CEM
Centre for Evaluation & Monitoring

@ProfCoe
#RUD2016

Investigating item quality using fit and other indicators

Robert Coe

Rasch User Group, Durham, 18 March 2016

Durham University

Quality

1. Utility for measurement
   a) Fit with measurement model (Rasch)
   b) Alignment with intended construct interpretations
2. Utility for learning
   a) Alignment with intended learning aims
   b) Value of diagnostic information
      i. For teachers
      ii. For students
   c) Reinforcement, retrieval

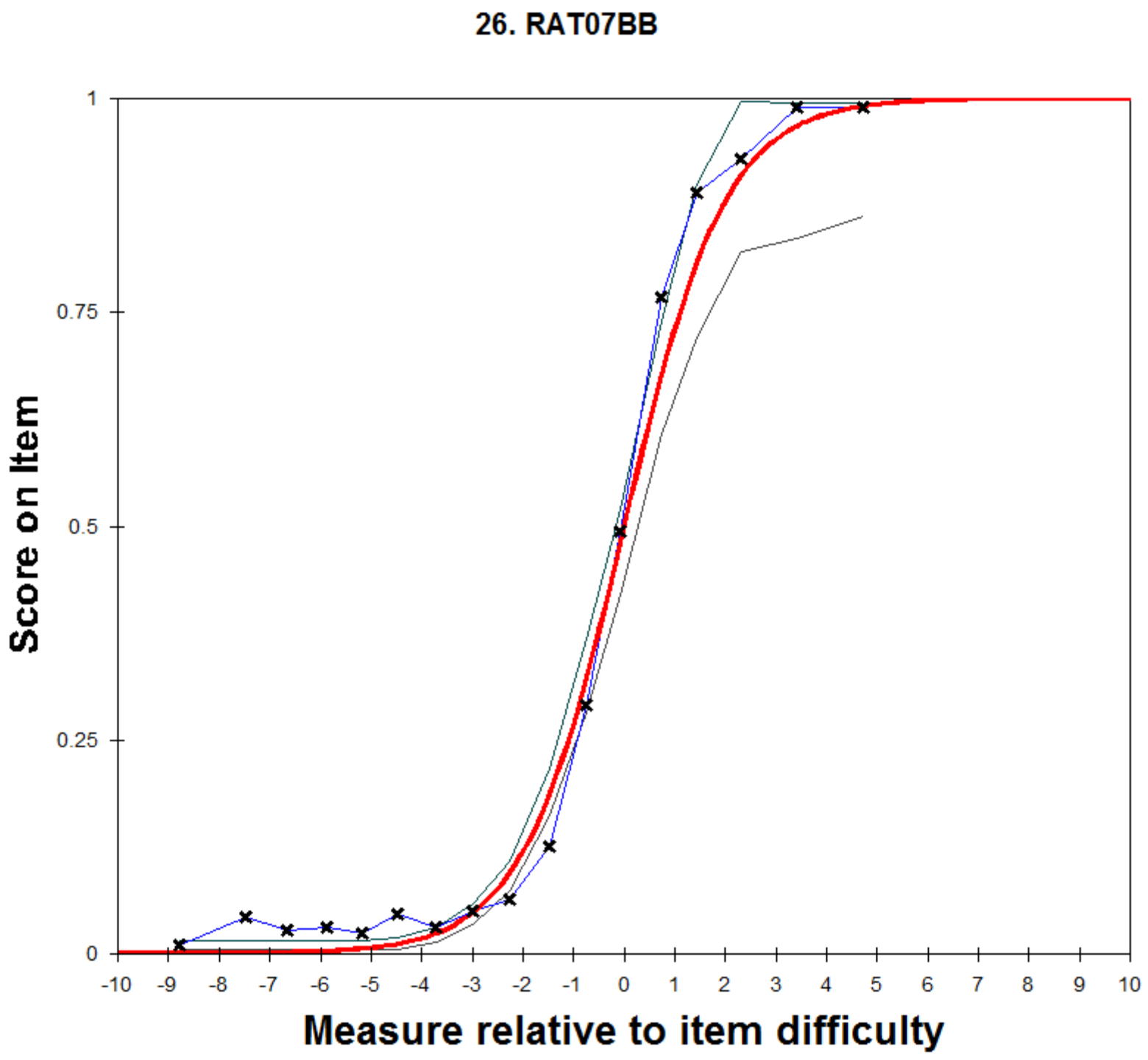
Model fit

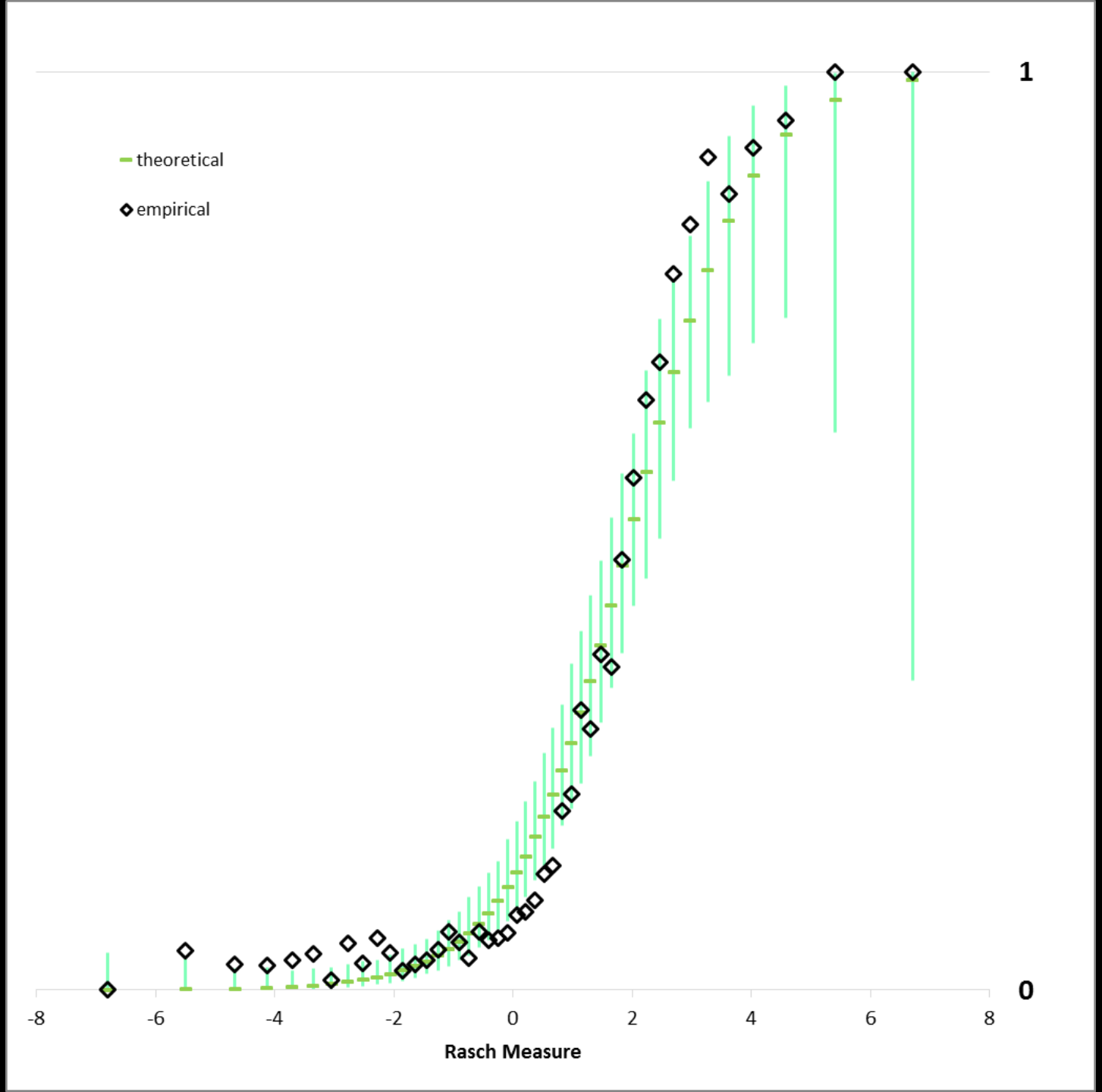
- INFIT/OUTFIT
- Discrimination
  - IRT parameter/index
  - Item-measure correlation
  - 27% rule (Kelley, 1939)
- H -coeff of homogeneity (Loevinger, 1948; Mokken, 1971; Mokken & Lewis, 1982)
- Other fit statistics?

Problems with INFIT/OUTFIT

Karabatsos (2000) JAppMeas

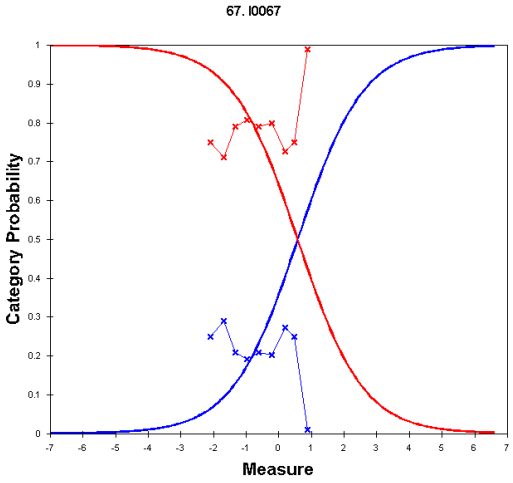
- 'Residual fit statistics' are confounded: parameters are estimated from data; fit stats test fit between data and parameters …
- Interpretation of INFIT/OUTFIT is sample dependent
- They do a poor job of identifying misfit


26. RAT07BB
Score on Item
1
0.75
0.5
0.25
0
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
Measure relative to item difficulty

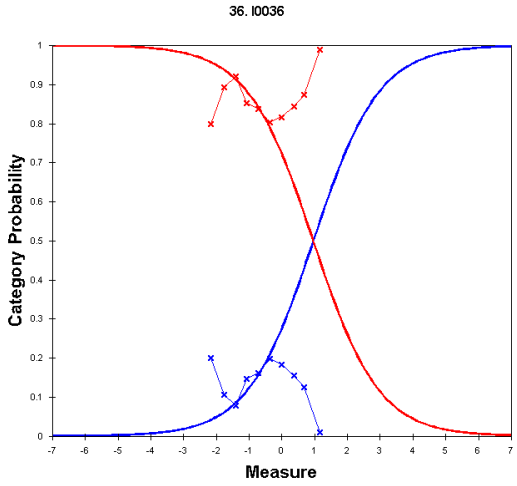

theoretical
empirical
1
0
-8
-6
-4
-2
0
2
4
6
8
Rasch Measure

Item Statistics	
Number of responses	7,685
Maximum score	1
Mean score on item	0.12
Item difficulty (Rasch measure)	1.98
INFIT (mean sq)	0.98
OUTFIT (mean sq)	3.80
IRT Discrimination parameter	0.92
Item-measure correlation (actual)	0.42
Item-measure correlation (expected)	0.46
Percent match model (observed)	91
Percent match model (expected)	90

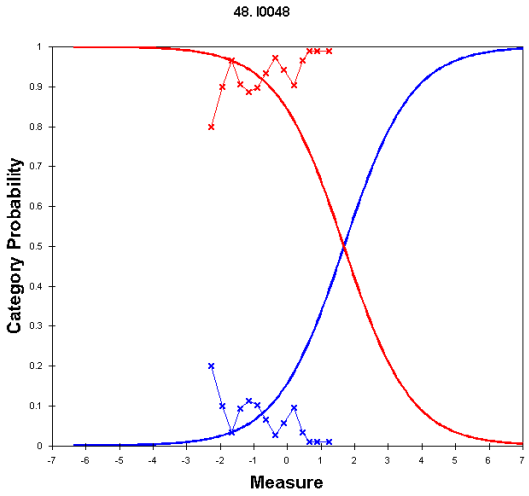
3. b) Infit and outfit indicate model fit?



infit 1.07, outfit 1.15



infit 1.04, outfit 1.08

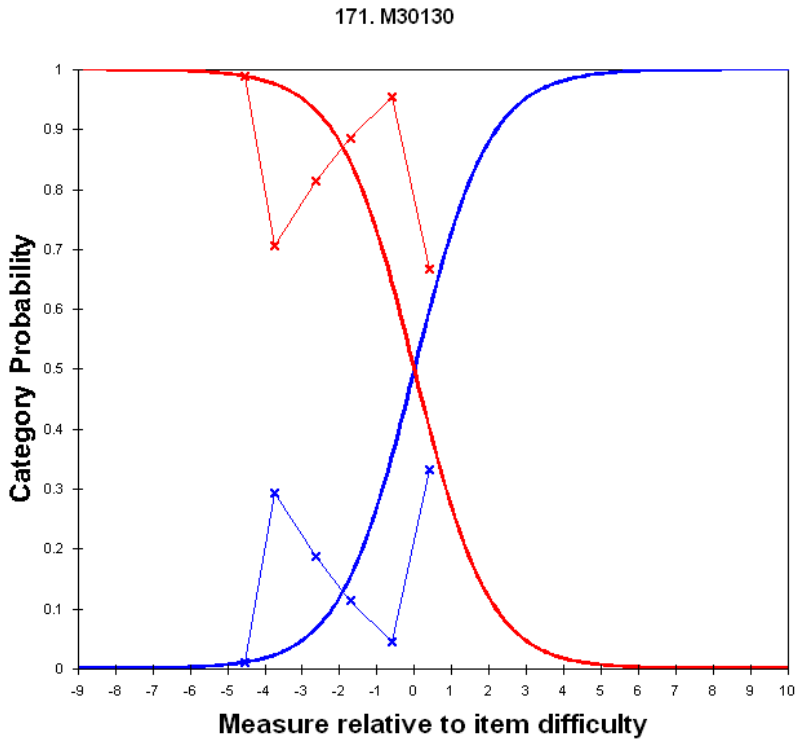


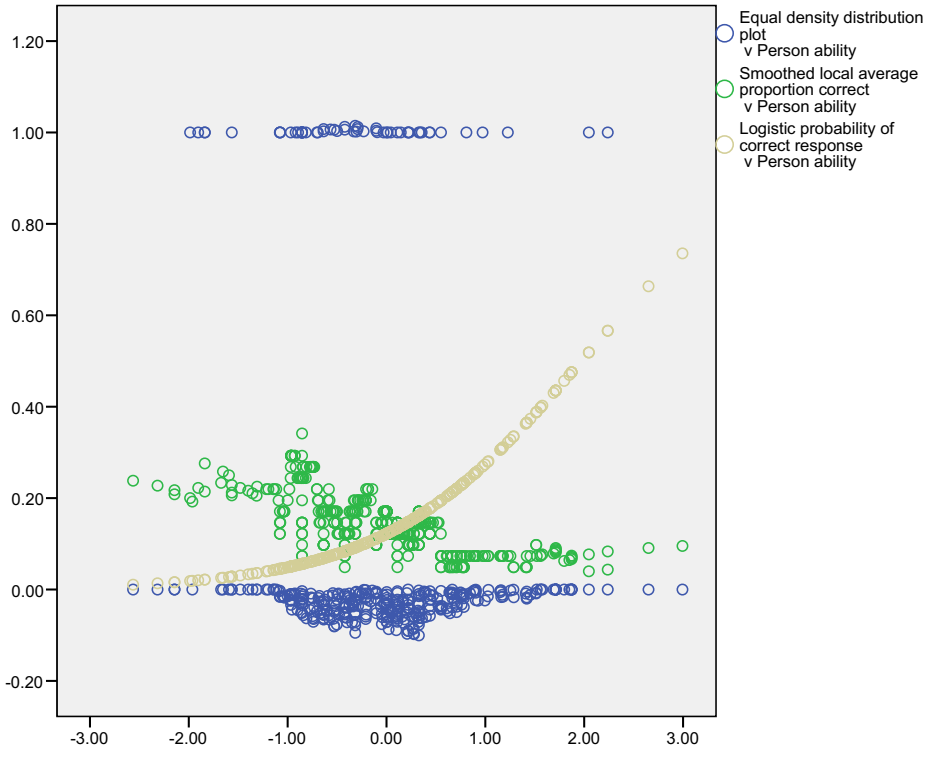
infit 1.06, outfit 1.27

Durham University

CEM Centre for Evaluation & Monitoring

WINSTEPS category probability curves can be misleading





Infit = 1.31

Durham University

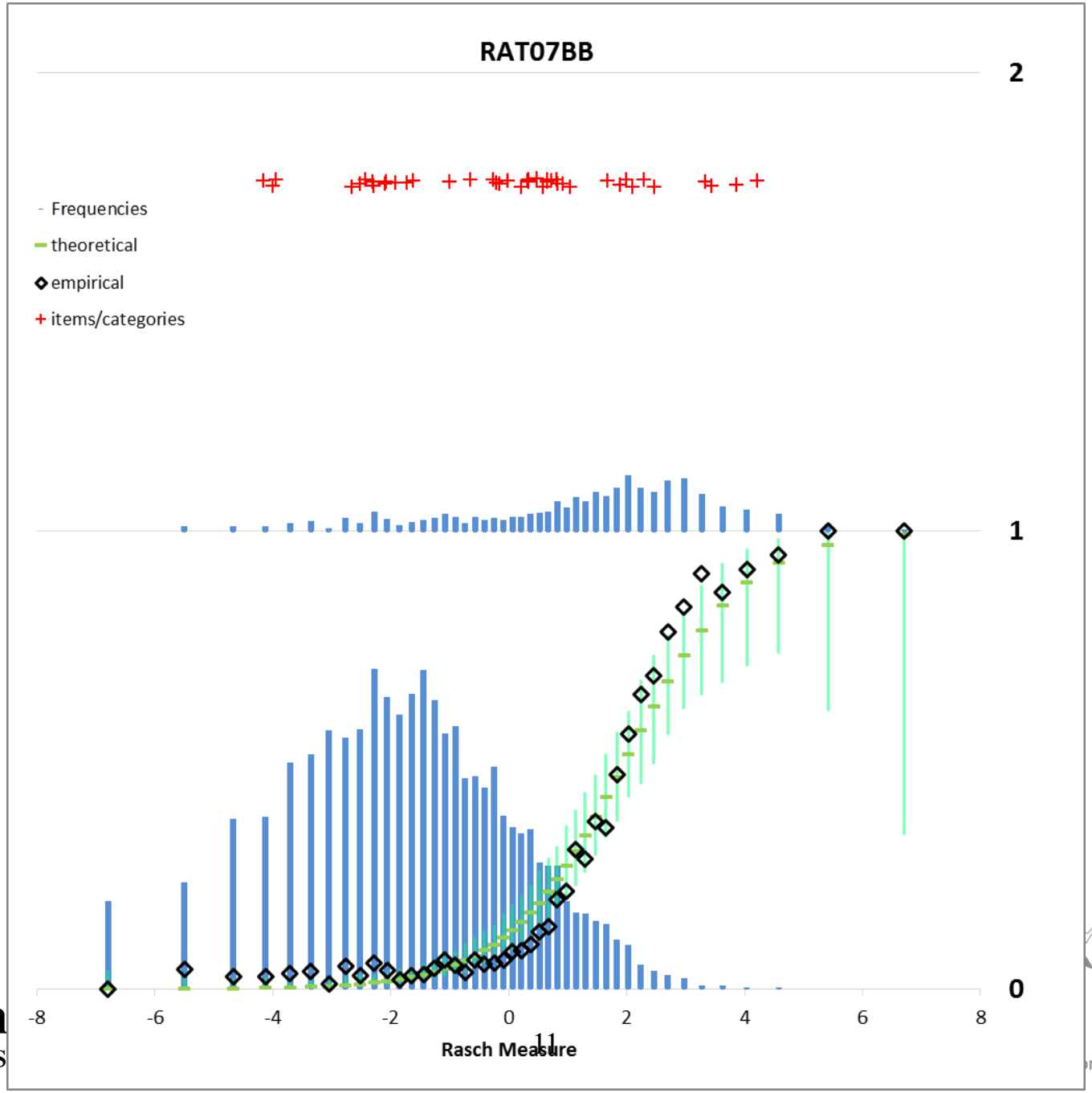
CEM Centre for Evaluation & Monitoring

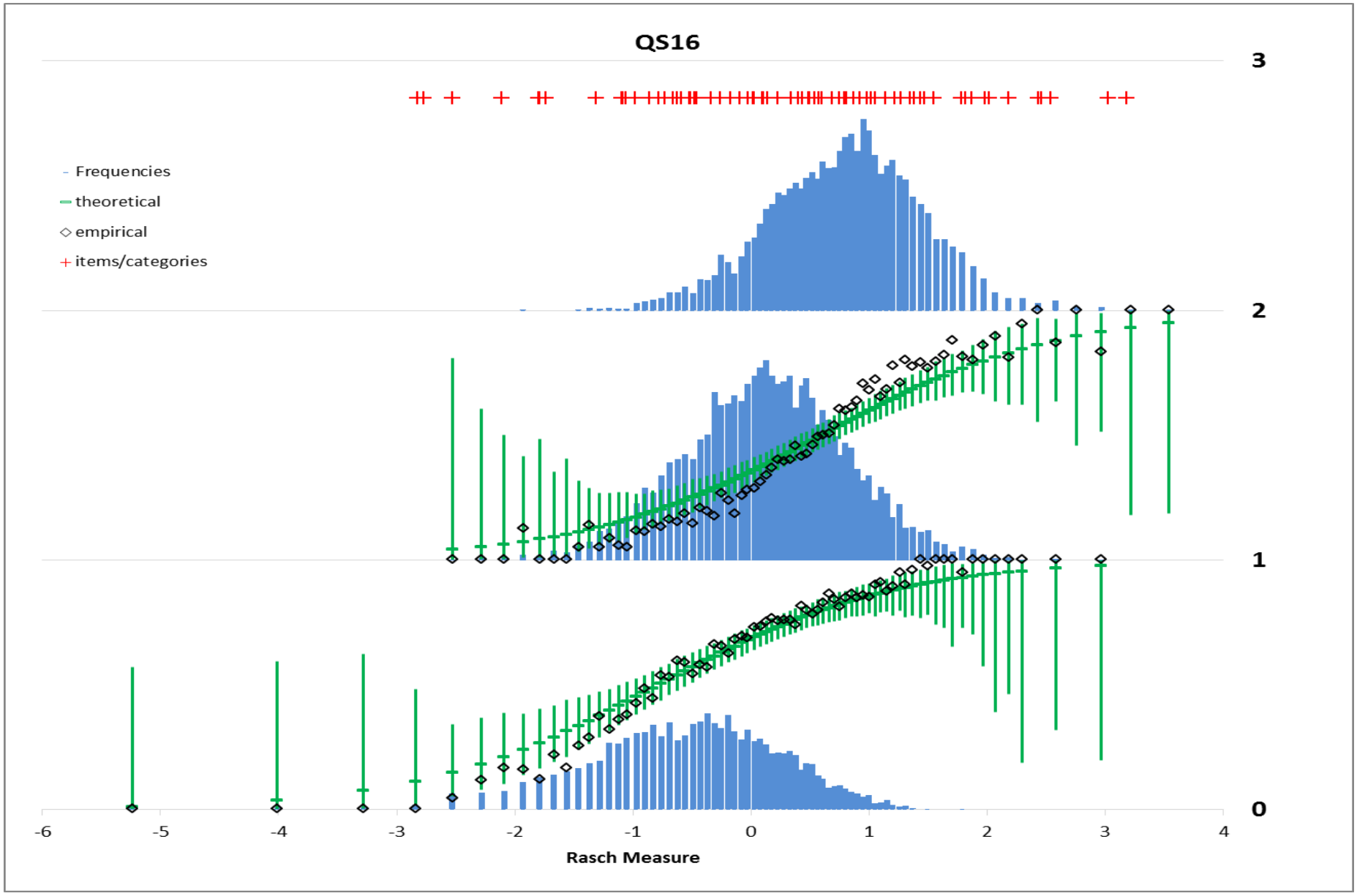
These data are from the analysis of a 20 item test of algebra, taken by around 600 candidates.

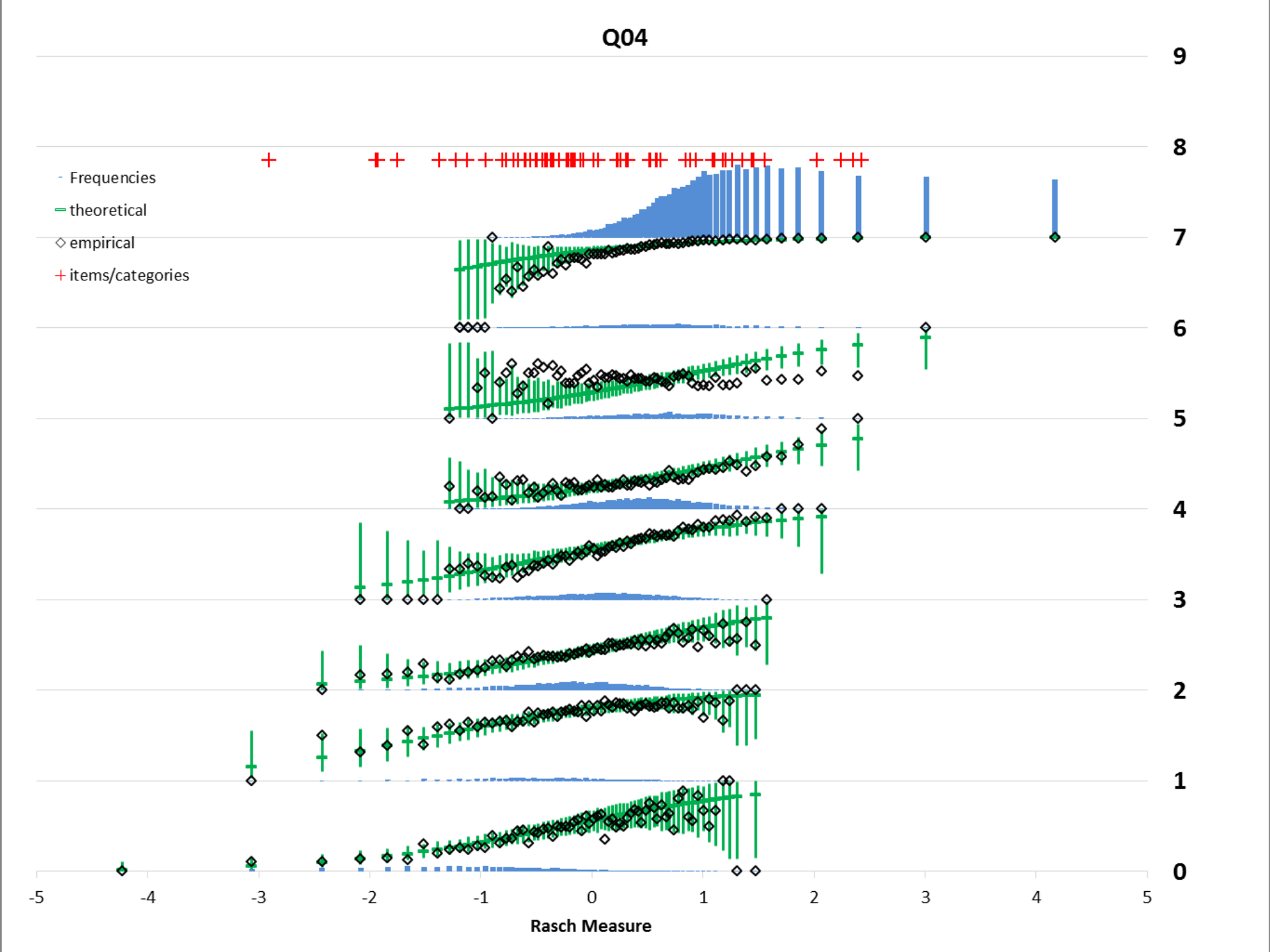
Two of the test items have been replaced with completely random responses. A further two items have had half their responses replaced with random data.

Can you identify the corrupted items?

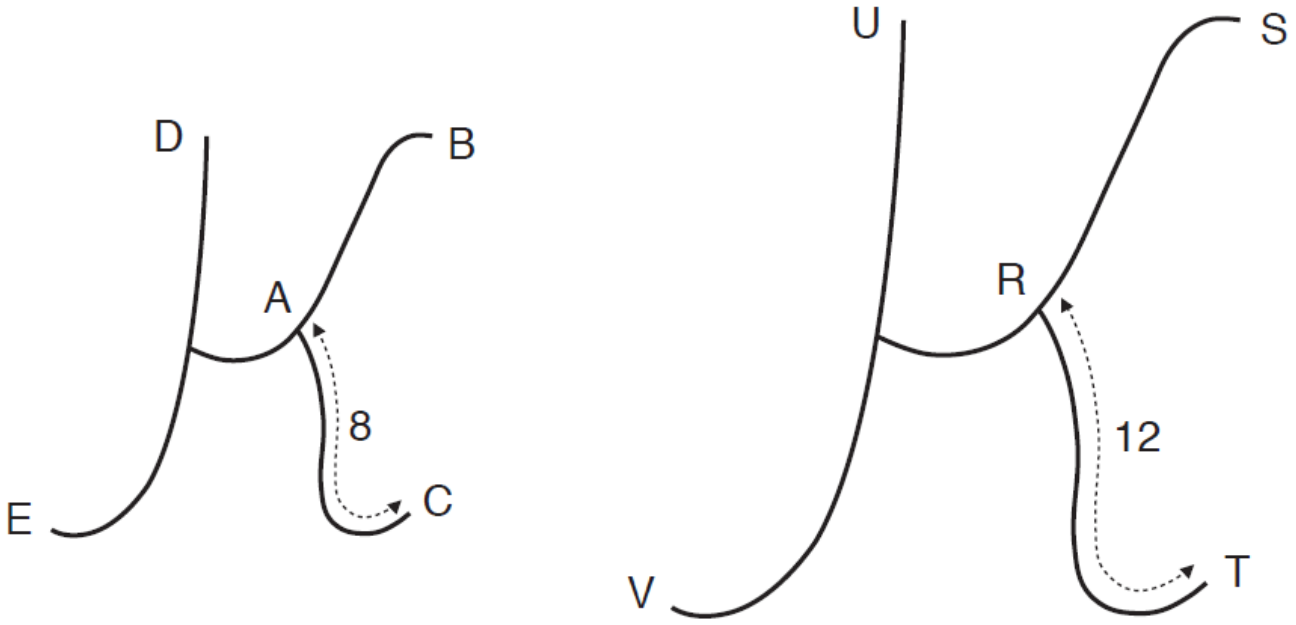
		alg01aa_r	alg02ab_r	alg03ab_r	alg04aa_r	alg04dd_r	alg04ff_r
		1	2	3	4	5	6
alg01aa_r	1	=	0.09	0.07	0.17	0.17	0.08
alg02ab_r	2	0.09	=	0.15	0.14	0.18	0.22
alg03ab_r	3	0.07	0.15	=	0.13	0.13	0.12
alg04aa_r	4	0.17	0.14	0.13	=	0.62	0.27
alg04dd_r	5	0.17	0.18	0.13	0.62	=	0.32
alg04ff_r	6	0.08	0.22	0.12	0.27	0.32	=
alg05aa_r	7	0.10	0.30	0.10	0.35	0.33	0.20
alg06aa_r	8	0.14	0.34	0.11	0.29	0.29	0.17
alg07cc_r	9	0.15	0.39	0.21	0.24	0.30	0.19
alg08@@_r	10	0.09	0.15	0.06	0.20	0.23	0.13
alg09cc_r	11	-0.02	0.05	0.03	0.02	0.05	-0.0
alg10aa_r	12	-0.02	0.03	0.11	-0.01	0.02	-0.0
alg11bb_r	13	-0.02	-0.02	0.02	-0.04	0.02	-0.0
alg13aa_r	14	0.08	0.17	0.08	0.28	0.30	0.16
alg13ee_r	15	0.06	0.23	0.30	0.14	0.14	0.11
alg13ii_r	16	0.11	0.19	0.26	0.12	0.17	0.10


RAT07BB
2
1
0
Frequencies
theoretical
empirical
items/categories
-8
-6
-4
-2
0
2
4
6
8
Rasch Measure


QS16
Frequencies
theoretical
empirical
items/categories
3
2
1
0
-6
-5
-4
-3
-2
-1
0
1
2
3
4
Rasch Measure


Q04
Frequencies
theoretical
empirical
items/categories
9
8
7
6
5
4
3
2
1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
Rasch Measure

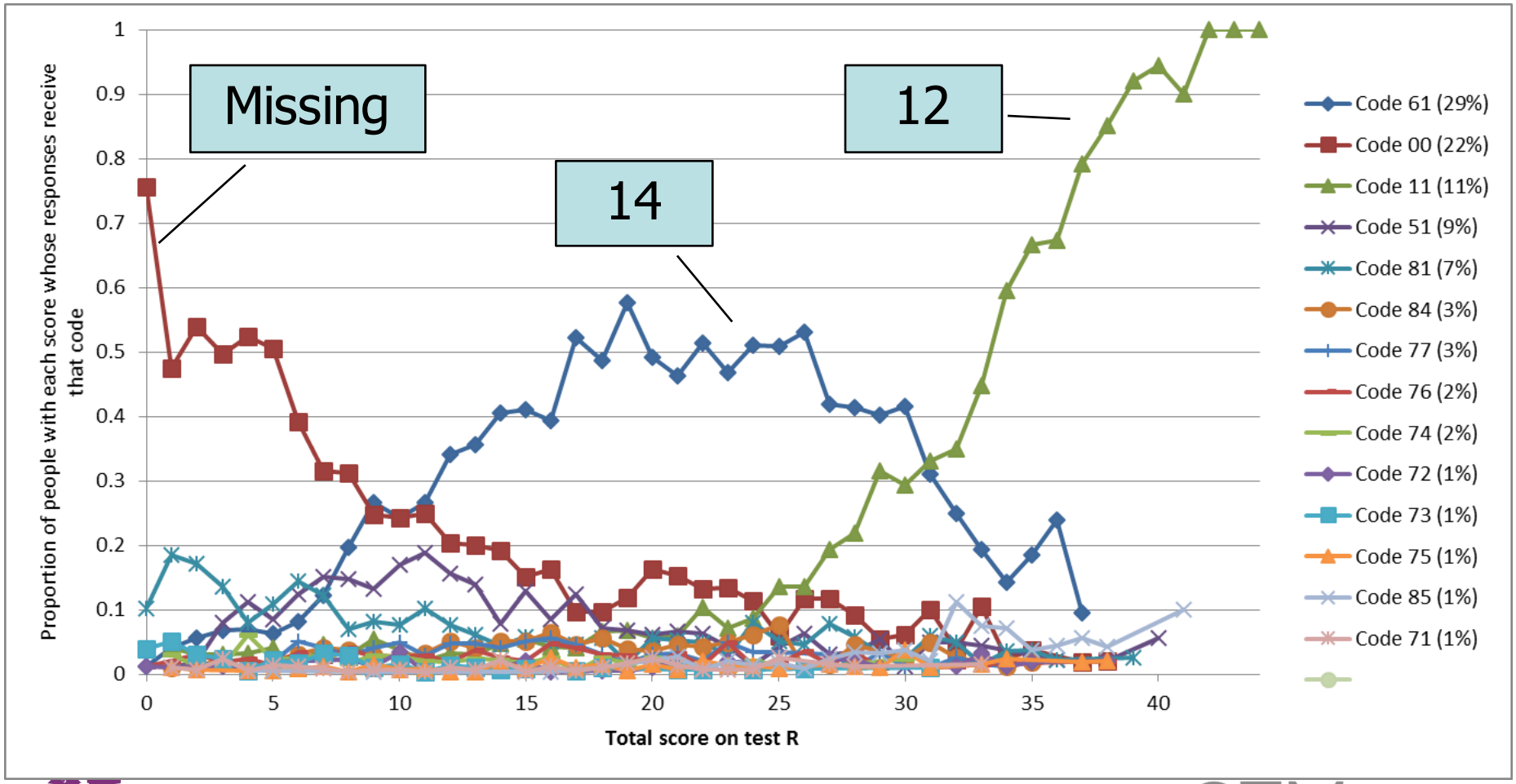
These two letters are the same shape. One is larger than the other.
AC is 8 units. RT is 12 units.



The curve AB is 9 units. How long is the curve RS ? . . . . . .

The curve UV is 18 units. How long is the curve DE ? . . . . . .

Enlargement 8→12, so ?→18



Algebra

4. **4 added to *n*** can be written as ***n* + 4**.
Add 4 onto each of these:

8	$n + 5$	$3n$
.....	.....	.....

n* multiplied by 4** can be written as **4*n.
Multiply each of these by 4:

8	$n + 5$	$3n$
.....	.....	.....

Durham University

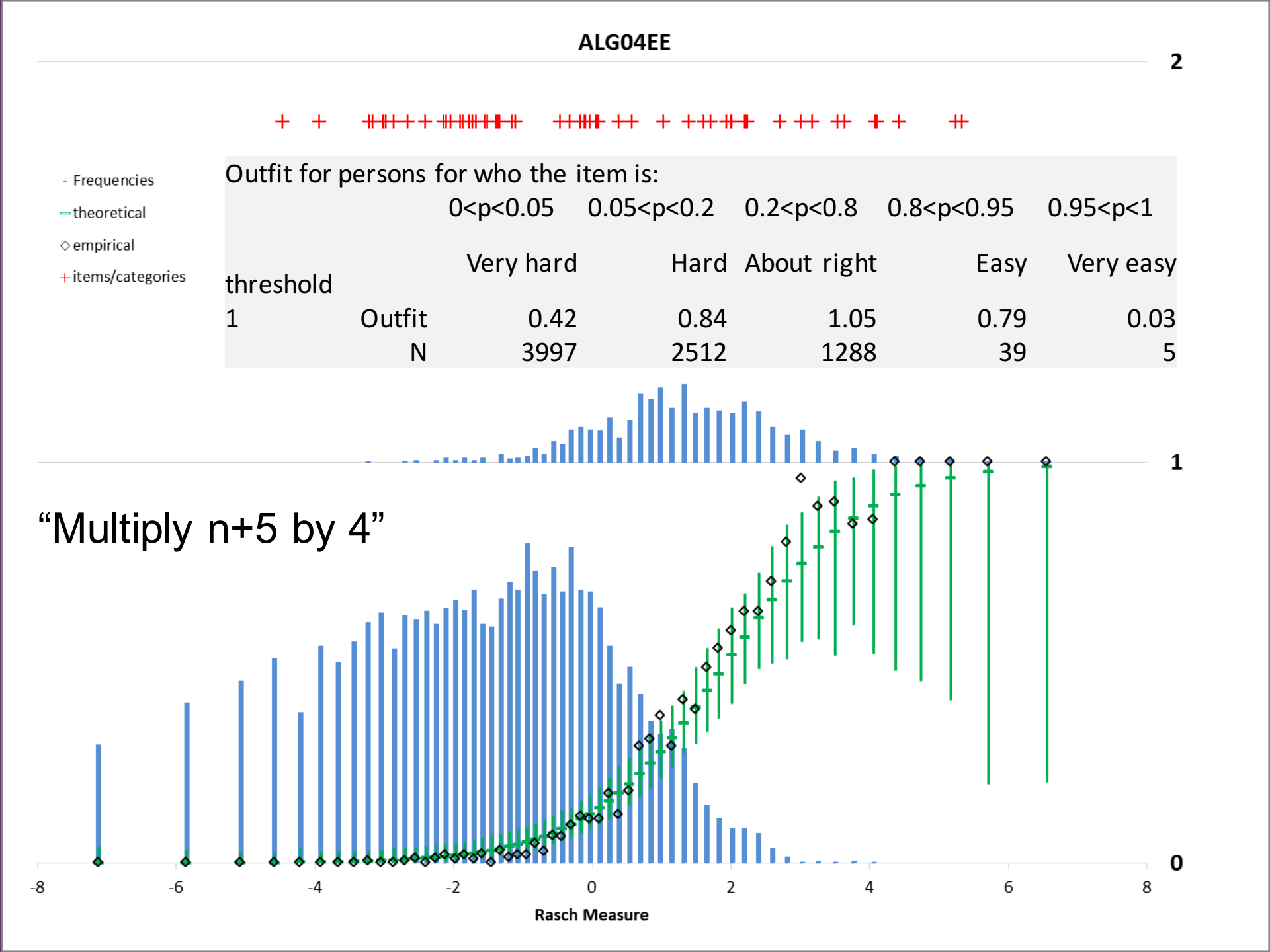
CEM Centre for Evaluation & Monitoring

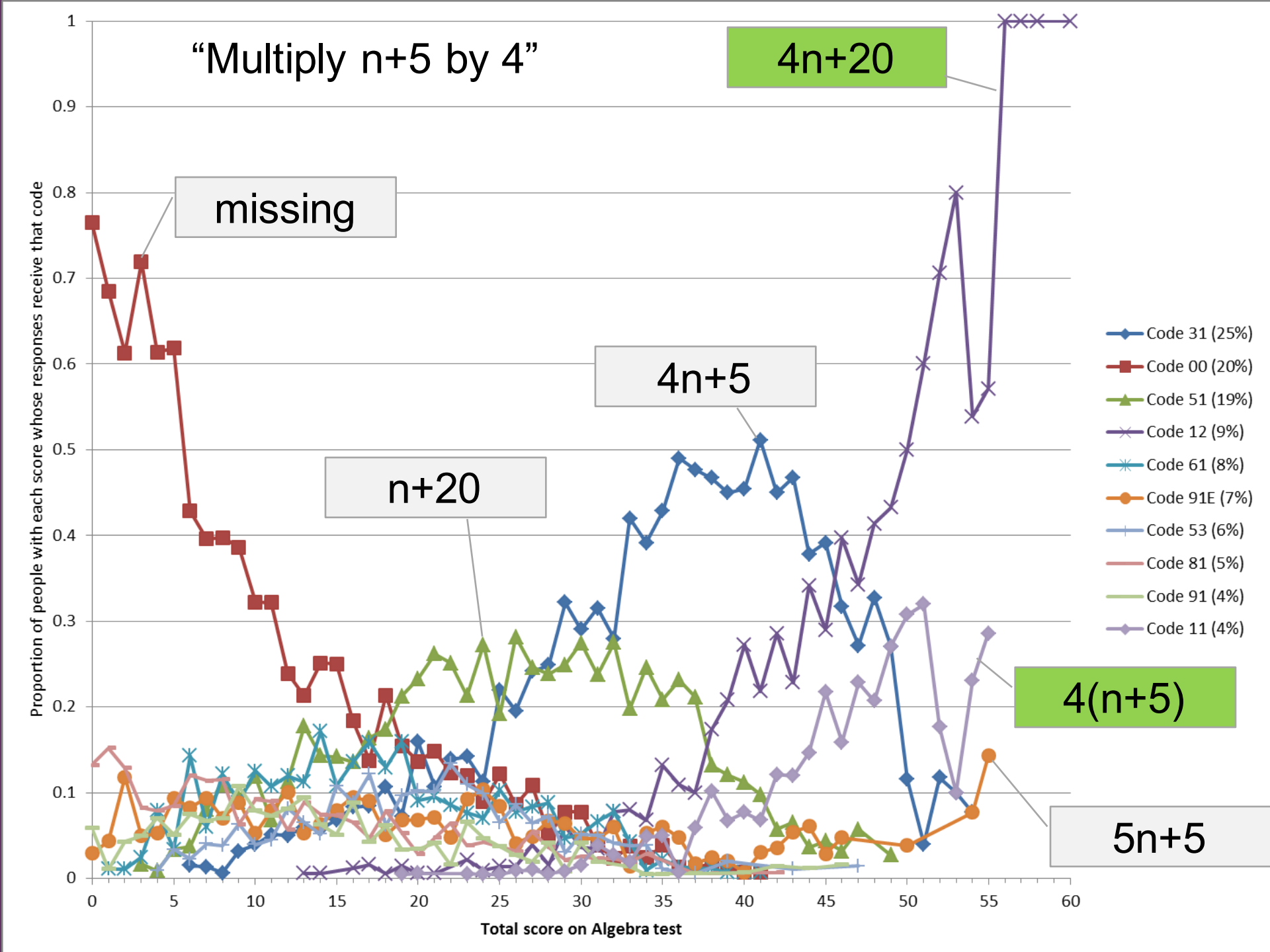
Question: ALG04EE

Mult n+5 by 4

Item Statistics

Number of responses	7,841
Maximum score	1
Mean score on item	0.11
Item difficulty (Rasch measure)	1.94
INFIT (mean sq)	0.92
OUTFIT (mean sq)	0.6682
IRT Discrimination parameter	1.0816
Item-measure correlation (actual)	0.4338
Item-measure correlation (expected)	0.3969
Percent match model (observed)	90.664
Percent match model (expected)	90.128


ALG04EE
2
1
0
Frequencies
theoretical
empirical
items/categories
Outfit for persons for who the item is:
0<p<0.05 0.05<p<0.2 0.2<p<0.8 0.8<p<0.95 0.95<p<1
threshold Very hard Hard About right Easy Very easy
1 Outfit 0.42 0.84 1.05 0.79 0.03
N 3997 2512 1288 39 5
"Multiply n+5 by 4"
-8 -6 -4 -2 0 2 4 6 8
Rasch Measure

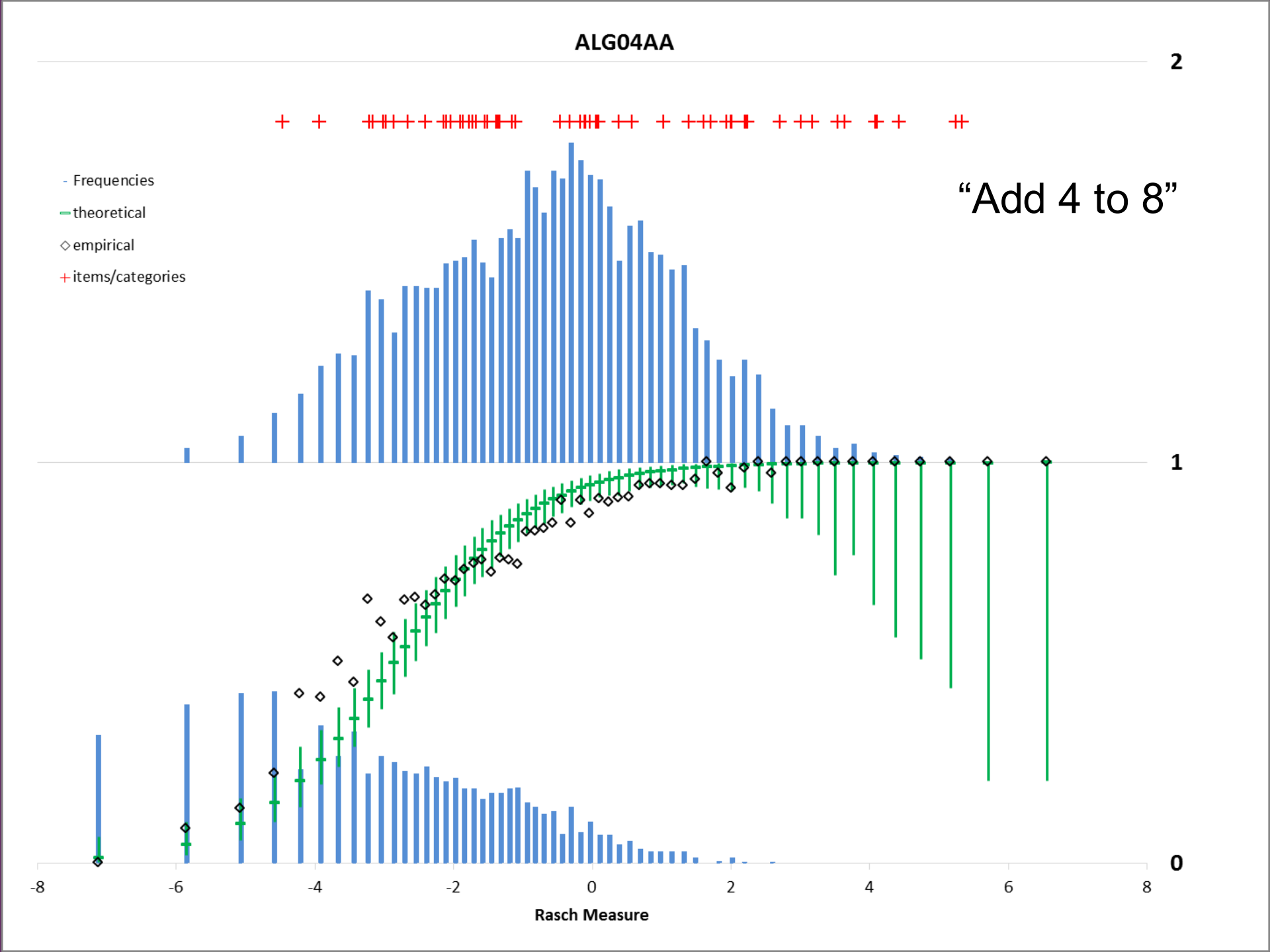

"Multiply n+5 by 4"
4n+20
missing
4n+5
n+20
4(n+5)
5n+5
Code 31 (25%)
Code 00 (20%)
Code 51 (19%)
Code 12 (9%)
Code 61 (8%)
Code 91E (7%)
Code 53 (6%)
Code 81 (5%)
Code 91 (4%)
Code 11 (4%)
Proportion of people with each score whose responses receive that code
Total score on Algebra test
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
35
40
45
50
55
60

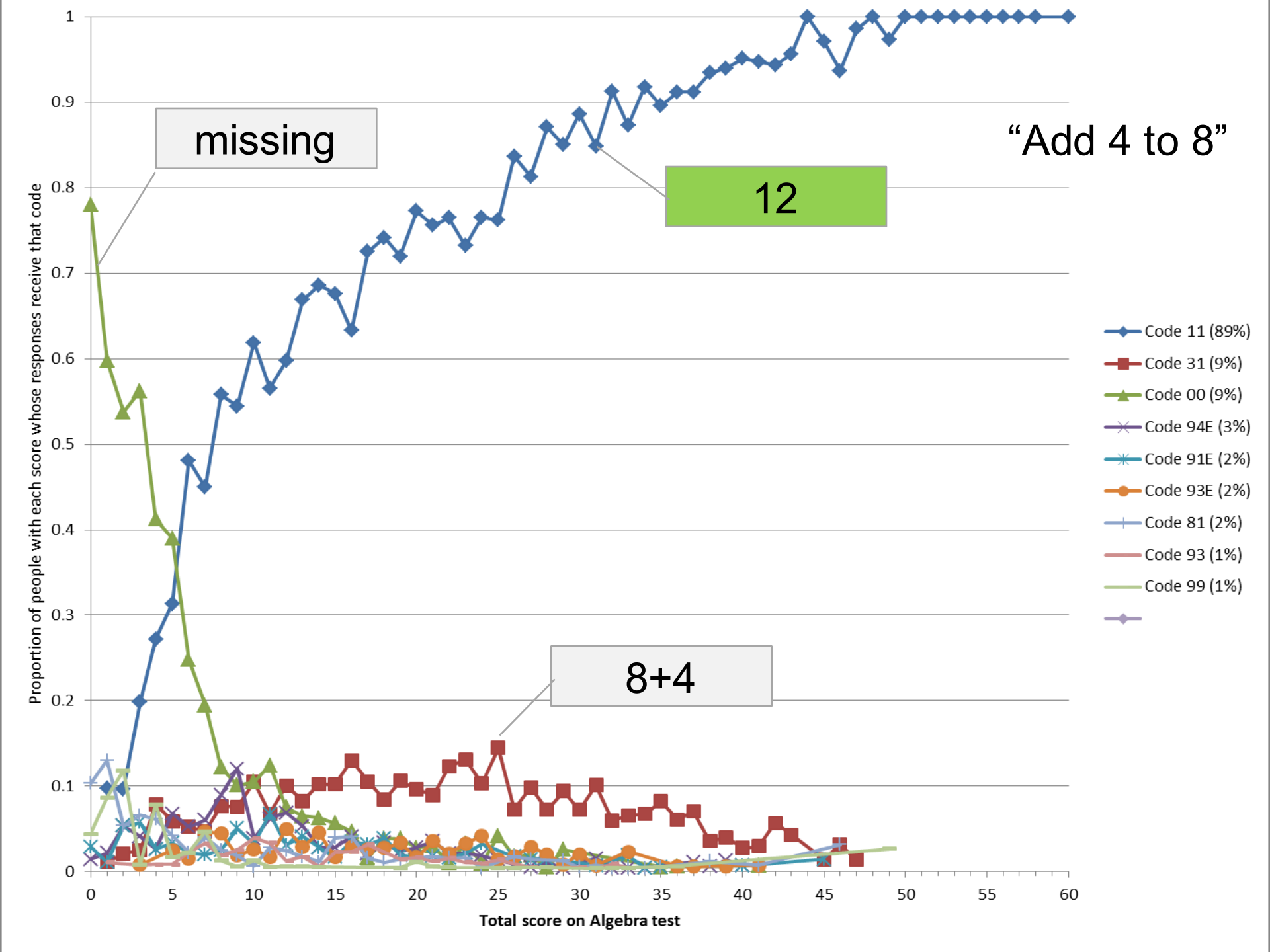
Question: ALG04AA

Add 4 to 8

Item Statistics

Number of responses	7,841
Maximum score	1
Mean score on item	0.74
Item difficulty (Rasch measure)	-2.86
INFIT (mean sq)	1.2693
OUTFIT (mean sq)	1.627
IRT Discrimination parameter	0.6577
Item-measure correlation (actual)	0.4636
Item-measure correlation (expected)	0.5731
Percent match model (observed)	78.117
Percent match model (expected)	83.017


ALG04AA
Frequencies
theoretical
empirical
items/categories
"Add 4 to 8"
2
1
0
-8
-6
-4
-2
0
2
4
6
8
Rasch Measure

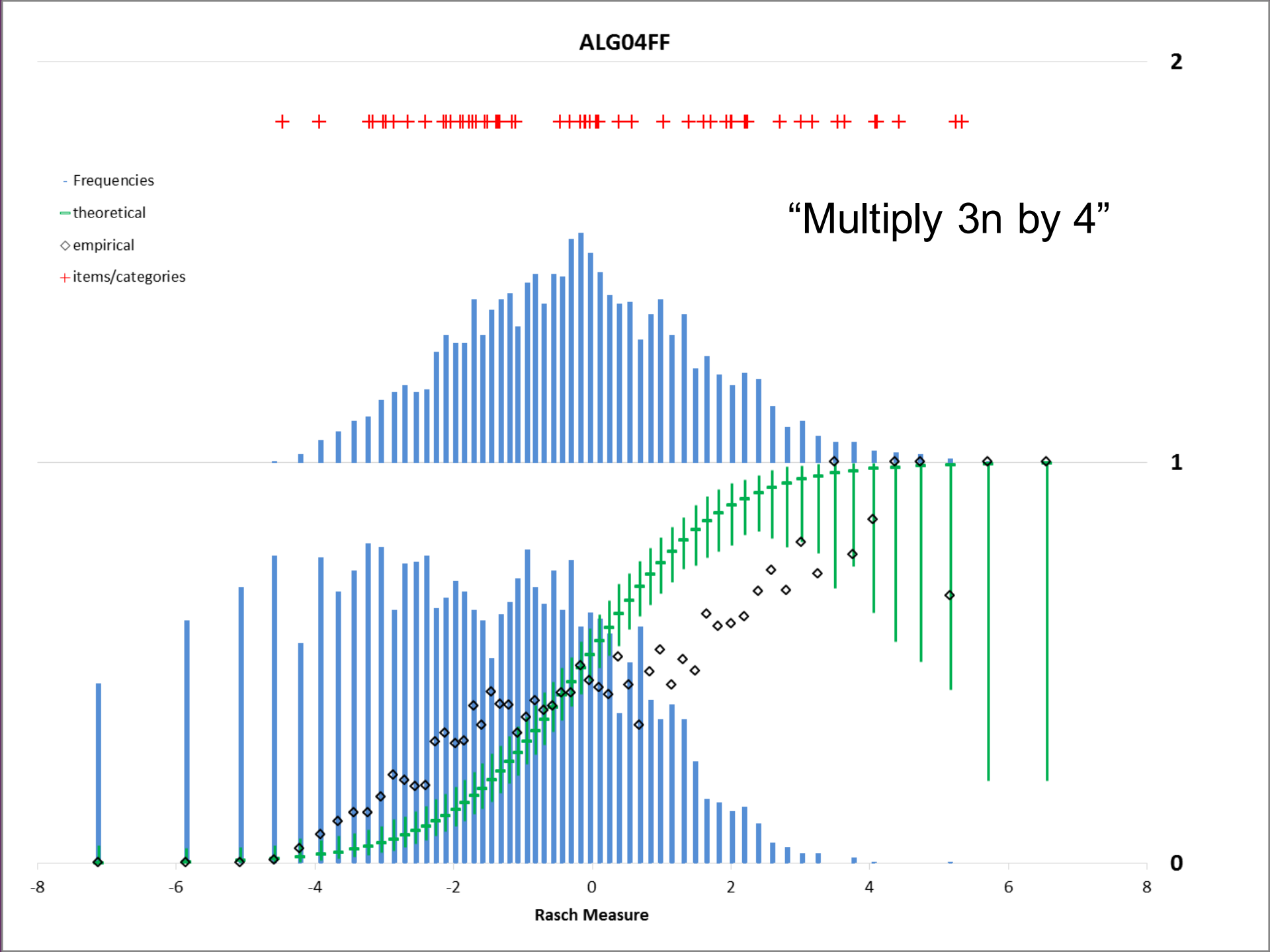

missing
12
8+4
"Add 4 to 8"
Proportion of people with each score whose responses receive that code
Total score on Algebra test
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
35
40
45
50
55
60
Code 11 (89%)
Code 31 (9%)
Code 00 (9%)
Code 94E (3%)
Code 91E (2%)
Code 93E (2%)
Code 81 (2%)
Code 93 (1%)
Code 99 (1%)

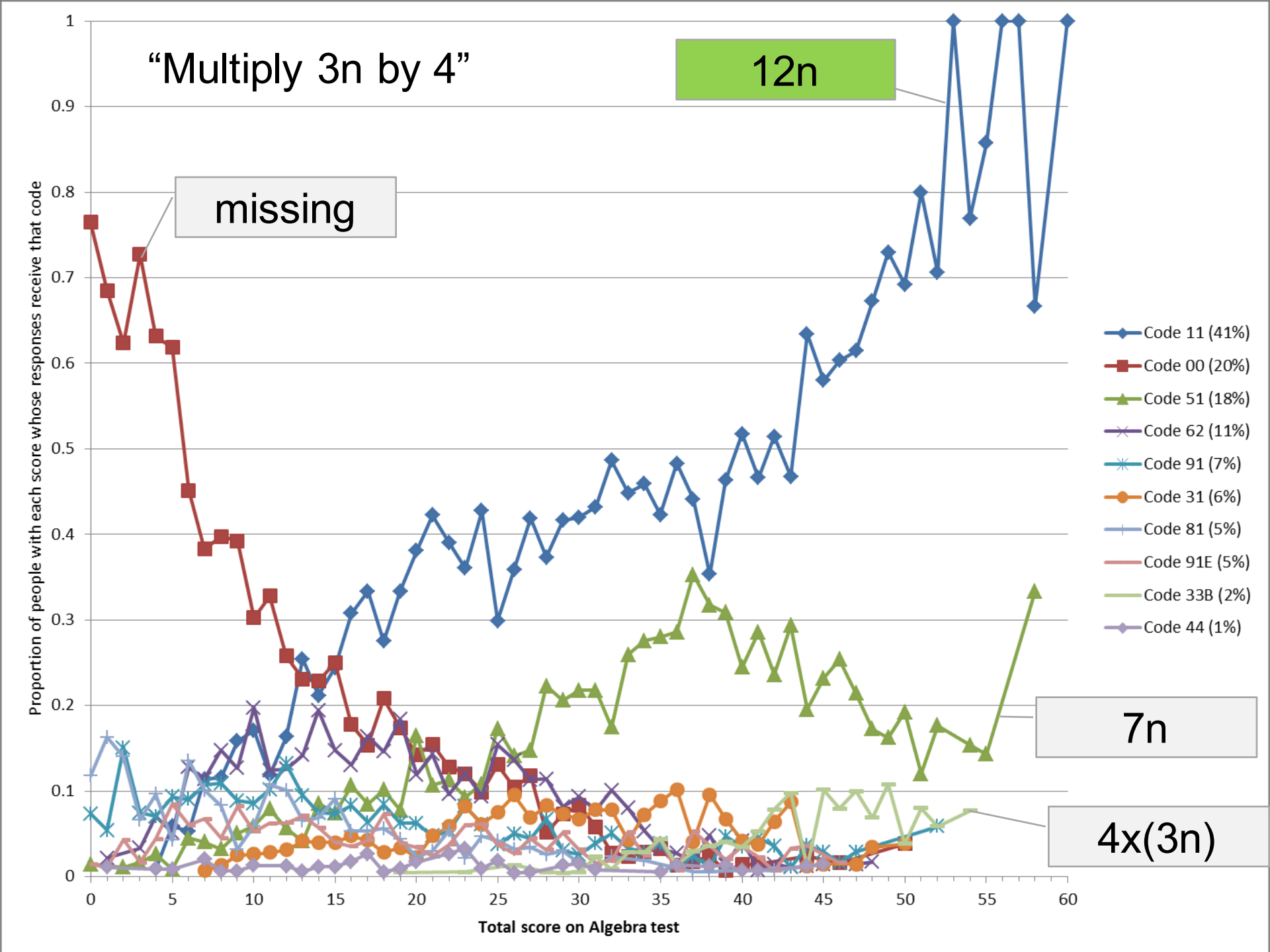
Question: ALG04FF

Mult 3n by 4

Item Statistics

Number of responses	7,841
Maximum score	1
Mean score on item	0.34
Item difficulty (Rasch measure)	-0.11
INFIT (mean sq)	1.5233
OUTFIT (mean sq)	1.8801
IRT Discrimination parameter	0.1397
Item-measure correlation (actual)	0.3335
Item-measure correlation (expected)	0.5512
Percent match model (observed)	65.338
Percent match model (expected)	78.106


ALG04FF
2
1
0
Frequencies
theoretical
empirical
items/categories
“Multiply 3n by 4”
-8
-6
-4
-2
0
2
4
6
8
Rasch Measure


"Multiply 3n by 4"
12n
missing
7n
4x(3n)
Proportion of people with each score whose responses receive that code
Total score on Algebra test
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
35
40
45
50
55
60
Code 11 (41%)
Code 00 (20%)
Code 51 (18%)
Code 62 (11%)
Code 91 (7%)
Code 31 (6%)
Code 81 (5%)
Code 91E (5%)
Code 33B (2%)
Code 44 (1%)