# INTRODUCTION

There is an increasing emphasis on statistics in the school curriculum, in line with the wide use of statistics in a variety of day-to-day situations (Australian Curriculum and Assessment Authority, 2010; Ministry of Education, 2007; National Council of Teachers of Mathematics (NCTM), 2000). The ubiquity of statistical information and the sophisticated displays are available in part because of the availability of technology. In schools, statistics teaching is supported by specialist software such as Tinkerplots (Konold & Miller, 2005) which has been designed to facilitate young students' statistical understanding by allowing them to play with data in meaningful ways. Projects such as the international Census At School (Royal Statistical Society Centre for Statistical Education, 2010) provide easy access to data that is generated by school-aged students and of immediate interest to them. Students enjoy these experiences (Carmichael & Hay, 2010) and produce relatively high level data analysis (Gil & Ben-Zvi, 2010) drawing sensible, although informal, inferences and explaining their findings in ways that make sense to them. These explanations and understanding are arguably more sophisticated than curriculum documents expect, although may not be recognized as emerging statistical understanding by statisticians.

Franklin etc al (2007) in the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) report provided detailed curriculum expectations at four levels. Technology use was explicit and embedded in their suggestions for investigations and assessment. The approaches to teaching and learning statistics suggested in this report provided a hierarchical framework in which higher-order thinking associated with conducting investigations and interpreting results was developed, supported by technology use. More specifically in relation to technology use Chance, Ben Zvi, Garfield and Medina (2007) provided an overview of the kinds of technology available in schools and colleges, and the implications of the use of technology for teaching. Technology can provide quick and easy ways to explore and summarise data, but can also become a "black box" which can impact on students' understanding. The goals for education need to change to acknowledge both the power and potential risks associated with technology use in statistics education, and alongside these shifts, assessment must also change.

Lesh (2000, p. 193) described how technology has produced an "explosion of representational media" that, although reducing the computational load, has "radically increased the interpretation and communication demands", comments echoed by Chance et al (2007). The period since Lesh was writing this has seen technology use in schools increase significantly, with new kinds of software, more interactivity, miniaturization in the form of graphing calculators, smartphones and MP3 players, use of data loggers and many other applications. Lesh argued that technology has changed the nature of the conceptual systems that are created both within education settings and in the wider world.

Such changes in conceptual demand are coherent with the goals for statistical education described by Garfield and Gal (1999). The goals enunciated by Garfield and Gal include the need to: Understand the purpose and logic of statistical investigations; Understand the process of statistical investigations; Master important procedural skills; Understand probability and chance; Develop interpretive skills and statistical literacy; Develop ability to communicate statistically; and Develop useful statistical dispositions. These goals have a strong focus on the interpretation and communication of information derived from statistical processes, and go beyond computational or graph drawing skills. Progress towards many of these new goals can be facilitated by the use of technology. It is expected that students in school today have access to sophisticated technology in the form of hand-held or portable devices, as well as personal computers. Using these tools students can manipulate and "play" with data more quickly and in ways that are not possible when the data are in hard copy formats, and which provide access to powerful ways of understanding statistics in line with the goals of Garfield and Gal. If, however, the nature of the cognitive constructs developed in classrooms has changed (Lesh, 2000) and the goals for statistical education have

shifted in focus towards inference and communication, there must be implications for assessment generally as well as assessment of statistics specifically.

Garfield and Chance (2000) recognised emerging challenges for assessment of statistical understanding. One of these was the impact of technology use. Summarising Garfield and Gal (1999), they suggested that ways need to be found of incorporating technology into assessment, and to ensure that students who were taught using technology were appropriately assessed when technology was not used in the assessment process. These comments raise issues for classrooms and for external assessment processes. Regardless of the nature of technology used, in any assessment process on which technology use has an impact steps must be taken to ensure that all students are able to use the technology sufficiently fluently so that it does not interfere with the assessment process (Pellegrino, winter 2002-3). More recently, Garfield and Ben-Zvi (2008) characterized the different goals for statistics education, and the differing demands of assessment, around the triad of statistical literacy, reasoning and thinking. They called for more use of alternative approaches to assessment, such as student projects and investigations, which can be facilitated and enhanced by appropriate technology use.

In the classroom, statistical understanding can be enhanced by the use of appropriate technology (Lipson, Francis & Kokonis, 2006). For example, in the middle-years of schooling specialist software *Tinkerplots* (Konold & Miller, 2005) provides an environment that allows students to "tinker" with the data using drag-and-drop approaches to creating data representations. Use of colour lets students explore multiple variables in one representation, and both discrete and continuous variables can be considered. The Tinkerplots interface uses "data cards" to display an individual record and also provides a graphical interactive interface that gives opportunities for students to explore, display, and summarise the data contained in the cards. Teachers must recognize the affordances offered by the technology and adjust their assessment expectations accordingly. This adjustment might include the nature of the assessment task and the criteria used to make judgements about students' achievement. Outside the classroom, external agencies, whether planning assessment for placement at the end of schooling or measuring educational achievement for monitoring processes, must consider both the nature of the learning that has occurred in technology-rich environments and also emerging approaches to using technology for assessment in large-scale testing programs. Issues associated with these in class assessment and large-scale external assessment, where technology is used as part of the assessment process and where it is not permitted, are considered in this paper, and suggestions made for future research and practice.

# ISSUES FOR CLASSROOM ASSESSMENT

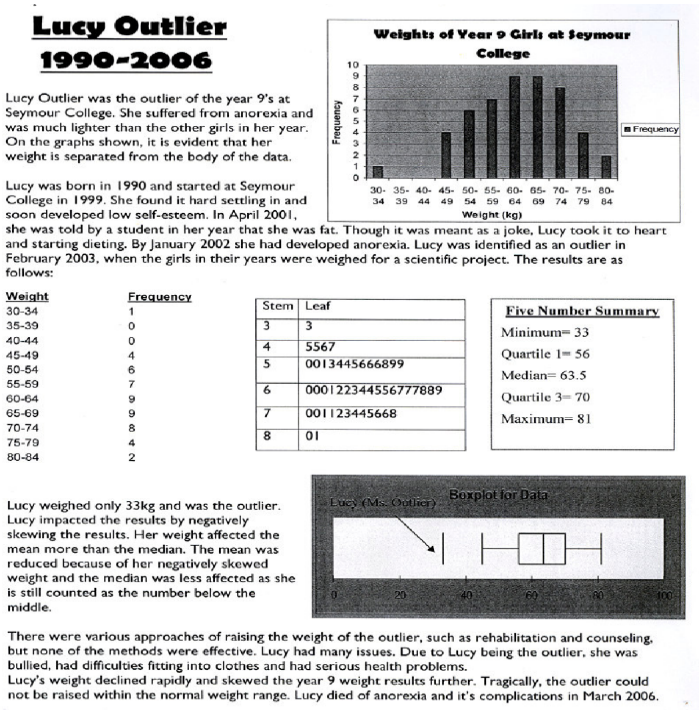
In classrooms there are two ways of considering the impact of technology on assessment. Jolliffe (1997) indicated that the ability to use computers effectively was one aspect of assessing statistics in the classroom that needed to be considered. In relation to such thinking, Fitzallen (2008), for example, used a pen-and-paper instrument to identify students' understanding of graphing and data representation that had been developed in a Tinkerplots environment, acknowledging the centrality of the technology use but limiting the assessment process to a traditional instrument rather than using the technology itself as part of the assessment. There are arguments in favour of such an approach. For example, students need to be able to recognize statistical representations produced and published using a variety of technological tools. Assessing understanding of graphical representations without technology use can indicate the extent to which students can transfer their technologically developed skills. In contrast, Lajoie (1997) took a different approach, describing the use of computers to enhance teaching and assessment, including tracking students' actual interactions with the computer using specialist screen-capture software to provide information to teachers about the ways in which their students were using technology tools. Although both of these approaches by Fitzallen and Lajoie aim to assess the use of technology, considering how students are interacting with the technology itself does not address aspects of statistical understanding, particularly the higher order goals described by Garfield and Gal (1999). In a transition period while students are less familiar with software packages, it may be appropriate to use pen-and-paper

assessment tools, but such traditional tools do not take account of Lesh's (2000) observations about the changes to conceptual systems that technology use creates.

The second way of thinking about technology is as a tool for developing and communicating statistical understanding, going beyond the production of different types of representation to the interpretation of these. By asking students to use technology to produce a report, for example, and applying some suitable framework for assessment criteria, judgements can be made about the quality of students' work produced with the support of technology of various types. Watson's (1997) Tier Model of statistical literacy is one such framework. The tier model has three levels of understanding: Use of terminology; Understanding terminology in context; and Thinking critically in context, including questioning claims made about data.

As an example of this approach to assessment in a technology-rich environment, where technology is a tool, one teacher in a research project, *StatSmart*, (Watson, Callingham & Donne, 2008) assessed students' understanding of the concept of outliers through the production of a poster. The students, all Grade 9 girls, were asked to collect data relevant to their peer group and to present an article for a teenage magazine as a biography of "Miss Outlier". They used technology to produce various displays of data and then interpreted these creatively in the context of a magazine article. In "Lucy Outlier" (Figure 1), the student explored data related to anorexia. Applying Watson's Tier Model to analyse her work, the student demonstrates clear achievement of the first two levels concerning use of terminology. She uses language such as "outlier", "mean", and "median", and explains how the "outlier" impacted on the distribution by "negatively skewing the results", indicating good use of terminology within the context of the data. She also appears to be making progress towards a critical understanding in her description of ways in which the outlier could be brought closer to the group's results.

Arguably the student could have produced a more sophisticated display but the article demonstrates an attempt to grapple with the statistical notion of an outlier in the context of a social setting of immediate interest to herself. This attempt is supported by the use of technology which has allowed her to produce representations of the data with minimal cognitive demands, so that she could consider the meanings behind the statistics. In this way, classroom assessment is using technology as a tool to scaffold higher levels of thinking which can be judged using a hierarchical or developmental framework such as Watson's Tier Model.



**Lucy Outlier**
**1990-2006**

Lucy Outlier was the outlier of the year 9's at Seymour College. She suffered from anorexia and was much lighter than the other girls in her year. On the graphs shown, it is evident that her weight is separated from the body of the data.



Lucy was born in 1990 and started at Seymour College in 1999. She found it hard settling in and soon developed low self-esteem. In April 2001, she was told by a student in her year that she was fat. Though it was meant as a joke, Lucy took it to heart and starting dieting. By January 2002 she had developed anorexia. Lucy was identified as an outlier in February 2003, when the girls in their years were weighed for a scientific project. The results are as follows:

| Weight | Frequency |
|---|---|
| 30-34 | 1 |
| 35-39 | 0 |
| 40-44 | 0 |
| 45-49 | 4 |
| 50-54 | 6 |
| 55-59 | 7 |
| 60-64 | 9 |
| 65-69 | 9 |
| 70-74 | 8 |
| 75-79 | 4 |
| 80-84 | 2 |

| Stem | Leaf |
|---|---|
| 3 | 3 |
| 4 | 5567 |
| 5 | 0013445666899 |
| 6 | 000122344556777889 |
| 7 | 001123445668 |
| 8 | 01 |

**Five Number Summary**
Minimum= 33
Quartile 1= 56
Median= 63.5
Quartile 3= 70
Maximum= 81

Lucy weighed only 33kg and was the outlier. Lucy impacted the results by negatively skewing the results. Her weight affected the mean more than the median. The mean was reduced because of her negatively skewed weight and the median was less affected as she is still counted as the number below the middle.



There were various approaches of raising the weight of the outlier, such as rehabilitation and counseling, but none of the methods were effective. Lucy had many issues. Due to Lucy being the outlier, she was bullied, had difficulties fitting into clothes and had serious health problems.
Lucy's weight declined rapidly and skewed the year 9 weight results further. Tragically, the outlier could not be raised within the normal weight range. Lucy died of anorexia and it's complications in March 2006.

*Figure 1.* Using technology to support assessment tasks that address understanding of statistical concepts.

Technology use, however, cannot compensate for poorly conceived tasks. The same principles that underpin all assessment must also apply when technology is used in classrooms either to produce work for assessment or as an integral part of the assessment itself. The Lucy Outlier example shows that rich teaching and learning tasks supported by technology use permit assessment *as* learning (Earl, 2003) when supported by the use of an appropriate framework against which teachers can make judgements (Sadler, 1998). In conceiving of assessment in this way, the assessment process becomes more coherent with the learning situation. As the goals and aims of statistical education have changed in light of technology use, the assessment expectations must also shift in response. Assessing statistical understanding cannot remain as a one-off, high stakes examination but has to incorporate the increased demands inherent in the goals of Garfield and Gal (1999).

At the classroom level, resources are needed to help teachers assess the higher order goals demanded by Garfield and Gal (1999). One approach might be to develop a generic set of standards, or scoring rubrics, based on a framework such as Watson's (1997) Tier Model or the later Statistical Literacy Hierarchy (Callingham & Watson, 2005; Watson & Callingham, 2003), or the GAISE levels (Franklin et al, 2005). Technology use could be incorporated into such standards, to include Jolliffe's (1997) expectation about considering computer use in assessment.

A somewhat different way of using the power of technology is that taken by the ARTIST project. ARTIST (Assessment Resource Tools for Improving Statistical Thinking) provides a database of validated items organized by topic and learning outcome. Teachers may use the data base to build tests appropriate to their context. Impressively, there are no purely computational items and both open-ended and multiple choice or forced choice item types are available (Garfield, del Mas & Chance, 2006). The topics are scaled and teachers are provided with these scales to use to make judgements about their students. This approach uses the technology to create meaningful assessments tailored by individual instructors to their specific needs. More of these kinds of resources are needed at all levels of schooling and in statistics teacher education.

# ISSUES FOR EXTERNAL ASSESSMENT AND EDUCATIONAL MEASUREMENT

Outside the classroom, a different set of challenges is faced by test developers and systems when technology use is incorporated in statistics teaching and learning. One such challenge is the use of technology itself as part of a large-scale assessment process; the second is assessing the nature of students' understanding when technology use has been central to its development. These two aspects of large scale assessment are considered in this section.

The impact of technology on large-scale assessment that is imposed by agencies outside the school is growing as computing facilities can handle larger amounts of data and have faster processing speeds. In many respects, however, the nature of the questions used has not changed. Such tests are delivered in traditional ways, using pencil-and-paper formats, but are machine scored. Figure 2 shows an example of a Grade 8 item from the 2003 Trends in International Mathematics and Science Study (TIMSS) (TIMSS & PIRLS International Study Centre, 2009). It addresses students' understanding of average but can be answered correctly by any student who can apply the algorithm. In terms of the goals identified by Garfield and Gal (1999), it assesses only "Master important procedural skills". It is a machine-scored item that is cheap and efficient to administer on a large-scale but does not address higher-order thinking or interpretive capabilities.

Joe had three test scores of 78, 76, and 74, while Mary had scores of 72, 82, and 74. How did Joe's average (mean) score compare with Mary's average (mean) score?

A. Joe's was 1 point higher.
B. Joe's was 1 point lower.
C. Both averages were the same.
D. Joe's was 2 points higher.
E. Joe's was 2 points lower.

*Figure 2*. Grade 8 item from TIMSS 2003.

Such an item could be made more conceptual by reversing it. For example if it were worded as shown in Figure 3, the question could not only address conceptual understanding of average, but also provide some diagnostic information. Choice A in this version suggests no understanding of average since all of Joe's scores are below Mary's, so that the mean could not be the same. Choice C could suggest confusion between the mean and the mode, and choice D could imply mistaking the mean for the median.

Joe and Mary did three tests each. The scores were all out of 100. Their average (mean) score was the same. Which one of the following choices could have been the scores they received?

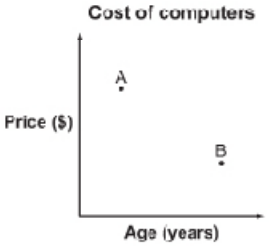
| | | |
|---|---|---|
| A. Joe: | 62, 82, 75 | Mary: 69, 78, 84 |
| B. Joe: | 78, 76, 74 | Mary: 72, 82, 74 |
| C. Joe: | 77, 77, 69 | Mary: 77, 67, 77 |
| D. Joe: | 65, 77, 83 | Mary: 71, 77, 86 |

*Figure 3*. Revised TIMSS item.

The Australian National Assessment Program – Numeracy and Literacy (NAPLAN) has a statement of minimum expectations at each grade level assessed. For Grade 9 students there is some acknowledgement of the interpretive goals for statistical understanding with the statement "students at the *minimum standard* at Year 9 can … summarise sample data from a population and make informal inferences in response to questions and hypotheses." (Australian Curriculum, Assessment and Reporting Authority (ACARA), 2010). A typical item illustrates the expectation and this is shown in Figure 4. The extent to which this "representative item" addresses the stated standard is arguable. The item requires some level of graph reading skill but there is little inference required and, in this example, no need to summarise any data. In relation to NAPLAN items, Nisbet (2010) commented "Overall, the data-handling test items focus on a limited range of knowledge and skills compared to what is expected to be taught in schools".

A shop sells new and used computers.

The graph shows the price of 2 similar computers and their age in years.




Which one of these statements is true?

- Computer B is older and less expensive than computer A.
- Computer A is newer and less expensive than computer B.
- Computer A is older and more expensive than computer B.
- Computer B is newer and more expensive than computer A.

*Figure 4*. Representative example of a statistics item from Grade 9 Numeracy test from http://www.naplan.edu.au/verve/_resources/Numeracy_Year_9_4.pdf

A Grade 9 NAPLAN item that considers summary data is shown in Figure 5. In this question, however, little more than counting is required if the meaning of the word "median" is understood. Students are not required to summarise data for themselves. Again it addresses only procedural skills and there is no attempt in any of the items shown to consider the meanings of these measures of central tendency. In addition, there is no acknowledgement of any use of technology with all items being part of calculator-free tests.

In a gym class, 29 students took turns jumping.

Pete recorded the height each student jumped.

Height (cm)

| | |
|---|---|
| 3 | 2 4 |
| 4 | 1 5 6 |
| 5 | 2 4 4 8 9 |
| 6 | 1 1 3 4 5 6 6 8 9 |
| 7 | 2 2 5 7 8 |
| 8 | 3 5 5 |
| 9 | 1 2 |

What is the median height?

A) 63 cm B) 64 cm C) 65 cm D) 66 cm

*Figure 5.* Grade 9 NAPLAN item.

At this point in time, it seems that large-scale assessment of the kind common across education systems is limited to procedural knowledge and lacks the capacity to address the kinds of statistical understandings that students are developing through the use of technology shown in the Lucy Outlier example (Figure 1). Technology use by students goes largely overlooked and the items used are very similar to those traditionally found in text books. The most likely reason for the lack of deep conceptual knowledge addressed is that external, large-scale tests are required to be cost effective (Nisbet, 2010).

Developing conceptual items may require more training of item writers, improved trialing or paneling with experts in a particular field. All of these alternatives need more time, and hence more money. The ARTIST resource (Garfield, del Mas & Chance, 2006), however, demonstrates that quality multiple-choice questions addressing higher-order goals for statistical education are possible. In addition, in large-scale assessment, of necessity facilities are limited to those most likely to be available. If students are to be assed using online resources, such as "Oxygen" or using technology to produce answers, then all students must have the same familiarity with the technological tools.

In contrast to the lower level items described in the previous section, consider the item called "Oxygen" from World Class Tests (World Class Arena, 2009) shown in Figure 6 and aimed at the same age group. To answer this correctly, students interact with the graph using a slider to manipulate two different variables to identify their effect on the rate of oxygen production. They answer a sequence of carefully designed questions designed to test their ability to make inferences based on data. Responses are collected in hard copy and scored using a scoring scheme.

This approach provides students the opportunity to manipulate multivariate data in a context that is relevant to them in light of discussions about climate change. It makes use of the power of technology to provide questions that are challenging and that address higher-order thinking, in line with Garfield and Gal's (1999) goals. The item has to be scored manually, however, which is a serious drawback in the context of large-scale testing because of the expense incurred.

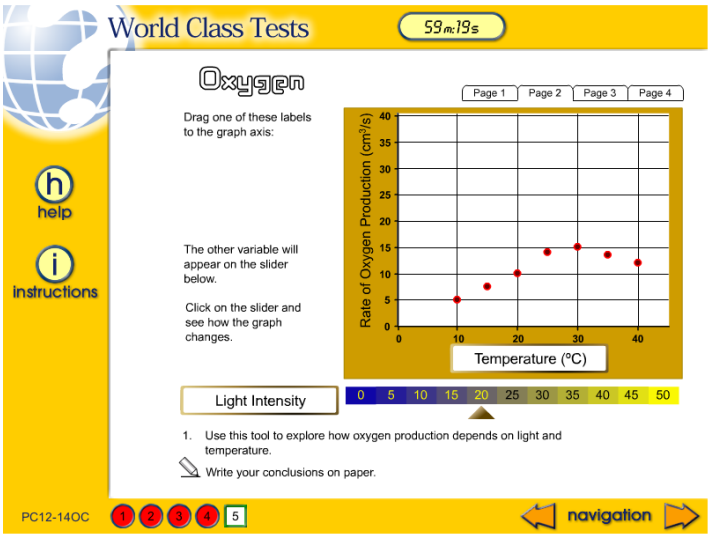



*Figure 6.* Oxygen item from World Class Arena text for 12-14 year-old students.

Of interest in the context of this discussion is that the TIMMS and NAPLAN items (Figures 2, 4 and 5) were designed to address the curriculum, whereas the World Class Arena Oxygen item (Figure 6) was not constrained by curriculum expectations. World Class Arena designers are freer to draw on technology to identify what it is possible for students to do rather than assess what is expected of students of a particular age. This observation raises issues about curriculum design, especially in an information-rich age where technology use enhances the development of statistical thinking. It may be that the curriculum itself limits teachers' potential to use technology for teaching and learning through low expectations of students' capacities, and that this limitation is reinforced by machine-scored assessment imposed by external agencies, setting up a self-perpetuating cycle.

Anecdotal evidence suggests that there is an additional unreported issue, at least in Australia. When NAPLAN data are returned at the school level the results are reported by curriculum strand. In general, students perform well on the statistics and probability aspects of the test and schools, as a result,

tend to focus their attention on those aspects of the curriculum where students are not achieving. These results, however, are based on items such as those shown in Figures 4 and 5, which test lower level procedural skills. Students can demonstrate much deeper understanding, such as shown in the Lucy Outlier example, but their statistical development will not be enhanced unless schools deliberately address the higher order goals. Such shifts in classroom practice are unlikely to occur while schools believe that their students perform well on statistics as measured in external assessments.

Another issue concerns the very different nature of the statistics assessment items shown here. In terms of mathematics tests, where items such as these are generally located, there is good evidence that despite the apparently different thinking skills that mathematics items draw on in the different domains, such as statistics or geometry, the underlying construct of mathematics ability can be treated as a single dimension (Burg, 2008). When items make use of the power of technology, however, such as the Oxygen item, it is not clear whether they are measuring an identical construct to those items delivered in a traditional test form, such as the TIMSS and NAPLAN items. It is possible that technology enhanced tests could be measuring two different dimensions of understanding statistics, or two totally unrelated constructs. Wilhelm and Schroeders (2008) indicated that equivalence testing, to determine whether tests delivered using different modes are measuring the same construct in the same way, is important in high stakes examinations where part or all of the examination is delivered in different formats. Often the delivery affects the results suggesting that the two formats, technology-based or paper-and-pencil machine scored, are likely to measure something slightly different. Wilhelm and Schroeders' study, however, was using technology such as smartphones. Few studies have been undertaken of the equivalence of instruments such as the examples indicated, and these are urgently needed, particularly in light of Lesh's (2000) comments.

Further complexity is added where computer adaptive testing is used. In this system, students answer questions and, on the basis of their responses, the computer provides the next item according to the ability of the student. These systems rely on having a large bank of validated items and are usually underpinned by Rasch measurement (Bond & Fox, 2007). One such system is the Lexile or Quantile Frameworks (Stenner, 2009) which automatically produce appropriate items for students and then place them onto a validated scale. At present, however, these systems rely on multiple choice questions similar to the TIMSS item shown in Figure 2. Systems are being developed, however, to score automatically open-ended written works, such as essays. These systems rely on sentence length and complexity, rather than a true analysis of the ideas presented, and, as such, would appear to have limited use at present for statistics education, where subtle differences in interpretation may indicate widely differing understanding. For example, when responding to a question "what does random mean", two responses made by students are "Picked without order" and "To just pick anything". The first of these answers is more sophisticated in its thinking but uses very much the same kind of language. At the present state of development most computer software would not have enough information on the basis of these two answers to distinguish them, but a trained rater can identify such fine distinctions. It remains to be seen whether these kinds of systems can ever be used for the types of short answer questions sometimes used to assess statistics, although there may be some possibility of using them to assess a statistical report, for example. The greatest limitation in this respect is the fact that systems such as the Lexile Framework rely on text analysis and do not appear to be able, at this point in time, to consider representations such as graphs or data tables. Maybe this will remain impossible.

In terms of technology, large-scale externally imposed assessment of statistical understanding seems to have some way to go. At present the predominant use of technology is to machine score items that address the lower level skills indicated by Garfield and Gal (1999). There appears to be no consideration of the impact of technology use on students' statistical understanding.

# DISCUSSION

Ben-Zvi (2000) indicated ways in which technology impacts specifically on statistics learning, echoing Lesh's (2000) view that technology is an agent for cognitive change. Garfield and Chance (2000)

listed some of the challenges in assessment for statistic educators and this list included the need to embrace the use of new technological tools. As long ago as 1997, Joliffe indicated that technology use must be considered when assessment is being developed. The increasing impact of technology on teaching and learning statistics, however, has not been matched by developments in assessment to the same degree. Many of the exemplary approaches, including World Class Arena and ARTIST tools, are used by limited numbers of educators for specific purposes but have not made a serious impact on mainstream assessment.

Assessment must provide effective feedback to students and teachers to improve learning outcomes (Black & Wiliam, 1998). Technology holds the promise of doing this in immediate and effective ways. To realize this potential, however, requires creative thinking on the part of teachers and systems. In addition, statistics teaching itself needs to have more emphasis not only in the mathematics curriculum, but in cross-curriculum contexts as well. At the classroom level it seems possible to achieve the goals of Garfield and Gal (1999) with the use of rich assessment *as* learning (Earl, 2003) tasks supported by rubrics or scoring guides based on hierarchical or developmental frameworks. The key issues would appear to be appropriate professional learning for teachers.

At the systems level, however, there are greater challenges for statistics education. Curriculum designers need to take heed of current research that students supported by appropriate learning technology can develop relatively sophisticated understanding of statistical ideas. They can draw informal inferences which become refined as they progress through school (Franklin et al, 2007; Watson & Callingham, 2003). As yet the tools to measure such understanding are not available for large-scale, machine-scored use, and until they are widely used, systems are unlikely to make the curriculum changes needed because current assessments indicate that students are performing well on statistics items.

In addition, serious thought needs to be given to Lesh's (2000) and Ben Zvi's (2000) contentions that using technology fundamentally changes the nature of the cognitive constructs developed. If this is so, then using traditional assessment approaches could disadvantage students who are increasingly learning via technology use. The nature of the changes to cognition needs to be identified and this is another area that needs further research. Given the claim that communication and interpretation demands are much greater in a technology-rich environment (Lesh, 2000), there is an urgent need for statistics educators at the school level to consider the different cognitive load and change teaching and assessment accordingly.

The challenges faced in assessing statistical understanding in a technological age are not trivial. It is a more than a decade since Garfield and Gal (1999) called for changes to statistics education and indicated the challenges for assessment that such changes would pose. During that period technology use has impacted on classrooms in a variety of ways but assessment of statistical understanding has not changed to match the new teaching developments. Perhaps it is time to reconsider the assessment challenge and develop new approaches that take account of both technology use by students and the power of technology to deliver quality assessment.

# REFERENCES


Australian Curriculum, Assessment and Reporting Authority (ACARA) (2010), "*Numeracy, Year 9 National Minimum Standards*," Sydney: Author. Accessed 12 December 2010 from http://www.naplan.edu.au/numeracy_year_9_national_minimum_standards.html

Ben-Zvi, D. (2000), "Toward Understanding the Role of Technological Tools in Statistical Learning," *Mathematical Thinking and Learning, 2*(1), 127-155.

Black, P., & Wiliam, D. (1998), "Assessment and Classroom Learning," *Assessment in Education*, March, 7-74.

Bond, T. G. & Fox, C. M. (2007), "*Applying the Rasch model. Fundamental measurement in the human sciences,"* (2nd Edition,) Mahwah, NJ: Lawrence Erlbaum.

Burg, S. (2008), "*An Investigation Of Dimensionality Across Grade Levels and Effects on Vertical Linking for Elementary Grade Mathematics Achievement Tests*," Paper presented at the annual meeting of National Council of Measurement in Education, New York City. Accessed 29 December from http://www.quantiles.com/Attachments/Burg%20NCME%202008.pdf

Callingham R. & Watson, J. M. (2005), "Measuring Statistical Literacy," *Journal of Applied Measurement, 6 (1)*, 29, 19-47.

Carmichael, C. & Hay, I. (2010), "Developmental Changes in Australian School Students' Interest for Statistical Literacy," in C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the 8th International Conference on the Teaching of Statistics*, *Ljubljana, Slovenia*, July, Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications.php?show=icots8

Chance, B. Ben-Zvi, D., Garfield, J. & Medina, E. (2007). "The Role of Technology in Improving Student Learning of Statistics," *Technology Innovations in Statistics Education, 1*(1) (26pp). Retrieved from http://escholarship.org/uc/item/8sd2t4rr

Earl, L. M. (2003), "*Assessment AS learning: Using Classroom Assessment to Maximise Student Learning*," Thousand Oaks, CA: Sage Publications.

Fitzallen, N. (2008). "Validation of nn an Assessment Instrument Developed for Eliciting Student Prior Learning in Graphing and Data Analysis," in M. Goos, R. Brown, & K. Makar (Eds.). *Proceedings of the 31st Annual Conference of the Mathematics Education Research Group of Australasia*. Brisbane: MERGA.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). "*Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-k–12 Curriculum Framework*," Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/education/gaise/

Garfield, J. B., & Ben-Zvi, D. (2008). "Assessment in statistics education (ch. 4)," in J.B. Garfield, D. Ben-Zvi, B. Chance, E. Medina, C. Roseth, & A. Zieffler (Eds.). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer.

Garfield, J., & Chance, B. (2000), "Assessment in Statistics Education: Issues and Challenges," *Mathematical Thinking and Learning, 2*(1), 99-125.

Garfield, J., & Gal, I. (1999), "Assessment and Statistics Education: Current Challenges and Directions," *International Statistical Review*, 67, 1–12.

Garfield, J., del Mas R., & Chance, B. (2006), "*Assessment Resource Tools for Improving Statistical Thinking (ARTIST),"* [website]. Retrieved from https://app.gen.umn.edu/artist/index.html

Gil, E. & Ben-Zvi, D. (2010), "Emergence of Reasoning about Sampling among Young Students in the Context of Informal Inferential Reasoning," in C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society*. *Proceedings of the 8th International Conference on the Teaching of Statistics, Ljubljana, Slovenia*, Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications.php?show=icots8

Jolliffe, F. (1997), "Issues in Constructing Instruments," in I. Gal & J.B. Garfield (Eds.) *The assessment challenge in statistics education* (pp. 191-204), Amsterdam: IOS Press & International Statistical Institute.

Konold, C., & Miller, C.D. (2005). "*TinkerPlots: Dynamic data exploration*," [Computer software] Emeryville, CA: Key Curriculum Press.

Lajoie, S.P. (1997). "Technologies for Assessing and Extending Statistical Learning," In I. Gal & J.B. Garfield (Eds.) *The Assessment Challenge in Statistics Education* (pp. 179-190), Amsterdam: IOS Press & International Statistical Institute.

Lesh, R. (2000). "Beyond Constructivism: Identifying Mathematical Abilities that are Most Needed for Success Beyond School in the Age of Information," *Mathematics Education Research Journal, 12*, 3, 177-195.

Lipson, K., Francis, G., & Kokonis S. (2006), "Developing a Computer Interaction to Enhance Student Understanding in Statistical Inference," in A. Rossman & B. Chance (Eds.) *Working cooperatively in statistics education. Proceedings of the 7th international conference on teaching statistics*. (CD ROM), Salvador, Bahia, Brazil: International Association for Statistical Education and International Statistical Institute.

Ministry of Education (2007), "*Draft Mathematics and Statistics Curriculum,"* Wellington, NZ: Author.

National Council of Teachers of Mathematics, (2000), "*Principles and Standards for School Mathematics*," Reston, VA: Author.

Nisbet, S. (2010). "National Testing of Data Handling in Years 3, 5 and 7 in Australia." In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS 8, July, 2010). Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications.php?show=icots8

Pellegrino, J. (winter 2002-3). "Knowing what students know", *Science and Technology, XIX*(2), 48-52.

Royal Statistical Society Centre for Statistical Education (2010), "*CensusAtSchool International"* [website], accessed 21 December 2010 at http://www.censusatschool.com/

Sadler, R. (1998), "Formative Assessment: Revisiting the Territory," *Assessment in Education: Policies, Principles and Practice, 5*, 77-84.

Stenner, A. J. (2009), "*Lexile framework*," plenary address to the Pacific Rim Objective Measurement Symposium (PROMS), July 28 2009, Institute of Education, Hong Kong.

TIMSS and PIRLS International Study Centre (2009), "*TIMSS Publically Released Items*," Boston: Author. Accessed 29 October 2009 from http://timssandpirls.bc.edu/

Watson, J.M. (1997), "Assessing Statistical Thinking," in I. Gal & J.B. Garfield (Eds.) *The Assessment Challenge in Statistics Education* (pp. 107-121), Amsterdam: IOS Press & International Statistical Institute.

Watson, J.M., & Callingham, R.A. (2003), "Statistical Literacy: A Complex Hierarchical Construct," *Statistics Education Research Journal, 2*(2), 3-46.

Watson, J.M., Callingham, R. & Donne, J. (2008), "Establishing Pedagogical Content Knowledge for Teaching Statistics," in C. Batanero, G. Burrill, C. Reading & A. Rossman (Eds.) *Challenges for Teaching and Teacher Education.* (Proceedings of the ICMI Study 18 and IASE 2008 Round Table Conference). Monterrey, Mexico: ICMI, IASE, ISI.

Wilhelm, O., & Schroeders, U. (2008, September), "*Traditional and Computerized Ability Measurement: Stressing Equivalence Versus Exploiting Opportunities*," International Research Workshop on the transition to computer-based assessment. Lessons learned from the PISA 2006 computer based assessment of science (CBAS) and implications for large scale testing. Reykjavik, 29 September–1 October.

World Class Arena (2009), "'*Oxygen', Interactive Problem Solving Item"*, London: GL Assessment. Accessed 29 October 2009 from http://testwise.worldclassarena.org/testwise/ondemand/welcome.html