Lowrie, Tom, Diezmann, Carmel M., & Kay, Russell (2011) The development of the graphics-decoding proficiency instrument. *Evaluation and Research in Education*, *24*(4), pp. 285-296.

The Development of the Graphics-Decoding Proficiency Instrument

Tom Lowrie

Research Institute for Professional Practice, Learning and Education,
Charles Sturt University, Wagga Wagga, Australia

Carmel M. Diezmann

Centre for Learning Innovation,
Queensland University of Technology, Australia

Russell Kay

Research Institute for Professional Practice, Learning and Education,
Charles Sturt University, Wagga Wagga, Australia

Corresponding author:
Professor Tom Lowrie
Charles Sturt University, Locked Bag 588, Wagga Wagga NSW 2678, AUSTRALIA
Electronic mail: tlowrie@csu.edu.au
Telephone: +61 2 6933 4328
Facsimile: +61 2 6933 2962

Notes on contributors

Professor Tom Lowrie is the Director of the Research Institute for Professional Practice, Learning and Education (RIPPLE) at Charles Sturt University. Professor Lowrie's recent research investigates the extent to which young children use spatial reasoning and visual imagery to solve mathematics problems.

Professor Carmel Diezmann is interested in how students learn to interpret the graphics embedded in mathematics assessment items and how this interpretation affects their performance. Professor Diezmann is the Assistant Dean (Research) in the Faculty of Education at Queensland University of Technology, Australia.

Russell Kay is an Adjunct Senior Lecturer in the Faculty of Education at Charles Sturt University. He has particular interests in the use of quantitative research methods and his recent research has been in the areas of student achievement and lecturer self-efficacy.

The Development of the Graphics-Decoding Proficiency Instrument

The Graphics-Decoding Proficiency (G-DP) instrument was developed as a screening test for the purpose of measuring students' (aged 8-11 years) capacity to solve graphics-based mathematics tasks. These tasks include number lines, column graphs, maps and pie charts. The instrument was developed within a theoretical framework which highlights the various types of information graphics commonly presented to students in large-scale national and international assessments. The instrument provides researchers, classroom teachers and test designers with an assessment tool which measures students' graphics decoding proficiency across and within five broad categories of information graphics. The instrument has implications for a number of stakeholders in an era where graphics have become an increasingly important way of representing information.

Keywords: information graphics, assessment items, assessment instrument, mathematics

Context

For many children in today's society, visual representations including graphic displays are common place—both in school contexts and out-of-school settings. As school systems attempt to provide learning opportunities for students to develop skills and processes that equip them to function in society, visual-and spatial-reasoning becomes increasingly important and valued. In recent years, there has been a considerable shift in the way students' knowledge is being assessed in schools (Lowrie & Diezmann, 2009), with an increased use of graphics in these assessment contexts. Graphics typically include graphs, maps, diagrams, charts, networks, and according to Bertin (1967/1983) are seen as visual representations for 'storing, understanding and communicating essential information' (p. 2).

Not only is there more graphical information in typical mathematics assessment items but the level of visual-spatial detail in these items has also increased. This is due, in part, to the fact that the graphics in tests and assessment instruments can be cost effectively produced in colour and with visual detail not able to be achieved in the recent past. As a result, students are required to interpret graphical displays with detailed patterns, shading and other retinal properties (Bertin, 1967/1983). However, this increase in detail can heighten the demand on the interpreter of the graphic. As Roth (2002) argued, in a visually-oriented society, greater attention must be given to the practices of reading, producing and understanding graphical representations. In terms of assessment, young students are less likely to encounter word-based problems in mathematics tests and are more likely to be required to decode tasks which have high graphics demand (Lowrie & Diezmann, 2009).

Graphical reasoning, and therefore the capacity to decode information in a task, is influenced by the nature of the task, the properties of the graphic, and the skills and experiences brought to the task (Brna, Cox, & Good, 2001). Kosslyn (2006) maintained that the graphical composition of a task included not only the actual graphic but also the information embedded within the task. For young students, the capacity to decode text-and graphic-based information is challenging in its own right, but as graphics become more detailed, additional demands are placed on students' decoding skills (Lowrie & Diezmann, 2011). Studies by Hittleman (1985) and Carpenter and Shah (1998) have shown that students find it challenging to move between text and graphics to the extent that it can disturb their thinking. Indeed, the graphic can often make the task more difficult to decode (Berends & van Lieshout, 2009; Elia, Gagatsis, & Demetriou, 2007; Schmidt-Weigand, Kohnert, & Glowalla, 2010). Hence, further research on how young students decode graphics is necessary, especially given the increased attention such tasks are afforded in curricula and testing

(Lowrie & Diezmann, 2009). Of particular concern is the large number of graphics used in everyday mathematical situations each with specific conventions of use (Harris, 1996).

The scope of the project

Background

This paper describes the development of an instrument designed to determine 8 to 11 year old students' proficiency in decoding graphics represented in mathematics tasks. The purpose of the instrument is to provide educators with an assessment tool which identifies students' level of proficiency when decoding specific types of graphics items. Specifically, the instrument provides classroom teachers with a tool which can be administered in either whole-class or individual settings.

Although a similar instrument has been developed for older children (see Diezmann & Lowrie, 2009a), few mathematics instruments are available for students of this age group (8-11 years). Moreover, to our knowledge, no other instruments have specifically included all types of graphics representations within a single instrument. In this sense the instrument serves as a screening device designed to identify overall strengths or weaknesses across all types of graphical representations and mark proficiencies or deficiencies in particular areas or domains. A graphical processing instrument is necessary given the fundamental changes in the way mathematics tasks are now presented and represented in curricula and test design at both national and international levels. Such an instrument also needs to consider that various types of graphics items have different structure, composition, balance and intent (Bertin, 1967/1983), both in terms of the actual graphic, and in relation to the graphic as part of an assessment item. As Kosslyn (2006) recognised, the actual graphic contains particular design features which help convey (and potentially hinder) information. Thus an effective screening test needs to incorporate a range of different graphical structures in order to comprehensively assess student decoding proficiency.

Theoretical framework

The theoretical framework which underpins the development of the instrument is based on the work of Bertin (1967/1983), who describes graphics in terms of information within the graphic, the properties of the system, and the underlying components that govern and combine these properties. One of the most important aspects of Bertin's work was the identification of the nature of graphics and the extent to which particular types of graphics differed from one another. Mackinlay (1999) built on aspects of Bertin's work by making more explicit the extent to which graphics could be classified in relation to the perceptual elements they contain, and the encoding techniques required to construct the graphic. Mackinlay classified graphics within six 'graphical languages', which make specific reference to the structure of the graphics in terms of how the graphic is encoded.

These six graphical languages are named Axis, Apposed-position, Map, Retinal-list, Connection, and Miscellaneous, respectively. Axis language items include line graphics represented on either a horizontal or vertical axis and typically include number lines. Apposed-position graphics encode information on two axes and often include line graphs or column graphs. Map language graphics are similar to Apposed-position items; however they have information encoded through the spatial location of marks. These items typically include road maps or topographic maps. In mathematics education contexts, students are expected to develop their understanding of maps by interpreting common grids and compass directions, and by following and providing directions for movement within and around map contexts. By contrast, Retinal-list graphics are not dependent on position or axis structure. These items often require the decoding of information through the translation or rotation of objects. Connection language encodes information by connecting a set of node objects with a set of link objects. This graphical language requires the interpretation of information in relation to connections and links between sets of objects (e.g., family trees and tennis draws). Due to the lack of Connection items in published tests, we included these items as part of the Miscellaneous language. This language encodes information with a variety of additional graphical techniques. These graphical languages include specifically tailored graphics such as pie charts and Venn diagrams. See the appendix for an example of the five categories of graphical languages.

With colleagues, Lowrie and Diezmann (e.g., Diezmann & Lowrie, 2009b; Lowrie & Diezmann, 2007) have incorporated Mackinlay's framework into mathematics education contexts suitable for primary-aged students. The rationale for the application of this framework was to provide an understanding of how students decode and process information graphics within mathematics teaching and assessment contexts. These studies have revealed distinct differences between males and females (in favour of males) on Axis and Map languages (Lowrie & Diezmann, 2011) and the extent to which the orientation of a graphic influences student performance (e.g., Lowrie, Diezmann & Logan, 2009). The current study extends this work through the establishment of an instrument which is suitable for younger students.

Method

Participants

A pilot instrument was administered to 273 students (M = 135, F = 138) from four primary schools throughout one of Australia's largest cities, in what would typically be described as middle-class metropolitan suburbs (less than 5% of the participants had English as a second language). Students' age ranged from 8 to 11 years (Median age = 9.4, Grades 3-5). The schools were randomly chosen from a convenience sample within a regional cluster of schools situated a practical distance from the University. The sample size represented 26% of the total population size of the cluster.

Procedure and Initial Analysis

The items were selected from state, national and international year-level mathematics tests that had been administered to students in their early years of primary school or to similarly aged students (e.g., Queensland School Curriculum Council, 2000a). Since all items for the instrument were sourced from high-stakes mathematics assessment instruments which were being used for national and international benchmarking, criterion validity for the respective items had been previously established. Moreover, the items within these tests were selected from age appropriate instruments and thus content validity was also achieved. In terms of construct validity, a panel of expert mathematics educators (N = 5) independently categorised 175 items within Mackinlay's (1999) graphical languages framework based on item structure and composition (reliability coefficient Cronbach's alpha 0.90 was achieved).

The research team approached schools in a region (randomly selected for investigation) with information about the nature of the study and the specific requirements for participation. Once ethics approval had been given at both University and Education Department levels, information packages were sent to parents or guardians via the schools. All participants had written consent from a parent or guardian to take part in the study, with students also completing a participant consent form. Researchers administered the instrument to primary-aged children (N = 273) in whole-class situations in the presence of the classroom teacher. The researchers administered the instruction protocol orally and explained the nature of the study. Participants were given one hour to complete the instrument and this was sufficient time for all students.

Descriptive, bivariate and Guttman scaling (Kline, 2005) analyses were undertaken to reduce the item pool. This was necessary to ensure that the items varied in complexity and required substantial levels of graphical decoding. This analysis sought to 1) identify items of varying difficulty; 2) produce items with moderate to high *within* language correlations; and 3) ensure that hybrid items, with moderate to high *across* language correlations, were avoided. These procedures resulted in a bank of 40 items (8 items from each of 5 graphical languages).

Results

Exploratory Factor Analyses

Five Exploratory Factor Analyses (EFA) were undertaken (on items within each of the respective languages) using tetrachoric correlation matrices. This procedure was undertaken to ensure that the sample size to number of variables ratio was adhered to (MacCallum, Widaman, Zhang, & Hong, 1999). The ratio of 34:1 for sample size to variable is well above the recommended 20:1 ratio. In addition, the sample size of 273 more than adequate given the number of factors produced. Since dichotomous (i.e., correct and incorrect) measures were used, analyses were conducted with tetrachoric correlations as they were more appropriate than Pearson or Spearman coefficients (Kaplan & Saccuzzo, 2008). Moreover, tetrachoric correlations overcome problems caused when items of differing difficulty are correlated. These correlations were produced using Systat 16 whilst the EFAs were conducted with SPSS 17. The principal component extraction method, with Oblimin rotation, resulted in the reduction of the original 40 items to 25 items. Thus, the instrument now comprised 25 items with five items in each of the five graphical language categories (see Appendix for an example of each of the five languages). This reduction was undertaken to select the items which were most closely related thus, the maximum amount of shared variance of items within each language was retained (see Table 1).

INSERT TABLE 1 ABOUT HERE

Given the different difficulty of items, it was necessary that their individual contributions reflect their importance in calculating the overall score. In order to achieve this outcome, items where weighted using principal component factor loadings to calculate aggregated scores within each language. The correlations between the aggregated language scores are displayed in Table 2. The distributions of the scores were within normal skewness and kurtosis values desirable for further analysis (since they fell within the -1 to +1 range, see Tabachnick & Fidell, 2001 for further explanation).

There were moderate correlations among most of the aggregated language measures. The Miscellaneous language was the most highly correlated with the other four languages. The Apposed-position items were moderately correlated with the Miscellaneous items and less so with the other three language items.

The means and standard deviations for student performance on each graphical language across grade are presented in Table 3. As anticipated due to maturation and experience factors, mean scores increased at each grade level for each of the five graphical languages. Mean score increases across grade (i.e., between Grade 3 and Grade 4 or Grade 4 and Grade 5) and language were relatively stable with most increases ranging from 18% to 27%. The three outlying mean score increases occurred for the Axis (32%), Apposed-position (11%) and Retinal (10%) between Grades 3 and 4.

INSERT TABLES 2 AND 3 ABOUT HERE

Further Analysis of the 25 Items

Descriptive analysis was undertaken to determine the hierarchy of difficulty for the five items within each language. Data concerning the proportion of correct responses (see Table 1) indicated a varying degree of difficulty across the five items within language, for all the graphical languages. The scaling properties for items within each of the five languages were assessed by using procedures derived for Guttman scaling (see Bernard, 2000) in order to establish the strength of the hierarchical scale within language. The coefficient of reproducibility is produced by calculating the number of instances a student correctly answers a more difficult item after failing to correctly solve a less difficult item. The proportion of these "errors" is then subtracted from 1 to yield the coefficient. The coefficients for each language on the combined data were Axis = 0.91; Apposed-position = 0.90; Retinal-list = 0.87; Map = 0.93; and Miscellaneous = 0.90 respectively. Guttman indicated that coefficients of 0.90 are desirable for establishing a hierarchy among items. Hence, within each language there is a sequence of five items of increasing difficulty which are largely predictive of future performance in that language. These results confirm the applicability of the instrument as a screening tool.

Cross-Validating the Instrument by use of Confirmatory Factor Analysis

A Confirmatory Factor Analysis (CFA) was conducted with those participants (N=199, 73% of the original cohort) who had completed each of the 25 items from the instrument in order to have a complete data set (i.e., all participants completed all questions). It was hypothesized that a single congeneric measurement model with one latent variable could adequately fit these data. This assumption is based on the fact that Mackinlay's (1999) theoretical framework describes a "**set** of primitive graphical languages" (p.74, emphasis added) which encompasses a collective set of graphics based representations. Using different scales (languages), the congeneric model assumes that each observed variable measures the same latent variable and is the least restrictive model for reliability estimation (Graham, 2006). Consequently, structural equation modeling (SEM) was undertaken to explore the relationship between decoding performance and performance across graphical languages. The fit indices (N=199; χ^2=5.45; df=5; p=.365; χ^2/df= 1.09 (p=.62); CFI= .99; NNFI = .98; RMSEA=.02 [90%CI=.00-.10]) for the model yielded a very good fit to the data.

In an evaluation of the model (also see Figure 1), the chi-square statistic associated with the p value, the comparative fit index (CFI), the nonnormed fit index (NNFI), and the root-mean-square error of approximation (RMSEA) are reported. The nonsignificant value of the chi-square statistic (p=.365) indicates a good fit; however, in order to test the sensitivity to sample size, a χ^2/df of less than 2 (Maruyama, 1998) should also be obtained. The strength of the model is also reinforced when CFI and NNFI indices are greater than .90 (Hoe, 2008)—both are above .98 in this model. Finally, the RMSEA value (.02) is less than the .05 recommended (Yuan, 2005) to ensure an absolute fit index for the model. Consequently, all GFI statistics indicate a good model fit (also see Table 4).

INSERT TABLE 4 ABOUT HERE

In terms of the five graphical language categories, the Miscellaneous language provided the most shared variance to the model (52%). Three of the other four categories, namely Axis, Retinal-list and Map, contributed similarly, with more than 40% shared variance. The Apposed-position language items contributed only 34% of the shared variance.

INSERT FIGURE 1 ABOUT HERE

Discussion

The main goal of this study was to develop an instrument which could serve as a screening measure for lower to middle primary-age students across several domains of graphics-decoding proficiency. The 25-item instrument comprised five items from each of the five graphical languages most commonly found in school mathematics. The component loadings from the EFA were strong for almost all of the 25 items, apart from one item in each of the Apposed-position and Retinal languages. The shared variance of the items for the respective languages ranged from 50% for the Axis language to only 35% for the Apposed-position language. Given the different representation (including different perspectives) and orientation (including vertical and horizontal orientations) of items within the five graphic-language structures, these loading proportions are both acceptable and indicate a good fit for 23 of the 25 items and validated our conceptual thinking regarding our graphical framework. Nevertheless, it is noteworthy that the five Apposed-position items contribute 35% of the variance to the language structure. Items within this language include various types of graphs (including line, bar and column graphs). Elsewhere (Lowrie & Diezmann, 2005), it is argued that relationships between items within this language are, at best, moderate since questions measure such a range of concepts. In this sense, these results provide further evidence that this language can be viewed as more varied and amorphous than the Axis, Map and Retinal domains. The values for R squared in Figure 1 indicate the complexity of these relationships so that whilst the languages may share and underlying commonality, they are clearly distinctive both conceptually and empirically.

The CFA tested the degree to which a single congeneric model of graphical ability could be seen as underlying performances in five language areas. From a conceptual standpoint, the results of the structural equation modelling indicate that lower – middle primary-aged students' graphics decoding proficiency comprises a number of interrelated decoding measures. These five graphical languages provide strong evidence of the interconnectedness of decoding ability within general graphics understanding.

The instrument can be used as a diagnostic measure of student *graphics decoding proficiency* in mathematics tasks. Thus, it can be used as a screening device to identify strengths and weaknesses of individual students across graphical domains. In addition, items from the instrument can be used to elicit students' sense making through one-to-one interviews. We suggest that the bank of items (i.e., the five items in a language) should be used to measure an individual's understanding of a graphical language; however, the instrument is not intended to be used as a general mathematics assessment tool, nor to provide precise measures of achievement in the particular content domains. Noteworthy, is the utility of the instrument which provides a means of rapid, diagnostic assessment across a range of graphical domains.

The Graphics-Decoding Proficiency (G-DP) instrument has the potential to provide a number of educational benefits for classroom teachers and test designers. The instrument provides teachers with the ability to monitor their students' performance and particularly to identify areas of weakness in students' decoding proficiency (in terms of graphical languages) rather than only recognising students' content-based knowledge which is typically identified in assessment tools. Therefore, the instrument along with appropriately tailored, one-on-one interviews provides insights into why and how individual students might perform seemingly erratically on graphically-rich mathematics tasks with similar mathematics content. This instrument has the added potential to identify the skills students might need to fully benefit from Learning Objects (software) that use dynamic and multiple graphics since all dynamic representations of graphics use a combination of the graphics representations presented in this instrument.

Conclusion

The results of the study provide insights into how students perform on tasks commonly found in national and international mass testing contexts. The Graphics-Decoding Proficiency instrument meets the original aims of the study with the statistical analysis supporting the robustness of the instrument. We acknowledge that the number of items in the instrument (i.e., 24) may lead to fatigue for some students, however, the separateness of the five languages within the instrument allow for administration to be undertaken over shorter time periods. In the future there is also scope for the items to be constructed in a digital form and to increase the sample to include more dynamic representations—we appreciate that the instrument can only be administered in a pencil-and-paper form at present. Future research on instrument development should broaden the instrument's scope to include items that can be implemented with younger and older students since testing practices across the globe are requiring students to complete tests that have high levels of graphics.

Acknowledgements

This research is supported by Australian Research Council funding (#DP0880751). Special thanks to Kim Woodland and Tracy Logan who assisted with the interpretation of data for this project and to Lindy Sugars for assistance with data collection.

Table 1.
Principal component analysis for items within languages.

Language	% shared variance	Item	Proportion correct	Component loading
Axis	49.99%	81	.83	.71
		76	.81	.64
		166	.58	.78
		46	.54	.70
		92	.50	.70
Apposed-position	35.45%	72	.84	.74
		129	.80	.26
		120	.79	.57
		149	.56	.77
		157	.45	.49
Retinal-list	38.17%	83	.68	.62
		73	.58	.57
		144	.57	.32
		8	.49	.68
		45	.44	.80
Map	46.39%	63	.85	.63
		145	.80	.76
		148	.71	.63
		117	.69	.81
		90	.49	.55
Miscellaneous	45.97%	96	.86	.76
		40	.65	.57
		150	.57	.68
		153	.54	.73
		164	.50	.64

Table 2.
Correlation matrix for the Aggregated Language Measures (N = 199).

Measures	1	2	3	4	5
1. Axis	1				
2. Apposed	.35*	1			
3. Retinal	.49*	.37*	1		
4. Map	.37*	.38*	.42*	1	
5. Misc.	.45*	.44*	.44*	.49*	1

*p<.01 (2-tailed)

Table 3.
Means (Standard Deviations) of students' performance on examples of the five graphical languages.

Graphical language	Grade 3 (n = 58)	Grade 4 (n = 58)	Grade 5 (n = 83)
Axis	1.75 (0.95)	2.32 (0.90)	2.76 (0.86)
Apposed-position	1.67 (0.67)	1.86 (0.52)	2.25 (0.62)
Retinal	1.40 (0.82)	1.54 (0.85)	1.96 (0.78)
Map	1.97 (0.89)	2.40 (0.86)	2.86 (0.67)
Miscellaneous	1.73 (0.85)	2.06 (0.92)	2.61 (0.77)

Table 4.
GFIs and other statistics for the tested model.

GFIs/other statistics	Result
Chi-square	5.45
df	5
p	.365
NFI	.978
TLI	.996
IFI	.988
CFI	.998

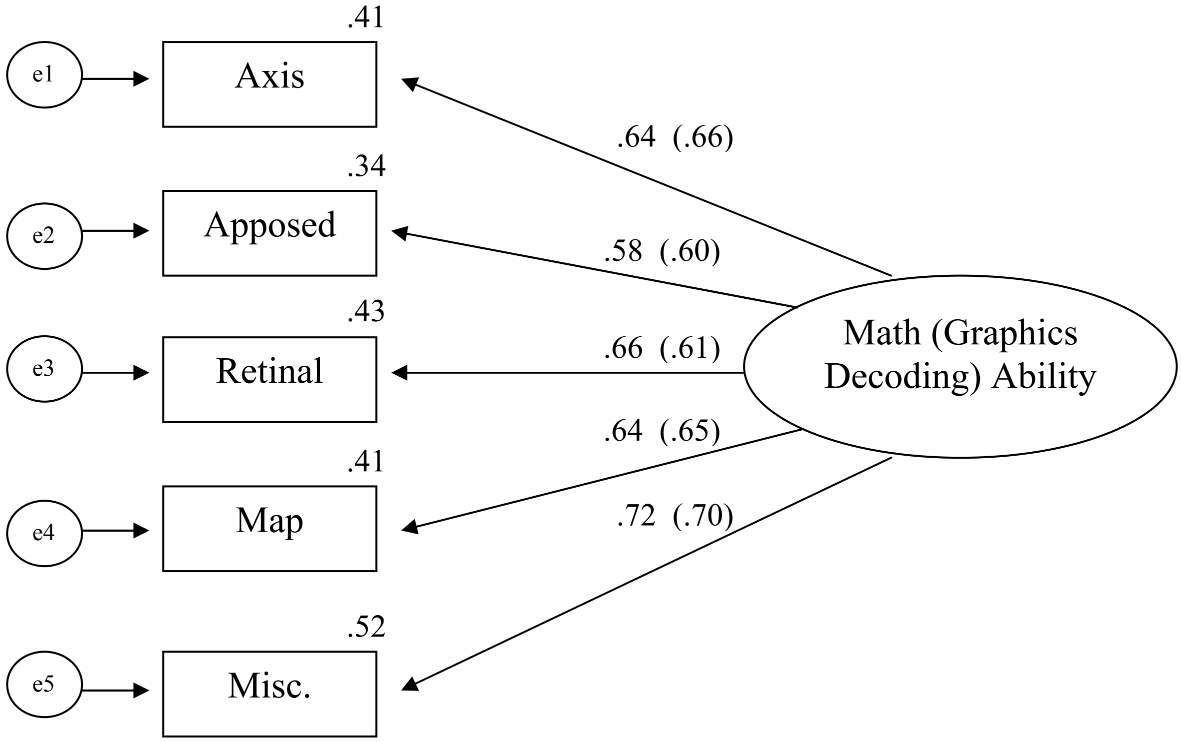


Figure 1. Graphics decoding ability confirmatory factor analysis model with unstandardised (and standardised) parameter estimates.

References

Berends, I.E., & van Lieshout, E.C.D.M. (2009). The effect of illustrations in arithmetic problem-solving: Effects of increased cognitive load. *Learning and Instruction, 19*, 345-353.

Bernard, H.R. (2000). *Social Research Methods: Quantitative and Qualitative Approaches*. Thousand Oaks, CA: Sage.

Bertin, J. (1983). *Semiology of graphics* (W. J. Berg, Trans.). Madison, WI: The University of Wisconsin Press. (Original work published 1967)

Brna, P., Cox, R., & Good, J. (2001). Learning to think and communicate with diagrams: 14 questions to consider. *Artificial Intelligence Review 15*, 115-134.

Carpenter, P.A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied, 4*, 75-100.

Diezmann, C.M., & Lowrie, T. (2009a). An instrument for assessing primary students' knowledge of information graphics in mathematics. *Assessment in Education: Principles, Policy and Practice, 16*(2), 131-147.

Diezmann, C. M., & Lowrie, T. (2009b). The role of fluency in a mathematics item with an embedded graphic: Interpreting a pie chart. *ZDM: The International Journal on Mathematics Education, 41*(5), 651-662.

Elia, I., Gagatsis, A., & Demetriou, A. (2007). The effects of different modes of representation on the solution of one-step additive problems. *Learning and Instruction, 17*, 658-672.

Graham, J.M. (2006). Congeneric and (essentially) Tau-Equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930-944.

Harris, R.L. (1996). *Information graphics: A comprehensive illustrated reference*. Atlanta, GA: Management Graphics.

Hittleman, D.R. (1985). A picture is worth a thousand words…if you know the words. *Childhood Education, 61-62*: 32-36.

Hoe, S.L. (2008). Issues and procedures in adopting Structural Equation Modeling technique. *Journal of Applied Quantitative Methods, 3*(1), 76-83.

Kaplan, R.M., & Saccuzzo, D.P. (2009). *Psychological testing: Principles, applications and issues* (2nd ed). Belmont, CA: Cengage Learning.

Kline, T.J.B. (2005). *Psychological testing: A practical approach to design and evaluation.* Thousand Oak, CA: Sage.

Kosslyn, S.M. (2006). *Graph design for the eye and mind*. New York: Oxford University Press.

Lowrie, T., & Diezmann, C.M. (2011). Solving graphics tasks: Gender differences in middle-school students. *Learning and Instruction, 21*, 109-125.

Lowrie, T., & Diezmann, C.M. (2009). National numeracy tests: A graphic tells a thousand words. *Australian Journal of Education, 53*(2), 141-158.

Lowrie, T., & Diezmann, C. M. (2007). Solving graphics problems: Student performance in the junior grades. *Journal of Educational Research, 100*(6), 369-377.

Lowrie, T., & Diezmann, C. (2005). Fourth-grade students' performance on graphical languages in mathematics. In H.L. Chick & J.L. Vincent (Eds.), *Proceedings of the 29th annual conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 265-272). Melbourne, Australia: PME.

Lowrie, T., Diezmann, C. M., & Logan, T. (2009). Gender differences in orientation on primary students' performance on items rich in graphics. In M. Tzekaki, M. Kaldrimidou, & C. Sakonidis (Eds.), *In search for theories in mathematics education* (Proceedings of the 33rd annual conference of the International Group for the Psychology of Mathematics Education, Vol. 4, pp. 33-39). Thessaloniki, Greece: PME.

MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84-99.

Mackinlay, J. (1999). Automating the design of graphical presentations of relational information. In S.K. Card, J.D. Mackinlay, & B. Schneiderman (Eds.), *Readings in information visualization: Using vision to think* (pp. 66-81). San Francisco, CA: Morgan Kaufmann.

Maruyama, G.M. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage.

Roth, W.-M. (2002). Reading graphs: Contributions to an integrative concept of literacy. *Journal of Curriculum Studies, 34*(2), 1-24.

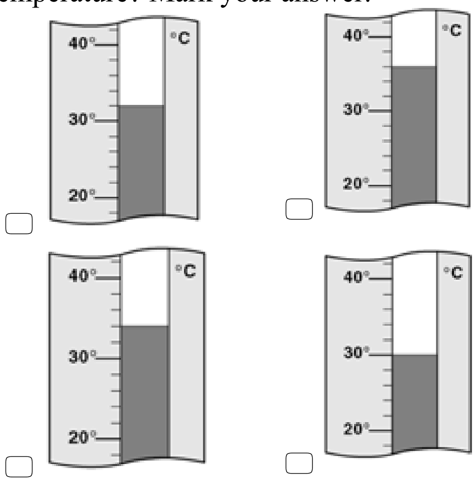
Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system- and self-paced instruction in multimedia learning. *Learning and Instruction, 20*(2), 100-110.

Tabachnick, B.G., & Fidell, L.S. (2001). Using multivariate statistics (4^{th} ed.). Boston: Allyn & Bacon.

Yuan, K. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research 40*, 115-148.

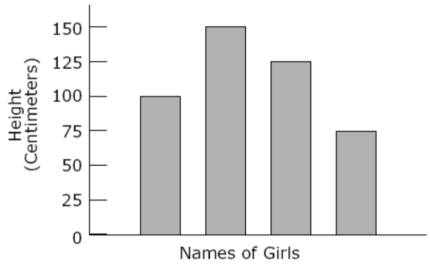
Appendix. Sample items from the five graphical languages.

The temperature on the soccer field was 34°C. Which thermometer shows this temperature? Mark your answer.



Item 1—Axis. Texas Education Agency. (2007). *Texas assessment of knowledge and skills: Grade 3* (p. 23). Texas: Author.

The graph shows the heights of four girls.

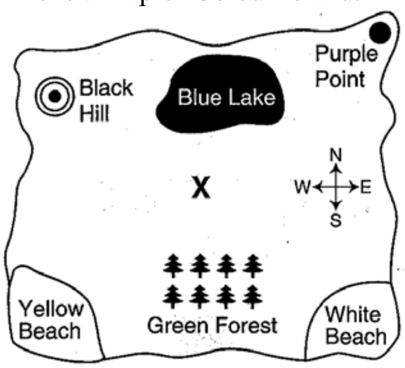


The names are missing from the graph. Debbie is the tallest. Amy is the shortest. Dawn is taller than Sarah. How tall is Sarah?

- ▢ 75 cm
- ▢ 100 cm
- ▢ 125 cm
- ▢ 150 cm

Item 2—Apposed-position. National Center for Education Statistics. (1995). *Complete TIMSS 4 mathematics items: Item number MO12126* (p. 79). Washington, DC: Author.

This is a map of Colour Island.

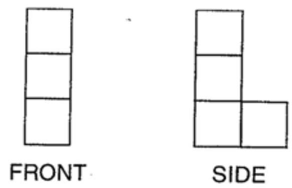


Pete the Pirate is standing at **X**. From there, he walks south and then west. Where could Pete the Pirate be now?

- ▢ Yellow Beach
- ▢ White Beach
- ▢ Purple Point
- ▢ Black Hill

Item 3—Map. Educational Assessment Australia. (2006). *International competitions and assessments for schools: Mathematics 2006: Grade 3 paper A* (p. 5). Sydney: University of New South Wales.

Sara made a model using small cubes. This is what it looked like from the front and from the side.



How many cubes did Sara use in her model?

- ▢ three
- ▢ four
- ▢ six
- ▢ seven

Item 4—Retinal-list. Australian Council for Educational Research. (n.d.). *Progressive Achievement Tests in mathematics: Test booklet 2B* (revised edn., p. 7). Camberwell, Victoria: Author.

Lani left for a holiday on Saturday 3rd January.

January						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

February						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28

She came back home on Monday 9th February.
How long was her holiday?

- ☐ about 1 week
- ☐ about 3 weeks
- ☐ about 4 weeks
- ☐ about 6 weeks

Item 5—Miscellaneous. New South Wales Department of Education and Training. (1999). *Basic skills testing program: Aspects of numeracy: Year 3* (p. 5). Sydney, Australia: Author.