}$ .
- (4) Recognising the relative effect of operations on numbers: this means that students know how the four basic operations affect the results. For example, knowing that multiplication does not always make the result bigger (e.g. the result of  $199 \times .9$  will be less than 199).
- (5) Judging the reasonableness of computational results: this means that one can use estimation strategies or do mental computation to judge the reasonableness of the results, such as knowing  $400 \times .249$  is close to 100 (because .249 is about 1/4)

and deciding that an answer of 9.96 is not reasonable, without written computation.

Because of the importance of number sense, many types of assessments have been used in mathematics education to examine students' number sense performance, the strategies used by students, and students' misconceptions, including paper-and-pencil tests (Markovits and Sowder 1994; Menon 2004), interviews (Markovits and Sowder 1994; Reys and Yang 1998; Yang 2005), and two-tier tests (Li and Yang 2010; Lin, Li and Yang, *forthcoming*). Each type of test has advantages and disadvantages in practical applications (Li and Yang 2010; Lin, Li and Yang, *forthcoming*; Reys 1994). On the one hand, paper-and-pencil tests can collect a lot of data at one time, but do not make it easy to detect, in any depth, students' understandings and misconceptions about number sense (Lin, Li and Yang, *forthcoming*; Reys, 1994). On the other hand, interviews can help researchers to investigate, in depth, students' thinking, misconceptions and possible factors that cause misconceptions (Reys, 1994; Yang 2005). However, they require a lot of time and many interviewers who need to be trained. These factors limit the practicality of interviews for collecting data from large samples (Li and Yang 2010). Thus, due to the small sample sizes, it can be difficult to make a generalisation from interview data.

In order to build on the strengths of paper-and-pencil tests and interviews, a two-tier test for number sense was developed by the authors (Li and Yang 2010; Yang and Li 2013) based on earlier studies in science education (Tan et al. 2002; Treagust 1988; Tsai and Chou 2002). In the number sense two-tier test (NSTTT), the first tier assesses students' responses to number sense-related questions and the second-tier test detects students' reasons for their choices in the first tier (Li and Yang 2010; Yang and Li 2013). Studies of the NSTTT show that the test can be used not only to investigate students' performance on number sense (answer tier), but also to detect students' thinking with regard to their choices and related misconceptions (reason tier) (Li and Yang 2010; Lin, Li and Yang, *forthcoming*; Yang and Li 2013). In addition, a strength of the NSTTT is that feedback and reinforcement are immediate (Li and Yang 2010; Lin, Li and Yang, *forthcoming*). Students can get feedback immediately, and review and revise their incorrect answers and reasons. Moreover, teachers can modify their teaching activities to address their students' misconceptions based on data from the NSTTT (Li and Yang 2010; Lin, Li and Yang, *forthcoming*).

However, the two-tier test does have a weakness, as described in several studies from science education (Caleon and Subramaniam 2010; Treagust 1988). Caleon and Subramaniam (2010) argued that two-tier tests cannot be used 'to segregate mistakes resulting from lack of knowledge from mistakes due to genuine [misconceptions]; and to distinguish correct answers based on guessing from correct answers based on genuine understanding' (314). They further suggested including confidence ratings in the two-tier tests. Confidence ratings concern how one estimates the accuracy of one's own performance (Stankov and Dolph 2000). Confidence ratings can help us understand whether students' correct answers are based only on guessing, or whether students' incorrect answers indicate a lack of some knowledge. Confidence ratings can also act as an index showing the strength of students' understanding and their familiarity with problems (Glenberg et al. 1987). More importantly, the strength or level of students' misconceptions can be identified using confidence ratings (Caleon and Subramaniam 2010). When students have stronger misconceptions (incorrect answers with confidence ratings of more than 80%), this can imply that they have more serious misconceptions. This

information about students' serious misconceptions can be quite significant and useful for teachers. By using the information, teachers may, therefore, be able to change their teaching plans to help students to overcome their misconceptions.

## Method

### *Development of the NS4TT instrument*

The NS4TT instrument for fifth graders (10–11 years old) used in the present study was based on the earlier NSTTT for fifth graders (Yang, Li, and Lin 2008). In the NSTTT, the first-tier test assessed children's responses to number sense-related questions and the second tier examined children's reasons for their choices in the first tier. To examine children's confidence levels for each response, the confidence rating scale was added below the answer and reason tiers of each item in the NSTTT as a four-point Likert scale: 0 = Just guessing; 1 = Unconfident; 2 = Confident; and 3 = Very confident. Therefore, the first tier (answer tier) of the test assessed students' responses to number sense-related questions; the second-tier (confidence rating) test detected students' confidence level for their choices in the first-tier; the third-tier assessed students' reasons for their choices in the first-tier; and the fourth-tier examined students' confidence level for their choices in the third-tier.

There were five components (F1, F2, F3, F4, and F5) in the NS4TT and each component contained eight items. The NS4TT thus included 40 items, in total, for fifth graders. F1 was defined as *Understanding the meanings of numbers and operations*. F2 was *Recognising the number size*. F3 was *Using multiple representations of numbers and operations*. F4 was *Recognising the relative effect of operations on numbers*. F5 was *Judging the reasonableness of computational results*. These five components are generally studied in number sense-related research because these components are usually thought of as significant factors that directly influence whether one can flexibly solve numerical problems (Jordan et al. 2010; McIntosh, Reys, and Reys 1992; NCTM 2000; Yang, Li, and Lin 2008). An example of an item from the NS4TT is shown in Figure 1.

In the reason tier, there were usually three types of reasons for the correct option in the answer tier. The first was a written method. The second was a number sense method which was the correct answer. The third was a common student misconception. However, for the incorrect options in the answer tier, only common misconceptions were provided, but the number sense method was not provided.

It is worth noting that our design in the reason tier was quite different from the usual design of such assessments in science education. That is, respondents could not see all the potential reasons in our design of the reason tier. Different sets of reasons were provided based on the respondents' choices in the answer tier. We made this adjustment to make the four-tier test more suitable for testing mathematics content. More specifically, the main issue of applying a four-tier test in mathematics content is that once an answer has been chosen (even if just guessing the answer), respondents can easily eliminate non-corresponding reasons for their answer choice as viable choices for the reason tier. Many of the reason options are specific calculation results and, thus, it can be easy for students to determine which reason belongs to which answer. For example, in Figure 1, if students choose .1280 as a correct answer, they are less likely to choose the reason '7.84 is about 8; 4.96 is about 5; the sum is about 13. The closest one is 12.8.' That reason is clearly applicable only to the answer 12.8, and thus the student may rule out that option.



Answer Tier	Question 4/ The test contains 40 questions			
	Question 4			
	Given $784+496=1280$ , what is the result of $7.84+4.96$ ?			
	<input type="radio"/> 0.1280	<input type="radio"/> 1.28	<input type="radio"/> 12.8	<input type="radio"/> 1280
	<input type="button" value="Submit"/>			114s Remaining

Confidence Rating for Answer	Question 4/ The test contains 40 questions			
	Question 4			
	Given $784+496=1280$ , what is the result of $7.84+4.96$ ?			
	Your choice is: 12.8			
	Confidence Rating			
	<input type="radio"/> Just Guessing	<input type="radio"/> Unconfident	<input type="radio"/> Confident	<input type="radio"/> Very Confident
	<input type="button" value="Submit"/>			84s Remaining

Reason Tier	Given $784+496=1280$ , what is the result of $7.84+4.96$ ?					
	Your choice is: 0.1280					
	My Reason for the Choice					
	<input type="radio"/> There are four decimals altogether, so the result should be four digits after the decimal point.					
	<input type="radio"/> 7.84 and 4.96 are small, so the sum would be small as well					
	68s Remaining					
	Given $784+496=1280$ , what is the result of $7.84+4.96$ ?					
	Your choice is: 1.28					
	My Reason for the Choice					
	<input type="radio"/> There is one integer in both of 7.84 and 4.96, so the answer would have one integer.					
<input type="radio"/> There are two decimals in both of 7.84 and 4.96, so the answer would have two decimals.						
68s Remaining						
Given $784+496=1280$ , what is the result of $7.84+4.96$ ?						
Your choice is: 12.8						
My Reason for the Choice						
<input type="radio"/> By calculation <table border="1" style="display: inline-table; vertical-align: middle;"> <tr><td>7.84</td></tr> <tr><td>+ 4.96</td></tr> <tr><td>12.80</td></tr> </table>				7.84	+ 4.96	12.80
7.84						
+ 4.96						
12.80						
<input type="radio"/> 7.84 is about 8; 4.96 is about 5; the sum is about 13. The closest one is 12.8						
<input type="radio"/> 7.84 and 4.96 have two decimals, so the sum of them is 12.80 and "0" could be eliminated						
68s Remaining						
Given $784+496=1280$ , what is the result of $7.84+4.96$ ?						
Your choice is: 1280						
My Reason for the Choice						
<input type="radio"/> They are all the same numbers to be added, so the answer is still 1280.						
<input type="radio"/> The decimal point has no influence on the answer, so it is still 1280						
68s Remaining						

Confidence Rating for Reason Tier	Question 4/ The test contains 40 questions			
	Question 4			
	Given $784+496=1280$ , what is the result of $7.84+4.96$ ?			
	Your choice is: 12.8			
	Confidence Rating for your Reason			
	<input type="radio"/> Just Guessing	<input type="radio"/> Unconfident	<input type="radio"/> Confident	<input type="radio"/> Very Confident
	<input type="button" value="Submit"/>			44s Remaining

Figure 1. An example item from the NS4TT (only one set of reasons in the reason tier can be seen by the student; it corresponds to the student's choice in the answer tier).

This situation is unlike that of the four-tier test for science content in which students may plausibly choose the same reason to explain several different answers. Therefore, if we followed the same design from science education in a mathematics education context (creating three or four reasons and showing them all at the same time), the reasons might end up in one-to-one correspondence with the answers (four answers and the corresponding reason for each answer). In this situation, one might expect that students could easily choose a corresponding reason for their choice in the answer tier without truly understanding the reasons. Because of this, we needed to create more than one reason for each option in the answer tier. However, this would result in eight or more options in the reason tier per item (each answer must have at least two different reasons). We did not think that items with more than eight options made a good design. Students might end up spending a lot of time responding to each item and eventually lose their patience for carefully thinking through their choices.

Therefore, in our design, students could only see the set of reasons that corresponded to their choice in the answer tier (usually two or three different reasons) and these corresponding options were based on students' most frequent misconceptions. Students were forced to focus on these compatible options in the reason tier instead of seeing all the possible options. We believed that allowing students to focus on fewer but more relevant reasons would provide more meaningful results.

### ***Participants***

Six classes in a public elementary school with a total of 195 fifth-grade students (10–11 years old) participated in this study at the end of the school year (i.e. the end of June). Among these 195 students, there were 96 girls and 99 boys. The school was located in the urban area of a middle-level city in the south of Taiwan. It is a middle-sized school that has 36 classes in total from first-grade to sixth-grade and all the fifth-grade students in this school participated in this study. These classes were normal groupings (the students were not sorted by their ability; each class consisted of both high and low mathematics performing students) and the majority of these students came from middle socio-economic status backgrounds. Generally, these students were at an average level of achievement in mathematics as compared with the typical attainment of pupils in other schools in this city.

### ***Procedures***

The NS4TT was administrated via an online environment in which students were asked to complete 40 items on computers individually in two sections. The items in the NS4TT were written in Chinese. It was the same as the language of instruction that students were used to in the school. Each section was 40-min long and there was a 10-min break between these two sections. There were no papers and pencils allowed for students during the research process, to test their number sense ability more effectively.

In order to ensure the validity of the instrument, we explained to students how to answer items in the online test system before they started to answer the items. We also clearly stated that no paper-and-pencil calculations were allowed during the test, and that students were encouraged to think more flexibly and efficiently and also to solve the problems from alternative and non-traditional strategies. In other words, we implicitly told students to solve the online test problems by number sense strategies.



### ***Ethics of the study***

To make sure all the participants in this study were not harmed or disadvantaged, the ethics of the study were considered. Before the students participated in the study, we briefly introduced our study to both students and their parents. We let them understand the necessary information and their rights in this study. It included where this study would take place; how long it would take; what they would be asked to do, and so on. In addition, the participants were also informed that the data would only be used for the academic purpose; their personal information would be protected; and they had the right to stop participating at any time during the research. All the students and parents were asked to sign a consent form to make sure that they were aware of the necessary information about their participation. Finally, the ethics of this project were also examined by the Ministry of Science and Technology in Taiwan.

### ***Data analysis***

All the students' responses were collected via computers and then analysed using SPSS 17 to calculate statistical results. In the SPSS analysis, each correct response was coded as 1 and each incorrect response was coded as 0 in the answer and reason tiers. Regarding students' responses in the confidence tiers, each response was coded from 0 to 3 based on the students' responses to the four-point Likert scale. In order to know whether students were more confident on the answer tier, a *t*-test was also conducted to examine the difference between confidence ratings in the answer and reason tiers.

We used a threshold of 35% of responses in a particular answer option and 21% of responses in a particular answer–reason option pair as criteria to evaluate whether students' misconceptions were *significant misconceptions* in this study. Caleon and Subramaniam (2010) defined a significant misconception as a particular incorrect answer or answer–reason pair which was chosen by 10% of students beyond the percentage who would have been expected to choose that answer or answer–reason pair by random choice. Therefore, in this study since usually an item had four options in the answer tier and a total of nine options in the reason tier (the correct answer option had three corresponding reason options; the incorrect answer options had two corresponding reason options), a significant misconception referred to cases where more than 35% of students chose a particular incorrect answer (25% + 10%) or more than 21% of students chose a particular answer–reason pair (11% + 10%).

Regarding the levels of significant misconceptions, if a significant misconception had a low confidence level which was below the median 1.5, the significant misconception was identified as a *spurious* significant misconception. If significant misconceptions had medium (between 1.5 and 1.8) or high confidence (1.8 and above), they were identified as *moderate* or *strong* significant misconceptions. These criteria for deciding confidence levels were adapted from Caleon and Subramaniam (2010) as well.

### ***Reliability and validity***

Cronbach's alpha ( $\alpha$ ) for F1, F2, F3, F4, and F5 are .90, .91, .91, .91, and .90, respectively. Cronbach's alpha for the total scale is .98. This showed that the NS4TT has a high acceptable level ( $\alpha > .90$ ) of internal consistency.

Regarding the content validity, the options (both answer options and response options) in the NS4TT came from earlier number sense studies (e.g. Yang, Li, and Lin

2008). Therefore, these options represented students' most frequent incorrect responses. In addition, during the development process (see Li and Yang 2010; Lin, Li and Yang, forthcoming), the instrument was reviewed by practising teachers, researchers and teacher educators who are experts in studying number sense to check whether it was accurate and relevant to the fifth grade. We also interviewed eight fifth-grade students individually. These students were requested to answer all 40 questions and share their opinions on these questions including wording, content and the reason for their responses.

We used the interview data to advance our item design in the NS4TT, particularly in the reason tier. For example, in answering the following item in the interview 'Without using paper-and-pencil, which number is larger: 7.2 or 7.1987?', many students responded 'There are 4 digits (1987) in the 7.1987, however, there is only one digit (2) in the 7.2. Therefore, 7.1987 is greater than 7.2'. The interview helped us capture one of the student misconceptions. We, therefore, created an option in the reason tier by using this misconception.

## Results

### *Student performance on the number sense four-tier diagnostic test*

Table 1 shows the mean percentages of correctness for both the answer and reason tiers, their corresponding mean confidence ratings, and *t*-test values between confidence ratings in the answer tier and reason tier. It shows that the mean percentage of correctness in the answer tier for each number sense component was between .44 and .53. The total mean percentage was .48.

Table 1 also shows that two components, *recognising the relative effect of operations on numbers* (F4) and *judging the reasonableness of computational results* (F5), are two relatively challenging components for students (.44 and .47 in the answer tier; below the average .48). This result accords with the earlier study of Yang, Li, and Lin (2008) that F4 and F5 were also the two most challenging components for fifth graders.

When looking at the reason tier in Table 1, all the correct percentages decreased and the total mean percentage of correctness reduced to .20. Similar to the result in the answer tier, F4 and F5 were the two most challenging components in the reason tier (.14 and .16 respectively).

Table 1. Mean values of correctness and confidence ratings for five domains of number sense.

	ANS (SD)		ACONF (SD)		RSN (SD)		RCONF (SD)		ACONF-RCONF
F1	.53	(.20)	1.84	(.56)	.26	(.20)	1.62	(.70)	7.83*
F2	.49	(.28)	1.92	(.60)	.19	(.18)	1.61	(.72)	9.80*
F3	.49	(.21)	1.79	(.59)	.27	(.20)	1.55	(.75)	7.96*
F4	.44	(.22)	1.85	(.58)	.14	(.16)	1.59	(.74)	7.95*
F5	.47	(.22)	1.90	(.58)	.16	(.17)	1.68	(.70)	7.22*
Total	.48	(.17)	1.86	(.55)	.20	(.15)	1.61	(.69)	9.26*

Notes: *N* = 195; ANS = Answer tier; ACONF = Answer tier confidence ratings; RSN = Reason tier; RCONF = Reason tier confidence ratings; F1 = Understanding the meanings of numbers and operations; F2 = Recognising the number size; F3 = Using multiple representations of numbers and operations; F4 = Recognising the relative effect of operations on numbers; F5 = Judging the reasonableness of computational results; ACONF-RCONF = *t*-test comparison between A-tier and R-tier confidence ratings.

\**p* < .01.

It should be noted that the correct percentage in the reason tier for *recognising the number size* (F2) fell dramatically from .49 to .19. Thus, the students' responses in the reason tier seemed to reveal that students were actually not as proficient in this component. When further examining why their scores decreased, we found that although many students could choose correct answers in the answer tier, many of them could not use a number sense-based method to support their choices. For example, consider the following item:

In a marathon competition, Doraemon has run  $\frac{4}{10}$  km and Nobida has run  $\frac{3}{5}$  km. Who has run further? (A) Nobida (B) Doraemon (C) no difference (D) cannot compare.

In this item, 68% of students chose the correct answer in the answer tier but only 16% of them decided the magnitude of these two fractions by using one half as a benchmark (categorised as the number sense-based method). Nineteen per cent of the students said  $\frac{3}{5}$  is larger because of the misconception 'the denominator is smaller; the fraction is always larger' and 21% of them solved the problem by finding the common denominator which was coded as a rule-based method in this study.

Although the response from students using the common denominator method was coded as rule-based method in this study, it must be noted that we did not mean that using the common denominator method is necessary to represent the non-number sense method in general. It depends on how it is used. If a student uses the common denominator method by, effectively, doing paper-and-pencil calculations, we tend to think the student uses rule-based method. However, if a student fluently uses the common denominator method (e.g. incorporate with mental computation or some other efficient ways), we might think the student uses a number sense method. However, for the Taiwanese students, when they used common denominator method, their solutions were mostly based on the paper-and-pencil calculations (see e.g. Yang 2005). This was the reason that we coded them as rule-based in this study.

It is important to note that students' performance in the answer tier in the domain of *recognising the number size* (F2) was quite consistent with the earlier study that students performed best in F2 (Yang, Li, and Lin 2008). However, in the present study, when performance in the reason tier was considered, we found that students actually had difficulty in this component. This shows that using the reason tier in the number sense test can help us examine students' performance below the surface. That is, students may still be able to answer the number sense item successfully without using a number sense-based method. For correct answers, the responses in the reason tier can help us distinguish number sense-based methods from non-number sense-based methods.

In addition, items in the component of *understanding the meanings of numbers and operations* (F1) and *using multiple representations of numbers and operations* (F3) may have been relatively easy for the students, since the percentages of correctness were higher than the average (.26, .27 > .20 in Table 1). Again, these results were somewhat different from the results in Yang, Li, and Lin (2008). In Yang et al., F1 and F3 were relatively difficult components for students, with percentages of correctness lower than the average.

### ***Higher confidence ratings for the answer tier than the reason tier***

Table 1 also reports the results of students' confidence ratings. The students' confidence ratings ranged from 1.79 to 1.92 with a total mean rating of 1.86 for the answer tier while the confidence ratings for the reason tier were from 1.55 to 1.68 with an average

rating of 1.61. Both ratings for the answer tier and reason tier are higher than the median (1.5) which implies that students are generally confident in their choices in both tiers (in our Likert scale, 1 = Unconfident and 2 = Confident; 1.86 and 1.61 are closer to confident than unconfident). The *t*-test results show that there were significant differences between the answer tier confidence ratings and the reason tier confidence ratings for each number sense component and for the total scale. This indicates that the students had significantly higher confidence ratings for the answer tier than for the reason tier.

The data in Table 1 also showed that the students had lower confidence ratings on the number sense component *using multiple representations of numbers and operations* (F3) in both the answer tier (1.79) and the reason tier (1.55) compared with the other components.

**Relationship between students’ confidence ratings in answer and reason tiers**

Table 2 shows the percentage of students who chose the same confidence ratings in both the answer and reason tiers, and the percentages of students who chose higher or lower confidence ratings in the answer tier than in the reason tier. About 68% of students gave equal confidence ratings in both tiers for each number sense component and whole number sense; about 24% of students had higher confidence ratings in the answer tier than in the reason tier for each number sense component and whole number sense; and about 8% of students thought they were more confident in the reason tier than in the answer tier for each number sense component and whole number sense.

**Difference in confidence ratings between correct and incorrect responses**

Table 3 reports the difference in confidence ratings between students who gave correct responses and incorrect responses in the answer and reason tiers. The results show that students were able to recognise what they knew and what they did not know. The delta values of confidence ratings for the students who gave correct responses and those who gave incorrect responses were all positive in both tiers across the five components and whole number sense. Particularly, in the reason tier, the delta values were higher than the values in the answer tier.

Examining each test item, we found that there were four test items in the answer tier showing negative delta values as shown in Table 4. This implies that the students had a higher level of misconceptions on these items. They probably believed that these problems were easy but were not aware their thinking on these problems was incorrect.

Table 2. Percentages of students who chose equal, higher and lower confidence ratings in relation to both tiers.

	ACONF = RCONF (%)	ACONF > RCONF (%)	ACONF < RCONF (%)
F1	68.53	22.50	8.97
F2	65.83	26.99	7.12
F3	66.86	24.29	8.85
F4	68.01	24.29	7.69
F5	69.04	22.18	8.78
Total	67.66	24.05	8.28

Notes: *N* = 195; ACONF = Answer tier confidence ratings; RCONF = Reason tier confidence ratings.

Table 3. The difference in confidence ratings between students who gave correct and incorrect responses.

	C-ACONF	IN-ACONF	$\Delta$ ANS	C-RCONF	IN-RCONF	$\Delta$ RSN
F1	2.00	1.67	.59	2.16	1.43	1.04
F2	2.12	1.74	.63	2.18	1.48	.97
F3	1.94	1.65	.49	2.16	1.32	1.12
F4	2.01	1.73	.48	2.24	1.49	1.01
F5	2.02	1.81	.36	2.10	1.61	.70
Total	2.02	1.72	.55	2.17	1.47	1.01

Notes: C-ACONF = Mean confidence for students who gave correct answers; IN-ACONF = Mean confidence for students who gave incorrect answers;  $\Delta$ ANS =  $\frac{(C-ACONF)-(IN-ACONF)}{SD \text{ of all confidence rating for the tier}}$ ; C-RCONF = Mean confidence for students who gave correct reasons; IN-RCONF = Mean confidence for students who gave incorrect reasons;  $\Delta$ RSN =  $\frac{(C-RCONF)-(IN-RCONF)}{SD \text{ of all confidence rating for the tier}}$ .

Table 4. Students' mean incorrect answer confidence ratings greater than mean correct answer confidence ratings.

ID	Problems	ANS	ACONF	C-ACONF	IN-ACONF	$\Delta$ ANS
F1-5	.71008 = $7 \times .1 + \square \times .001 + 8 \times .00001$ , $\square = ?$ (A) 0 (B) 1 (C) 10* (D) 100	.53	1.73	1.62	1.86	-.26
F4-2	Given $784 + 496 = 1280$ , what is the answer of $7.84 + 4.96$ ? (A) .1280 (B) 1.28 (C) 12.8* (D) 1280	.55	2.10	2.09	2.11	-.03
F4-3	Regarding the calculation of $1234 \div 5 \times 6$ , which of the following is correct? (A) $1234 \div 5 \times 6 = 1234 \div (5 \times 6)$ (B) $1234 \div 5 \times 6 = 1234 \times 6 \div 5$ * (C) $1234 \div 5 \times 6 = 1234 \div 6 \times 5$ (D) $1234 \div 5 \times 6 = 5 \times 6 \div 1234$	.28	1.78	1.70	1.81	-.13
F5-8	What could be the general classroom floor area? (A) 9 m <sup>2</sup> (B) 90 m <sup>2</sup> * (C) 900 m <sup>2</sup> (D) 9000 m <sup>2</sup>	.39	1.78	1.78	1.78	-.01

Notes: \* indicates the correct answer; ANS = Answer tier; ACONF = Answer tier confidence ratings; C-ACONF = Mean confidence for students who gave correct answers; IN-ACONF = Mean confidence for students who gave incorrect answers;  $\Delta$ ANS =  $\frac{(C-ACONF)-(IN-ACONF)}{SD \text{ of all confidence rating for the tier}}$ ; F1-5 = Item#5 in the *understanding the meanings of numbers and operations* (F1); The items were written in Chinese, but for the convenience, they are translated into English in this paper.

Table 4 shows that item F1-5 had the most negative delta value (-.26). This item is about place value in decimals. Students may easily have been distracted into putting 1 in the square because '1' is the number after '7' in the decimal .71008. However, these students did not recognise that the '1' denoted one hundredth instead of one thousandth. It is also worth mentioning that even though the value of C-ACONF was relatively low (1.62 in Table 4, below the average value of 2.02 in Table 3), the percentage of correctness for this item was high (.53 in Table 4). This seems to indicate that some students were not so sure about their correct choice. More importantly, many students believed their incorrect answers for this item were correct.

Table 5. Students' significant misconceptions in the answer tier.

ID	Problems	ANS	N	MIC (%)	ACONF
F1-6	12345674 is an eight-digit number. Considering the place value, how many times greater is the value represented by the 4 on the left than the value represented by the 4 on the right (A) 1 (B) $\frac{1}{10,000}$ (C) 1000 (D) 10,000*	(C)	69	35.4	1.87
F4-7	How many digits will the product have when multiplying a two-digit number by a two-digit number? (A) must be two-digit numbers (B) must be three-digit numbers (C) must be four-digit numbers (D) could be three- or four-digit numbers*	(C)	70	35.9	1.99
F4-8	Without doing calculations, which one is larger? (A) $\square + 819 = 2006$ , (B) $\square - 819 = 2006$ , (C) $\square \times 819 = 2006$ (D) $\square \div 819 = 2006$ ?	(C)	85	43.6	1.80
F5-8	What could be the area of a typical classroom floor? (A) 9 m <sup>2</sup> (B) 90 m <sup>2</sup> * (C) 900 m <sup>2</sup> (D) 9000 m <sup>2</sup>	(C)	76	39.0	1.70

Notes: \* indicates the correct answer; MIC = Misconception; Significant misconceptions = Incorrect answers % > 35%; Strong significant misconceptions = Significant misconceptions with the confidence ratings > 1.8; Moderate significant misconceptions = Significant misconceptions with the confidence ratings between 1.5 and 1.8; ANS column represents the wrong answer given by the students with the misconception.



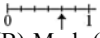
Students' significant misconceptions

Table 5 reports the students' significant misconceptions in the answer tier (misconception % > 35% was defined as a significant misconception). F1-6 and F4-7 are related to the concept of place value; F4-8 is related to understanding the effect of operations, and F5-8 is related to students' number sense of real-life situations. Within these four items, F1-6 and F4-7 showed strong significant misconceptions (ACONF > 1.8), and F4-8 and F5-8 showed moderate significant misconceptions by their confidence ratings (1.5 < ACONF ≤ 1.8). In addition, there were no significant misconceptions identified in the F2 and F3 domains.

Table 6 shows students' significant misconceptions in the reason tier. For many items, students chose the correct answers but could not justify their answers with a number sense-based method. The most common significant misconception for those who could choose the correct answer was the rule-based approach. That is, even though some problems had asked students not to use calculations, students still heavily relied on the calculation when seeking a justification for their answers.

In addition, Table 6 also shows some common misconceptions for students. For example, in item F2-2, around one fifth of the students chose 'When the denominator is smaller, the fraction is larger,' and in item F4-7, around one fifth of the students chose 'Because multiplication always makes the result bigger'. Students also had difficulties in understanding the visual representations. In items F3-1 and F3-5, around one fourth of the students chose incorrect reasons for their answers. This indicated that these students did not have a conceptual understanding of the number lines/visual representations used in these items. In item F5-1, the misconception was about choosing an inappropriate benchmark for estimation.

Table 6. Students' significant misconceptions in the reason tier.

ID	Problems	ANS	Reason	N	MIC (%)	RCONF
F1-4	Use 9, 8, 6, 4, 1 (each can only be used once) to form the smallest five-digit number you can. Which digit should be put in the ones place? (A) 1 (B) 4 (C) 8 (D) 9*	(A)	Because 1 is the smallest number within 9, 8, 6, 4, 1	44	22.6	1.75
F1-8	Two watermelons exactly the same size are divided by five children. How much should each child get? (A) $\frac{1}{2}$ (B) $\frac{1}{5}$ (C) $\frac{2}{5}$ * (D) 2.5	(C)*	$25 = \frac{2}{5}$	71	36.4	2.04
F2-1	Determine which of $\frac{12}{8}$ and $\frac{12}{9}$ is larger  (A) $\frac{12}{8}$ * (B) $\frac{12}{9}$ (C) no difference (D) cannot compare	(A)*	By finding the common denominator, $\frac{12}{8} = \frac{108}{72}$ and $\frac{12}{9} = \frac{96}{72}$ , so $\frac{12}{8}$ is larger	42	21.5	1.95
F2-2	Determine which of $\frac{14}{24}$ and $\frac{8}{18}$ is larger  (A) $\frac{14}{24}$ * (B) $\frac{8}{18}$ (C) no difference (D) cannot compare	(B)	When the denominator is smaller, the fraction is larger	44	22.6	1.95
F2-4	Determine which of $\frac{3}{4}$ and $\frac{5}{6}$ is larger  (A) $\frac{3}{4}$ (B) $\frac{5}{6}$ * (C) no difference (D) cannot compare	(B)*	$\frac{3}{4} = \frac{18}{24}$ , $\frac{5}{6} = \frac{20}{24}$ , therefore, $\frac{3}{4} > \frac{5}{6}$	54	27.7	2.33
F3-1	Which of the diagrams does not represent $\frac{1}{4}$ of the whole rectangle? 	(A)*	Each area of the three shaded regions in the rectangle is not equal. Therefore, it cannot be $\frac{1}{4}$ of the whole rectangle	47	24.1	1.80
F3-5	Which of the following statements about 0.4 is true?  John: if the whole circle is one, then the shaded parts are 0.4  Mark: the arrow indicates 0.4  (A) John (B) Mark (C) Both of them are correct (D) Both of them are incorrect*	(C)	There are four sectors shaded and from 0, you count four to the current position in the number line	53	27.2	1.86
F4-2	Given $784 + 496 = 1280$ , what is the result of $7.84 + 4.96$ ?  (A) .1280 (B) 1.28 (C) 12.8* (D) 1280	(C)*	By calculation: $\begin{array}{r} 7.84 \\ + 4.96 \\ \hline 12.80 \end{array}$	70	35.9	2.11
F4-3	Calculating the problem: $1234 \div 5 \times 6$ . Whose procedure is correct?	(A)	By the mnemonic 'Please Excuse My Dear Aunt Sally', we should do multiplications prior to divisions	55	28.2	1.73

(Continued)



Table 6. (Continued).

ID	Problems	ANS	Reason	N	MIC (%)	RCONF
	(A) Ming: $1234 \div 5 \times 6 = 1234 \div (5 \times 6)$ (B) Hua: $1234 \div 5 \times 6 = 1234 \times 6 \div 5^*$ (C) Mei: $1234 \div 5 \times 6 = 1234 \div 6 \times 5$ (D) Yu: $1234 \div 5 \times 6 = 5 \times 6 \div 1234$					
F5-1	Without calculation, which of the following estimations best predicts the result of '249 $\times$ 4?'	(C)	Because 249 can round to 300 so $300 \times 4$ is 1200	49	25.1	1.92
	(A) less than 1000* (B) equal to 1000 (C) greater than 1000 (D) cannot determine					
F5-1	Without calculation, which of the following estimations best predicts the result of '249 $\times$ 4?'	(A)*	Because of the algorithm: $\begin{array}{r} 249 \\ \times 4 \\ \hline 996 \end{array}$	85	43.6	2.33
	(A) less than 1000* (B) equal to 1000 (C) greater than 1000 (D) cannot determine					
F5-2	There are 12 candies in a bag. Momojan and her five classmates divide two bags of candies evenly. How many 'bags' can each one get?	(C)	Because two bags of candies are divided to five people	49	25.1	1.78
	(A) $\frac{12}{5}$ bags, (B) 4 bags, (C) $\frac{2}{5}$ bags (D) $\frac{5}{6}$ bags*					

Notes: \*indicates the correct answer (but the reason is not for the number sense method); MIC = Misconception; Significant misconceptions in answer-reason tier = Incorrect reason% > 21%; Strong significant misconceptions = Significant misconceptions with the confidence ratings > 1.8; Moderate significant misconceptions = Significant misconceptions with the confidence ratings between 1.5 and 1.8; Reason column represents the non-number-sense-based methods.

Discussion

This study shows that the online NS4TT can be used not only to examine students' number sense performance and the relationship of confidence ratings between the answer tier and reason tier, but also to diagnose students' significant misconceptions. The findings of this study showed that the students' average percentage correctness on the number sense test was about 48%, and only 20% of the students applied number sense-based methods to justify their answers. This finding is consistent with earlier studies which showed that Taiwanese students' number sense performance was lower than 50% on the answer tier and lower than 25% on the reason tier (Li and Yang 2010; Yang, Li, and Lin 2008).

In addition, the results show that these students had a significantly higher confidence rating on the answer tier than the reason tier for each number sense component and for the whole number sense test. This implies that students may consider the problems in the answer tier and the reason tier to be two different levels of problems. Problems in

the reason tier may have more cognitive demand because they are related to students' metacognition. This result is consistent with the finding of Caleon and Subramaniam (2010).

Comparing the students' confidence with respect to the different number sense components, the students had the lowest confidence rating in the answer tier and reason tier on the component of *using multiple representations of numbers and operations* (F3). This implies that the students were not so familiar with (or had difficulties with) pictorial representation problems (e.g. items containing number lines or using circles to represent fractions). Earlier studies (e.g. Yang, Li, and Lin 2008) showed that students performed better on the component F3, but these studies did not show their relatively low confidence in this component. Since F3 is not the most challenging component for the students, we conjecture that many of the students may be able to solve the problems but that they are not so familiar with these problems (as mentioned earlier, students' confidence is related to familiarity with problems). Support for this conjecture may be found in the analysis of textbooks in Asian countries. Studies showed that textbooks in Asian countries had fewer visual representation problems in comparison to those in western countries (Cai 1995; Cai and Lester 2005; Huang and Cai 2011; Jiang, Hwang, and Cai 2014).

About two-thirds of the students in this study had equal confidence ratings in both tiers. About one-fourth of the students had higher confidence ratings on the answer tier than on the reason tier. Less than one-tenth of the students had higher confidence ratings on the reason tier than on the answer tier.

According to Caleon and Subramaniam (2010), students in the first group (with equal confidence ratings in both tiers) tend to believe that the answers and reasons are interconnected (even with different difficulty levels). Students in the second group (with higher confidence ratings on the answer tier than on the reason tier) may have an easier time finding the answers but have difficulties giving justifications. The third group of students (with higher confidence ratings on the reason tier than on the answer tier) may reflect two situations. The first situation is that reason options may be easier to understand for some students, although this is not typical for many students. The second situation is that some students may have chosen 'Just guessing' for their confidence level in the answer tier and confident or very confident in the reason tier, but their 'Just guessing' in the answer tier meant that they solved the problems by themselves instead of directly applying methods learned from textbooks or teachers.

The findings showed the many students either had equal confidence ratings in both tiers or had confidence ratings in the answer tier. This seems reasonable, since most students either thought the answer tier and reason tier were interrelated or believed the answer tier was easier. The findings were also consistent with the findings in the earlier studies that problems in the reason tier were more challenging for students (e.g. Caleon and Subramaniam 2010).

When further compared to the findings of Caleon and Subramaniam (2010), the students in this study had a slightly higher tendency to think of answers and reasons as interrelated (68% in this study; 57% in Caleon and Subramaniam). The percentage of students who reported higher confidence in the answer tier were about 10% fewer than reported by Caleon and Subramaniam (22% in this study; 33% in Caleon and Subramaniam) but the percentages of students in the third group were roughly the same (8% in this study; 10% in Caleon and Subramaniam).

Moreover, the delta values of confidence ratings between the students who gave correct responses and those who gave incorrect responses in the reason tier ( $\Delta$ RSN)

were higher than the corresponding delta values for the answer tier for the five components and the whole number sense scale ( $\Delta\text{RSN} > \Delta\text{ANS}$ ). This indicates that the reason tier has better discrimination which seems again to echo the observation that the reason tier is more challenging for students. This may also be further evidence that students' responses in the reason tier are better for detecting their number sense.

Furthermore, this study shows that the four-tier online test is capable of identifying several significant student misconceptions in both the answer tier and reason tier. Earlier studies have shown that Taiwanese students did not perform well on number sense tests (e.g. Yang, Li, and Lin 2008). However, these studies did not provide further information about the level of students' misconceptions. By using the confidence ratings in this study, we can specify the degree of their misconceptions. That is, a higher confidence level implies that the student holds the misconception with greater tenacity and the student is more certain of the correctness of the misconception. In this case, it may be more difficult to correct the misconception. We believe that this information will be helpful for researchers and teachers to understand students' misconceptions further. More importantly, with this information in hand, then teachers may be better able to correct and rectify students' misconceptions.

## Conclusion

This study showed that the online four-tier test is able to give diagnostic information about students' number sense performance, their confidence ratings in both the answer tier and reason tier, and their misconceptions. In particular, the strength of students' misconceptions can be detected via the online four-tier test. There were, in total, 16 significant misconceptions identified, most of which were at the strong level. One factor that may explain the high confidence ratings for these significant misconceptions is that many students selected the rule-based method option as the best answer. This indicates that these students tend to use rule-based methods to solve numerical problems. This result is consistent with earlier studies in which Taiwanese students were inclined to use written methods to solve problems (Yang 2005; Yang, Li, and Lin 2008).

This study also showed that Taiwanese students do not perform well on number sense but have relatively high confidence ratings. If we only consider performance on the answer tier, the students' total mean score is .48 which implies that students can, on average, answer 48% of the items correctly. When performance on the reason tier is included (i.e. a correct answer for the answer tier and a number sense-based method for the reason tier), the total mean score drops to .20. This result implies that, on average, students used number sense strategies to obtain the correct answer for 20% of the questions, and they used rule-based strategies (or some other non-number-sense method) to obtain the correct answer for about 28% of the questions. When taking confidence ratings into account, the mean scores in the two tiers are accompanied by confidence ratings of about 62 and 53% (calculated by  $\frac{1.86}{3}$  and  $\frac{1.61}{3}$ , where 3 is the highest confidence rating). This means the students had high levels of confidence in their overall performance, despite the fact that performance was generally low.

Although generalisations of the present study may be limited due to the small sample size and the representativeness of the sample, we suggest that the findings do provide some insights into this area.

First, not only can the NS4TT be used to understand student's reasons for an answer (other than content knowledge) and the strength of their misconceptions, but the reason tier, in particular, is effective for detecting students' number sense. For example, earlier

studies have shown that students perform well in the component of *recognising the number size* (F2). However, in F2, our reason tier further shows that students tend to use non-number sense-based methods. We would be less likely to find this if we used a traditional number sense test.

Second, the NS4TT can be used to obtain students' confidence ratings about the correctness of their chosen options. This helps to identify students' 'true understanding' (correct answers with high confidence ratings). For example, we were able to know that even with the high correctness for the answer tier, the students may not have been so familiar with using number sense-based methods for the problems based on their confidence ratings.

We use the term 'true understanding' to reflect the confidence level on a respondent's correct answer. When we say that someone has true understanding, it means this student is highly confident about his or her correct answer. This is the ability related to 'evaluating' from a metacognition perspective. Along these lines, if students chose a different way of thinking and reasoning (non-number sense methods), it did not necessarily imply that the students do not have true understanding. However, we probably could say these students did not prefer to use the number sense methods.

Third, the NS4TT is better than two-tier number sense test at glean quantitative and qualitative information at the same time. That is, in addition to quickly collecting students' responses and their reasons for their responses, the NS4TT adds information about the students' confidence level.

Finally, this study provides a method for designing a four-tier test in mathematics education. In our design, the test has different sets of reasons in the reason tier for corresponding answers in the answer tier. This may provide an approach to how a four-tier test can be successfully applied in mathematics education. As we have mentioned above, in mathematics education, an option in the reason tier is usually used to explain a specific answer (e.g. a reason for explaining the choice of .1280). Students can easily recognise which reason is used to explain which answer (e.g. choose .1280 as the correct answer and then easily see which reason was specifically written to go with .1280). Therefore, we provide a set of reasons for each answer and respondents only see corresponding reasons based on the answer they have chosen. This saves time by allowing students to focus on the corresponding reasons.

### ***Limitations of this study***

Several limitations existed in this study:

- (1) Collection of interview data is very time-consuming, but it does allow a wider range of expression and a deeper analysis of students' thinking. In the form of multiple-choice items in this study, the validity is limited.
- (2) To make any assessment fair, it is essential that the students knew the rules and understood the task before the test. Therefore, we told students not to use calculations in the test and helped them understand that the test was to test their number sense (we did this implicitly, not directly telling them) before the students started to work on these items. However, it is possible that some students may still have thought that this was a test and they should answer the problems by calculations even though they knew how to solve the problems by the number sense methods.

- (3) The amount of reading required might affect students' performance in the test. To minimise this threat to validity, we tried to make all the items have a similar amount of reading. However, sometimes it was not very easy to do so because we wanted to involve different types of problems and we had to make sure each problem was clearly stated. It is also possible that students with a high level of mathematical skill and a lower level of literacy may be disadvantaged by the presentation, even if the volume of reading was similar across test items (Metsisto 2005).
- (4) The items in this study did not have an 'other' category to allow students to write down their other strategies. This is because one of the goals of developing this online number sense test was to help teachers or researchers quickly obtain statistical reports from the test. In addition, it may be not easy for some students to work out whether their approach to the solution can be categorised in one of the options offered. These students may actually categorise their responses as many different 'other' responses, whereas, in fact, their responses have to be categorised in one of the options offered. Not providing an 'other' option might force them to choose the closest one to their 'real' categorisation.

Nevertheless, our items were created based on our long-term work in this area (e.g. Li and Yang 2010; Lin, Li and Yang, *forthcoming*; Yang 2005; Yang and Li 2013). The research thus far has involved a lot of participants and interview data. Therefore, we believe these items are relatively mature. That is, most students' possible responses were probably already represented in the options. However, it is still the case that having no open-ended options is a limitation of the study.

## Funding

This paper is a part of a research project supported by the Ministry of Science and Technology, Taiwan [grant number MOST 104-2511-S-415-002]. Any opinions expressed here are those of the authors and do not necessarily reflect the views of the Ministry of Science and Technology, Taiwan.

## References

- Batanero, C., and E. Sanchez. 2005. "What is the Nature of High School Students' Conceptions and Misconceptions about Probability?" In *Exploring Probability in School*, edited by G. Jones, 40 vols, 241–266. New York: Springer US.
- Berch, D. B. 2005. "Making Sense of Number Sense: Implications for Children with Mathematical Disabilities." *Journal of Learning Disabilities* 38 (4): 333–339.
- Cai, J. 1995. "A Cognitive Analysis of U.S. and Chinese Students' Mathematical Performance on Tasks Involving Computation, Simple Problem Solving, and Complex Problem Solving." *Journal for Research in Mathematics Education Monograph Series* 7: i–151.
- Cai, J., and F. K. Lester. 2005. "Solution Representations and Pedagogical Representations in Chinese and U.S. Classrooms." *The Journal of Mathematical Behavior* 24 (3–4): 221–237.
- Caleon, I. S., and R. Subramaniam. 2010. "Do Students Know What They Know and What They Don't Know? Using a Four-Tier Diagnostic Test to Assess the Nature of Students' Alternative Conceptions." *Research in Science Education* 40 (3): 313–337. doi:[10.1007/s11165-009-9122-4](https://doi.org/10.1007/s11165-009-9122-4).
- Cetin-Dindar, A., and O. Geban. 2011. "Development of a Three-Tier Test to Assess High School Students' Understanding of Acids and Bases." *Procedia-Social and Behavioral Sciences* 15: 600–604.

- Chinn, C. A., and W. F. Brewer. 1993. "The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction." *Review of Educational Research* 63 (1): 1–49.
- Clement, J. 1982. "Algebra Word Problem Solutions: Thought Processes Underlying a Common Misconception." *Journal for Research in Mathematics Education* 13: 16–30.
- Dehaene, S. 2001. "Precis of the Number Sense." *Mind and Language* 16: 16–36.
- Dunphy, E. 2007. "The Primary Mathematics Curriculum: Enhancing Its Potential for Developing Young Children's Number Sense in the Early Years at School." *Irish Educational Studies* 26 (1): 5–25.
- Dyson, N. I., N. C. Jordan, and J. Glutting. 2013. "A Number Sense Intervention for Low-Income Kindergartners at Risk for Mathematics Difficulties." *Journal of Learning Disabilities* 46 (2): 166–181.
- Glenberg, A. M., T. Sanocki, W. Epstein, and C. Morris. 1987. "Enhancing Calibration of Comprehension." *Journal of Experimental Psychology: General* 116 (2): 119.
- Hirsch, L. S., and A. M. O'Donnell. 2001. "Representativeness in Statistical Reasoning: Identifying and Assessing Misconceptions." *Journal of Statistics Education* 9 (2): 1–22.
- Huang, R., and J. Cai. 2011. "Pedagogical Representations to Teach Linear Relations in Chinese and U.S. Classrooms: Parallel or Hierarchical?" *The Journal of Mathematical Behavior* 30 (2): 149–165.
- Jiang, C., S. Hwang, and J. Cai. 2014. "Chinese and Singaporean Sixth-Grade Students' Strategies for Solving Problems about Speed." *Educational Studies in Mathematics* 87 (1): 1–24.
- Jordan, N. C., J. Glutting, and C. Ramineni. 2010. "The Importance of Number Sense to Mathematics Achievement in First and Third Grades." *Learning and Individual Differences* 20 (2): 82–88.
- Jordan, N. C., C. Ramineni, and M. W. Watkin. 2010. "Validating a Number Sense Screening Tool for Use in Kindergarten and First Grade: Prediction of Mathematics Proficiency in Third Grade." *School Psychology Review* 39 (2): 181–195.
- Li, M. N., and D. C. Yang. 2010. "Development and Validation of a Computer-Administered Number Sense Scale for Fifth-Grade Children in Taiwan." *School Science and Mathematics* 110 (4): 220–230.
- Lin, Y.-C., M. N. Li, and D. C. Yang. (forthcoming). "Diagnosing Students' Misconceptions in Number Sense via a Web-Based Two-Tier Test." *Eurasia Journal of Mathematics, Science & Technology Education*.
- Markovits, Z., and J. T. Sowder. 1994. "Developing Number Sense: An Intervention Study in Grade 7." *Journal for Research in Mathematics Education* 25 (1): 4–29.
- McIntosh, A., B. J. Reys, and R. E. Reys. 1992. "A Proposed Framework for Examining Basic Number Sense." *For the Learning of Mathematics* 12 (3): 2–8.
- Menon, R. 2004. "Elementary School Children's Number Sense." *International Journal for Mathematics Teaching and Learning*. <http://www.cimt.plymouth.ac.uk/journal/default.htm>.
- Metsisto, D. 2005. "Reading in the Mathematics Classroom." In *Literacy Strategies for Improving Mathematics Instruction*, edited by J. M. Kenney, E. Hancewicz, L. Heuer, D. Metsisto, and C. L. Tuttle, 9–23. Alexandria, VA: Association for Supervision and Curriculum Development.
- National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: NCTM.
- Reys, B. J. 1994. "Promoting number sense in middle grades." *Teaching Mathematics in the Middle School* 1: 114–120.
- Reys, R. E., and D. C. Yang. 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.
- Sood, S., and A. K. Jitendra. 2007. "A Comparative Analysis of Number Sense Instruction in Reform-Based and Traditional Mathematics Textbooks." *The Journal of Special Education* 41 (3): 145–157.
- Sowder, J. T. 1992. "Estimation and number sense." In *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics*, edited by G. A. Douglas, 371–389. New York: Macmillan Publishing.
- Sreenivasulu, B., and R. Subramaniam. 2013. "University Students' Understanding of Chemical Thermodynamics." *International Journal of Science Education* 35 (4): 601–635.

- Stankov, L., and J. Crawford. 1997. "Self-Confidence and Performance on Tests of Cognitive Abilities." *Intelligence* 25 (2): 93–109.
- Stankov, L., and B. Dolph. 2000. "Metacognitive Aspects of Test-Taking and Intelligence." *Psychologische Beitrage* 42 (2): 213–227.
- Tan, K. C. D., N. K. Goh, L. S. Chia, and D. F. Treagust. 2002. "Development and Application of a Two-Tier Multiple Choice Diagnostic Instrument to Assess High School Students' Understanding of Inorganic Chemistry Qualitative Analysis." *Journal of Research in Science Teaching* 39 (4): 283–301.
- Tobin, K. G., and W. Capie. 1981. "The Development and Validation of a Group Test of Logical Thinking." *Educational and Psychological Measurement* 41 (2): 413–423.
- Treagust, D. F. 1988. "Development and Use of Diagnostic Tests to Evaluate Students' Misconceptions in Science." *International Journal of Science Education* 10 (2): 159–169.
- Tsai, C. C., and C. Chou. 2002. "Diagnosing Students' Alternative Conceptions in Science." *Journal of Computer Assisted Learning* 18 (2): 157–165.
- Verschaffel, L., B. Greer, and De Corte, E. 2007. "Whole Number Concepts and Operations." In *Second Handbook of Research on Mathematics Teaching and Learning*, edited by F. Lester Jr., 557–628. Charlotte, NC: Information Age.
- Yang, D. C. 2005. "Number Sense Strategies Used by 6th-Grade Students in Taiwan." *Educational Studies* 31 (3): 317–333.
- Yang, D. C., and M. N. Li. 2013. "Assessment of Animated Self-Directed Learning Activities Modules for Children's Number Sense Development." *Journal of Educational Technology and Society* 16 (3): 44–58.
- Yang, D. C., M. N. Li, and C. I. Lin. 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.
- Zakay, D., and J. Glicksohn. 1992. "Overconfidence in a Multiple-Choice Test and Its Relationship to Achievement." *Psychological Record* 42 (4): 519–525.