A TECHNOLOGY-BASED STATISTICAL REASONING ASSESSMENT TOOL IN DESCRIPTIVE STATISTICS FOR SECONDARY SCHOOL STUDENTS

Shiau Wei Chan, Zaleha Ismail
Faculty of Education
Universiti Teknologi Malaysia
81310 UTM, Johor, Malaysia
shiauweichan@yahoo.com or swchan4@live.utm.my

ABSTRACT
The focus of assessment in statistics has gradually shifted from traditional assessment towards alternative assessment where more attention has been paid to the core statistical concepts such as center, variability, and distribution. In spite of this, there are comparatively few assessments that combine the significant three types of statistical reasoning (reasoning about center, spread, and distribution) with information technology in the context of secondary school students. Hence, this paper intends to discuss the development and initial validation of a technology-based statistical reasoning assessment tool that has been created based on a previously developed statistical reasoning framework. This framework has been useful in evaluating students' statistical reasoning levels in task-based interviews. The assessment tool formulated through this study will be used to refine and validate the initial statistical reasoning framework. There are five tasks in this instrument and each item is labeled according to four key constructs. The technological tool that has been used in solving tasks is dynamic mathematics software. This technology-based statistical reasoning assessment tool can be applied for further investigation.
Keywords: alternative assessments; statistical reasoning

INTRODUCTION

Assessment is an important part in the teaching and learning process (Dikli, 2003; Usun, 2003; Jamil, 2012) which can provide a clearer picture on what the students have learnt and problems they encountered (Akkya, Karakirik, & Durmus, 2005). Nevertheless, most instructors tend to employ traditional assessments in the classroom; hence they can only gather information about what students know and can do rather than what is really going on in the classroom (Bayram, 2005). Traditional assessments such as true-false tests and multiple choice tests do not give an apparent picture of students' performance and the efficacy of the teaching method adopted. Furthermore, in the traditional statistics classroom, instructors are likely to use textbooks, chalkboard, and paper-and-pencil activities. They only focus on computation skills, routine rules, and memorization of formulas (Qian, 2011). Teachers are thus unable to guide students to reason statistically and they only promote procedural understanding rather than conceptual understanding of statistical concepts (Garfield, delMas & Chance, 2007). Therefore, an alternative assessment is needed for three different reasons. First, to assess the conceptual and meaningful understanding of students. Second, to place more emphasis on the learning procedure rather than the product, and finally, to stimulate more effective learning and teaching practice (Durmus & Karakirik, 2005).

Beginning in the 1990s, the focus of statistics education has slowly moved from traditional assessment to alternative assessment that includes statistical literacy, statistical reasoning, and statistical thinking. The use of information technology provides numerous opportunities to formulate more successful assessment (Jamil, 2012). This is also supported by Csapo, Ainley, Benett, Latour & Law (2010) who claimed that the integration of information technology in assessment has gradually become imperative nowadays. Such innovative assessment tools can promote pedagogical innovation and curriculum reform rather than retaining its traditional function, which is to support the statistical reasoning of students (Chan & Ismail, 2012). To date, a few types of statistical reasoning assessments are being used in statistics education, notably the statistical reasoning assessment (SRA), Comprehensive of Assessment of Outcomes in a First Statistics Course (CAOS), Assessment Resource Tools for Improving Statistical Thinking (ARTIST), and so on. However, these assessments do not involve the use of particular software to assess statistical reasoning. Moreover, some topics are inappropriate for Malaysian secondary school students because they are not in the syllabus, for instance topics concerning correlation and causation in SRA. Therefore, this study filled this gap by formulating a new statistical reasoning assessment tool to suit Malaysian secondary school students, with particular emphasis on descriptive statistics.

Furthermore, unlike previous efforts, elements of GeoGebra spreadsheet have been integrated into the assessment so that the developed assessment can become a technology-based statistical reasoning assessment tool. GeoGebra is dynamic mathematics software which combines the features of a spreadsheet, dynamic geometry software, and computer algebra systems. It is open source software and can be freely downloaded from

the website (http://www.geogebra.org/cms/en/) (Hohenwarter & Lavicza, 2007; Hohenwarter & Preiner, 2007). By using the spreadsheet, the users are able to observe the changes in relationships before and after the figures alteration by moving, stretching and shrinking the figure. The utilization of GeoGebra software provides dynamic visualization which can develop users' understanding of statistical concepts (Boz, 2005). Although emphasis on 'big ideas' or central statistical ideas like center, distribution and variability in teaching and learning statistics has become increasingly obvious nowadays (Garfield & Ben-Zvi, 2007; Garfield & Ben-Zvi, 2008), most students continue to perceive these ideas as exclusive concepts. Therefore, to foster students' understanding about the relationships between these central statistical ideas, three types of statistical reasoning have been incorporated into the technology-based statistical reasoning assessment tool developed through this study, i.e. reasoning about center, spread, and distribution. This paper thus intends to discuss this assessment tool which has been used to characterize and assess students' statistical reasoning across four key constructs and five levels of reasoning.

STATISTICAL REASONING

According to Garfield and Chance (2000), statistical reasoning is described as "the way people reason with statistical ideas and make sense of statistical information. It involves making interpretations based on sets of data, or statistical summaries of data. Students need to be able to combine ideas about data and chance, which leads to making inferences and interpreting statistical results (p. 101)". Meanwhile, Lovett (2001) stated that statistical reasoning involves the use of statistical ideas and tools to summarize and draw assumptions besides making conclusions from the data. Martin (2009), on the other hand, defined statistical reasoning as "forming conclusions and judgments according to the data from observation studies, experiments or sample surveys" (p. 13).

As mentioned before, three types of statistical reasoning were integrated into this assessment tool, namely reasoning about center, spread, and distribution. Reasoning about center concerns data analysis that involves mean, mode, and median. Besides that, reasoning about spread involves range, interquartile range, variance, and standard deviation. Reasoning about distribution entails interpreting a compound structure comprised of reasoning about features such as center, spread, skewness, density, and outliers as well as other concepts such as causality, chance, and sampling (Pfannkuch & Reading, 2006).

In general, statistical reasoning, thinking, and literacy are unique domains, but two instructional perspectives have been formed to describe how these three outcomes are related to each other. Some instructional activities, if viewed from different instructional perspectives, may enhance students' understanding in two or more domains. The first perspective is that statistical literacy provides the foundation to develop the basic knowledge and skills needed to foster statistical thinking and reasoning. Some content of statistical reasoning, thinking, and literacy overlap, but some are independent (delMas, 2002). Another instructional perspective suggests that statistical literacy contains all the learning outcomes. It implies that statistical reasoning and thinking are subsets of statistical literacy and thus do not have their own independent content (delMas, 2002).

delMas (2002) also proposed that test or task items can be used to assess certain domains and perhaps the same item can assess more than one domain. To demonstrate this, he listed out the typical words that are used in tests or tasks that can differentiate statistical reasoning, statistical thinking, and statistical literacy. For instance, to develop statistical literacy, students are often required to identify an example that represents a certain concept, describe a graph, and translate and interpret the results. To enhance statistical reasoning, instructors can ask the students to explain how or why the findings are produced as they are. Meanwhile, to promote statistical thinking, students can be required to apply their knowledge to authentic questions, to review and assess the design and conclusions for studies, or to summarize information from the classroom to new circumstances.

Essentially, statistical reasoning and thinking are exploited interchangeably in some studies. However, there are other studies that distinguish statistical reasoning and statistical thinking. For example, Wild and Pfannkuch (1999) demonstrated this through a model constructed for statistical thinking. delMas (2004) also stated that we can distinguish statistical reasoning from statistical thinking by referring to the methods used by the respondents in solving a task. For instance, someone who possesses statistical reasoning ability can give an explanation for the findings and conclusion. Another person who has statistical thinking skills, on the other hand, is able to apply statistical understanding and processes while solving the task.

Some statisticians assert that statistics is basically an independent subject from general mathematics (e.g. Gal & Garfield, 1997; Cobb & Moore, 1997; Moore & Cobb, 2000; Rossman, Chance, & Medina, 2006; Garfield & Ben-Zvi, 2008). Gal and Garfield (1997) and Rossman, Chance, and Medina (2006) claimed that there are some differences between statistics and mathematics in terms of context, measurement issues, data collection, and

reasoning methods. Context is described as the foundation of meaning and foundation for the analysis of findings, which is vital in statistics when interpreting data and drawing conclusions. However, context might or might not play a role in mathematics (delMas, 2004). Besides that, measurement and data collection are also more crucial in statistics than mathematics as statistics generally depends on valid measurement and data collection. In mathematics, accurate measurement is not necessary and rough measurement is accepted (Gattuso & Ottaviani, 2011). Furthermore, mathematics involves deductive reasoning where a conclusion is made rationally based on definitions and axioms while statistics involves inductive reasoning where the conclusion may be vague but is still acceptable and legitimate. Moreover, Gal and Garfield (1997) stated that while statistics is often undefined, mathematics is more accurate. Nonetheless, mathematics is barely a set of procedures in statistics and there is no single mathematical solution for statistics.

Mathematical reasoning refers to reasoning about patterns as mathematics is considered the study of patterns. It is about certainty and proof within given hypothesis (Gal & Garfield, 1997). According to delMas (2004), mathematical reasoning and statistical reasoning are almost the same, but there are some discrepancies that will lead to different kinds of mistakes, especially when students solve highly abstract tasks. For mathematics, there are fewer tendencies to apply real-world context to the tasks. In contrast, real world context is emphasized in statistics (Cobb & Moore, 1997). Diverse types of statistical instruction are required to enhance students' understanding of statistics ideas and processes as students respond in different ways to statistics compared to mathematics. This also indicates that to teach statistics more effectively and efficiently, instructors should concentrate less on theory and computations and focus more on data and concepts (Rossman, Chance, & Medina, 2006).

INITIAL STATISTICAL REASONING FRAMEWORK

This initial statistical reasoning framework is very important in our study as it forms the basis of this technology-based statistical reasoning assessment tool. It was first formulated to characterize and assess students' statistical reasoning levels in descriptive statistics based on Garfield's (2002) model of statistical reasoning. There are five levels of statistical reasoning embedded in this framework, i.e. idiosyncratic, verbal, transitional, procedural and integrated process reasoning. At the idiosyncratic reasoning level, students know some of the statistical words and symbols, but tend to capitalize on them without totally understanding them, and so the meaning itself is most often inaccurate. Consequently, the students may combine them with unconnected information. At the verbal reasoning level, students have the verbal understanding of some concepts, but they cannot relate them to the actual behavior. To put it in another way, students can give and pick the true definition, but they do not understand the concepts completely. In addition, they may be able to discriminate the dimension of a statistical concept or process accurately, but do not know the procedure to combine them in order to reach the transitional reasoning level. At the procedural reasoning level, students are able to determine the dimensions of statistical concepts or procedures correctly, but are incapable of integrating them completely. Once the students have a complete understanding of statistical procedures and can confidently organize the rules and behavior, it can be said that the students have achieved the integrated process reasoning level (Garfield, 2002).

On the other hand, the four key constructs in this technology-based statistical reasoning assessment tool are describing data; organizing and reducing data; representing data; and analyzing and interpreting data based on the framework of Jones, Thornton, Langrall, Mooney, Perry & Putt (2000). Describing data involves accurate reading of raw data or data demonstrated in charts, tables, or graphs (Jones et al., 2000). It combines the reading of data from the studies of Curcio (1981, 1987) and Curcio and Artz (1997). In the study of Jones et al. (2000), four processes were put forth including reading data representations, demonstrating awareness of essential graphing conventions, identifying when different displays represent the same data, and assessing different displays of the same data. In terms of describing data, Mooney (2002) identified the existence of four sub-processes, namely demonstrating consciousness of exhibited features, distinguishing similar data in various data depictions, assessing the efficacy of data depiction in data presentation, and recognizing components of data values. For the initial framework of this study, only three sub-processes were used in describing data and are as shown in Table 1. These sub-processes consist of extracting and generating information from the data or graph; showing awareness of the displayed attributes of graphical representation; and recognizing the general features of the graphical representation. For the first sub-process, the students have to extract and generate explicit information while reading the data displays. They ought to be aware of the displayed attributes of graphical representation, which is composed of graphical conventions (e.g., title and axis labels) related to the second sub-process. This sub-process is identical to the first sub-process of Mooney (2002). Furthermore, the third sub-process is new to the framework where students need to identify the general features of the graphical representation including shape, center, and spread. By integrating these three features together, students will recognize them as a whole entity rather than isolated concepts (Garfield and Ben-Zvi, 2007).

Table 1: Describing data

Level / Construct	Level 1 Idiosyncratic	Level 2 Verbal	Level 3 Transitional	Level 4 Procedural	Level 5 Integrated Process
Describing Data	Does not extract and generate the idiosyncratic or relevant information from the data or graph	Extracts and generates some information from the data or graph verbally, but are ambiguous or unclear	Extracts and generates one or two dimensions of the information from the data or graph	Extracts and generates the information from the data or graph correctly	Extracts and generates the information from the data or graph completely
	Does not show awareness to the displayed attributes of graphical representation	Shows awareness to the displayed attributes of graphical representation orally, but partly correct	Shows little awareness to the displayed attributes of graphical representation	Shows some awareness to the displayed attributes of graphical representation	Shows complete awareness to the displayed attributes of graphical representation
	Does not recognize the general features of the graphical representation	Recognizes the general features of the graphical representation in words, but partly accurate	Recognizes one or two general features of the graphical representation	Recognizes the general features of the graphical representation accurately	Recognizes the general features of the graphical representation completely

Organizing and reducing data involves arranging, classifying, or merging data into a summary form (Moore, 1997) and requires the measurements of center and spread (Jones et al., 2000). The study of Jones et al. (2000) has four sub-processes related to this key construct: (1) categorizing and arranging data; (2) identifying the information that might be lost in the restructuring of data; (3) explaining data in terms of typicality or representativeness; and (4) portraying data in terms of spread. Mooney (2002), on the other hand, only introduced three sub-processes - categorizing and arranging data; expressing data with measures of center; and delineating the variability of data. Similar to Jones et al. (2000), Groth (2003) also distinguished four sub-processes for organizing and reducing data, i.e., applying measures of dispersion, utilizing measures of center, arranging sets of raw data, and distinguishing the outcomes of data conversion upon center and spread. This study only utilized three sub-processes for the initial framework (Table 2), notably organizing the data into a computer system; reducing the data using the measure of center, either by calculation or aided by technology; and reducing the data using the measure of spread, either by calculation or aided by technology. These three sub-processes are unique in the sense that they involve the utilization of information technology, an aspect that has been neglected in previous studies. The students are required to organize the data into the computer system rather than doing it manually. For the second and third sub-processes, the students have to reduce their data using measures of center and spread in two ways – manual and automated calculation. The latter is done by using the computer. After the students have performed the manual calculation, they have to check the answers against the answers calculated using the computer.

Table 2: Organizing and reducing data

Level / Construct	Level 1 Idiosyncratic	Level 2 Verbal	Level 3 Transitional	Level 4 Procedural	Level 5 Integrated Process
Organizing and Reducing Data	Unable to organize the data into a computer system	Provides oral statements when organizing the data into a computer system, but only partly correct	Organizes the data into a computer system with major mistakes	Organizes the data into a computer system with minor mistakes	Organizes the data into a computer system in the right way
	Unable to reduce the data using the measures of center, either by calculation or aided by technology	Reduces the data using the measures of center in words, either by calculation or aided by technology but only accurate to some extent	Reduces the data using the measures of center with major errors, either by calculation or aided by technology	Reduces the data using the measures of center with minor errors, either by calculation or aided by technology	Reduces the data using the measures of center completely, either by calculation or aided by technology
	Unable to reduce the data using the measures of spread, either by calculation or aided by technology	Reduces the data using the measures of spread orally, either by calculation or aided by technology but only accurate to some extent	Reduces the data using the measures of spread with major faults, either by calculation or aided by technology	Reduces the data using the measures of spread with minor faults, either by calculation or aided by technology	Reduces the data using the measures of spread completely, either by calculation or aided by technology

The third key construct is representing data and encompasses presenting data in a graphical form, which means that the process requires basic conventions related to the presentations (Jones et al., 2000). Moreover, the authors have recognized two sub-processes for representing data: (1) completing a partially created data representation; and (2) producing representations to signify different organizations of a data set. In this regard, Mooney (2002) has also put forth three sub-processes to present data, i.e., creating a data depiction for a given set of data; finishing an incompletely created atypical data depiction; and constructing an interchangeable data depiction. Only three sub-processes are applied in this initial framework as revealed in Table 3. The processes include demonstrating the data sets graphically using the computer, identifying different representations for the same data set, and judging the effectiveness of two different representations for the same data. Undeniably, the execution of this key construct also demands the use of information technology. In the first sub-process, the students are required to graphically present the data set using the GeoGebra software. This sub-process encourages the students to learn and interact actively using the computer as they drag the figures dynamically and learn to present the data set using a variety of graphical presentations (e.g., from a histogram to a box plot and stem and leaf plot). The second sub-process, i.e., identifying the different representations for the same data set, is similar to the second sub-process of describing data in the study of Mooney (2002). The third sub-process is also identical to the third sub-process of describing data in the same study. Unlike earlier studies, this study does not just assess the process of constructing graphs but tries to make sense of the created graph to enhance sophisticated reasoning about representing data (Friel, Curcio & Bright, 2001).

Table 3: Representing data

Level / Construct	Level 1 Idiosyncratic	Level 2 Verbal	Level 3 Transitional	Level 4 Procedural	Level 5 Integrated Process
Representing Data	Demonstrates the data sets graphically using the computer without precise display	Provides verbal statements when demonstrating the data sets graphically using the computer, but only partially correct	Demonstrates the data sets graphically using the computer with major errors	Demonstrates the data sets graphically using the computer with minor errors	Demonstrates the data sets graphically using the computer with a valid display
	Does not identify the different representations for the same data set	Identifies the different representations for the same data set in words, but only partially correct	Identifies one or two aspects of the different representations for the same data set	Identifies the different representations for the same data set in the correct way	Identifies the different representations for the same data set in a complete and comprehensive way
	Does not judge the effectiveness of two different representations for the same data set	Judges the effectiveness of two different representations for the same data set orally, but only partially correct	Judges one or two elements of the effectiveness of two different representations for the same data set	Judges the effectiveness of two different representations for the same data set accurately	Judges the effectiveness of two different representations for the same data set completely

Lastly, analyzing and interpreting data entails recognizing trends, patterns, and formulating deductions or presumptions from the data (Jones et al., 2000). It consists of reading between the data and reading beyond the data (Curcio, 1987). Jones et al. (2000) introduced two sub-processes for analyzing and interpreting data: (1) comparing and combining data; and (2) extrapolating and making predictions from the data. Additionally, three sub-processes were employed in Mooney's (2002) study for analyzing and interpreting data, i.e., comparing between data displays and data sets; comparing within the data displays or data sets; and making inferences from a given data display or data set. Groth (2003) recognized eight sub-processes - exploring sample means; contrasting univariate data sets; determining atypical points in a tabular data set; interpolating within bivariate data; making multiplicative comparisons; explaining bivariate relationships; finding out atypical points in a graphical bivariate data set; and extrapolating from bivariate data. In this study, only three sub-processes were chosen (Table 4), i.e., making comparisons within the same data set; making comparisons between two different data sets; and making predictions, inferences or conclusions from the data or graphs. The first and second sub-processes are equivalent to the first and second sub-processes of Mooney's (2002) study. Making comparisons within the same data set is the first sub-process where students ought to compare the same data set. In addition, students have to compare two different data sets for the second sub-process. Finally, they have to make predictions, inferences, or conclusions from the data or graphs in the third sub-process. This is also similar to the third sub-process from the study of Mooney (2002), which involves making inferences from the data or graph. The process of making predictions is somewhat similar to the second sub-process from the study of Jones et al.

(2000). Another element, making conclusion, is new and does not exist in earlier studies. This has now been included as it is imperative for the students to know how to summarize data or graphs while solving tasks.

Table 4: Analyzing and interpreting data

Level / Construct	Level 1 Idiosyncratic	Level 2 Verbal	Level 3 Transitional	Level 4 Procedural	Level 5 Integrated Process
Analyzing and Interpreting Data	Does not make comparisons within the same data sets	Makes some comparisons within the same data sets verbally, but are incomplete	Makes one or two comparisons within the same data sets	Makes the comparisons within the same data sets correctly	Makes the comparisons within the same data sets completely
	Does not make comparisons between two different data sets	Makes comparisons between two different data sets in words, but are somewhat incorrect	Makes one or two comparisons between two different data sets	Makes comparisons between two different data sets accurately	Makes comparisons between two different data sets completely
	Does not make prediction, inference or conclusion from the data or graphs	Makes prediction, inference or conclusion from the data or graphs in words, but are incomplete	Makes one or two prediction, inference or conclusion from the data or graphs	Makes prediction, inference or conclusion from the data or graphs in the appropriate way	Makes prediction, inference or conclusion from the data or graphs in a complete and comprehensive way

METHODOLOGY

Instrument Development

After developing the initial statistical reasoning framework, this technology-based statistical reasoning assessment tool was constructed to refine and validate the initial statistical reasoning framework. The topics of descriptive statistics covered in this assessment tool are measures of central tendency and measures of variability. There are five tasks in this assessment tool with 56 items altogether. Every item is associated with the sub-processes of four main constructs as indicated in Tables 5 to 8.

Table 5: Examples of Items in the sub-processes for describing data

Constructs	Code	Sub-processes	Items
Describing data	D1	Extracting and generating information from the data or graph	1) What are the highest and lowest amount of protein (in grams) for various fast food sandwiches? 2) Write the name of the feature at each of the labels on the five-number summary of the box plot and record the values from the computer.
	D2	Showing awareness to the displayed attributes of graphical	1) What does this graph tell you?

		representation	
	D3	Recognizing the general features of the graphical representation	1) Describe the distribution of the graph with respect to its shape, center and variability.

Table 6: Examples of Items in the sub-processes for organizing and reducing data

Constructs	Code	Sub-processes	Items
Organizing and reducing data	O1	Organizing the data into a computer system	1) Organize the data into GeoGebra spreadsheet.
	O2	Reducing the data using the measures of center, either by calculation or aided by technology	1) What is the mean of the graph? Explain how. 2) What is the mode of the graph? Explain how. 3) What is the median of the graph? Explain how.
	O3	Reducing the data using the measures of spread, either by calculation or aided by technology	1) What is the range of the graph? Explain how. 2) What is the interquartile range of the graph? Explain how. 3) What is the standard deviation of the graph? Explain how.

Table 7: Examples of Items in the sub-processes for representing data

Constructs	Code	Sub-processes	Items
Representing data	R1	Demonstrating the data sets graphically using the computer	1) Draw the graph using GeoGebra dynamic worksheet by dragging the red circle. Tick the check box of Show histogram, Show mean and Show median. 2) Drag the red circle to draw the new histogram. 3) Construct a frequency polygon using GeoGebra spreadsheet. 4) Represent the data in another way. 5) Construct a box plot for each set of data. 6) Construct a stem and leaf plot for each set of data.
	R2	Identifying the different representations for the same data set	1) Describe how the box plot is related to its matching histogram.
	R3	Judging the effectiveness of two different representations for the same data	2) Which graph do you think represents the data better, the histogram or the box plot? Explain why.

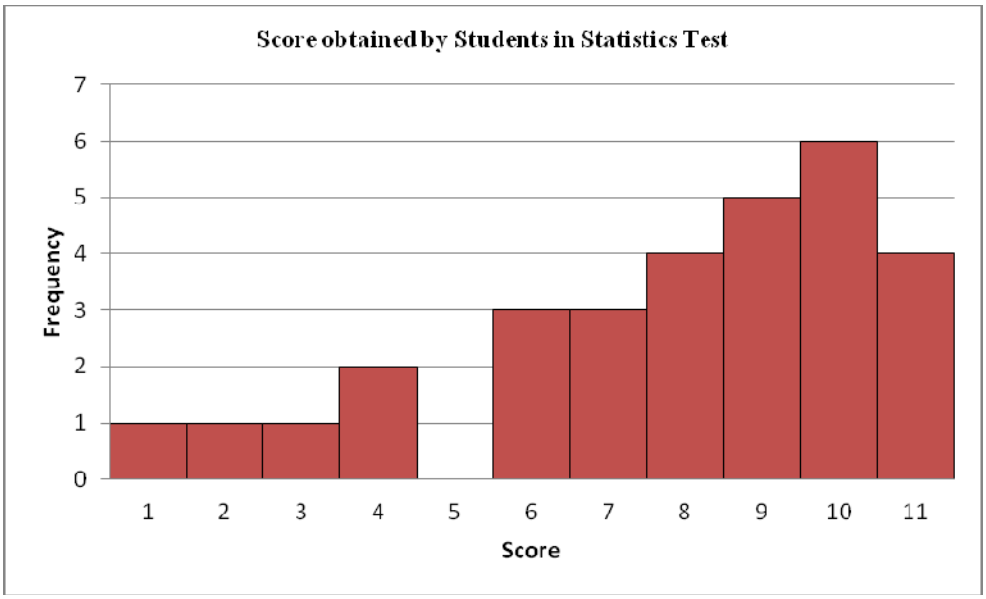
Table 8: Examples of Items in the sub-processes for analyzing and interpreting data

Constructs	Code	Sub-processes	Items
Analyzing and interpreting data	A1	making comparisons within the same data set	1) Compare your answer in Question 2, and 4 with the values shown on the computer. If the answers are different, explain why. 2) Compare the results in question 15 with question 14. What do you observe? Explain why.

			3) Compare the answer you predicted in Question 3 to the value shown on the computer. If the answers are different, explain why.
	A2	making comparisons between two different data sets	1) Compare the distribution of both box plots with respect to shape, center and variability. 2) Compare the distribution of both stem and leaf plots with respect to shape, center and variability.
	A3	making prediction, inference or conclusion from the data or graphs	1) Which measures of center is the most suitable to be used to represent the score obtained by students? Explain why. 2) Which measures of spread is the most suitable to be used to represent the score obtained by students? Explain why. 3) Predict which data set has greater variability, Malaysia or Taiwan. Explain why. 4) Make a conclusion from the data of unemployment rates of males and females. 5) Are there any similarities or differences between the two graphs produced on the computer? Explain.

To evaluate the usefulness of this assessment tool, the participating students are required to solve five tasks in the task-based interview sessions. Students will use the GeoGebra software as the technological tool. During the task-based interview phase, the researcher will interview the students one-by-one and the interview sessions are both video-taped and audio-taped. Both recordings of the interview protocols are transcribed into written form and then tabulated and coded. Subsequently, the data obtained will be used to refine and validate the initial statistical reasoning framework.

Task 1



1) What does this graph tell you?
2) What is the mean of the graph? Explain how.
3) What is the mode of the graph? Explain how.
4) What is the median of the graph? Explain how.
5) Draw the graph using GeoGebra dynamic worksheet by dragging the red circle. Tick the check box of Show histogram, Show mean and Show median.
6) Compare your answer in Question 2, and 4 with the values shown on the computer. If the answers are different, explain why.
7) What is the range of the graph? Explain how.
8) What is the interquartile range of the graph? Explain how.
9) What is the standard deviation of the graph? Explain how.
10) Tick the check box of Show IQR and Show Std Dev.
11) Compare your answer in Question 8 and 9 with the values shown on the computer. If the answers are different, explain why.
12) Describe the distribution of the graph with respect to its shape, center and variability.

Another set of new scores obtained by students from a different class are as follows:

13) Drag the red circle to draw the new histogram.
14) Record the values of mean, median, interquartile range, and standard deviation from the computer.

Two students who each obtained a score of 1 are added to the graph.

15) Record the values of mean, median, interquartile range, and standard deviation from the computer.
16) Compare the results in question 15 with question 14. What do you observe? Explain why.
17) Which measures of center is the most suitable to be used to represent the score obtained by students? Explain why.
18) Which measures of spread is the most suitable to be used to represent the score obtained by students? Explain why

Figure 1: Task 1

Task 1 requires students to explore ungrouped data. In Question 1, students have to obtain the information from the histogram. Furthermore, they need to understand and use the concepts of mean, mode, and median of ungrouped data in Questions 2, 3, 4, 5 and 6. As for Questions 7, 8, 9, 10 and 11, the students should understand and use the concepts of range, interquartile range, and standard deviation for ungrouped data. Moreover, they ought to understand how the concepts of center, spread and distribution are related to each other in Question 12.

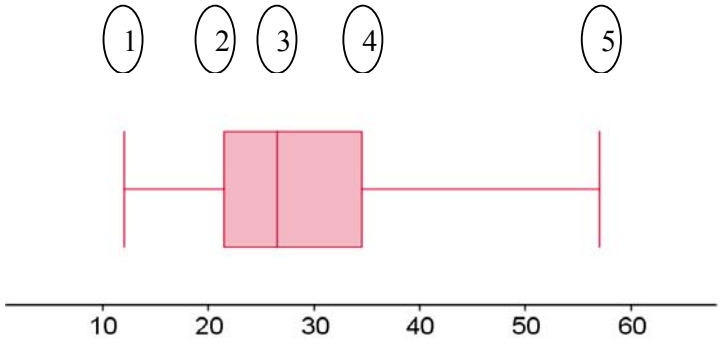
Questions 13, 14, 15 and 16 ask students to determine an outlier in the data set. Meanwhile, Question 17 and 18 require students to identify the most suitable measures of center and spread for the given data.

Task 2

The data below indicate the amount of protein (in grams) for various fast food sandwiches (Source: The Doctor's Pocket Calorie, Fat, and Carbohydrate Counter, 2002).

23	30	20	27	44	26	35	20	29	29
25	15	18	27	19	22	12	26	34	15
27	35	26	43	35	14	24	12	23	31
40	35	38	57	22	42	24	21	27	33

1) What are the highest and lowest amount of protein (in grams) for various fast food sandwiches?
2) Organize the data into GeoGebra spreadsheet.
3) Construct a frequency polygon using GeoGebra spreadsheet.
4) Record the values of the mean, median and standard deviation from the computer.
5) Describe the distribution of the graph in terms of its shape, center and variability.
6) Represent the data in another way.
7) Write the name of the feature at each of the labels on the five-number summary of the box plot and record the values from the computer.



No	Five-number summary	Value
1		
2		
3		
4		
5		

8) Describe how the box plot is related to its matching histogram.
9) Which graph do you think represents the data better, the histogram or the box plot? Explain why.

Figure 2: Task 2

In Task 2, students are asked to investigate the grouped data obtained from the raw data in Question 1. In this part, the students have to organize, present and interpret data in a frequency polygon for the grouped data in Questions 2, 3 and 4. In addition, for Question 5, the students should understand how the concepts of center,

spread and distribution are related to each other. Furthermore, Questions 6 and 7 need students to present and interpret data in box plots where the students have to match the box plot to the histogram in Questions 8 and 9.

Task 3

The following data shows the yearly instant noodle consumption (in millions of packets) for Malaysia and Taiwan from year 2002 to 2007 (Source: Global Oils & Fats Business Magazine, 2009)

Country	2002	2003	2004	2005	2006	2007
Malaysia	7.4	8.2	8.7	8.9	10.6	11.8
Taiwan	9.4	10.0	9.5	8.9	8.7	8.5

1) What are the highest and lowest yearly instant noodle consumption (million packets) for Malaysia?
2) What are the highest and lowest yearly instant noodle consumption (million packets) for Taiwan?
3) Predict the instant noodle consumption for Malaysia and Taiwan in 2008. Explain why.
4) Predict which data set has greater variability, Malaysia or Taiwan. Explain why.
5) Organize the data into GeoGebra spreadsheet.
6) Construct a box plot for each set of data.
7) Record the mean, standard deviation, minimum value, first quartile, median, third quartile, and maximum value for each of the data set.
8) Compare the distribution of both box plots with respect to shape, center and variability.
9) Make a conclusion from the data of instant noodle consumption for Malaysia and Taiwan.

Figure 3: Task 3

Students can then compare the box plots generated from the two data sets in Questions 1 and 2 in Task 3. In addition, they also need to make a prediction from two data sets in Questions 3 and 4. Questions 5, 6 and 7 require students to organize, present and interpret two data sets in the box plots. In Question 8, students are required to relate the concepts of center, spread and distribution to compare the two data sets. Then, they have to make a conclusion from the data in Question 9.

Task 4

A survey was conducted on a sample of people from a country in 1995. The data demonstrated the percentage of males and females who were unemployed (Source: New York Times Almanac).

Males					Females				
1.5	6.6	5.6	0.3	7.7	7.0	6.8	5.6	0.5	9.4
4.1	3.1	4.6	6.0	6.6	3.0	3.4	6.5	8.7	8.0
9.6	4.4	5.2	6.0	8.7	7.7	5.3	4.6	7.2	5.9
9.8	5.9	3.1	5.6	2.2	9.2	8.8	3.2	8.6	3.3
4.6	5.6	1.9	8.8		5.0	8.6	3.7	8.0	

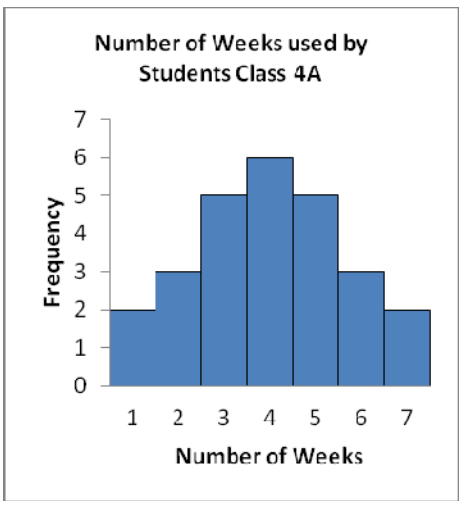
1) What are the highest and lowest percentage of unemployed males?
2) What are the highest and lowest percentage of unemployed females?
3) Organize the data into GeoGebra spreadsheet.
4) Construct a stem and leaf plot for each set of data.
5) Compare the distribution of both stem and leaf plots with respect to shape, center and variability.
6) Make a conclusion from the data of unemployment rates of males and females.

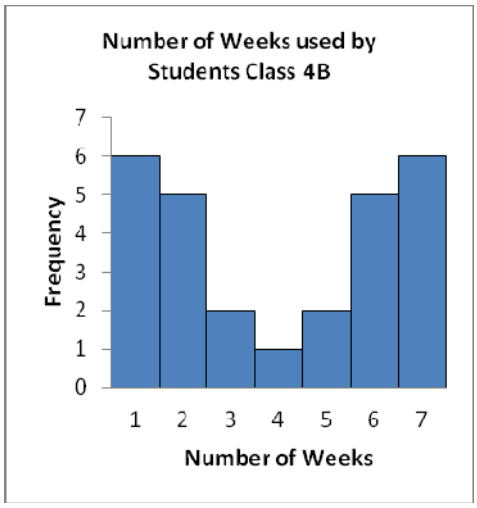
Figure 4: Task 4

For Task 4, the students can compare the stem and leaf plots drawn from the two data sets in Questions 1 and 2. They have to then organize and present the two data sets in stem and leaf plots in Questions 3 and 4. Next, Question 5 asks students to relate the concepts of center, spread and distribution to compare two data sets. Lastly, in Question 6, they need to make a conclusion from the data.

Task 5

The following graphs illustrate the number of weeks used by the students from class 4A and 4B to finish reading a storybook.





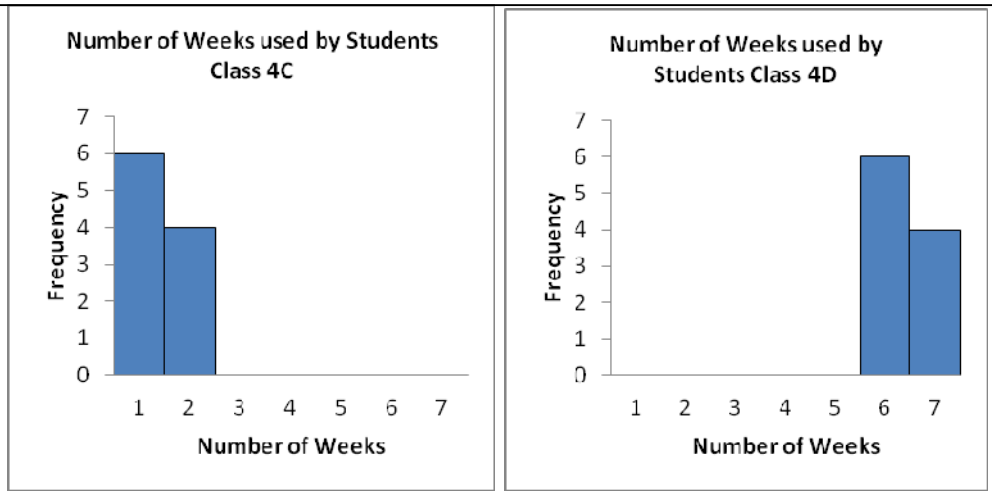
1) What are the highest and lowest number of weeks used by the students from class 4A to finish reading a storybook?
2) What are the highest and lowest number of weeks used by the students from class 4B to finish reading a storybook?
3) Predict which class has the lower standard deviation. Explain why.
4) Drag the red circle on the GeoGebra dynamic worksheet to create the histograms for Class 4A and Class 4B. Tick the check box of Show Std Dev.
5) Compare the answer you predicted in Question 3 to the value shown on the computer. If the answers are different, explain why.
6) Are there any similarities or differences between the two graphs produced on the computer? Explain.

The teacher did a survey of the number of weeks used by the students from class 4A and 4B to finish reading a storybook during the school holidays. The following data indicated the results of the survey.

Week	1	2	3	4	5	6	7
Class 4A	3	3	3	3	3	3	3
Class 4B	0	3	5	6	4	3	0

7) Predict which class has the larger standard deviation. Explain why.
8) Drag the red circle on the GeoGebra dynamic worksheet to create the histograms. Tick the check box of Show Std Dev.
9) Compare the answer you predicted in Question 7 to the value shown on the computer. If the answers are different, explain why.
10) Are there any similarities or differences between the two graphs produced on the computer? Explain.

The graphs below show the number of weeks used by the students from class 4C and 4D to finish reading a storybook.



11) Predict which class has the larger standard deviation. Explain why.
12) Drag the red circle on the GeoGebra dynamic worksheet to create the histograms for Class 4C and Class 4D. Tick the check box of Show Std Dev.
13) Compare the answer you predicted in Question 11 to the value shown on the computer. If the answers are different, explain why.
14) Are there any similarities or differences between the two graphs produced on the computer? Explain.

Figure 5: Task 5

Task 5 requires students to explore histograms. In Questions 1 and 2, they ought to obtain the information from two histograms. This is followed by making a prediction from two histograms in Questions 3, 5, 7, 9, 11 and 13. Finally, they need to present and interpret the data in histograms for two data sets in Questions 4, 6, 8, 10, 12, and 14.

Tasks Validation

Content validity was carried out to determine the matching degree between the content and domain being measured (Gay, Mills & Airasian, 2009). The tasks of this technology-based statistical reasoning assessment tool had been validated by three experts, a crucial step which ensures that the items can evaluate the students' statistical reasoning level. The cooperation was carried out via electronic mail. The instrument was not validated concurrently by all experts, but was reviewed by one expert and amended accordingly before it was sent to the next expert. These three experts are lecturers from foreign universities that have published significantly influential works in the field of statistical reasoning. Expert A is an associate professor from the University of Minnesota, USA, with extensive experience in the field. He has taught statistics to university students for more than 20 years and has published countless papers about statistical reasoning in refereed journals, book chapters, and conference proceedings. Expert B is an associate professor from Illinois State University, USA, with years of teaching experience in statistics as well. He was actively involved in the development of models for statistical reasoning. Expert C is a senior lecturer from the University of New England, Australia who has numerous publications on statistical reasoning such as reasoning about sampling, reasoning about variation, informal inferential reasoning, and so forth. All experts contributed valuable views and suggestions to the constructed tasks other than helping to verify the accuracy of the English words used. Appropriate corrections were then made. Since this instrument is in dual language (English and Malay), two lecturers who are excellent in Malay helped to verify the language accuracy.

Tasks Reliability

Inter-rater agreement was sought to confirm the reliability of this instrument (Slavin, 2007). Two raters were involved in statistics; both of them are lecturers from local universities and are proficient in statistics and mathematics education. Rater A is an associate professor from Universiti Teknologi Malaysia and has 15 years of teaching experience in statistics and mathematics. The rater's field of specialization is in advanced mathematical thinking and problem solving. Meanwhile, rater B is a senior lecturer from the same university who has extensive teaching experience in statistics and mathematics subjects as well. He was a lecturer in the

Islamic Azad University, Iran, before joining the current university. The researcher tabulated the four constructs, sub-processes, and items before both raters were asked whether they agree or disagree. This was done by either giving a (√) or (X). Both raters were requested to judge the appropriateness of the items under the four constructs within a two week period before an in-depth discussion was held. Then, the percentage of agreement was calculated based on their judgment.

DISCUSSION

The validity and reliability of the technology-based statistical reasoning assessment tool had been measured. The three experts who validated the instrument had commented on the strengths and weaknesses of the instrument. Concerning instrument strength, expert A mentioned that there were some good items in this assessment tool. In addition, expert A also pointed out that it is acceptable to have both statistical literacy and statistical reasoning items in the instrument as some content is interconnected and sometimes statistical reasoning is the subset of statistical literacy (delMas, 2002). Expert B stated that there were two good questions to assess statistical reasoning, i.e., 'Describe the distribution of the graph with respect to its shape, center, and variability' and 'Which graph do you think represents the data better, the histogram or the box plot? Explain why.' Expert C found this instrument interesting and is looking forward to reading the published results.

For the weaknesses of the instrument, expert A recommended to change Question 1 ('What does this graph tell you?') in Task 1 to 'What can you tell me about the statistics test scores from this graph?' as he did not understand what the question wanted. However, no changes were made to this question; we expect the students to give answers on the display features of the graph such as the title and axis label and not merely on the statistics test scores. Besides that, expert A was confused about one of the questions in Task 2 – 'Justify your conclusion for the data', so this question was then eliminated from the instrument to avoid confusion on behalf of the students. Expert B said that there were too many questions focused around the GeoGebra computer program rather than statistical reasoning. Therefore, two sub-processes of representing data which concerned procedural steps in GeoGebra software were changed to make room for better judgment on the students' reasoning level. One of the sub-processes for representing data was unchanged because procedural steps like drawing or constructing a graph are needed in order to carry out the subsequent reasoning step.

The three experts also gave some recommendations to improve the instrument. For instance, expert A suggested that the question in Task 4 be changed so that the data can be more robust. Nonetheless, the researcher kept the question as the data was obtained from a practical source. Not only that, expert A also suggested that the questions in Task 5 which are related to asking the students to create a graph using the GeoGebra software be modified before identifying the minimum and maximum value and estimating the standard deviation. This suggestion was partly accepted as it is essential that the students compare the similarities and differences between the two graphs in terms of the value of mean, median, standard deviation, and interquartile range. This step can only be done after they have estimated the value of standard deviation. Moreover, expert B requested to have a question that entails comparing quiz scores of the first class with those of the second class. Such a question was not inserted as this would confuse the students in terms of the order of the classes. Expert B also suggested the addition of two more questions into Task 3 and Task 4, which are 'Predict the noodle consumption for 2008 and explain why' and 'What conclusion can you draw from the data about the unemployment rates of men and women'. This suggestion was accepted and thus the two questions were included in the instrument. Expert C proposed adding the phrase, 'you decided on this value' behind the 'Explain how'. However, the researcher felt that 'Explain how' can be understood easily and it is a typical phrase for assessing statistical reasoning. To sum up, the views, comments, responses, and recommendations given by these three experts were encouraging and helpful.

The degree of reliability of this instrument was manually calculated at the first stage in terms of percentage of agreement between the two raters. This was done by dividing the number of times both raters mutually agreed on a certain item by the number of possible observations. The computed percentage of agreement was 96.4 %. The same results were then analyzed using SPSS software and the output was as indicated in Table 9. The result was consistent with the manual calculation, i.e., 96.4 %. According to Boyatzis (1998), stability of a measure of consistency between the judgments of two rates can only be established if the percentage of agreement is at least 70 %. Therefore, it can be concluded that the inter-rater reliability for this assessment tool is reasonably consistent. Since the instrument has strong validity and reliability, it is highly recommended that this instrument be used not only at the secondary school level, but also at the university level.

Table 9: Percentage of agreement

Reviewer_difference

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-1	2	3.6	3.6	3.6
	0	54	96.4	96.4	100.0
	Total	56	100.0	100.0	

CONCLUSIONS

A technology-based statistical reasoning assessment tool has been developed in order to assess and characterize students' statistical reasoning across four key constructs and five levels of reasoning as well as to refine and validate the initial statistical reasoning framework. This assessment tool will be tested empirically in a task-based interview session. It is probable that this newly developed assessment tool will promote students' conceptual understanding of statistical concepts, thus leading them to reason statistically. In future studies, instructors and researchers can make use of this assessment tool to assess students' statistical reasoning level in terms of different races, gender, country, educational background, and so forth. Further investigations require not just improvement to the framework but also to disseminate tools and methods more extensively beyond students studying statistics.

ACKNOWLEDGEMENT

We appreciatively acknowledge the assistance and guidance given by Associate Professor Robert C. delMas, Associate Professor Edward Mooney, and Dr Chris Reading to review and validate the content of this technology-based statistical reasoning assessment tool. Besides, we are also thankful to Associate Professor Dr Yudariah Mohammad Yusof and Dr Hamidreza Kashefi as raters.

REFERENCES

Akkaya, R., Karakirik, E., & Durmus, S. (2005). A computer assessment tool for concept mapping. *The Turkish Online Journal of Educational Technology*, *4*(3), 3-6.

Bayram, S. (2005). Software mapping assessment tool documenting behavioral content in computer interaction: Examples of mapped problems with KID PIX program. *The Turkish Online Journal of Educational Technology*, *4*(2), 7-17.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. London: SAGE Publication.

Boz, N. (2005). Dynamic visualization and software environments. *The Turkish Online Journal of Educational Technology*, *4*(1), 26-32.

Chan, S.W., & Ismail, Z. (2012). The role of information technology in developing students' statistical reasoning. *Procedia-Social and Behavioral Sciences*, *46*, 3660-3664.

Cobb, G. W., & Moore, D. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*, 801-823.

Csapó, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2010). *Draft white paper 3: Technological issues for computer-based assessment.* Assessment and Teaching of 21st Century Skills, The University of Melbourne, Australia. Retrieved from http://atc21s.org/wp-content/uploads/2011/11/3-Technological-Issues.pdf

Curcio, F. R. (1981). *The effect of prior knowledge, reading and mathematics achievement, and sex on comprehending mathematical relationships expressed in graphs*. Brooklyn, NY: St. Francis College.

Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graph. *Journal for Research in Mathematics Education*, 18(5), 382-393.

Curcio, F. R., & Artz, A. F. (1997). Assessing students' statistical problem solving behaviors in small-group setting. In I. Gal & J. B. Garfield (Eds.) *The assessment challenge in statistics education* (pp. 123-138). Amsterdam: IOS Press.

delMas, R.C. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education*, 10(3). Retrieved from http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html

delMas, R.C. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 79-95). The Netherlands: Kluwer Academic Publishers.

Dikli, S. (2003). Assessment at a distance: Traditional vs. alternative assessments. *The Turkish Online Journal of Educational Technology*, *2*(3), 13-19.

Durmus, S., & Karakirik, E. (2005). A computer assessment tool for structural communication grid. *The Turkish Online Journal of Educational Technology*, 4(4), 3-6.

Friel, S.N., Curcio, F.R., and Bright, G.W. (2001). Making Sense of Graphs: Critical Factors influencing Comprehension and Instructional Applications. *Journal for Research in Mathematics Education*, *32*(2), 124-158.

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1-13). Amsterdam, The Netherlands: IOS Press.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, 10 (3). Retrieved from http://www.amstat.org/publications/jse/v10n3/garfield.html

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, *75*(3), 372-396.

Garfield, J., delMas, R.C., & Chance, B. (2007). Using Students' Informal Notions of Variability to Develop an Understanding of Formal Measures of Variability. In Lovett, M.C., and Shah, P. (Eds.) *Thinking with Data* (pp. 87-116). New York: Lawrence Erlbaum.

Garfield, J. B., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning*, *2*(1&2), 99-125.

Garfield, J. B., & Ben-Zvi, D. (2008). Research on teaching and learning statistics. In J. B.Garfield & D. Ben-Zvi (Eds.) *Developing students' statistical reasoning: Connecting research and teaching practice* (pp. 21-43). Springer.

Garfield, J. B., & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review*, *67*(1), 1-12.

Gattuso, L., & Ottaviani, M.G. (2011). Complementing mathematical thinking and statistical thinking in school mathematics. In C. Batanero, G. Burrill & C. Reading (Eds.), *Teaching statistics in school mathematics-Challenges for teaching and teacher education*: *A Joint ICMI/IASE Study* (pp. 121-132). Springer.

Gay, L. R., Mills, G. E., & Airasian, P. (2009). *Educational research: Competencies for analysis and applications*. New Jersey: Pearson Education International.

Groth, R. E. (2003). *Development of a high school statistical thinking framework*. Unpublished doctoral dissertation, Illinois State University.

Hohenwarter, M., & Lavicza, Z. (2007). Mathematics teacher development with ICT: towards an International GeoGebra Institute. In D. Küchemann (Ed.), *Proceedings of the British Society for Research into Learning Mathematics*. *27*(3). University of Northampton, UK: BSRLM.

Hohenwarter, M., & Preiner, J. (2007). Dynamic Mathematics with GeoGebra. *Journal for Online Mathematics and its Applications*, Volume 7. March 2007. Article ID 1448.

Jamil, M. (2012). Perceptions of university students regarding computer assisted assessment. *The Turkish Online Journal of Educational Technology*, *11*(3), 267-275.

Jones, G. A., Thornton, C.A., Langrall, C. W., Mooney, E.S., Perry, B., & Putt, I. A. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning*, *2*, 269-307.

Lovett, M. (2001). A collaborative convergence on studying reasoning processes: A case study in statistics. In D. Klahr & S. Carver (Eds.). *Cognitive and instruction*: *Twenty-five years of progress* (pp. 347-384). Mahwah, NJ: Lawrence Erlbaum.

Martin, G. (2009). *Focus in school mathematics: Reasoning and sense making*. National Council of Teachers of Mathematics.

Mooney, E.S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning*, *4(*1), 23-63.

Moore, D., & Cobb, G. (2000). Statistics and mathematics: Tension and cooperation. *The American Mathematical Monthly*, *107,* 615-630.

Moore, D.S. (1997). *Statistics: Concepts and controversies* (4th ed.). New York: Freeman.

Pfannkuch, M., & Reading, C. (2006). Reasoning about distribution: A complex process. *Statistics Education Research Journal*, *5*(2), 4-9.

Qian, G. (2011). Teaching, Learning and Retention of Statistical Ideas in Introductory Statistics Education. *European Journal of Pure and Applied Mathematics*, *4*(2), 103-116.

Rossman, A. J., Chance, B. L., & Medina, E. (2006). Some important comparisons between statistics and mathematics, and why teachers should care. In G.F. Burrill (ed.) *Thinking and reasoning about data and chance: Sixty-eighth NCTM yearbook* (pp. 323-334). Reston, VA: National Council of Teachers of Mathematics.

Slavin, R. E. (2007). *Educational research in an age of accountability*. United States of America: Pearson Education, Inc.

Usun, S. (2003). Advantages of computer based educational technologies for adult learners. *The Turkish Online Journal of Educational Technology*, *2*(4), 3-9.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-265