# SMART::Test <br> Solving Linear Equations <br> Development and Student Results Technical Report (2022) 

Vicki Steinle Beth Price<br>Kaye Stacey

# Melbourne Graduate School of Education University of Melbourne, Australia 

Contents
Section 1: Introduction ..... 3
1.1 Introduction to the SMART::test system ..... 3
1.2 Publications relating to SMART::tests ..... 5
Section 2: Data Collection ..... 7
2.1 Sample ..... 7
2.2 Details of items ..... 9
2.3 Rubrics for allocating students to stages ..... 13
Section 3: Confirmation of similar items: Versions $1 \& 2$ ..... 15
Section 4: Distribution of stages ..... 18
4.1 Overall stage distributions for Versions P, $1 \& 2$ ..... 18
4.2 Stage distributions by year level for Versions $1 \& 2$ ..... 19
Section 5: Distributions of responses: CR\%, IR \%, OR\% ..... 20
5.1 Overall test results ..... 20
5.2 Results by item ..... 21
Section 6: Accuracy Analyses: Three Facilities ..... 26
Section 7: Error Analyses: Item Response Summary ..... 32
Appendix 1: Screenshots of Versions P, $1 \& 2$ ..... 39
Appendix 2: Accuracy Coding ..... 41
Appendix 3: Three Facilities and the Facility Range ..... 47
Appendix 4: The Item Position Effect ..... 49
Appendix 5: Support for Versions 1 and 2 being parallel ..... 50
Appendix 6: Illustration of learning: Subsets T1 and T2 of Version 2 ..... 52
Appendix 7: Variation in score better explained by stage than by year level ..... 53

## Section 1: Introduction

### 1.1 Introduction to the SMART::test system

This Technical Report provides examples of the methods of analysis that were carried out to identify and investigate the types of errors made by students as they attempted to solve linear equations. The equations were presented to students as a SMART::test (Specific Mathematics Assessments that Reveal Thinking www.smartvic.com).

These online tests provide teachers with formative assessment of student learning in over 60 very specific mathematics topics, giving automated diagnostic feedback to teachers. Within seconds, the report to teachers gives information on each student's stage of learning and any common errors or misconceptions that each showed in the topic being tested.

In addition to the diagnoses, the testing system provides teachers with explanations about the cognitive hurdles in this topic and about the common errors that students make. In other words, as well as teachers finding out what individual students can and cannot do, information is provided about why they might be experiencing a particular difficulty. The report to teachers also includes some suggestions for assisting students to overcome these difficulties. This information is provided so that the teacher can plan their lessons using the given diagnostic information. Another of our goals for this system is to inform teachers of relevant research findings about student learning in mathematics with the predicted outcomes being increasing teachers' mathematical pedagogical content knowledge and improving student outcomes, as illustrated by Figure 1.1.


Figure 1.1: How teachers use the SMART::tests system and intended outcomes
The SMART::test that generated the 'Solving linear equations' data reported here attempts to identify the strategy that students use when solving linear equations, and to alert the teacher to difficulties exhibited by students. The 'developmental stages' are built on our research, as well as research by others, into students' strategies conducted either by interview or by examining students' written answers.

Three versions of the 'Solving linear equations' test are detailed in this report. Version 1 was developed from an earlier pilot version (Version P), and Version 2 was created to be parallel to Version 1, but with cosmetic changes to make the tests look different to discourage students seated next to each other in a classroom from copying. These changes were:

- using different colours,
- using a different pronumeral ( $a$ in Version 1 and $n$ in Version 2),
- swapping the positions of a few equations,
- modifying the coefficients and/or constants while keeping the equation structure the same.

Most of the teachers who chose to use the tests, as they were interested to see the results of their own students, provided class time for students to complete the tests, while some allowed students to complete outside of class. Hence, we cannot be sure that students completed this test 'under test conditions'; the time allowed for students to complete the test is not known; and we expect that, like other tests, some students do not complete the test as they run out of time. Students who submit the test without answering any items are removed from the data; it is likely that they planned to complete the test but were interrupted and completed it later.

Initial work to develop some SMART::tests and the automated diagnostic module was funded by the Australian Research Council (through Linkage Project Grant Supporting personalised learning in secondary schools through the use of specific mathematics assessments that reveal thinking (2008 - 2010)), Department of Education and Early Childhood Development (Victoria) and the Catholic Education Office Melbourne. The SMART::tests have been adapted in various commercialised products, in which the authors have no ongoing financial stake. The research has been approved by the University of Melbourne, Australia. Approval number 1239210.2. Title: Improvement of the SMART maths online testing system. Researchers: Prof K. C. Stacey, Dr L. M. Ball, Dr V. A. Steinle, Dr E. Gvozdenko, Mrs E. A. Price.

### 1.2 Publications relating to SMART::tests

Steinle, V., Stacey, K., \& Price, E. (accepted). Beyond Accuracy: A process for analysis of constructed responses in large datasets and insights into students' equation solving.

Steinle, V., Stacey, K., \& Price, E. (in preparation). Assessing students’ strategies for solving linear equations in an automated online testing environment.

Stacey K., Steinle V., Price B., \& Gvozdenko, E. (2018). Specific Mathematics Assessments that Reveal Thinking: An Online Tool to Build Teachers' Diagnostic Competence and Support Teaching. In T. Leuders, K. Philipp, \& J. Leuders (Eds.), Diagnostic Competence of Mathematics Teachers. Mathematics Teacher Education, Vol. 11, (pp. 241 - 261). Springer, Cham. https://doi.org/10.1007/978-3-319-66327-2_13 https://minerva-access.unimelb.edu.au/handle/11343/247768

Stacey, K., Steinle, V., Price, B., \& Gvozdenko, E. (2017). Fit in Algebra? Mach den smart-test. MatheWelt 202, pp. 1 - 16. Seelze, Germany: Friedrich Verlag GmbH. (Translators Judith Blomberg \& Maike Abshagen).

Barzel, B., \& Holzapfel, L. (2017). Strukturen also Basis der Algebra. Mathematik Lehren, 202, 2-8
Akhtar, Z., \& Steinle, V. (2017). The Prevalence of the 'letter as object' misconception in junior secondary students. In A. Downton, S. Livy \& J. Hall (Eds.), 40 years on: We are still learning! (Proceedings of the 40th Annual Conference of the Mathematics Education Research Group of Australasia, pp. 77 - 84). Melbourne: MERGA. https://merga.net.au/Public/Publications/Annual_Conference_Proceedings/2017_MERGA_annual_conference_proceedings.aspx

Rule, V. (2017). A case study of the influence of diagnostic information on a teacher's planning for a Year 8 algebra lesson. [Master's thesis, University of Melbourne]. https://minervaaccess.unimelb.edu.au/handle/11343/213997

McKee, S. J. (2017). Using teacher capacity to measure improvement in key elements of teachers' mathematical pedagogical content knowledge. [Doctoral thesis, University of Melbourne] https://minerva-access.unimelb.edu.au/handle/11343/123566

Price, B., Stacey, K., Steinle, V., \& Gvozdenko, E. (2014). Using percentages to describe and calculate change. In J. Anderson, M. Cavanagh, \& A. Prescott (Eds.), Curriculum in focus: Research guided practice (Proceedings of the 37th annual conference of the Mathematics Education Research Group of Australasia pp. 517 - 524). Sydney: MERGA. https://www.merga.net.au/Public/Public/Publications/Annual_Conference_Proceedings/2014_MERGA_CP.aspx

Quenette, J. (2014). Diagnostic testing and changes to teaching practice in Year 9 mathematics classes. [Master's thesis, University of Melbourne]. https://minerva-access.unimelb.edu.au/handle/11343/43027

Guzman, M. A. (2014). The smart test system: teachers' views about this formative assessment for mathematics. [Master's thesis, University of Melbourne]. https://minerva-access.unimelb.edu.au/handle/11343/44090

Stacey, K., Steinle, V., Gvozdenko, E., \& Price, B. (2013). SMART online formative assessments for teaching mathematics, Curriculum \& Leadership Journal, Vol. 11, Issue 20. http://www.curriculum.edu.au/leader/smart_online_formative_assessments_for_teaching_ma,36822.html?issueID=12826

Stacey, K. (2013). Bringing research on students' understanding into the classroom through formative assessment. In V. Steinle, L. Ball \& C. Bardini (Eds.), Mathematics education: Yesterday, today and tomorrow (Proceedings of the 36th annual conference of the Mathematics Education Research Group of Australasia, pp. 13 - 20). Melbourne: MERGA. https://www.merga.net.au/Public/Public/Publications/Annual_Conference_Proceedings/2013_MERGA_CP.aspx

Price, B., Stacey, K., Steinle, V., Gvozdenko, E. (2013). SMART online assessments for teaching mathematics. Mathematics Teaching (Issue 235, pp. 10-15).

Akhtar, Z., \& Steinle, V. (2013). Probing students' numerical misconceptions in school algebra. In V. Steinle, L. Ball \& C. Bardini (Eds.), Mathematics education: Yesterday, today and tomorrow (Proceedings of the 36th annual conference of the Mathematics Education Research Group of Australasia, pp. 36 - 43). Melbourne:

MERGA.
https://www.merga.net.au/Public/Public/Publications/Annual Conference Proceedings/2013 MERGA CP.aspx
Akhtar, Z. \& Steinle, V. (2013). Letter as object misconception in junior secondary school algebra. In A. M. Lindmeier \& A. Heinze (Eds.) Proceedings of 37th Conference of the International Group for the Psychology of Mathematics Education. (Vol. 5, p. 4) Kiel, Germany: PME. http://www.igpme.org/publications/current-proceedings/

Stacey, K., Price, B., \& Steinle, V. (2012). Identifying stages in a learning hierarchy for use in formative assessment - the example of line graphs. In J. Dindyal, L. P. Cheng \& S. F. Ng (Eds.), Mathematics education: Expanding horizons (Proceedings of the 35th annual conference of Mathematics Education Group of Australasia, pp. 393 - 400). Adelaide: MERGA. https://merga.net.au/Public/Publications/Annual_Conference_Proceedings/2012_MERGA_CP.aspx

Steinle, V., \& Stacey, K. (2012). Teachers' Views of using an on-line, formative assessment system for Mathematics. Pre-proceedings. 12th International Congress on Mathematical Education Topic Study Group 33, July 2012, (pp. 6721 - 6730). COEX, Seoul, Korea. https://minervaaccess.unimelb.edu.au/handle/11343/247767

Price, B., Stacey, K. Steinle, V., Chick, H., Gvozdenko, E. (2011). Getting SMART About Assessment for Learning in 2011. Reflections, 36 (3), pp. 3 - 7.

Baratta, W., Price, E., Stacey, K., Steinle, V., \& Gvozdenko, E. (2010). Percentages: The effect of problem structure, number complexity and calculation format. In L. Sparrow, B. Kissane \& C. Hurst (Eds.), Shaping the future of mathematics education (Proceedings of 33rd annual conference of the Mathematics Education Research Group of Australasia, pp. 61 - 68). Fremantle: MERGA. https://merga.net.au/Public/Publications/Annual_Conference_Proceedings/2010_MERGA_CP.aspx

Stacey, K., Steinle, V., Wu, M., Pierce, R., \& Giri, J. (2010, 29 Nov-2 Dec). Evaluating automated processes for revealing students' mathematical thinking. Unpublished symposium presentation. Annual Conference of Australian Association of Research in Education, University of Melbourne, Melbourne. https://minerva-access.unimelb.edu.au/handle/11343/247766

Steinle, V., Gvozdenko, E., Price, B., Stacey, K., \& Pierce, R. (2009). Investigating students' numerical misconceptions in algebra. In R. Hunter, B. Bicknell \& T. Burgess (Eds.), Crossing divides (Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia, pp. 491 498). Wellington:

MERGA. https://merga.net.au/Public/Publications/Annual_Conference_Proceedings/2009_MERGA_CP.aspx

Stacey, K., Price, B., Steinle, V., Chick, H., \& Gvozdenko, E. (2009, Sept 28-Oct 1). SMART Assessment for Learning. [Conference presentation]. International Society for Design and Development in Education, Cairns, Australia. https://minerva-access.unimelb.edu.au/handle/11343/247762

Price, B., Stacey, K., Steinle, V., Chick, H., \& Gvozdenko, E. (2009). Getting SMART about Assessment for Learning. In D. Martin, T. Fitzpatrick, R. Hunting, D. Itter, C. Lenard, T. Mills \& L. Milne (Eds.), Proceedings of 2009 Annual Conference of the Mathematical Association of Victoria. (pp. 174-181). MAV: Melbourne. https://minerva-access.unimelb.edu.au/handle/11343/247761

## Section 2: Data Collection

### 2.1 Sample

- Table 2.1 shows that the Version P data presented in this report comes from students who completed the online test in a two-year period (February 2014 to December 2015)
- Similarly, Version 1 data comes from students who completed the online test in a three-year period, while Version 2 data comes from a four-year period to increase the size of the sample of this less-used version.
- Note that as this system does not identify students, it is not possible to link students from one test to another and it is possible that some students might have contributed data more than once over the multi-year periods.
- While the system does not track individual students, our best estimates of teacher use indicate that about 1 in 5 teachers use both Version 1 and Version 2 at the same time (presumably before teaching and randomly assigned to Version 1 or 2); of the 4 in 5 teachers who choose to use only Version 1 at that time, $10 \%$ of their students complete Version 2 later (presumably after teaching).

Table 2.1: Years when test version used \& data collected

| Features | Version P | Version 1 | Version 2 |
| :--- | :---: | :---: | :---: |
| Module | $\bmod 67^{\wedge}$ | $\bmod 224$ | $\bmod 226$ |
| Screens | $5 \& 6$ | both | both |
| Number of items | 11 | 14 | 14 |
| Years in use | $2011-2016$ | 2016 -current | 2016 -current |
| Years for this data | $2014-2015$ | $2016-2018$ | $2016-2019$ |
| Total number of test submissions* | 1209 | 3041 | 792 |
| Number of Australian test submissions | 1152 | 3010 | 734 |
| $\%$ Australian test submissions | $95 \%$ | $99 \%$ | $93 \%$ |

${ }^{\wedge}$ Module 67 consisted of 6 screens; the first 4 were on a different task (matching equations to word problems, see screenshots later) and do not otherwise contribute to this report.

* only test submissions with at least one response were analysed; for Version P, this means at least one response to the items on screens 5 \& 6 .
- Table 2.2 shows for Version P, for example, that there was a total of 1152 studentsubmissions from Years 8, 9 and 10 (ages 14 to 16) over a 2-year period.
- For simplicity, student-submissions will henceforth be referred to as students.
- The sample for Version 2 (a total of 734 students over a 4-year period) has been split into two subsets; 393 students who are expected to be completing their first test in a given year (T1) and 341 students who are expected to be completing their second test (T2).
- As the system does not track students, the breakdown into T1 and T2 subsets was accomplished by considering the teachers. Teachers who used Version 2 with a class, in the window 1 to 4 weeks after their first use, were identified; students in these classes were allocated to T2. (The data analysis in this report indicates that this process has been successful; see Appendices 5 and 6.)
- Table 2.2 shows that the percentage distributions for Versions P, $1 \& 2$ are strikingly similar; the dominant groups are Years 8 and 9 , and there is a split of about 50:50 into Year 7/8 and Year $9 / 10$. This allows us to compare results for these samples later.
- Similarly, with Version 2 subsets (T1 \& T2) in Table 2.3

Table 2.2: Year levels (ages) of students who used Versions P, 1 and 2

| Year | Approximate | Number |  |  | Percentage |  |  |
| :---: | :---: | ---: | ---: | ---: | ---: | ---: | ---: |
|  |  | Version | Version | Version | Version | Version | Version |
|  |  | P | 1 | 2 | P | 1 | 2 |
| Year 7 | 13-year-olds | 0 | 284 | 41 | $0 \%$ | $9 \%$ | $6 \%$ |
| Year 8 | 14-year-olds | 520 | 1284 | 338 | $45 \%$ | $43 \%$ | $46 \%$ |
| Year 9 | 15-year-olds | 451 | 1074 | 252 | $39 \%$ | $36 \%$ | $34 \%$ |
| Year 10 | 16-year-olds | 181 | 368 | 103 | $16 \%$ | $12 \%$ | $14 \%$ |
|  | Total | 1152 | 3010 | 734 | $100 \%$ | $100 \%$ | $100 \%$ |

Table 2.3: Year levels (ages) of students who used Version 2 subsets: T1 and T2

| Year | Approximate | Number |  | Percentage |  |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: |
| level | age | Subset T1 | Subset T2 | Subset T1 | Subset T2 |
| Year 7 | 13-year-olds | 40 | 1 | $10 \%$ | $0 \%$ |
| Year 8 | 14-year-olds | 155 | 183 | $39 \%$ | $54 \%$ |
| Year 9 | 15-year-olds | 134 | 118 | $34 \%$ | $35 \%$ |
| Year 10 | 16-year-olds | 64 | 39 | $16 \%$ | $11 \%$ |
|  | Total | 393 | 341 | $100 \%$ | $100 \%$ |

- Teachers who use these tests to inform their teaching (i.e., as formative assessment) are likely to provide different work for students who are assessed as already having mastered the topic (i.e., Stage 4 in this test) and are unlikely to ask these students to sit another test. Note that if these Stage 4 students were asked to complete a second test, (which would be the case if the data was collected as part of an intervention study, for example) then the resulting distribution of stages in T2 would be higher for Stage 4 and lower for Stages 0, 1, 2 and 3.
- This Australian data comes from over 30 schools, taught by more than 180 teachers. While this is a large sample, it is a convenience, rather than representative, sample. Over $95 \%$ of the students were from Victoria and automated records of start and finish times showed that the average time that students were engaged with a test was about 15 minutes.


### 2.2 Details of items

- Excluding other external issues (such as test administration conditions), we see item response being influenced by Item Characteristics and Student Characteristics, as illustrated in Figure 2.1.


Figure 2.1: Item Response as determined by Item and Student and Characteristics. Note: $\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d} \in \mathrm{Z}^{+}$

- Screenshots are provided of all versions in Appendix 1
- To minimise the need for cross-referencing, the equation and the solution are provided when items are mentioned; for example, $[3 a+8=23 \mid 5]$ indicates the equation $3 a+8=23$ has solution 5.
- Table 2.4 provides a summary of the items in the 3 versions
- Colour coding is used to show items which were used in more than one version:
- 4 uncoloured items from Version $P$ were retired
- 7 coloured items from Version P were reused in Version 1 (but different positions)
- Version 2: Items were created which were intended to be parallel to those in Version 1 (using $n$ instead of $a$ ), including:
- identical items Q3 \& Q4, (swapped positions)
- minor edits to Q5 \& Q6 (same solutions and swapped positions)
- identical items Q7 \& Q8

Table 2.4: Summary of items in Versions P, 1 \& 2

|  | Version P | Version 1 | Version 2 |
| :---: | :---: | :---: | :---: |
| Q1 | $[4 n+9=37 \mid 7]$ | $[3 a+8=23 \mid 5]$ | $[4 n+11=23 \mid 3]$ |
| Q2 | $[4(n-3)=21 \mid 8.25]$ | $[4 a+9=37 \mid 7]$ | $[3 n+5=26 \mid 7]$ |
| Q3 | $[5 n-2=3 n+6 \mid 4]$ | $[5 a+7=15 \mid 1.6]$ | $[8 n+3=16 \mid 1.625]$ |
| Q4 | $[12 n+2=8 n+15 \mid 3.25]$ | $[8 a+3=16 \mid 1.625]$ | $[5 n+7=15 \mid 1.6]$ |
| Q5 | $[3 a+8=23 \mid 5]$ | $[8 a+5=3 a+14 \mid 1.8]$ | $[11 n+3=7 n+16 \mid 3.25]$ |
| Q6 | $[4 a-7=16 \mid 5.75]$ | $[12 a+2=8 a+15 \mid 3.25]$ | $[9 n+3=4 n+12 \mid 1.8]$ |
| Q7 | $[14-2 a=8 \mid 3]$ | $[7 a-11=2 a-4 \mid 1.4]$ | $[7 n-11=2 n-4 \mid 1.4]$ |
| Q8 | $[6-5 a=15 \mid-1.8]$ | $[12-11 a=5-a \mid 0.7]$ | $[12-11 n=5-n \mid 0.7]$ |
| Q9 | $[9 a+3=7 a+15 \mid 6]$ | $[7 a-2=16 \mid 18 / 7]$ | $[5 n-1=16 \mid 3.4]$ |
| Q10 | $[8 a+5=3 a+14 \mid 1.8]$ | $[14-2 a=8 \mid 3]$ | $[15-2 n=9 \mid 3]$ |
| Q11 | $[12-11 a=5-a \mid 0.7]$ | $[3 a+6+2 a=7 \mid 0.2]$ | $[2 n+4+3 n=5 \mid 0.2]$ |
| Q12 |  | $[a+2) / 5=3 \mid 13]$ | $[(n+1) / 5=3 \mid 14]$ |
| Q13 |  | $[9 / 3+1=5 \mid 12]$ | $[n / 4+3=8 \mid 20]$ |
| Q14 |  | $[4(a-3)=21 \mid 8.25]$ | $[5(n-2)=8 \mid 3.6]$ |

Table 2.5: Version P: Categorisation of 11 equations into Groups PA, PB, PC and PD

| Item details |  |  | Equation structure |  |  |  |  | Solution features |  | Label |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Pronumeral Features |  |  | Explicit Operations | Implicit Operations |  |  |  |
|  |  |  | Number of occurrences | Location: Left, Right | Sign of Coefficient |  |  | Solution | Z/R |  |
| Q1 | 3031_1 | $4 n+9=37$ | 1 | 1,0 | + | + | $\times$ | 7 | $Z^{+}$ | PD1 |
| Q2 | 3031_2 | $4(n-3)=21$ | 1 | 1,0 | + | - | ${ }^{\text {\# }}$ | 8.25 | $R^{+}$ | PD2 |
| Q3 | 3031_3 | $5 n-2=3 n+6$ | 2 | 1,1 | + + | -+ | $\times \times$ | 4 | $Z^{+}$ | PD3 |
| Q4 | 3031_4 | $12 n+2=8 n+15$ | 2 | 1,1 | + + | + + | $\times \times$ | 3.25 | $R^{+}$ | PD4 |
| Q5 | 3104_1 | $3 a+8=23$ | 1 | 1,0 | + | + | $\times$ | 5 | $Z^{+}$ | PA1 |
| Q6 | 3104_2 | $4 a-7=16$ | 1 | 1,0 | + | - | $\times$ | 5.75 | $R^{+}$ | PB1 |
| Q7 | 3104_3 | $14-2 a=8$ | 1 | 1,0 | - | - | $\times$ | 3 | $Z^{+}$ | PA2 |
| Q8 | 3104_4 | $6-5 a=15$ | 1 | 1,0 | - | - | $\times$ | -1.8 | $R^{-}$ | PB2 |
| Q9 | 3104_5 | $9 a+3=7 a+15$ | 2 | 1,1 | + + | + + | $\times \times$ | 6 | $Z^{+}$ | PD5 |
| Q10 | 3104_6 | $8 a+5=3 a+14$ | 2 | 1,1 | + + | + + | $\times \times$ | 1.8 | $R^{+}$ | PC1 |
| Q11 | 3104_7 | $12-11 a=5-a$ | 2 | 1,1 | -- | -- | $\times \times$ | 0.7 | $R^{+}$ | PC2 |

\# students may or may not recognise that brackets indicate multiplication

- Q8 (PB2) is the only item with a negative solution- this item was not used in later versions as computation with negative numbers includes an extraneous factor, when the aim is to identify solving strategies, although computation with negative numbers, (E.g. in Q11, $-11 a+a$ $=-10 a$ or $-12 a)$ occurs on the way to some positive solutions.

Table 2.6: Version 1: Categorisation of 14 equations into Groups A, B, C and D

| Version 1 Item details |  |  | Equation structure |  |  |  |  | Solution features |  | Label |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Pronumeral Features |  |  | Explicit Operations | Implicit Operations |  |  |  |
|  |  |  | Number of occurrences | Location: Left, Right | Sign of Coefficient |  |  | Solution | Z/R |  |
| Q1 | 3105_1 | $3 a+8=23$ | 1 | 1, 0 | + | + | $\times$ | 5 | $Z^{+}$ | A1 |
| Q2 | 3105_2 | $4 a+9=37$ | 1 | 1, 0 | + | + | $\times$ | 7 | $Z^{+}$ | A2 |
| Q3 | 3105_3 | $5 a+7=15$ | 1 | 1, 0 | + | + | $\times$ | 1.6 | $R^{+}$ | B1 |
| Q4 | 3105_4 | $8 a+3=16$ | 1 | 1, 0 | + | + | $\times$ | 1.625 | $R^{+}$ | B2 |
| Q5 | 3106_1 | $8 a+5=3 a+14$ | 2 | 1, 1 | + + | + + | $\times \times$ | 1.8 | $R^{+}$ | C1 |
| Q6 | 3106_2 | $12 a+2=8 a+15$ | 2 | 1,1 | + + | + + | $\times \times$ | 3.25 | $R^{+}$ | C2 |
| Q7 | 3106_3 | $7 a-11=2 a-4$ | 2 | 1,1 | + + | - - | $\times \times$ | 1.4 | $R^{+}$ | D1 |
| Q8 | 3106_4 | $12-11 a=5-a$ | 2 | 1,1 | - - | - - | $\times \times$ | 0.7 | $R^{+}$ | D2 |
| Q9 | 3107_1 | $7 a-2=16$ | 1 | 1, 0 | + | - | $\times$ | 18/7 | $R^{+}$ | E1 |
| Q10 | 3107_2 | $14-2 a=8$ | 1 | 1, 0 | - | - | $\times$ | 3 | $Z^{+}$ | E2 |
| Q11 | 3107_3 | $3 a+6+2 a=7$ | 2 | 2, 0 | + + | + + | $\times \times$ | 0.2 | $R^{+}$ | E3 |
| Q12 | 3107_4 | $(a+2) / 5=3$ | 1 | 1, 0 | + | + | $\stackrel{\text { - }}{ }$ | 13 | $Z^{+}$ | E4 |
| Q13 | 3107_5 | $a / 3+1=5$ | 1 | 1, 0 | + | + | $\stackrel{\text { - }}{ }$ | 12 | $Z^{+}$ | E5 |
| Q14 | 3107_6 | $4(a-3)=21$ | 1 | 1, 0 | + | - | × ${ }^{\text {\# }}$ | 8.25 | $R^{+}$ | E6 |

\# students may or may not recognise that a vinculum indicates division or that brackets indicate multiplication

- Q9 (E1) is the only item with a solution which, in fraction form, has a denominator of 7, and in decimal form, is a repeating decimal.
- Analyses in this report shows that this item is more difficult than we expected \& is not parallel to Q9 (E1) in Version 2.

Table 2.7: Version 2: Categorisation of 14 equations into Groups A, B, C and D

| Version 2 Item details |  |  | Equation structure |  |  |  |  | Solution features |  | Label |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Pronumeral Features |  |  | Explicit Operations | Implicit Operations |  |  |  |
|  |  |  | Number of occurrences | Location: Left, Right | Sign of Coefficient |  |  | Solution | Z/R |  |
| Q1 | 3108_1 | $4 n+11=23$ | 1 | 1, 0 | + | + | $\times$ | 3 | $Z^{+}$ | A1 |
| Q2 | 3108_2 | $3 n+5=26$ | 1 | 1, 0 | + | + | $\times$ | 7 | $Z^{+}$ | A2 |
| Q3 | 3108_3 | $8 n+3=16$ | 1 | 1, 0 | + | + | $\times$ | 1.625 | $R^{+}$ | B1 |
| Q4 | 3108_4 | $5 n+7=15$ | 1 | 1, 0 | + | + | $\times$ | 1.6 | $R^{+}$ | B2 |
| Q5 | 3109_1 | $11 n+3=7 n+16$ | 2 | 1,1 | + + | + + | $\times \times$ | 3.25 | $R^{+}$ | C1 |
| Q6 | 3109_2 | $9 n+3=4 n+12$ | 2 | 1,1 | + + | + + | $\times \times$ | 1.8 | $R^{+}$ | C2 |
| Q7 | 3109_3 | $7 a-11=2 a-4$ | 2 | 1,1 | + + | - | $\times \times$ | 1.4 | $R^{+}$ | D1 |
| Q8 | 3109_4 | $12-11 a=5-a$ | 2 | 1,1 | - - | - - | $\times \times$ | 0.7 | $R^{+}$ | D2 |
| Q9 | 3110_1 | $5 n-1=16$ | 1 | 1, 0 | + | - | $\times$ | 3.4 | $R^{+}$ | E1 |
| Q10 | 3110_2 | $15-2 n=9$ | 1 | 1, 0 | - | - | $\times$ | 3 | $Z^{+}$ | E2 |
| Q11 | 3110_3 | $2 n+4+3 n=5$ | 2 | 2, 0 | + + | + + | $\times \times$ | 0.2 | $R^{+}$ | E3 |
| Q12 | 3110_4 | $(n+1) / 5=3$ | 1 | 1, 0 | + | + | $\stackrel{\text { - }}{ }$ | 14 | $Z^{+}$ | E4 |
| Q13 | 3110_5 | $n / 4+3=8$ | 1 | 1, 0 | + | + | $\square^{\#}$ | 20 | $Z^{+}$ | E5 |
| Q14 | 3110_6 | $5(n-2)=8$ | 1 | 1, 0 | + | - | ${ }^{\text {\# }}$ | 3.6 | $R^{+}$ | E6 |

\# students may or may not recognise that a vinculum indicates division or that brackets indicate multiplication

Table 2.8: Version P: Rubric for allocating students to stages based on scores $(0,1,2)$ on groups of equations.

| Stages | Description of stages | Groups of equations^ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Gp PA | Gp PB | Gp PC | Gp PD |
| Stage 0 | Not yet at Stage 1 | 0, 1 | - | - | - |
| Stage 1 | Students can solve simple linear equations that are easy to solve by guessing | 2 | 0,1 | - | - |
| Stage 2 | $\ldots$ and can solve linear equations with more difficult solutions so that a systematic method such as unwinding (backtracking) is required; | 2 | 2 | 0,1 | - |
| Stage 3 | $\ldots$ and can solve linear equations with pronumerals on both sides and noninteger solutions, so that they need to be solved by "doing the same to both sides" (algebraic manipulation) | 2 | 2 | 2 | - |

- indicates that the score in this group is not considered in the rubric
${ }^{\wedge}$ Items in each group:
- $\quad \mathrm{Gp} \mathrm{PA}=\mathrm{Q} 5[3 a+8=23 \mid 5] \& \mathrm{Q} 7[14-2 a=8 \mid 3]$
- $\mathrm{Gp} \mathrm{PB}=\mathrm{Q} 6[4 a-7=16 \mid 5.75] \&[6-5 a=15 \mid-1.8]$
- $\mathrm{Gp} \mathrm{PC}=\mathrm{Q} 10[8 a+5=3 a+14 \mid 1.8] \& \mathrm{Q} 12[12-11 a=5-a \mid 0.7]$
- $\mathrm{Gp} \mathrm{PD}=\mathrm{Q} 1,[4 n+9=37 \mid 7], \mathrm{Q} 2[4(n-3)=21 \mid 8.25]$, Q3 [5n-2=3n+6|4], Q4 [12n+2=8n+15|3.25] \& Q9 $[9 a+3=7 a+15 \mid 6]$

Issues with rubric in Version P:

1) Long test with Screens 1 to 4 (non-SLE items) then Screens 5 and 6 (SLE)
2) Rubric items were located on Screen 6; those who answered some of Screen 5 but then omitted (OR) the rubric items, were allocated to Stage 0 . Analysis of item accuracy for the students in Stage 0 showed that they had higher than expected facility on the items in Screen 5. So, some students were in Stage 0 due to IPE rather than lack of skills.
3) Each group consisted of a pair of items with different structure (ie non-homogeneous); one with positive coefficient(s) \& the other with negative coefficient(s) (the second being much harder than the first).

- The original hurdle was $2 / 2$ so that students needed to demonstrate success on the harder item as well as the easier item, but this is a very high standard for items which are numerical constructed response (we do use high hurdles for MC items)
- If any allowance was made for "careless slips" then this lowered the hurdle considerably; students with a score of 1 in a group (typically because they could do only the easier item) were moved up through the stages
- Outcome: we decided to use a group of 2 homogenous items (with a hurdle of $>0$ ) in Versions $1 \& 2$.

Table 2.9: Versions $1 \& 2$ : Rubric for allocating students to stages based on scores $(0,1,2)$ on groups of items

| Stage | Description of stages | Groups of items^ |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | A | B | C | D | E |
| 0 | Not yet at Stage 1 | 0 |  | - | - |  |
| 1 | Students can solve simple linear equations that are easy to solve by repeated substitution; | 1,2 | 0 | - | - | - |
| 2 | ...and can solve linear equations with more difficult solutions so that a systematic method such as unwinding is required; | 1,2 | 1,2 | 0 | - | - |
| 3 | ...and can solve linear equations (involving addition only) with pronumerals on both sides and non-integer solutions, so that they need to be solved by algebraic manipulation; | 1,2 | 1,2 | 1,2 | 0 | - |
| 4 | ...and can further solve linear equations with pronumerals on both sides and non-integer solutions involving subtraction. | 1,2 | 1,2 | 1,2 | 1,2 | - |

- indicates that the score in this group is not considered in the rubric
${ }^{\wedge}$ Items in each group: Version 1
- Group A: A1 $[3 a+8=23 \mid 5] \&$ A2 $[4 a+9=37 \mid 7]$
- Group B: B1 $[5 a+7=15 \mid 1.6] \&$ B2 [ $8 a+3=16 \mid 1.625]$
- Group C: C1 [8a+5=3a+14|1.8]\& C2 [12a+2=8a+15|3.25]
- Group D: D1 [7a-11=2a-4|1.4] \& D2 [12-11a=5-a|0.7]
- Group E: E1 [7a-2 = 16| 18/7], E2 [14-2a=8|3], E3 [3a+6+2a=7|0.2], $\mathrm{E} 4\left[{ }^{(a+2)} / 5=3 \mid 13\right], \mathrm{E} 5\left[{ }^{a} / 3+1=5 \mid 12\right] \& \mathrm{E} 6[4(a-3)=21 \mid 8.25]$
${ }^{\wedge}$ Items in each group: Version 2
- Group A: A1 $[4 n+11=23 \mid 3]$ \& A2 $[3 n+5=26 \mid 7]$
- Group B: B1*[5n+7=15|1.6] \& B2*[8n+3=16|1.625]
- Group C: C1*[9n+3=4n+12|1.8]\& C2*[11n+3=7n+16|3.25]
- Group D: D1 [7n-11 = 2n-4|1.4] \& D2 [12-11n=5-n|0.7]
- Group E: E1 [5n-1 = 16|3.4], E2 [15-2n=9|3], E3 [2n+4+3n=5|0.2], $\mathrm{E} 4\left[{ }^{(n+1)} / 5=3 \mid 14\right]$, E5 $\left[{ }^{n} / 4+3=8 \mid 20\right] \&$ E6 $[5(n-2)=8 \mid 3.6]$


## Section 3: Confirmation of similar items: Versions 1 \& 2

- Rather than focusing only on the pairs within Groups A to D, a more data-driven approach was used; accuracy responses for all 91 item pairs $(14 \times 13 \div 2)$ were generated.
- The 91 combinations were ranked on the percentage of students who responded the same (i.e., a score of 2 indicates both correct and a score of 0 means neither correct), which is the bolded column in the tables.
- Note that items which are positioned close together are more likely to be similar due to the IPE (see Appendix 4).
- The top three item pairs for both Version 1 and 2 are (A1, A2), (B1, B2) and (C1, C2); about $90 \%$ of students, in both versions, responding the same way to these item pairs.
- The pair (D1, D2) was ranked lower in both versions; with $85 \%$ ( $81 \%$ ) of students responding the same way in Version 1 (Version 2).
- These results confirm that:
- Groups A, B and C are homogenous and hence the rubric hurdle of a non-zero score is appropriate
- Group D is not as homogenous and hence the rubric hurdle of a non-zero score is not appropriate
- The positions of (C1, D1) and (C2, D1) are slightly above (D1, D2) in both versions, meaning that D1 is closer to Group C than to D2; students who could solve equations with pronumerals on each side involving addition (Group C), were only slightly affected by the subtraction in D1 (involving subtracting constants).
- an improvement to the test would be to replace D1 by an item more similar to D2 (negative coefficients), possibly moving D1 to Group E (as a "flag" item for the Stage 3 students), subject to space.

Table 3.1: Version 1: Percentage of students responding the same to item pairs: Top 11 and last (rank 91)

| Rank | Item X | Item Y | Accuracy combinations |  |  |  | Total | $\begin{gathered} \text { Same } \\ (\text { Score }=2,0) \end{gathered}$ |  | Different (Score = 1) |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | $(1,1)$ | $(1,0)$ | $(0,1)$ | $(0,0)$ |  | n | \% | n | \% |
| 1 | A1 $[3 \mathrm{a}+8=23 \mid 5]$ | A2 $[4 \mathrm{a}+9=37 \mid 7]$ | 2408 | 162 | 102 | 338 | 3010 | 2746 | 91\% | 264 | 9\% |
| 2 | C1 $18 \mathrm{a}+5=3 \mathrm{a}+14 \mid 1.8]$ | $\mathrm{C} 2[12 \mathrm{a}+2=8 \mathrm{a}+15 \mid 3.25]$ | 1139 | 210 | 81 | 1580 | 3010 | 2719 | 90\% | 291 | 10\% |
| 3 | B1 [5a $+7=15 \mid 1.6]$ | B2 [8a $+3=16 \mid 1.625]$ | 1800 | 227 | 83 | 900 | 3010 | 2700 | 90\% | 310 | 10\% |
| 4 | C2 [12a+2 $=8 a+15 \mid 3.25]$ | D1 [7a-11 = 2a-4\|1.4] | 947 | 273 | 91 | 1699 | 3010 | 2646 | 88\% | 364 | 12\% |
| 5 | $\mathrm{E} 4\left[{ }^{(a+2)} / 5=3 \mid 13\right]$ | E5 $\left[\left.\frac{a}{3}+1=5 \right\rvert\, 12\right]$ | 1159 | 248 | 189 | 1414 | 3010 | 2573 | 85\% | 437 | 15\% |
| 6 | $\mathrm{C} 1[8 a+5=3 a+14 \mid 1.8]$ | D1 [7a-11 $=2 a-4 \mid 1.4]$ | 969 | 380 | 69 | 1592 | 3010 | 2561 | 85\% | 449 | 15\% |
| 7 | D1 $[7 a-11=2 a-4 \mid 1.4]$ | D2 $12-11 a=5-a \mid 0.7]$ | 684 | 354 | 100 | 1872 | 3010 | 2556 | 85\% | 454 | 15\% |
| 8 | E3 $[3 a+6+2 a=7 \mid 0.2]$ | E6 $[4(a-3)=21 \mid 8.25]$ | 846 | 298 | 191 | 1675 | 3010 | 2521 | 84\% | 489 | 16\% |
| 9 | $\mathrm{C} 2[12 a+2=8 a+15 \mid 3.25]$ | $\mathrm{E} 3[3 a+6+2 a=7 \mid 0.2]$ | 916 | 304 | 228 | 1562 | 3010 | 2478 | 82\% | 532 | 18\% |
| 10 | D1 [7a-11 $=2 a-4 \mid 1.4]$ | $\mathrm{E} 3[3 a+6+2 a=7 \mid 0.2]$ | 824 | 214 | 320 | 1652 | 3010 | 2476 | 82\% | 534 | 18\% |
| 11 | C2 [12a+2=8a+15\|3.25] | D2 [12-11a=5-a\|0.7] | 716 | 504 | 68 | 1722 | 3010 | 2438 | 81\% | 572 | 19\% |
| 91 | $\mathrm{A} 1[3 a+8=23 \mid 5]$ | D2 [12-11a=5-a\|0.7] | 750 | 1820 | 34 | 406 | 3010 | 1156 | 38\% | 1854 | 62\% |

Table 3.2: Version 2: Percentage of students responding the same to item pairs: Top 10 and last (rank 91)

| Rank | Item X | Item Y | Accuracy combinations |  |  |  | Total | $\begin{gathered} \text { Same } \\ (\text { Score }=2,0) \end{gathered}$ |  | Different <br> (Score = 1) |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | $(1,1)$ | $(1,0)$ | $(0,1)$ | $(0,0)$ |  | n | \% | n | \% |
| 1 | A1 $[4 n+11=23 \mid 3]$ | A2 $[3 n+5=26 \mid 7]$ | 614 | 25 | 23 | 72 | 734 | 686 | 93\% | 48 | 7\% |
| 2 | C1* $[9 n+3=4 n+12 \mid 1.8]$ | C2* [11n+3=7n+16\|3.25] | 361 | 31 | 30 | 312 | 734 | 673 | 92\% | 61 | 8\% |
| 3 | B1* $[5 n+7=15 \mid 1.6]$ | B2* $[8 n+3=16 \mid 1.625]$ | 466 | 50 | 30 | 188 | 734 | 654 | 89\% | 80 | 11\% |
| 4 | $\mathrm{C} 1 *[9 n+3=4 n+12 \mid 1.8]$ | D1 [7n-11 $=2 n-4 \mid 1.4]$ | 307 | 84 | 21 | 322 | 734 | 629 | 86\% | 105 | 14\% |
| 5 | C2* $[11 n+3=7 n+16 \mid 3.25]$ | D1 [7n-11 $=2 n-4 \mid 1.4]$ | 302 | 90 | 26 | 316 | 734 | 618 | 84\% | 116 | 16\% |
| 6 | B2* $[8 n+3=16 \mid 1.625]$ | E1 $[5 n-1=16 \mid 3.4]$ | 415 | 101 | 31 | 187 | 734 | 602 | 82\% | 132 | 18\% |
| 7 | B1* $[5 n+7=15 \mid 1.6]$ | E1 $[5 n-1=16 \mid 3.4]$ | 402 | 94 | 44 | 194 | 734 | 596 | 81\% | 138 | 19\% |
| 8 | $\mathrm{E} 4\left[{ }^{(n+1)} / 5=3 \mid 14\right]$ | E5 $[n / 4+3=8 \mid 20]$ | 311 | 81 | 59 | 283 | 734 | 594 | 81\% | 140 | 19\% |
| 9 | A1 $[4 n+11=23 \mid 3]$ | B2* [8n+3=16\|1.625] | 507 | 132 | 9 | 86 | 734 | 593 | 81\% | 141 | 19\% |
| 10 | D1 $[7 n-11=2 n-4 \mid 1.4]$ | D2 [12-11n=5-n\|0.7] | 213 | 115 | 27 | 379 | 734 | 592 | 81\% | 142 | 19\% |
| 91 | A1 $[4 n+11=23 \mid 3]$ | D2 [12-11n=5-n\|0.7] | 233 | 406 | 7 | 88 | 734 | 321 | 44\% | 413 | 56\% |

## Section 4: Distribution of stages

### 4.1 Overall stage distributions for Versions P, 1 \& 2



Figure 4.1: Distribution of Stages 0 to 3 for Version P


Figure 4.2: Distribution of Stages 0 to 4 for Version $1 \&$ Version 2 subsets T1 and T2.

### 4.2 Stage distributions by year level for Versions 1 \& 2

- Figure 4.3 shows the breakdown by year level of the distribution of stages.
- There are only three samples with less than 60 students [Year 7 within T1 ( $\mathrm{n}=40$ ) and T2 $(\mathrm{n}=1)$, and Year 10 within T2 $(\mathrm{n}=39)$ ]; results for these groups need to be considered carefully.
- As expected, in all three samples there is a general increasing trend in the percentage of students in Stage 4 from Year 7 to Year 10; 20\% to $57 \%$ (Version 1) and $30 \%$ to $56 \%$ (T1) and, for T2, $49 \%$ to $82 \%$ for Year 8 to Year 10.
- We note the similarity of Year 8 and Year 9 in Version 1 (both have $33 \%$ Stage 4) and a decrease from $30 \%$ to $17 \%$ for Year 7 to 8 in T1.


Figure 4.3: Distribution of stages by year level for Versions 1 and 2 (subsets T1 and T2)

## Section 5: Distributions of responses: CR\%, IR\%, OR\%

### 5.1 Overall test results

- Figure 5.1 shows that, for Version P, about half the responses are correct, and the incorrect and omitted responses are split 50:50
- This indicates that there is a good match between tests and students; very high or very low CR\% would indicate that the tests were too easy or too hard, respectively.
- The distributions are almost identical for Version 1 and Version 2: T1. Considering (i) the similar structure of the equations in Versions 1 and 2 and (ii) the similar sample composition, this provides initial evidence that the two test versions are parallel (see Appendix 5 for more detail).
- In contrast, $66 \%$ of the responses by students in the T2 sample (who completed Version 2) are CR, which is consistent with these students completing their test after teaching (see Appendix 6 for more detail).


Figure 5.1: Overall Percentage Distribution of Responses, for all versions


Figure 5.2: Version P ( $\mathrm{n}=1152$ ): Percentages correct (CR), incorrect (IR) and omitted (OR) $\wedge 4$ items on Screen 1 in common with Linsell (2010) ${ }^{1}$, 7 items in common with Version 2

[^0]

Figure 5.3: Version 1 ( $\mathrm{n}=3010$ ): Percentages correct (CR), incorrect (IR) and omitted (OR) * 7 items common to Version P, ^ 3 items common with Linsell (2010) ${ }^{2}$, \# includes 2.5 ( $\mathrm{n}=88$ )

[^1]

Figure 5.4: Version 2 ( $\mathrm{n}=734$ ): Percentages correct (CR), incorrect (IR) and omitted (OR) * 4 items swapped position from Version 1 to 2


Figure 5.5: Version 2-T1 (n=393): Percentages correct (CR), incorrect (IR) and omitted (OR) * 4 items swapped position from Version 1 to 2


Figure 5.6: Version 2-T2 ( $\mathrm{n}=341$ ): Percentages correct (CR), incorrect (IR) and omitted (OR) * 4 items swapped position from Version 1 to 2

## Section 6: Accuracy Analyses: Three Facilities

Three facilities are used throughout this document and described fully in Appendix 3.

- Facility I: Percentage of correct responses on an item, from all the students.
- Facility II: Percentage of correct responses on an item, from those students who are considered to have been administered the item. (PISA uses this method for item calibration where students who finish the test early are removed from later item statistics.)
- Facility III: Percentage of correct responses on an item, from only the students who provided a response.
- We define the Facility Range as [Facility I, Facility III] $=\left[\frac{n(C R)}{n(C R)+n(I R)+n(N R)}, \frac{n(C R)}{n(C R)+n(I R)}\right]$ and note that Facility II is within this range.


## Versions 1 \& 2:

- This accuracy data confirms that the item groups A, B, C, D get progressively harder (lower CR\%).
- Within Groups A, B and C, the facility of the two items is extremely close (as intended), although the second item in each group has a lower facility than the first.
- There are two reasons for this; firstly, the numerical features of the solution vary (more decimal places) and secondly, it is generally found that later items in a test have lower facility (IPE) hence item position is included in Figure 2.1.

| Version P ( $\mathrm{n}=1152$ ) |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | 0\% | 20\% |  | \% | 60\% |  |  | 100\% |
| $\begin{gathered} \text { Q2 } \\ \text { Q } \end{gathered}$ | $\begin{aligned} & 1[4 n+ \\ & (n-3)= \\ & n-2= \\ & 2=8 n+ \\ & 5[3 a+8 \\ & 4 a-7= \\ & 7[14- \\ & 6-5 a= \\ & +3=7 \\ & -5=3 a \\ & -11 a=5 \end{aligned}$ |  |  | $\mathrm{y} \mathrm{I}$ | Facility | Range |  |  |  |  |
| Item | $\mathrm{N}(\mathrm{CR})$ | Number students |  |  | Facilities |  |  | Standard Error^ |  |  |
|  |  | I | II | III | I | II | III | I | II | III |
| Q1 | 969 | 1152 | 1152 | 1134 | 84\% | 84\% | 85\% | 1\% | 1\% | 1\% |
| Q2 | 501 | 1152 | 1152 | 1038 | 43\% | 43\% | 48\% | 1\% | 1\% | 2\% |
| Q3 | 661 | 1152 | 1116 | 932 | 57\% | 59\% | 71\% | 1\% | 1\% | 1\% |
| Q4 | 448 | 1152 | 1069 | 822 | 39\% | 42\% | 55\% | 1\% | 2\% | $2 \%$ |
| Q5 | 809 | 1152 | 1052 | 935 | 70\% | $77 \%$ | 87\% | 1\% | 1\% | 1\% |
| Q6 | 546 | 1152 | 939 | 857 | 47\% | 58\% | 64\% | 1\% | 2\% | 2\% |
| Q7 | 541 | 1152 | 899 | 838 | 47\% | 60\% | 65\% | 1\% | 2\% | 2\% |
| Q8 | 293 | 1152 | 855 | 750 | 25\% | $34 \%$ | 39\% | 1\% | $2 \%$ | 2\% |
| Q9 | 440 | 1152 | 778 | 669 | 38\% | 57\% | 66\% | 1\% | 2\% | 2\% |
| Q10 | 372 | 1152 | 678 | 632 | 32\% | 55\% | 59\% | 1\% | 2\% | 2\% |
| Q11 | 206 | 1152 | 642 | 592 | 18\% | 32\% | 35\% | 1\% | 2\% | 2\% |

Figure 6.1: Version P: Three facilities \& SE, Facility Ranges = [Facility I, Facility III],
${ }^{\wedge}$ Standard Error $\left.=\sqrt{ }[p(1-\mathrm{p}) / \mathrm{n})\right]$

| Version $1(\mathrm{n}=3010)$ |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | 0\% | 20\% |  | \% | 60\% | 80 |  | 100\% |
| $\begin{array}{r} \mathrm{B} \\ \mathrm{C} 1 \\ \mathrm{C} 2[12 \\ \mathrm{D} 1 \\ \mathrm{D} 2 \end{array}$ | A1 [3a <br> A2 [4a <br> $1[5 a+$ <br> $[8 a+3=$ $\begin{aligned} & a+5=3 a \\ & +2=8 a \\ & a-11= \\ & 2-11 a= \end{aligned}$ <br> [7a-2 <br> E2 [14 $3 a+6+$ $4[(a+2)$ <br> E5 [a/3 $[4(a-3)$ | $\begin{aligned} & 8=23 \mid \\ & 9=37 \mid \\ & =15 \mid \\ & 16 \mid 1.62 \\ & +14 \mid 1 \\ & 15 \mid 3.2 \\ & a-4 \mid 1 \\ & 5-a \mid 0 \\ & 16 \mid 18 \\ & 2 \mathrm{a}=8 \\ & \mathrm{a}=7 \mid 0 \\ & 5=3 \mid \\ & 1=5 \mid \\ & 21 \mid 8.2 \end{aligned}$ |  |  |  |  |  |  |  |  |
|  |  | Num | er stud |  |  | acilitie |  | Stan | $\operatorname{ard}$ E | rror^ |
| m | N(CR) | I | II | III | I | II | III | I | II | III |
| A1 | 2570 | 3010 | 3010 | 2984 | 85\% | 85\% | 86\% | 1\% | 1\% | 1\% |
| A2 | 2510 | 3010 | 3010 | 2957 | 83\% | 83\% | 85\% | 1\% | 1\% | 1\% |
| B1 | 2027 | 3010 | 2993 | 2775 | 67\% | 68\% | 73\% | 1\% | 1\% | 1\% |
| B2 | 1883 | 3010 | 2919 | 2694 | 63\% | 65\% | 70\% | 1\% | 1\% | 1\% |
| C1 | 1349 | 3010 | 2879 | 2201 | 45\% | 47\% | 61\% | 1\% | 1\% | 1\% |
| C2 | 1220 | 3010 | 2646 | 2048 | 41\% | 46\% | 60\% | 1\% | 1\% | 1\% |
| D1 | 1038 | 3010 | 2575 | 1940 | 34\% | 40\% | 54\% | 1\% | 1\% | 1\% |
| D2 | 784 | 3010 | 2535 | 1850 | 26\% | $31 \%$ | 42\% | 1\% | 1\% | 1\% |
| E1 | 1378 | 3010 | 2501 | 2286 | 46\% | 55\% | 60\% | 1\% | 1\% | 1\% |
| E2 | 1354 | 3010 | 2431 | 2260 | 45\% | 56\% | 60\% | 1\% | 1\% | 1\% |
| E3 | 1144 | 3010 | 2342 | 1921 | 38\% | 49\% | 60\% | 1\% | 1\% | 1\% |
| E4 | 1407 | 3010 | 2173 | 2009 | 47\% | 65\% | 70\% | $1 \%$ | 1\% | 1\% |
| E5 | 1348 | 3010 | 2098 | 1996 | 45\% | 64\% | 68\% | 1\% | 1\% | 1\% |
| E6 | 1037 | 3010 | 2050 | 1914 | 34\% | 51\% | 54\% | 1\% | 1\% | 1\% |

Figure 6.2: Version 1: Three facilities \& SE, Facility Ranges $=[$ Facility I, Facility III], ${ }^{\wedge}$ Standard Error $\left.=\sqrt{ }[p(1-p) / n)\right]$, \# includes $2.5(n=88)$


Figure 6.3: Version 2: Three facilities \& SE, Facility Ranges $=[$ Facility I, Facility III], * 4 items swapped position from Version 1 to 2, $\wedge$ Standard Error $=\sqrt{ }[p(1-p) / n)]$

Version 2-T1 ( $\mathrm{n}=393$ )
$0 \% 10 \% ~ 20 \% ~ 30 \% ~ 40 \% ~ 50 \% ~ 60 \% ~ 70 \% ~ 80 \% ~ 90 \% ~$

-Facility I Facility Range

| Item | N(CR) | Number students |  |  |  | Facilities |  |  | Standard Error^ |  |  |
| :--- | ---: | :--- | :--- | :--- | :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | I | II | III | I | II | III | I | II | III |  |
| A1 | 333 | 393 | 393 | 391 | $85 \%$ | $85 \%$ | $85 \%$ | $2 \%$ | $2 \%$ | $2 \%$ |  |
| A2 | 327 | 393 | 393 | 388 | $83 \%$ | $83 \%$ | $84 \%$ | $2 \%$ | $2 \%$ | $2 \%$ |  |
| B2* | 241 | 393 | 390 | 359 | $61 \%$ | $62 \%$ | $67 \%$ | $2 \%$ | $2 \%$ | $2 \%$ |  |
| B1* | 234 | 393 | 377 | 358 | $60 \%$ | $62 \%$ | $65 \%$ | $2 \%$ | $2 \%$ | $3 \%$ |  |
| C2* | 169 | 393 | 375 | 287 | $43 \%$ | $45 \%$ | $59 \%$ | $2 \%$ | $3 \%$ | $3 \%$ |  |
| C1* | 165 | 393 | 349 | 275 | $42 \%$ | $47 \%$ | $60 \%$ | $2 \%$ | $3 \%$ | $3 \%$ |  |
| D1 | 138 | 393 | 344 | 259 | $35 \%$ | $40 \%$ | $53 \%$ | $2 \%$ | $3 \%$ | $3 \%$ |  |
| D2 | 119 | 393 | 342 | 251 | $30 \%$ | $35 \%$ | $47 \%$ | $2 \%$ | $3 \%$ | $3 \%$ |  |
| E1 | 199 | 393 | 338 | 318 | $51 \%$ | $59 \%$ | $63 \%$ | $3 \%$ | $3 \%$ | $3 \%$ |  |
| E2 | 171 | 393 | 333 | 302 | $44 \%$ | $51 \%$ | $57 \%$ | $3 \%$ | $3 \%$ | $3 \%$ |  |
| E3 | 160 | 393 | 324 | 272 | $41 \%$ | $49 \%$ | $59 \%$ | $2 \%$ | $3 \%$ | $3 \%$ |  |
| E4 | 168 | 393 | 304 | 285 | $43 \%$ | $55 \%$ | $59 \%$ | $2 \%$ | $3 \%$ | $3 \%$ |  |
| E5 | 178 | 393 | 294 | 270 | $45 \%$ | $61 \%$ | $66 \%$ | $3 \%$ | $3 \%$ | $3 \%$ |  |
| E6 | 154 | 393 | 281 | 265 | $39 \%$ | $55 \%$ | $58 \%$ | $2 \%$ | $3 \%$ | $3 \%$ |  |

Figure 6.4: Version 2-T1: Three facilities \& SE, Facility Ranges $=[$ Facility I, Facility III],

* 4 items swapped position from Version 1 to 2, $\wedge$ Standard Error $=\sqrt{ }[p(1-p) / n)]$


Figure 6.5: Version 2-T2: Three facilities \& SE, Facility Ranges = [Facility I, Facility III], * 4 items swapped position from Version 1 to 2, ${ }^{\wedge}$ Standard Error $\left.=\sqrt{ }[p(1-p) / n)\right]$

## Section 7: Error Analyses: Item Response Summary

- Table 7.1 gives precise coding rules, then tables that follow provide the top four errors to each item in each test version

Table 7.1: Precise Coding Rules for Errors

| Error <br> Code | Label | Coding Rules |
| :---: | :---: | :---: |
| AS | Addition as Subtraction* | For $\mathrm{a} x+\mathrm{b}=\mathrm{c}$, response $\left({ }^{\mathrm{c}+\mathrm{b}} / \mathrm{a}\right)$ is consistent with solving $\mathrm{a} x-\mathrm{b}=\mathrm{c}$ instead. This is rare in this dataset. Specific example of a PM for Gps A, B \& E. |
| DM | Division as Multiplication* | For $x / \mathrm{a}+\mathrm{b}=\mathrm{c}$, response $\left({ }^{\mathrm{c}-\mathrm{b}} / \mathrm{a}\right)$ is consistent with solving $\mathrm{a} x+\mathrm{b}=\mathrm{c}$ instead. |
| FS | First Step | For $\mathrm{ax}+\mathrm{b}=\mathrm{c}$, response ( $\mathrm{c}-\mathrm{b}$ ) is consistent with solving $x+\mathrm{b}=\mathrm{c}$ instead. This response is also obtained by carrying out only the first step in solving $\mathrm{a} x+\mathrm{b}=\mathrm{c}$. |
| HW | Half-way | For non-integer $\mathrm{CR}=\mathrm{K} . \mathrm{X}$, then HW $=$ K. 5 |
| IC | Ignore unknowns and Compare | For $\mathrm{a} x+\mathrm{b}=\mathrm{c} x+\mathrm{d}$, response is consistent with ignoring unknowns (or setting value of $x$ to 1) and comparing totals on each side, giving responses $\pm((\mathrm{a}+\mathrm{b})-(\mathrm{c}+\mathrm{d}))$. (Similar to MA error, but not used consistently by MA students) |
| MA | Multiplication as Addition | For $\mathrm{a} x+\mathrm{b}=\mathrm{c}$, response (c-b-a) is consistent with solving $\mathrm{a}+x+\mathrm{b}=\mathrm{c}$ instead. Usually called Conjoin error. |
| MC | Miscalculation | Miscalculation when using a correct procedure for integer $\mathrm{CR}=\mathrm{K}$, (e.g., misremembered multiplication fact). $\mathrm{MC}=\{\mathrm{K}-1, \mathrm{~K}+1\}$. |
| NI | Near Integer | For non-integer $\mathrm{CR}=\mathrm{K} . \mathrm{X}$, then $\mathrm{NI}=\{\mathrm{K}, \mathrm{K}+1\}$ (i.e., Floor or Ceiling of CR). Likely to be students using the substituting strategy not looking beyond whole numbers. |
| NN | Near Number | Composite code: For non-integer $\mathrm{CR}=\mathrm{K} . \mathrm{X}, \mathrm{NN}=\mathrm{NI} \cup \mathrm{HW}$ $=\{\mathrm{K}, \mathrm{K}+1 / 2, \mathrm{~K}+1\}$ |
| OO | Order of Operations | For items involving brackets and fractions, response is consistent with completing correct steps in wrong order. |
| PM | Permutations | Composite code: For $\pm \mathrm{ax} \pm \mathrm{b}= \pm \mathrm{c} x \pm \mathrm{d}$, collection of incorrect responses <br> $\pm(\mathrm{b} \pm \mathrm{d}) / \pm(\mathrm{a} \pm \mathrm{c})$ or reciprocal. PM errors result from various combinations of using incorrect inverse operation (,+- ) and Reciprocal error. |
| RC | Reciprocal | When reaching $\mathrm{m} x=\mathrm{n}$, writes $x=\mathrm{m} / \mathrm{n}$. |
| RV | Reverse | For $\mathrm{b}-\mathrm{ax}=\mathrm{c}$, response $\left({ }^{\mathrm{c}+\mathrm{b}} / \mathrm{a}\right)$ is consistent with solving $\mathrm{a} x-\mathrm{b}=\mathrm{c}$ instead. |
| SA | Subtraction as Addition* | For $\mathrm{a} x-\mathrm{b}=\mathrm{c}$, response $(\mathrm{c}-\mathrm{b} / \mathrm{a})$ is consistent with solving $\mathrm{a} x+\mathrm{b}=\mathrm{c}$ instead (similarly with $\mathrm{b}-\mathrm{ax}=\mathrm{c}$ ). Specific example of a PM for Gps A, B \& E. |

Note: $\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d} \in \mathrm{Z}^{+}$to show students' perspective.
*There are 4 cases of "use given operation instead of inverse operation": SA and DM are more common than AS and MD (MD is rare so it isn't listed above)

Table 7.2: Version P ( $\mathrm{n}=1152$ ): Item Response Summary

| Item details | Rank | Response | Freq | IR\% | Likely Calculation | Best explanation |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Q1 | CIR1 | 24 | 23 | 14\% | 37-9-4 | MA |
| $4 n+9=37$ | CIR2 | 28 | 18 | 11\% | 37-9 | FS |
| $\mathrm{CR}=7$ | CIR3 | 8 | 13 | 8\% | CR + 1 | MC |
| CR: 969 (84\%) | CIR4 | 6 | 11 | 7\% | CR-1 | MC |
| OR: 18 (2\%) | Misc. | - | 100 | 61\% |  |  |
| IR: 165 (14\%) | Total |  | 165 | 100\% |  |  |
| Q2 | CIR1 | 6 | 174 | 32\% | $(21+3) / 4$ | OO |
| $4(n-3)=21$ | CIR2 | 8 | 54 | 10\% | Floor (8.25) | NI |
| $\mathrm{CR}=8.25$ | CIR3 | 2.25 | 43 | 8\% | (21-12)/4 | SA |
| CR: 501 (43\%) | CIR4 | 10 | 22 | 4\% |  |  |
| OR: 114 (10\%) | Misc. | - | 244 | 45\% |  |  |
| IR: 537 (47\%) | Total |  | 537 | 100\% |  |  |
| Q3 | CIR1 | 2* | 37 | 14\% | $(6-2) /(5-3)$ | PM |
| $5 n-2=3 n+6$ | CIR2 | 3* | 33 | 12\% | CR-1 | MC |
| $\mathrm{CR}=4$ | CIR3 | 1* | 30 | 11\% |  |  |
| CR: 661 (57\%) | CIR4 | 8 | 23 | 8\% | $6+2$ | FS |
| OR: 220 (19\%) | Misc. | - | 148 | 55\% |  |  |
| IR: 271 (24\%) | Total |  | 271 | 100\% |  |  |
| Q4 | CIR1 | 3* | 70 | 19\% | Floor (3.25) | NI |
| $12 n+2=8 n+15$ | CIR2 | 4* | 25 | 7\% | Ceiling (3.25) | NI |
| $\mathrm{CR}=3.25,{ }^{13} / 4$ | CIR3 | 8 | 22 | 6\% |  |  |
| CR: 448 (39\%) | CIR4 | 6 | 23 | 6\% |  |  |
| OR: 330 (29\%) | Misc. | - | 234 | 63\% |  |  |
| IR: 374 (32\%) | Total |  | 374 | 100\% |  |  |
| Q5 | CIR1^ | 3* | 18 | 14\% |  |  |
| $3 a+8=23$ | CIR1^ | 12 | 18 | 14\% | 23-8-3 | MA |
| $\mathrm{CR}=5$ | CIR3^ | 15 | 13 | 10\% | 23-8 | FS |
| CR: 809 (70\%) | CIR3^ | 4* | 13 | 10\% | CR-1 | MC |
| OR: 217 (19\%) | Misc. | - | 64 | 51\% |  |  |
| IR: 126 (11\%) | Total |  | 126 | 100\% |  |  |
| Q6 | CIR1 | 6 | 82 | 26\% | Ceiling (5.75) | NI |
| $4 a-7=16$ | CIR2 | 2.25 | 50 | 16\% | (16-7)/4 | SA |
| $\mathrm{CR}={ }^{23} / 4,5.75$ | CIR3 | 5* | 19 | 6\% | Floor (5.75) | NI |
| CR: 546 (47\%) | CIR4 | 7 | 15 | 5\% |  |  |
| OR: 295 (26\%) | Misc. | - | 145 | 47\% |  |  |
| IR: 311 (27\%) | Total |  | 311 | 100\% |  |  |
| Q7 | CIR1 | 11 | 84 | 28\% | $(8+14) / 2$ | RV |
| $14-2 a=8$ | CIR2 | -3 | 46 | 15\% | $(8-14) / 2$ | SA |
| $\mathrm{CR}=3$ | CIR3 | 4* | 26 | 9\% | 14-2-8\& CR + 1 | MA \& MC |
| CR: 541 (47\%) | CIR4 | 6 | 19 | 6\% | 14-8 | FS |
| OR: 314 (27\%) | Misc. | - | 122 | 41\% |  |  |


| IR: 297 (26\%) | Total |  | 297 | 100\% |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Q8 | CIR1 | 1.8 | 101 | 22\% | $(15-6) / 5$ | SA |
| $6-5 a=15$ | CIR2 | 4.2 | 72 | 16\% | $(15+6) / 5$ | RV |
| $\mathrm{CR}=-1.8$ | CIR3 | 4* | 31 | 7\% | 15-6-5 | $\mathrm{SA}+\mathrm{IC}(11,15)$ |
| CR: 293 (25\%) | CIR4 | 5* | 22 | 5\% |  |  |
| OR: 402 (35\%) | Misc. | - | 231 | 51\% |  |  |
| IR: 457 (40\%) | Total |  | 457 | 100\% |  |  |
| Q9 | CIR1 | 4* | 24 | 10\% |  |  |
| $9 a+3=7 a+15$ | CIR2 | 5* | 23 | 10\% | CR-1 | MC |
| $\mathrm{CR}=6$ | CIR3 | 7 | 22 | 10\% | $\mathrm{CR}+1$ | MC |
| CR: 440 (38\%) | CIR4 | 3* | 17 | 7\% |  |  |
| OR: 483 (42\%) | Misc. | - | 143 | 62\% |  |  |
| IR: 229 20(\%) | Total |  | 229 | 100\% |  |  |
| Q10 | CIR1 | 4* | 31 | 12\% | $(3+14)-(8+5)$ | $\mathrm{IC}(13,17)$ |
| $8 a+5=3 a+14$ | CIR2 | 2* | 23 | 9\% | Ceiling (1.8) | NI |
| $\mathrm{CR}={ }^{9} / 5,1.8$ | CIR3 | 3* | 20 | 8\% |  |  |
| CR: 372 (32\%) | CIR4 | 6* | 16 | 6\% |  |  |
| OR: 520 (45\%) | Misc. | - | 170 | 65\% |  |  |
| IR: 260 (23\%) | Total |  | 260 | 100\% |  |  |
| Q11 | CIR1 | -0.7 | 29 | 8\% | $(12-5) /(1-11)$ | PM |
| $12-11 a=5-a$ | CIR2 | 7/12 | 28 | 7\% | $(12-5) /(11+1)$ | PM |
| $\mathrm{CR}={ }^{7} / 10,0.7$ | CIR3 | 4* | 26 | 7\% | $(12-11)-5$ | $\mathrm{IC}(1,5)$ |
| CR: 206 (18\%) | CIR4 | 1.7 | 25 | 6\% | $(12+5) /(11-1)$ | PM |
| OR: 560 (49\%) | Misc. | - | 278 | 72\% |  |  |
| IR: 386 (34\%) | Total |  | 386 | 100\% |  |  |

* small integer responses (1-5) could be guesses by some students, ^ tied ranks

Floor ( x ) is greatest integer $\leq \mathrm{x}$, and Ceiling ( x ) is least integer $\geq \mathrm{x}$

Table 7.3: Version 1 ( $\mathrm{n}=3010$ ): Item Response Summary

| Item details | Rank | Response | Freq | IR\% | Likely Calculation | Best explanation |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| A1 | CIR1 | 12 | 124 | 30\% | 23-8-3 | MA |
| $3 x+8=23$ | CIR2 | 15 | 108 | 26\% | 23-8 | FS |
| $\mathrm{CR}=5$ | CIR3 | 3* | 32 | 8\% |  |  |
| CR: 2570 (85\%) | CIR4 | 4* | 29 | 7\% | CR-1 | MC |
| OR: 26 (1\%) | Misc. |  | 121 | 29\% |  |  |
| IR: 414 (14\%) | Total |  | 414 | 100\% |  |  |
| A2 | CIR1 | 24 | 125 | 28\% | 37-9-4 | MA |
| $4 x+9=37$ | CIR2 | 6 | 63 | 14\% | CR - 1 | MC |
| $\mathrm{CR}=7$ | CIR3 | 28 | 51 | 11\% | 37-9 | FS |
| CR: 2510 (83\%) | CIR4 | 8 | 32 | 7\% | CR + 1 | MC |
| OR: 53 (2\%) | Misc. |  | 176 | 39\% |  |  |
| IR: 447 (15\%) | Total |  | 447 | 100\% |  |  |
| B1 | CIR1 | 3* | 226 | 30\% | 15-7-5 | MA |
| $5 x+7=15$ | CIR2 | 8 | 78 | 10\% | 15-7 | FS |
| $\mathrm{CR}=8 / 5,1.6$ | CIR3 | 2* | 74 | 10\% | Ceiling (1.6) | NI |
| CR: 2027 (67\%) | CIR4 | 1.5 | 36 | 5\% | half-way 1\&2 | HW |
| OR: 235 (8\%) | Misc. |  | 334 | 45\% |  |  |
| IR: 748 (25\%) | Total |  | 748 | 100\% |  |  |
| B2 | CIR1 | 5* | 197 | 24\% | 16-3-8 | MA |
| $8 x+3=16$ | CIR2 | 2* | 79 | 10\% | Ceiling (1.625) | NI |
| $\mathrm{CR}={ }^{13} / 8,1.625$ | CIR3 | 13 | 53 | 7\% | 16-3 | FS |
| CR: 1883 (63\%) | CIR4 | 1.5 | 50 | 6\% | half-way 1\&2 | HW |
| OR: 316 (10\%) | Misc. |  | 432 | 53\% |  |  |
| IR: 811 (27\%) | Total |  | 811 | 100\% |  |  |
| C1 | CIR1 | 2* | 118 | 14\% | Ceiling (1.8) | NI |
| $8 x+5=3 x+14$ | CIR2 | 4* | 69 | 8\% | $(3+14)-(8+5)$ | $\mathrm{IC}(13,17)$ |
| $\mathrm{CR}={ }^{9} / 5,1.8$ | CIR3 | 3* | 63 | 7\% |  |  |
| CR: 1349 (45\%) | CIR4 | 5* | 45 | 5\% |  |  |
| OR: 809 (27\%) | Misc. |  | 557 | 65\% |  |  |
| IR: 852 (28\%) | Total |  | 852 | 100\% |  |  |
| C2 | CIR1 | 3* | 87 | 11\% | Floor (3.25) | NI |
| $12 x+2=8 x+15$ | CIR2 | 4* | 54 | 7\% | Ceiling (3.25) | NI |
| $\mathrm{CR}={ }^{13} / 4,3.25$ | CIR3 | 2* | 52 | 6\% |  |  |
| CR: 1220 (41\%) | CIR4 | 7 | 42 | 5\% |  |  |
| OR: 962 (32\%) | Misc. |  | 593 | 72\% |  |  |
| IR: 828 (28\%) | Total |  | 828 | 100\% |  |  |
| D1 | CIR1 | 3* | 143 | 16\% | $(11+4) /(7-2)$ | PM |
| $7 x-11=2 x-4$ | CIR2 | 2* | 71 | 8\% | $(11-7)-(4-2) \&$ Ceiling (1.4) | $\mathrm{IC}(4,2) \& N \mathrm{I}$ |
| $\mathrm{CR}={ }^{7} / 5,1.4$ | CIR3 | 4* | 66 | 7\% |  |  |
| CR: 1038 (34\%) | CIR4 | -3 | 52 | 6\% | $(-11-4) /(7-2)$ | PM |
| OR: 1070 (36\%) | Misc. |  | 570 | 63\% |  |  |


| IR: 902 (30\%) | Total |  | 902 | 100\% |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| D2 | CIR1 | 1* | 73 | 7\% | Ceiling (0.7) | NI |
| $12-11 x=5-x$ | CIR2 | $10 / 7$ | 65 | 6\% | $(11-1) /(12-5)$ | RC |
| $\mathrm{CR}={ }^{7} / 10,0.7$ | CIR2 | $7 / 12$ | 65 | 6\% | $(12-5) /(11+1)$ | PM |
| CR: 784 (26\%) | CIR4 | 4* | 57 | 5\% | (12-11)-5 | $\mathrm{IC}(1,5)$ |
| OR: 1160 (39\%) | Misc. |  | 806 | 76\% |  |  |
| IR: 1066 (35\%) | Total |  | 1066 | 100\% |  |  |
| E1 | CIR1 | 2* | 357 | 39\% | (16-2)/7 \& Floor (2.57) | SA \& NI |
| $7 x-2=16$ | CIR2 | 11 | 68 | 7\% | $16+2-7$ | MA |
| $\mathrm{CR}={ }^{18} / 7,2.57$ | CIR3 | 3* | 50 | 6\% | Ceiling (2.57) | NI |
| CR: 1378 (46\%) | CIR4 | 7 | 41 | 5\% | 16-2-7 | MA+SA |
| OR: 724 (24\%) | Misc. |  | 392 | 43\% |  |  |
| IR: 908 (30\%) | Total |  | 908 | 100\% |  |  |
| E2 | CIR1 | 11 | 179 | 20\% | $(8+14) / 2$ | RV |
| $14-2 x=8$ | CIR2 | 4* | 119 | 13\% | 14-2-8 | MA |
| $\mathrm{CR}=3$ | CIR3 | -3 | 112 | 12\% | $(8-14) / 2$ | SA |
| CR: 1354 (45\%) | CIR4 | 6 | 70 | 8\% | 14-8 | FS |
| OR: 750 (25\%) | Misc. |  | 426 | 47\% |  |  |
| IR: 906 (30\%) | Total |  | 906 | 100\% |  |  |
| E3 | CIR1 | 1* | 111 | 14\% | 7-6 | FS |
| $3 x+6+2 x=7$ | CIR2 | 5* | 102 | 13\% | $(3+2) /(7-6)$ | RC |
| $\mathrm{CR}=1 / 5,0.2$ | CIR3 | 4* | 50 | 6\% | $3+6+2-7$ | $\mathrm{IC}(11,7)$ |
| CR: 1144 (38\%) | CIR4 | 2* | 48 | 6\% |  |  |
| OR: 1089 (36\%) | Misc. |  | 466 | 60\% |  |  |
| IR: 777 (26\%) | Total |  | 777 | 100\% |  |  |
| E4 | CIR1 | 5* | 69 | 11\% | $(3-2) \times 5$ | OO(1) |
| $(x+2) / 5=3$ | CIR2 | 3* | 63 | 10\% |  |  |
| $\mathrm{CR}=13$ | CIR3 | 1* | 54 | 9\% |  |  |
| CR: 1407 (47\%) | CIR4 | 2.6 | 42 | 7\% | $3-2 / 5$ | $\mathrm{OO}(2)$ |
| OR: 1001 (33\%) | Misc. |  | 374 | 62\% |  |  |
| IR: 602 (20\%) | Total |  | 602 | 100\% |  |  |
| E5 | CIR1 | 14 | 106 | 16\% | 5×3-1 | OO |
| $x / 3+1=5$ | CIR2 | 4* | 66 | 10\% | 5-1 | FS |
| $\mathrm{CR}=12$ | CIR3 | 2* | 62 | 10\% | $(5+1) / 3$ | DM + AS |
| CR: 1348 (45\%) | CIR4 | 4/3 | 59 | 9\% | $(5-1) / 3$ | DM |
| OR: 1014 (34\%) | Misc. |  | 355 | 55\% |  |  |
| IR: 648 (22\%) | Total |  | 648 | 100\% |  |  |
| E6 | CIR1 | 6 | 143 | 16\% | $(21+3) / 4$ | OO |
| $4(x-3)=21$ | CIR2 | 2.25 | 70 | 8\% | $(21-12) / 4$ | SA |
| $\mathrm{CR}={ }^{33} / 4,8.25$ | CIR3 | 8 | 52 | 6\% | Floor (8.25) | NI |
| CR: 1037 (34\%) | CIR4 | 7 | 43 | 5\% |  |  |
| OR: 1096 (36\%) | Misc. |  | 569 | 65\% |  |  |
| IR: 877 (29\%) | Total |  | 877 | 100\% |  |  |

* small integer responses (1-5) could be guesses by some students

Floor( x ) is greatest integer $\leq \mathrm{x}$, and Ceiling( x ) is least integer $\geq \mathrm{x}$

Table 7.4: Version 2 ( $\mathrm{n}=734$ ): Item Response Summary (CIR not given for groups <10)

| Item details | Rank | Response | Freq | IR\% | Likely Calculation | Best explanation |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| A1 | CIR1 | 8 | 36 | 39\% | 23-11-4 | MA |
| $4 n+11=23$ | CIR2 | 12 | 15 | 16\% | 23-11 | FS |
| $\mathrm{CR}=3$ | CIR3 | 2* | 10 | 11\% |  |  |
| CR: 639 (87\%) |  |  |  |  |  |  |
| OR: 2 (0\%) | Misc. |  | 32 | 34\% |  |  |
| IR: 93 (13\%) | Total |  | 93 | 100\% |  |  |
| A2 | CIR1 | 18 | 28 | 30\% | 26-5-3 | MA |
| $3 n+5=26$ | CIR2 | 21 | 10 | 11\% | 26-5 | FS |
| CR = 7 |  |  |  |  |  |  |
| CR: 637 (87\%) |  |  |  |  |  |  |
| OR: 5 (1\%) | Misc. |  | 54 | 59\% |  |  |
| IR: 92 (13\%) | Total |  | 92 | 100\% |  |  |
| B2* | CIR1 | 5* | 52 | 30\% | 16-3-8 | MA |
| $8 n+3=16$ | CIR2 | 2* | 19 | 11\% | Ceiling (1.625) | NI |
| $\mathrm{CR}={ }^{13} / 8,1.625$ | CIR3 | 13 | 10 | 6\% | 16-3 | FS |
| CR: 516 (70\%) |  |  |  |  |  |  |
| OR: 44 (6\%) | Misc. |  | 93 | 53\% |  |  |
| IR: 174 (24\%) | Total |  | 174 | 100\% |  |  |
| B1* | CIR1 | 3* | 44 | 23\% | 15-7-5 | MA |
| $5 n+7=15$ | CIR2 | 2* | 25 | 13\% | Ceiling (1.6) | NI |
| $\mathrm{CR}={ }^{8} / 5,1.6$ | CIR3 | 8 | 16 | 8\% | 15-7 | FS |
| CR: 496 (68\%) | CIR4 | 5/8 | 13 | 7\% | 5/(15-7) | RC |
| OR: 45 (6\%) | Misc. |  | 95 | 49\% |  |  |
| IR: 193 (26\%) | Total |  | 193 | 100\% |  |  |
| C2* | CIR1 | 4* | 30 | 16\% | Ceiling (3.25) | NI |
| $11 n+3=7 n+16$ | CIR2 | 3* | 19 | 10\% | Floor (3.25) | NI |
| $\mathrm{CR}={ }^{13} / 4,3.25$ | CIR3 | 2* | 12 | 6\% |  |  |
| CR: 396 (54\%) | CIR4 | 9 | 11 | 6\% | $(7+16)-(11+3)$ | $\mathrm{IC}(14,23)$ |
| OR: 146 (20\%) | Misc. |  | 120 | 63\% |  |  |
| IR: 192 (26\%) | Total |  | 192 | 100\% |  |  |
| C1* | CIR1^ | 4* | 17 | 9\% | $(4+12)-(9+3)$ | $\mathrm{IC}(12,16)$ |
| $9 n+3=4 n+12$ | CIR1^ | 3* | 17 | 9\% |  |  |
| $\mathrm{CR}={ }^{9} / 5,1.8$ | CIR3 | 2* | 13 | 7\% | Ceiling (1.8) | NI |
| CR: 391 (53\%) | CIR4 | 1* | 12 | 7\% |  |  |
| OR: 161 (22\%) | Misc. |  | 123 | 68\% |  |  |
| IR: 182 (25\%) | Total |  | 182 | 100\% |  |  |
| D1 | CIR1 | 3* | 39 | 18\% | $(11+4) /(7-2)$ | PM |
| $7 n-11=2 n-4$ | CIR2 | 2* | 16 | 7\% | $(11-7)-(4-2) \&$ Ceiling(1.4) | $\mathrm{IC}(4,2) \& \mathrm{NI}$ |
| $\mathrm{CR}={ }^{7} / 5,1.4$ | CIR3^ | 1* | 11 | 5\% | Floor (1.4) | NI |
| CR: 328 (45\%) | CIR3^ | -3 | 11 | 5\% | $(-11-4) /(7-2)$ | PM |
| OR: 184 (25\%) | Misc. |  | 145 | 65\% |  |  |


| IR: 222 (30\%) | Total |  | 222 | 100\% |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| D2 | CIR1 | 7/12 | 33 | 11\% | $(12-5) /(11+1)$ | PM |
| $12-11 n=5-n$ | CIR2 | -0.7 | 23 | 8\% | $(12-5) /(-11+1)$ | PM |
| $\mathrm{CR}=7 / 10,0.7$ | CIR3 | 5* | 19 | 7\% |  |  |
| CR: 240 (33\%) | CIR4 | $-7 / 12$ | 18 | 6\% | $(-12+5) /(11+1)$ | PM |
| OR: 203 (28\%) | Misc. |  | 198 | 68\% |  |  |
| IR: 291 (40\%) | Total |  | 291 | 100\% |  |  |
| E1 | CIR1 | 3* | 66 | 36\% | $(16-1) / 5$ \& Floor (3.4) | SA \& NI |
| $5 n-1=16$ | CIR2 | 3.2 | 12 | 6\% | $17 / 5=3 \mathrm{rem} 2$ | fraction error |
| $\begin{aligned} & \mathrm{CR}={ }^{17} / 5,3.4 \\ & \mathrm{CR}: 446(61 \%) \end{aligned}$ |  |  |  |  |  |  |
| OR: 103 (14\%) | Misc. |  | 107 | 58\% |  |  |
| IR: 185 (25\%) | Total |  | 185 | 100\% |  |  |
| E2 | CIR1 | 12 | 55 | 22\% | $(9+15) / 2$ | RV |
| $15-2 n=9$ | CIR2 | -3 | 46 | 19\% | $(9-15) / 2$ | SA |
| $\mathrm{CR}=3$ | CIR3 | 6 | 27 | 11\% | 15-9 | FS |
| CR: 357 (49\%) | CIR4 | 4* | 17 | 7\% | 15-2-9 | MA |
| OR: 130 (18\%) | Misc. |  | 102 | 41\% |  |  |
| IR: 247 (34\%) | Total |  | 247 | 100\% |  |  |
| E3 | CIR1 | 1* | 46 | 21\% | 5-4 | FS |
| $2 n+4+3 n=5$ | CIR2 | 5* | 32 | 15\% | $(2+3) /(5-4)$ | RC |
| $\mathrm{CR}=1 / 5,0.2$ | CIR3 | 3* | 13 | 6\% |  |  |
| CR: 352 (48\%) | CIR4 | $2^{*}$ | 13 | 6\% |  |  |
| OR: 168 (23\%) | Misc. |  | 110 | 51\% |  |  |
| IR: 214 (29\%) | Total |  | 214 | 100\% |  |  |
| E4 | CIR1 | 2* | 24 | 13\% |  |  |
| ${ }^{(n+1)} / 5=3$ | CIR2 | 2.8 | 21 | 11\% | $3-1 / 5$ | OO |
| $\mathrm{CR}=14$ | CIR3 | 4* | 20 | 11\% |  |  |
| CR: 392 (53\%) | CIR4 | 3* | 15 | 8\% |  |  |
| OR: 155 (21\%) | Misc. |  | 107 | 57\% |  |  |
| IR: 187 (25\%) | Total |  | 187 | 100\% |  |  |
| E5 | CIR1 | 29 | 38 | 20\% | $4 \times 8-3$ | OO |
| $n / 4+3=8$ | CIR2 | 5* | 23 | 12\% | 8-3 | FS |
| $\mathrm{CR}=20$ | CIR3 | 2* | 17 | 9\% |  |  |
| CR: 370 (50\%) | CIR4 | $5 / 4$ | 16 | 8\% | $(8-3) / 4$ | DM |
| OR: 174 (24\%) | Misc. |  | 96 | 51\% |  |  |
| IR: 190 (26\%) | Total |  | 190 | 100\% |  |  |
| E6 | CIR1 | 2* | 35 | 19\% | $(8+2) / 5$ | OO |
| $5(n-2)=8$ | CIR2 | 5* | 19 | 11\% |  |  |
| $\mathrm{CR}=18 / 5,3.6$ | CIR3 | 3* | 15 | 8\% | Floor (3.6) | NI |
| CR: 365 (50\%) | CIR4 | 4* | 10 | 6\% | Ceiling (3.6) | NI |
| OR: 189 (26\%) | Misc. |  | 101 | 56\% |  |  |
| IR: 180 (25\%) | Total |  | 180 | 100\% |  |  |

* small integer responses (1-5) could be guesses by some students, $\wedge$ tied ranks

Floor ( x ) is greatest integer $\leq x$, and Ceiling ( x ) is least integer $\geq \mathrm{x}$

## Appendix 1: Screenshots of Versions P, 1 \& 2

- Only the last two screens (SLE: Solving Linear Equations) of Version P are relevant to this report; the first four screens involve a different task: choosing an equation to match a word problem.
- Versions $1 \& 2$ contain only SLE items to reduce the length of the tests (in Version P, of the 1237 students who gave a response to at least one item in the test, 298 ( $24 \%$ ) did not respond to any item on Screen 6)

| Screen $1: \# 3112$ |  |
| :--- | :--- | :--- |

Figure A1.1: Screenshots from Version P

| Version 1 | Version 2 |
| :---: | :---: |
| Screen 1: \# 3105 | Screen 1: \# 3108 |
| Solve each of the following equations using a pen and paper. Type your answer to each in the space provided. A calculator is provided for your use. | Solve each of the following equations using a pen and paper. Type your answer to each in the space provided. A calculator is provided for your use. |
| $3 a+8=23$ <br> $4 a+9=37$ <br> so $a=$ $\square$ so $a=$ $\square$ | $4 n+11=23$ $3 n+5=26$ <br> so $n=$ $\square$ so $n=$ $\square$ |
| $5 a+7=15$ $8 a+3=16$ <br> so $a=$ $\square$ so $a=$ $\square$ | $8 n+3=16$ $5 n+7=15$ <br> so $n=$ $\square$ so $n=$ $\square$ |
| Click here if you need a calculator | Click here if you need a calculator |
| Screen 2: \# 3106 | Screen 2: \# 3109 |
| Solve each of the following equations using a pen and paper. Type your answer to each in the space provided. A calculator is provided for your use. | Solve each of the following equations using a pen and paper. Type your answer to each in the space provided. A calculator is provided for your use. |
| $8 a+5=3 a+14$ $12 a+2=8 a+15$ <br> so $a=$ $\square$ so $a=$ $\square$ | $11 n+3=7 n+16$ $9 n+3=4 n+12$ <br> so $n=$ $\square$ so $n=$ $\square$ |
| $7 a-11=2 a-4$ $12-11 a=5-a$ <br> so $a=$ $\square$ so $a=$ $\square$ | $7 n-11=2 n-4$ $12-11 n=5-n$ <br> so $n=$ $\square$ so $n=$ $\square$ |
| Click here if you need a calculator | Click here if you need a calculator |
| Screen 3: \# 3107 | Screen 3: \# 3110 |
| Solve each of the following equations using a pen and paper. Type your answer to each in the space provided. <br> A calculator is provided for your use. $7 a-2=16$ $14-2 a=8$ <br> so $a=$ $\square$ $\text { so } a=$ $\square$ | Solve each of the following equations using a pen and paper. Type your answer to each in the space provided. A calculator is provided for your use. <br> $5 n-1=16$ $15-2 n=9$ <br> so $n=$ $\square$ so $n=$ $\square$ |
| $3 a+6+2 a=7$ $\frac{a+2}{5}=3$ <br> so $a=$ $\square$ $\text { so } a=$ $\square$ | $2 n+4+3 n=5$ $\frac{n+1}{5}=3$ <br> so $n=$ $\square$ so $n=$ $\square$ |
| $\frac{a}{3}+1=5$ $4(a-3)=21$ <br> so $a=$ $\square$ $\text { so } a=$ $\square$ | $\frac{n}{4}+3=8 \quad 5(n-2)=8$ <br> so $n=$ $\square$ so $n=$ $\square$ |
| Click here if you need a calculator |  |

Figure A1.2: Screenshots from Versions 1 \& 2

## Appendix 2: Accuracy Coding

- Frequency distribution of responses marked CR, for each item in each test
- The responses are sorted by decreasing frequency, except for Q9 (E1) in Version 1 Table A2.1: Version P: Responses marked correct on each item


Table A2.2: Versions 1 \& 2: Responses marked correct on each item

| Version 1 |  | Version 2 |  |
| :---: | :---: | :---: | :---: |
| $\mathrm{A} 1[3 a+8=23 \mid 5]$ |  | $\mathrm{A} 1[4 n+11=23 \mid 3]$ |  |
| Responses | Freq | Responses | Freq |
| 5 | 2570 | 3 | 639 |
| Total | 2570 | Total | 639 |
|  |  |  |  |
| A2 [4a+9=37]7] |  | A2 [3n+5=26\|7] |  |
| Responses | Freq | Responses | Freq |
| 7 | 2510 | 7 | 635 |
| Total | 2510 | 21/3 | 2 |
|  |  | Total | 637 |
|  |  |  |  |
| B1 [5a+7 = 15 \|1.6] |  | $\mathrm{B} 1 *[5 n+7=15 \mid 1.6]$ |  |
| Responses | Freq | Responses | Freq |
| 1.6 | 1826 | 1.6 | 314 |
| $8 / 5$ | 168 | 8/5 | 164 |
| $13 / 5$ | 13 | $13 / 5$ | 12 |
| 1.60 | 8 | 1/6 | 2 |
| 1,6 | 3 | $16 / 10$ | 1 |
| 1.61 | 2 | 1.600 | 1 |
| $\sim 1.6$ | 1 | 16/10 | 1 |
| $16 / 10$ | 1 | 8 over 5 | 1 |
| 1 and 3 fifths | 1 | Total | 496 |
| 1 and 3/5 | 1 |  |  |
| 1.6 or 8/5 | 1 |  |  |
| 1.6\& | 1 |  |  |
| 1r3 | $1^{\wedge}$ |  |  |
| Total | 2027 |  |  |
| $\wedge$ student repeatedly used this notation |  |  |  |
| B2 [8a+3=16\|1.625] |  | B2* [8n+3 = 16\|1.625] |  |
| Responses | Freq | Responses | Freq |
| 1.625 | 1428 | 1.625 | 271 |
| 13/8 | 156 | 13/8 | 172 |
| 1.6 | 138 | 1.6 | 40 |
| 1.63 | 82 | $15 / 8$ | 12 |
| 1.62 | 56 | 1.63 | 12 |
| $15 / 8$ | 9 | 1.62 | 3 |
| 15/8 | 3 | 163/100 | 1 |
| 1.60 | 2 | 1 and 5/8 OR 13/8 | 1 |
| 1 and 5 eighths | 1 | 1/6 | $1^{\wedge}$ |
| 1 and 5/8 | 1 | 1/625 | 1 |
| 1,625 | 1 | 1[.6] | 1 |
| 1.. 625 | 1 | 13 over 8 | 1 |
| 1.1.625 | 1 | Total | 516 |
| 1.6.25 | 1 |  |  |


| 13 over 8 | 1 |  |  |
| :---: | :---: | :---: | :---: |
| $13 / 8$ or 1.625 | 1 |  |  |
| 1 r 5 | $1^{\wedge}$ |  |  |
| Total | 1883 |  |  |
| ${ }^{\wedge}$ student repeatedly used this notation |  | $\wedge$ student repeatedly used this notation |  |
| $\mathrm{C} 1[8 a+5=3 a+14 \mid 1.8]$ |  | $\mathrm{C} 1 *[9 n+3=4 n+12 \mid 1.8]$ |  |
| Responses | Freq | Responses | Freq |
| 1.8 | 1144 | 1.8 | 234 |
| 9/5 | 186 | $9 / 5$ | 146 |
| $14 / 5$ | 8 | $14 / 5$ | 9 |
| 18/10 | 3 | 1/8 | $1^{\wedge}$ |
| 1 and 4/5 | 1 | 9 over 5 | 1 |
| $1 . .8$ | 1 | Total | 391 |
| 1.8/1 | 1 |  |  |
| 1.8 | 1 |  |  |
| 1.80 | 1 |  |  |
| 1.800 | 1 |  |  |
| 9 over 5 | 1 |  |  |
| 1.8 | 1 |  |  |
| Total | 1349 |  |  |
|  |  | $\wedge$ student repeatedly used this notation |  |
| $\mathrm{C} 2[12 a+2=8 a+15$ \|3.25] |  | C2* $[11 n+3=7 n+16 \mid 3.25]$ |  |
| Responses | Freq | Responses | Freq |
| 3.25 | 1010 | 3.25 | 234 |
| 13/4 | 167 | 13/4 | 145 |
| 3.2 | 14 | $31 / 4$ | 10 |
| 3.3 | 14 | 3.2 | 4 |
| $31 / 4$ | 8 | 3/25 | 1 |
| 3.20 | 2 | 13 over 4 | 1 |
| 13 over 4 | 1 | 31/4 | 1 |
| 3 and 1/4 | 1 | Total | 396 |
| $3.25 \quad 3.25$ | 1 |  |  |
| 3.25/1 | 1 |  |  |
| 3r 1 | $1^{\wedge}$ |  |  |
| Total | 1220 |  |  |
| $\wedge$ student repeatedly used this notation |  |  |  |
| D1 [7a-11 $=2 a-4 \mid 1.4]$ |  | D1 [7n-11 $=2 n-4 \mid 1.4]$ |  |
| Responses | Freq | Responses | Freq |
| 1.4 | 854 | 1.4 | 189 |
| $7 / 5$ | 173 | $7 / 5$ | 129 |
| $12 / 5$ | 4 | $12 / 5$ | 7 |
| 14/10 | 2 | 14/10 | 1 |
| 1 and 2/5 | 1 | 1/4 | 1 |
| 1.4] | 1 | 7 over 5 | 1 |
| 1.40 | 1 | Total | 328 |
| 1r 2 | $1^{\wedge}$ |  |  |




- Table A2.2 continued next page

| Version 1: E1 [7a-2 = 16\| 18/7] | (not ranked by freq) |
| :---: | :---: |
| Responses | Freq |
| 18 over 7 | 1 |
| 18.7 | $1^{\wedge}$ |
| 18/7 | 210 |
| 18/7 or 2.57 | 1 |
| $24 / 7$ | 1 |
| $24 / 7$ | 13 |
| 24 r 7 | 1 |
| 2 and 4/7 | 2 |
| 2.5 | 1\# |
| 2.5 | 87\# |
| 2.56 rounded | 1 |
| 2.57 | 589 |
| 2.57 2dp | 1 |
| 2.57 (Rounded because the number was recurring) | 1 |
| 2.57 (rounded) | 1 |
| 2.57/1 | 1 |
| 2.57@ | 1 |
| 2.570 | 1 |
| 2.571 | 90 |
| 2.57123 | 1 |
| 2.5713 | 2 |
| 2.5714 | 26 |
| 2.57142... | 1 |
| 2.571425671 | 1 |
| 2.571428 | 16 |
| 2.571428... | 3 |
| 2.5714285 | 1 |
| 2.57142857 | 8 |
| 2.571428571 | 67 |
| 2.5714285714 | 3 |
| 2.571428571428 | 1 |
| 2.57142857142857 | 2 |
| 2.571428571428571 | 6 |
| 2.5714285714285714285714285714286 | 1 |
| 2.5714285714286 | 1 |
| 2.571428571429 | 2 |
| 2.57142857143 | 13 |
| 2.571429 | 2 |
| 2.57143 | 28 |
| 2.5715 | 2 |
| 2.572 | 3 |
| 2.6 | 184 |
| Total | 1378 |
| $\wedge$ student repeatedly used this notation, \# 2.5 ( $\mathrm{n}=88$ ) was upgraded from IR to CR after the analysis of errors was complete |  |

## Appendix 3: Three Facilities and the Facility Range

- There are two simple methods to calculate item facility.
- The first method determines the percentage of correct responses from all the students which assumes that the only reason for a student to omit an item (i.e., OR) is that the student does not have the skills or knowledge to correctly answer the item.
- In contrast, the second method determines the facility by determining the percentage of correct responses from those students who answered the item, hence not making this assumption about the lack of skills or knowledge for the OR students.
- We note that the first method is a lower limit for a range of facilities (as it is based on the worst-case scenario that $0 \%$ of the OR group have the required skills or knowledge on that item) and that the second method is an upper limit (as it is based on the best-case scenario that the proportion of students in the OR group with the required skills or knowledge on that item is the same proportion as for the students who answered the item ${ }^{3}$ ).
- For item calibration, PISA uses a method that that provides a point estimate (which lies between the upper and lower limits above). Before the PISA item facility is determined, students' responses are first considered; a sequence of blanks that includes the last item indicates a student who has finished the test early. For a student with such a sequence of length $k$, the last $k-1$ items are not considered to be 'administered' and hence do not count in the statistics for these items. Note that the PISA facility for earlier items are closer to the facility lower limit (as more of these blanks are coded as incorrect) and for later items, closer to facility upper limit (as more of these blanks are coded as not administered).
- Rather than choosing one of these facilities, we provide details of all three, labelled in increasing order:
- Facility I (based on all students),
- Facility II (result of PISA item calibration method) and
- Facility III (based on students who answered).

We define the Facility Range as [Facility I, Facility III]
where Facility $\mathrm{I}=\frac{n(C R)}{n(C R)+n(I R)+n(O R)}$ or $\mathrm{CR} \%$
and Facility III $=\frac{n(C R)}{n(C R)+n(I R)}$
and $n(C R)+n(I R)+n(O R)$ is the total number of responses
and we note that Facility II is within this range.

[^2]Equation 1 (linking Facility III to Facility I) can be derived from above:


Similarly, Equation 2 (the size of the Facility Range) can be derived:

Facility III - Facility I = Facility I $\times$ scale factor $B$, where $B=\frac{0 R \%}{1-\text { OR } \%} \ldots \ldots . . . . . . . .$. . Equation (2)
noting that scale factor $\mathrm{B}=$ scale factor $\mathrm{A} \times \mathrm{OR} \%$
.Equation (3)

Table A3.1: Effect of varying OR\% on (i) Facility III and (ii) the size of the Facility Range

| OR\% | scale factor A <br> $=\frac{1}{1-\text { OR } \%}$ | Facility III | scale factor B <br> =Ax OR\% | Facility Range |
| :---: | :---: | :---: | :---: | :---: |
| 0.00 | 1.000 | = Facility I | 0.000 | $=0$ |
| 0.05 | 1.053 |  | 0.053 |  |
| 0.10 | 1.111 |  | 0.111 |  |
| 0.15 | 1.176 |  | 0.176 |  |
| 0.20 | 1.250 |  | 0.250 |  |
| 0.25 | 1.333 |  | 0.333 |  |
| 0.30 | 1.429 |  | 0.429 |  |
| 0.35 | 1.538 |  | 0.538 |  |
| 0.40 | 1.667 |  | 0.867 |  |
| 0.45 | 1.818 |  | 1.000 | $=$ Facility I |
| 0.50 | 2.000 | $=2 \times$ Facility I |  |  |

## Appendix 4: The Item Position Effect

- It is generally found that later items in a test have lower facility (e.g., Zeller et al. 2017) ${ }^{4}$, hence we include item position as one of the Item Characteristics in Figure 2.1.
- We noted this phenomenon when identical equations were moved between Version P and Version 1. Moving [ $4 a+9=37 \mid 7]$ from Q1 (Version P) to Q2 (Version 1) made little difference to the facility ( $84 \%$ and $83 \%$, respectively). In contrast, moving [ $3 a+8=23 \mid 5]$ from Q5 (Version P) to Q1 (Version 1) made a large difference to the facility ( $70 \%$ and $85 \%$, respectively).
- Another calculation of facility (based on only those who answered the item) gave results which were almost identical; $87 \%$ (Version P) and $86 \%$ (Version 1).
- The figure below shows these results graphically; the white bars show Facility I (CR\%), the total of white and black bars is Facility III (based on students who answered), and the black bars show the Facility Range.
- Inspection of the RHS of four bars in this figure confirmed that it was appropriate to group these two items into Group A in Version 1.
- Note that the SE are small (1\%)


## Facility Ranges for two items in Versions P \& 1

$0 \% 10 \% ~ 20 \% ~ 30 \% ~ 40 \% ~ 50 \% ~ 60 \% ~ 70 \% ~ 80 \% ~ 90 \% ~ 100 \% ~$

```
[4n+9 = 37 | 7], Q1 in Version P
[4a+9=37|7], Q2 in Version 1
[3a+8=23|5],Q1 in Version 1
[3a+8=23|5],Q5 in Version P
```



```
\([4 \mathrm{a}+9=37 \mid 7]\), Q2 in Version 1
```



```
Facility I \(\quad\) Facility Range
```

| Item | Facilities |  |  | Standard Error $^{\wedge}$ |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | I | II | III | I | II | III |
| $[4 \mathrm{n}+9=37 \mid 7]$, Q1 in Version P | $84 \%$ | $84 \%$ | $85 \%$ | $1 \%$ | $1 \%$ | $1 \%$ |
| $[4 \mathrm{a}+9=37 \mid 7]$, Q2 in Version 1 | $83 \%$ | $83 \%$ | $85 \%$ | $1 \%$ | $1 \%$ | $1 \%$ |
| $[3 \mathrm{a}+8=23 \mid 5]$, Q1 in Version 1 | $85 \%$ | $85 \%$ | $86 \%$ | $1 \%$ | $1 \%$ | $1 \%$ |
| $[3 \mathrm{a}+8=23 \mid 5]$, Q5 in Version P | $70 \%$ | $77 \%$ | $87 \%$ | $1 \%$ | $1 \%$ | $1 \%$ |

Figure A4.1: Facility I (white bars), Facility III (the total of the white and black bars), and the Facility Range (black bars). $\left.{ }^{\wedge} \mathrm{SE}=\sqrt{ }[\mathrm{p}(1-\mathrm{p}) / \mathrm{n})\right]$

[^3]Appendix 5: Support for Versions 1 and 2 being parallel

- using Version 2, T1 (only) as comparison


Figure A5.1: Percentage distribution of responses for Version 1 ( $\mathrm{n}=3010$ ) and Version 2: Subset T 1 ( $\mathrm{n}=393$ )

* 4 items swapped position from Version 1 to Version 2

| Version $1(\mathrm{n}=3010)$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| $\begin{array}{r} \mathrm{A} 1[3 \mathrm{a}+8=23 \mid 5] \\ \mathrm{A} 2[4 \mathrm{a}+9=37 \\ \mathrm{B} 1[5 \mathrm{a}+7=15 \mid 1.6] \\ 32[8 \mathrm{a}+3=16 \mid 1.625] \end{array}$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| $\begin{array}{r} \mathrm{C} 1[8 \mathrm{a}+5=3 \mathrm{a}+14 \mid 1.8] \\ \mathrm{C} 2[12 \mathrm{a}+2=8 \mathrm{a}+15 \mid \\ \mathrm{D} 1[7 \mathrm{a}-11=2 \mathrm{a}-4 \\ \mathrm{D} \\ \mathrm{D} 2[12-4] \\ \text { D }[12-11 \mathrm{a}=5-\mathrm{a} \\ \hline \end{array}$ |  |  |  |  |  |  |  | $\begin{array}{r} \mathrm{C} 2^{*}[11 \mathrm{n}+3=7 \mathrm{n}+16 \mid 3.25] \\ \mathrm{C} 1 *[9 \mathrm{n}+3=4 \mathrm{n}+12 \\ \mathrm{D} 1[7 \mathrm{n}-11=2 \mathrm{n}-4 \\ \mathrm{D} 2[1.4] \\ \mathrm{D} 2[12-11 \mathrm{n}=5-\mathrm{n} \\ \hline \end{array}$ |  |  |  |  |  |  |  |  |
| $\begin{array}{r} \text { E1 } \\ \text { E3 } 3 \text { } \\ \text { E4 } \\ E \\ \text { E6 } \end{array}$ | $\begin{aligned} & a-2= \\ & 2[14-2 \\ & +6+2 a \\ & (a+2) / 5 \\ & {[a / 3+1} \\ & a-3)= \end{aligned}$ |  |  |  |  |  |  | E3 | $\begin{aligned} & {[5 \mathrm{n}-} \\ & 2[15 \\ & +4+ \\ & {[(\mathrm{n}+1} \\ & 5[\mathrm{n} / 4 \\ & {[5(\mathrm{n}-} \end{aligned}$ | $\begin{aligned} & =16 \\ & n=5 \\ & =5 \\ & =3 \\ & =8 \\ & =8 \\ & =8 \end{aligned}$ | $\begin{aligned} & 4] \\ & 3] \end{aligned}$ |  |  |  |  |  |
| $\square$ Facility I © Facility Range |  |  |  |  |  |  |  | $\square$ Facility I ©Facility Range |  |  |  |  |  |  |  |  |
| Item | $\mathrm{N}(\mathrm{CR})$ | Number students |  |  | Facilities |  |  | Item | $\mathrm{N}(\mathrm{CR})$ | Number students |  |  | Facilities |  |  |  |
|  |  | I | II | III | I | II | III |  |  | I | II | III | I | II | III |  |
| A1 | 2570 | 3010 | 3010 | 2984 | 85\% | 85\% | 86\% | A1 | 333 | 393 | 393 | 391 | 85\% | 85\% | 85\% |  |
| A2 | 2510 | 3010 | 3010 | 2957 | 83\% | 83\% | 85\% | A2 | 327 | 393 | 393 | 388 | 83\% | 83\% | 84\% |  |
| B1 | 2027 | 3010 | 2993 | 2775 | 67\% | 68\% | 73\% | B2* | 241 | 393 | 390 | 359 | 61\% | 62\% | 67\% |  |
| B2 | 1883 | 3010 | 2919 | 2694 | 63\% | 65\% | 70\% | B1* | 234 | 393 | 377 | 358 | 60\% | 62\% | 65\% |  |
| C1 | 1349 | 3010 | 2879 | 2201 | 45\% | 47\% | 61\% | C2* | 169 | 393 | 375 | 287 | 43\% | 45\% | 59\% |  |
| C2 | 1220 | 3010 | 2646 | 2048 | 41\% | 46\% | 60\% | C1* | 165 | 393 | 349 | 275 | 42\% | 47\% | 60\% |  |
| D1 | 1038 | 3010 | 2575 | 1940 | 34\% | 40\% | 54\% | D1 | 138 | 393 | 344 | 259 | 35\% | 40\% | 53\% |  |
| D2 | 784 | 3010 | 2535 | 1850 | 26\% | 31\% | 42\% | D2 | 119 | 393 | 342 | 251 | 30\% | 35\% | 47\% |  |
| E1 | 1378 | 3010 | 2501 | 2286 | 46\% | 55\% | 60\% | E1 | 199 | 393 | 338 | 318 | 51\% | 59\% | 63\% |  |
| E2 | 1354 | 3010 | 2431 | 2260 | 45\% | 56\% | 60\% | E2 | 171 | 393 | 333 | 302 | 44\% | 51\% | 57\% |  |
| E3 | 1144 | 3010 | 2342 | 1921 | 38\% | 49\% | 60\% | E3 | 160 | 393 | 324 | 272 | 41\% | 49\% | 59\% |  |
| E4 | 1407 | 3010 | 2173 | 2009 | 47\% | 65\% | 70\% | E4 | 168 | 393 | 304 | 285 | 43\% | 55\% | 59\% |  |
| E5 | 1348 | 3010 | 2098 | 1996 | 45\% | 64\% | 68\% | E5 | 178 | 393 | 294 | 270 | 45\% | 61\% | 66\% |  |
| E6 | 1037 | 3010 | 2050 | 1914 | 34\% | 51\% | 54\% | E6 | 154 | 393 | 281 | 265 | 39\% | 55\% | 58\% |  |

Figure A5.2: Facility Ranges for Version $1(\mathrm{n}=3010)$ and Version 2: Subset T1 ( $\mathrm{n}=393$ )

* 4 items swapped position from Version 1 to Version 2


## Appendix 6: Illustration of learning: Subsets T1 and T2 of Version 2

- CR: each item in subset T2 (RHS) has higher facility than in T1 (LHS)
- OR: overall, only $10 \%$ in T2 cf $22 \%$ in T1, so half the omitted responses after teaching

| Percentage Distribution of Responses: <br> Version 2-T1 ( $\mathrm{n}=393$ ) | Percentage Distribution of Responses: Version 2-T2 ( $\mathrm{n}=341$ ) |
| :---: | :---: |
| 0\% $20 \% \quad 40 \% \quad 60 \% \quad 80 \% \quad 100 \%$ | 0\% $20 \% \quad 40 \% \quad 60 \% \quad 80 \% \quad 100 \%$ |
|  | $\mathrm{A} 1[4 \mathrm{n}+11=23 \mid 3] \quad 90 \%$ 退 $\quad 100$ |
| $\mathrm{A} 2[3 \mathrm{n}+5=26 \mid 7] \quad 83 \%$ 16\% | $\mathrm{A} 2[3 \mathrm{n}+5=26 \mid 7] \quad 91 \%$ 9 ${ }^{\text {cha }}$ |
| $B 2 *[8 n+3=16 \mid 1.625] \quad 61 \% \quad 30 \%$ 9\% | $\mathrm{B} 2 *[8 \mathrm{n}+3=16 \mid 1.625]$ 81\% |
|  | $\mathrm{B} 1 *[5 \mathrm{n}+7=15 \mid 1.6]$ 20\% ${ }^{\text {c }}$ |
|  |  |
|  |  |
| D1 $[7 \mathrm{n}-11=2 \mathrm{n}-4 \mid 1.4] \quad 35 \% \quad 31 \% \quad 34 \%$ |  |
| $\mathrm{D} 2[12-11 \mathrm{n}=5-\mathrm{n} \mid 0.7] \quad 30 \%$ 34\% | D2 $[12-11 \mathrm{n}=5-\mathrm{n} \mid 0.7] \quad 35 \%$ 47\% |
| E1 $[5 n-1=16 \mid 3.4]$ 年 $5 \%$ 30\% | E1 $[5 n-1=16 \mid 3.4]$ 12\% 5 |
|  |  |
| E3 $[2 \mathrm{n}+4+3 \mathrm{n}=5 \mid 0.2]$ $41 \%$ $28 \%$ $31 \%$ | $\mathrm{E} 3[2 \mathrm{n}+4+3 \mathrm{n}=5 \mid 0.2] \quad 56 \%)$ |
| $\mathrm{E} 4[(\mathrm{n}+1) / 5=3 \mid 14] \quad 43 \% \quad 30 \% \quad 2 \pi \%$ | $\mathrm{E} 4[(\mathrm{n}+1) / 5=3 \mid 14] \quad 66 \% \quad 21 \% \quad 14 \%$ |
| E5 [n/4 + 3 = 8\| 20$] \quad 45 \%$ 23\% 31\% |  |
| E6 $[5(n-2)=8 \mid 3.6] \quad 39 \% \quad 28 \% \quad 33 \%$ | E6 $[5(\mathrm{n}-2)=8 \mid 3.6] \quad 62 \% \quad 20 \% \quad 18 \%$ |
| $\begin{aligned} & ■ \text { Correct Response }(\mathrm{CR}) \% \quad \text { Incorrect Response (IR)\% } \\ & \square \text { Omitted Response }(\mathrm{OR}) \% \end{aligned}$ | $\begin{aligned} & \square \text { Correct Response (CR)\% } \quad \text { Incorrect Response (IR)\% } \\ & \square \text { Omitted Response (OR)\% } \end{aligned}$ |

Figure A6.1: Percentage distribution of responses for Version 2, Subset T1 (n=393) and Subset T2 (n=341)

* indicates items swapped position from Version 1


## Appendix 7: Variation in score better explained by stage than by year level

- The table below shows some of the outputs from the linear regressions performed on the data from Versions 1 and 2
- When the set of items being considered involves the rubric items (Rows $1 \& 2$ ), the stages explain $80 \%$ to $90 \%$ of the variation in total score (in both versions).
- This high \% is expected because Stage is constructed from scores on the rubric items
- When the rubric items are not included (Row 3), the stages explain about $50 \%$ to $60 \%$ of the variation in total score, compared to year level, which explains only $2 \%$ to $3 \%$
- Row 4 shows that there is very little relationship between stages and year levels
- We conclude that the reported stages give useful information to the teacher about the understanding of their students

Table A7.1: Summary of Linear Regression output with Versions 1 and 2

| Row | Setails of dependent | Expression of dependent |  | Version 1 |  | Version 2 |
| :--- | :--- | :--- | ---: | ---: | ---: | ---: | ---: |
|  | variable | variable as a function of |  |  |  |  |
|  |  | independent variable | $\mathrm{R}^{2}$ | F | $\mathrm{R}^{2}$ | F |
| 1 | 14 items (include 8 | Score/14=f(year) | $4 \%$ | $<.001$ | $4 \%$ | $<.001$ |
|  | rubric items which | Score/14=f(stage) | $80 \%$ | $<.001$ | $86 \%$ | $<.001$ |
|  | determine stage) | Score/14=f(year, stage) | $81 \%$ | $<.001$ | $86 \%$ | $<.001$ |
| 2 | 8 rubric items which | Score/8=f(year) | $3 \%$ | $<.001$ | $5 \%$ | $<.001$ |
|  | determine stage | Score/8=f(stage) | $91 \%$ | $<.001$ | $92 \%$ | $<.001$ |
|  |  | Score/8=f(year, stage) | $91 \%$ | $<.001$ | $92 \%$ | $<.001$ |
| 3 | 6 non-rubric items | Score/6=f(year) | $3 \%$ | $<.001$ | $2 \%$ | $<.001$ |
|  |  | Score/6=f(stage) | $49 \%$ | $<.001$ | $61 \%$ | $<.001$ |
|  |  | Score/6=f(year, stage) | $49 \%$ | $<.001$ | $61 \%$ | $<.001$ |
| 4 | Stage | Stage=f(year) | $2 \%$ | $<.001$ | $5 \%$ | $<.001$ |


[^0]:    ${ }^{1}$ Linsell, C. (2010). Secondary numeracy project students' development of algebraic knowledge and strategies. Findings from the New Zealand numeracy development projects 2009, 100-116.

[^1]:    ${ }^{2}$ Linsell, C. (2010). Secondary numeracy project students' development of algebraic knowledge and strategies. Findings from the New Zealand numeracy development projects 2009, 100-116.

[^2]:    ${ }^{3}$ We have found a few exceptions to this "reasonable" assumption!

    - Example 1: In a drag/drop question where individual "cards" were to be taken to one of three locations (YES, NO or UNSURE) we found some high-scoring students who omitted certain items (i.e. didn't move certain cards). They had "selected" only the cards that they wanted to move to YES.
    - Example 2: In a multiple-choice question (YES/NO) with multiple parts, we found some high-scoring students who omitted certain items (i.e., didn't select NO).
    - In both examples of multi-part questions, these students have not followed the given instructions, which are equivalent to 'answer all parts'. Our 'pattern recognition scripts' detect such behaviour.

[^3]:    ${ }^{4}$ Zeller, F., Reiß, S., \& Schweizer, K. (2017). Is the item-position effect in achievement measures induced by increasing item difficulty?. Structural Equation Modeling: A Multidisciplinary Journal, 24(5), 745-754.

