SNSA November norming study analysis report

DOCUMENT CONTROL

Theme	Standardisation
PRODUCT TITLE	SNSA data analysis plan
PRODUCT CODE	AR03
DOCUMENT No.	AR03-13

VERSION HISTORY

Date	Version	Comments	Author	Approved by	Date Approved
08/02/18	0.1	draft	Redacted - exemption applies		
16/2/18	0.2	edits			
21/2/18	0.3	further edits			
22/2/18	0.4	further edits			
27/2/18	0.5	further edits			
27/2/18	0.6	final edits			

Contents

Redacted - 4 lines - exemption applies

Introduction - Norm referencing

Norm referencing is an alignment of outcomes against distributions of achievement for population reference groups.

For SNSA, the immediate purpose of norm referencing is to provide an overall picture of Scottish learners' assessment results in numeracy, reading and writing skills against which assessment outcomes of individual children, and of subgroups such as classes or year groups in a school, or children in different SIMD groups, can be measured.

Sampling

With respect to establishing population norms, the ACER proposal envisaged the establishment of norms for each of the domains and year levels at a fixed time point during the school year. Furthermore our proposal was that the norming study be based on a probabilistic sample of learners. By definition, a probabilistic survey is one where each learner in the population has a *known, non-zero chance* of inclusion in the sample.

It was agreed to conduct two norming studies in the first year of SNSA implementation: one in November 2017, and one in March 2018. The chosen option for the norming studies had the following key features:

- Providing a collection of data from an adaptive online assessment administration in November 2017 and March 2018; and
- A systematic randomised selection of children and young people who complete the adaptive assessment in specified testing windows.

An equal probability sample of 800 learners across all schools in Scotland was selected for the norming study for each assessment domain (P1: Literacy and Numeracy, P4/P7/S3: Reading, Numeracy and Writing) and year group (8800 in total for November and 8800 in total for March 2018). Participants were sampled systematically from a list of all learners in Scotland using a 'random start, constant interval' selection method. The list was systematically organised (sorted) by important auxiliary variables to ensure implicit stratification of the sample by those variables. Local Authority (LA), SIMD, age and gender were agreed as stratification variables for the sampling of learners in the four year levels.

Data analyses for the November norming study

The November norming study was completed with a very satisfactory overall response rate (85%). The response data was representative of all LAs, with only one LA, East Ayrshire, having a large proportion of sampled schools not taking part (57%). This could have been due to a relatively late initial assignment of assessment forms to this particular LA.

The psychometrics team conducted several initial analysis procedures based on the data collected from the study. The analyses included a review of the data input from the Horizon system as well as some additional data cleaning and processing to prepare the data sets for the psychometric review.

The norming study covered four different years (P1, P4, P7 and S3). A number of different domains were covered in the study: Numeracy, (early) literacy, reading, and writing (grammar and punctuation, and spelling). Appendix 1 shows all forms indicating the assessed domains at each year level, and the numbers of assigned and assessed learners.

The data summary report is attached as Appendix 1.

Following the initial database verification checks, the psychometrics team performed a scaling review of the November norming study data conducting an IRT-based analysis. The scaling review included the following steps:

1. *Scottish calibrations by year level and domain* (P1 L/N, P4 R/N/W, P7 R/N/W, S3 R/N/W): Item fit, item statistics and Item Characteristic Curves (ICC) were reviewed to assess the psychometric quality of assessment items with regard to the Scottish context.
2. *Comparisons of Scottish and pre-calibrated item parameters* (P1 L/N, P4 R/N/W, P7 R/N/W, S3 R/N/W):
   We reviewed the differences between standardised Scottish (norming study) and pre-calibrated item parameters per year level and domain, and produced graphical displays (scatterplots with confidence interval lines). This enabled us to identify potential differences in item functioning between the underlying calibrations of the current delivery system and those based on the Scottish norming study.
3. *Review of differential item functioning* (P1 R/N, P4 R/N/W, P7 R/N/W, S3 R/N/W):
   The review of the occurrence of differential item functioning (DIF) by gender using an IRT model was conducted to investigate measurement equivalence of assessment items across gender groups in the Scottish context.

Item maps summarising the information from the calibrations are attached as Appendices 2-12.

In addition to the scaling review, we conducted a review of adaptive design. This involved a review of admissible paths (ABD, ABE, ABF, ACD, ACE, and ACF) to obtain percentages of learners in each path as well as mean scores and the distributions of assessment scores within each path in order to confirm the appropriateness of the adaptive design for the Scottish context.

The branching rules were also reviewed in order to check the extent to which they can be further optimised for the Scottish context.

The tables with the cluster summaries are attached as Appendices 13-16.

Conclusions from the November norming study

This section reports on the conclusions from the data analyses of the November norming study.

Table 1 Assessments reliability results

Form	Reliability
P1 numeracy	0.829
P1 literacy	0.782
P4 numeracy	0.865
P4 reading	0.897
P4 writing	0.892
P7 numeracy	0.887
P7 reading	0.876
P7 writing	0.841
S3 numeracy	0.897
S3 reading	0.884
S3 writing	0.816

The first analysis that was conducted was to ensure the reliability of the assessments. The summary of the results presented in Table 1 shows that the reliability of all assessments forms was high (more than 0.75). The reliability measure used in the analyses is based on item response modelling and indicates the degree to which items measure the same latent construct (i.e. the respective assessment domain). Reliabilities above 0.7 are typically considered as satisfactory and above 0.8 as excellent.

The second analysis conducted was focused on targeting – ensuring that the items used are appropriate for the learners' proficiency.

Appendices 2-12 show the Item Variable Maps for all forms which provides a mapping of Rasch item difficulty and learner capacity estimates. On the right hand side of the map, item difficulty parameters (indicating where a learner has a 0.5 probability of giving a correct response to an item) are shown on a logit scale according to the estimates of their difficulty, from the easiest (at the bottom of the map) to the most difficult (at the top of the map).

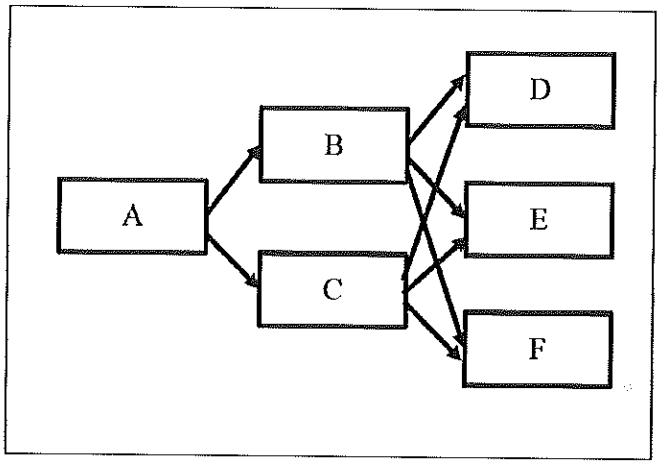
Analyses of the data from the learners in the norming study showed a wide distribution of achievement within the assessments. The analyses indicate that the items used in the study covered the range of capacity levels for learners within each domain and at each corresponding year level.

The Item Variable Maps show that while generally the distribution of learners' performances and the distribution of item difficulties are broadly at a similar level, there is an indication that there were somewhat more SNSA items matching the lower levels than the higher levels of learner capacity. The targeting differed somewhat across year levels and domains, with the closest match being between item difficulties and measured capacities at P1. Given that there was no prior information about targeting (except from the item trial in February 2017 based on convenience samples), achieving an ideal match between item difficulties and learner capacities has been a challenge since the inception of SNSA. We believe that the results generally indicate that each of the assessment forms is broadly at an appropriate level of difficulty for its corresponding group of learners, but that it is desirable to add more difficult items in the future, in order to make assessments more challenging and obtain more information about higher levels in each assessment domain.

Further analyses focused on the adaptive nature of the assessments, and reviewed the pathways within each assessment. Figure 1 illustrates the adaptive assessment design, and the branching process.

Figure 1 Adaptive assessment design

Adaptive assessment design



As illustrated in Figure 1, there are several possible paths that each learner can take, depending on how well they are performing in the assessment. Each cluster of items and each pathway is very carefully designed so that the whole assessment, for each learner, provides coverage of the required skills, as well as being matched to the demonstrated attainment of the learner as he or she progresses through the assessment.

As part of the analysis, the review percentages of learners in each path, as well as mean scores and the distributions of assessment scores within each path, were obtained, in order to confirm the appropriateness of the adaptive design for the Scottish context. The results of the pathways analysis are presented in Appendices 13-16.

The results show that the adaptive system works very well in distinguishing between learners with more and less capacity: hardly any learners were found in the 'non-expected' paths from the more difficult middle cluster to the easiest one in the third position, or from the easier middle cluster to the most difficult one at the end. Thus, it can be concluded that the algorithm determining learners' capacity at the end of the starting cluster (A) leading them one to the most appropriate next testlet works well, and that there are hardly any cases indicating a misclassification of learners' capacity.

The higher percentages in the more difficult clusters (though at different levels across year levels and domains) provide support for the previous finding regarding the targeting of the assessment: learners are doing relatively well on the current assessments as the majority of learners are being presented with the more difficult clusters. Therefore the assessments would benefit from the increase in the overall level of difficulty within each cluster.

Next steps

Data from analysis will be used to inform two key areas:

1) It will be used in conjunction with data from the Long Scale Equating Study and March Norming Study to inform the construction of the long scale to be used in reporting for the 2018-19 school year onwards.

2) Data will be used to inform discussion between ACER and SG around the target difficulty of the assessments to be constructed for the next school year. Assessments for the next school year will be constructed from a pool of items comprising items from the current live assessments and items from the current in-test trial clusters which are deemed to be performing appropriately. ACER recognises that it may not be appropriate to move to the 'perfect' capacity model for the assessments for the next school year for several reasons:
    a. noticeable change in difficulty between the first and second years of the assessments could be negatively received by teachers,
    b. it is possible that the item pool may not contain a sufficient number of higher capacity items, given the views of previous question review panels in rejecting some items on the grounds they were too difficult for the target year group.

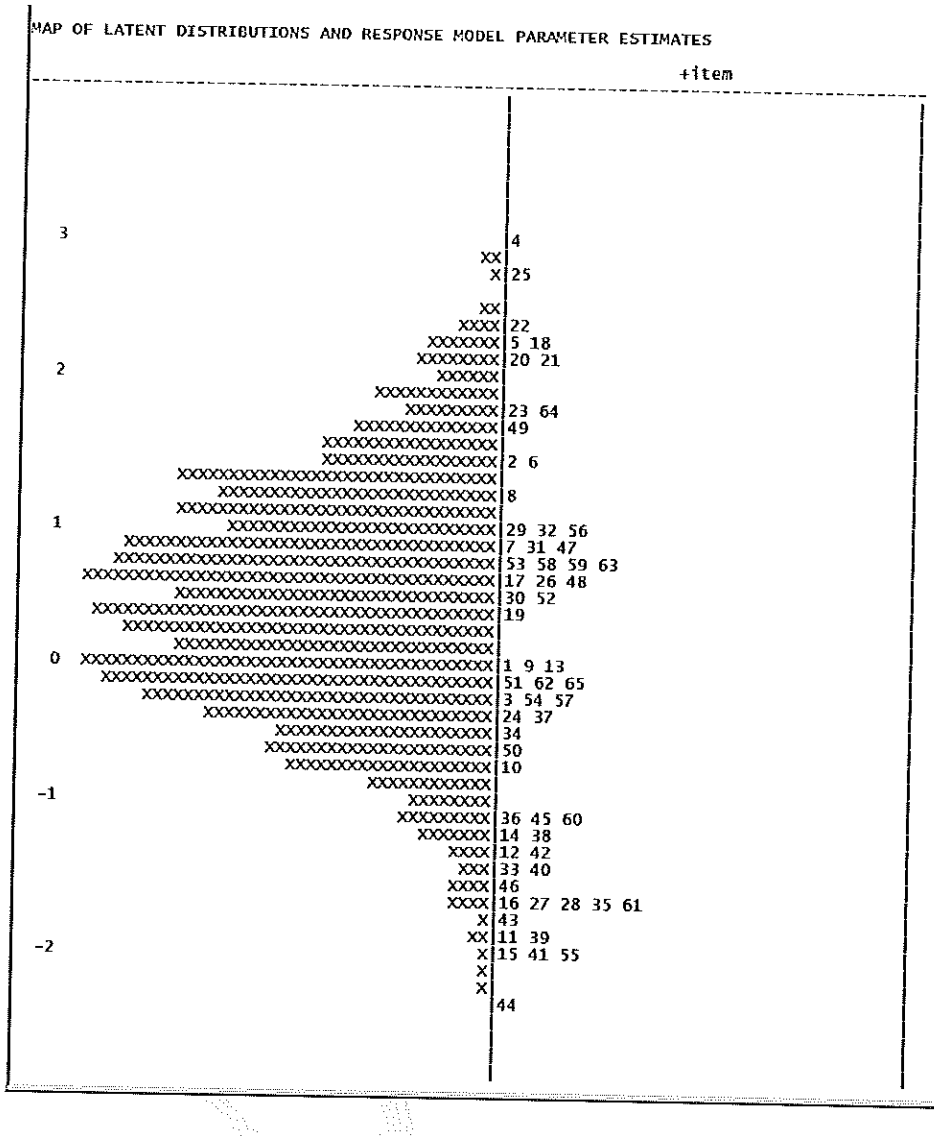
Appendices

Redacted - 4 lines - exemption applies

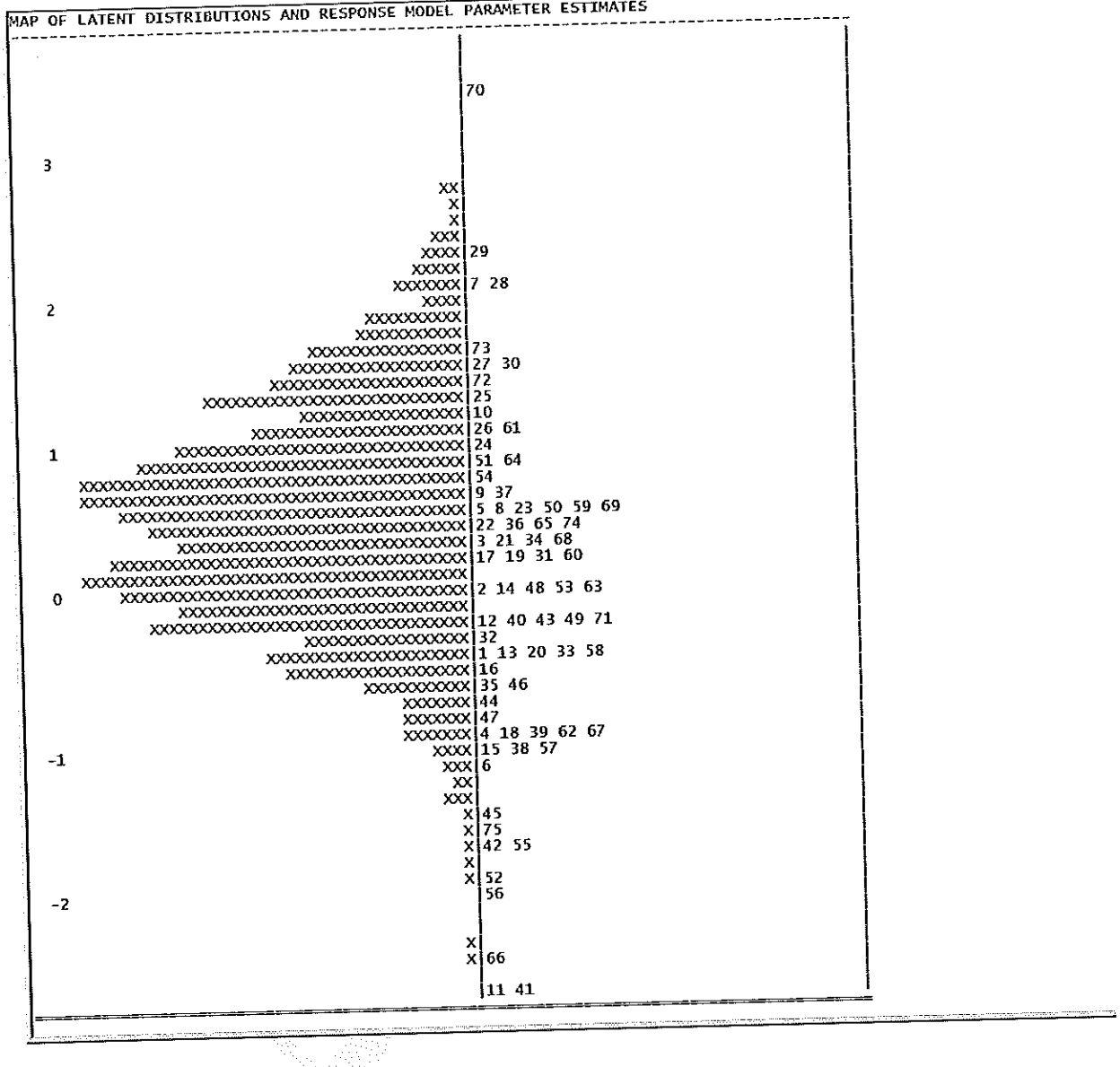
Appendix 1 – Data summary report

NNS Assessment	NNS # assigned	NNS # sat	% of total # assigned
P1 Numeracy	800	688	86
P1 Literacy	800	696	87
P4 Numeracy	800	701	88
P4 Reading	800	705	88
P4 Writing	800	699	87
P7 Numeracy	800	698	87
P7 Reading	800	698	87
P7 Writing	800	707	88
S3 Numeracy	800	638	80
S3 Reading	800	613	77
S3 Writing	800	595	74
Total	**8800**	**7438**	**85**

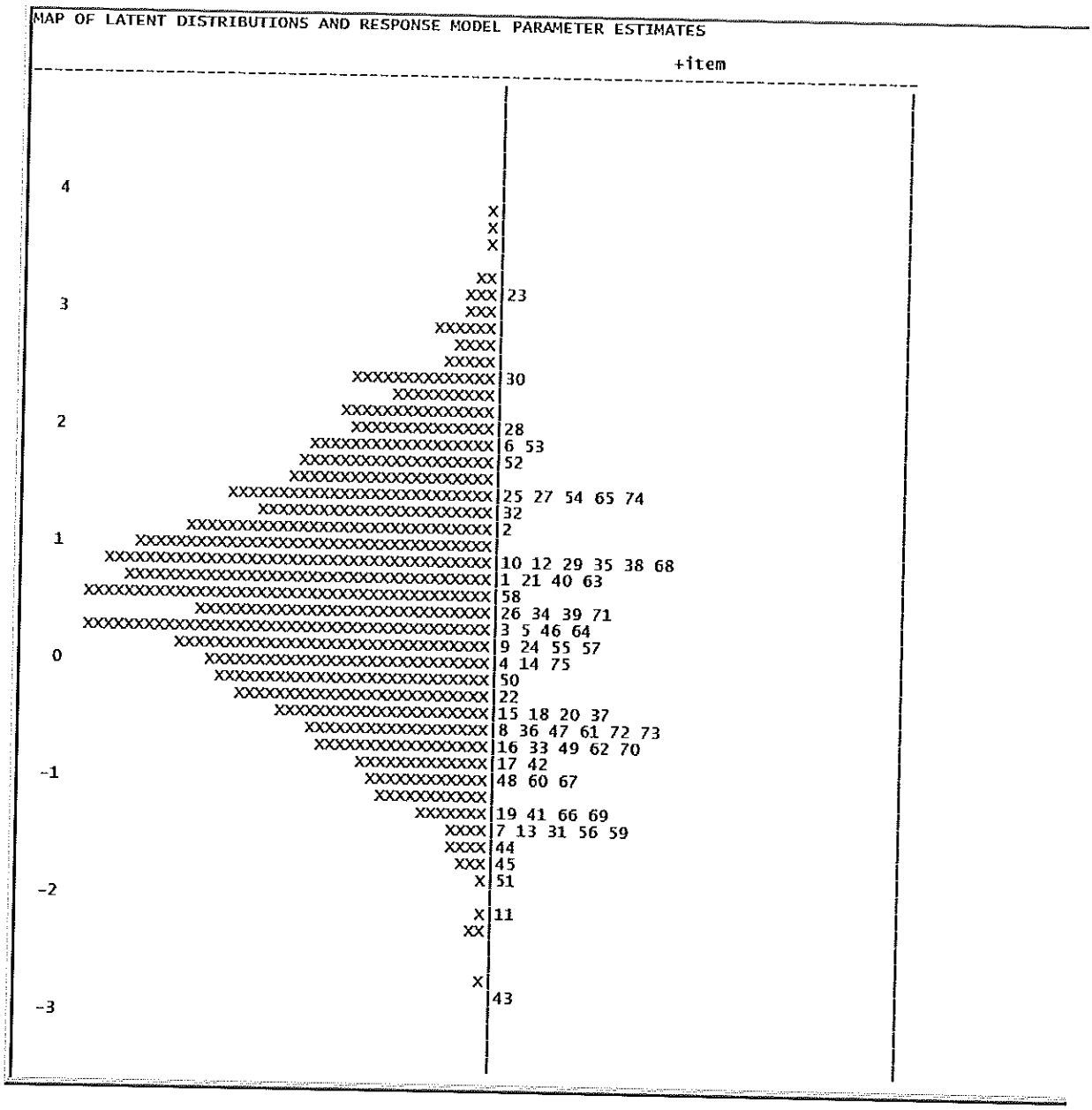
Appendix 2 – Item Map for P1 Numeracy



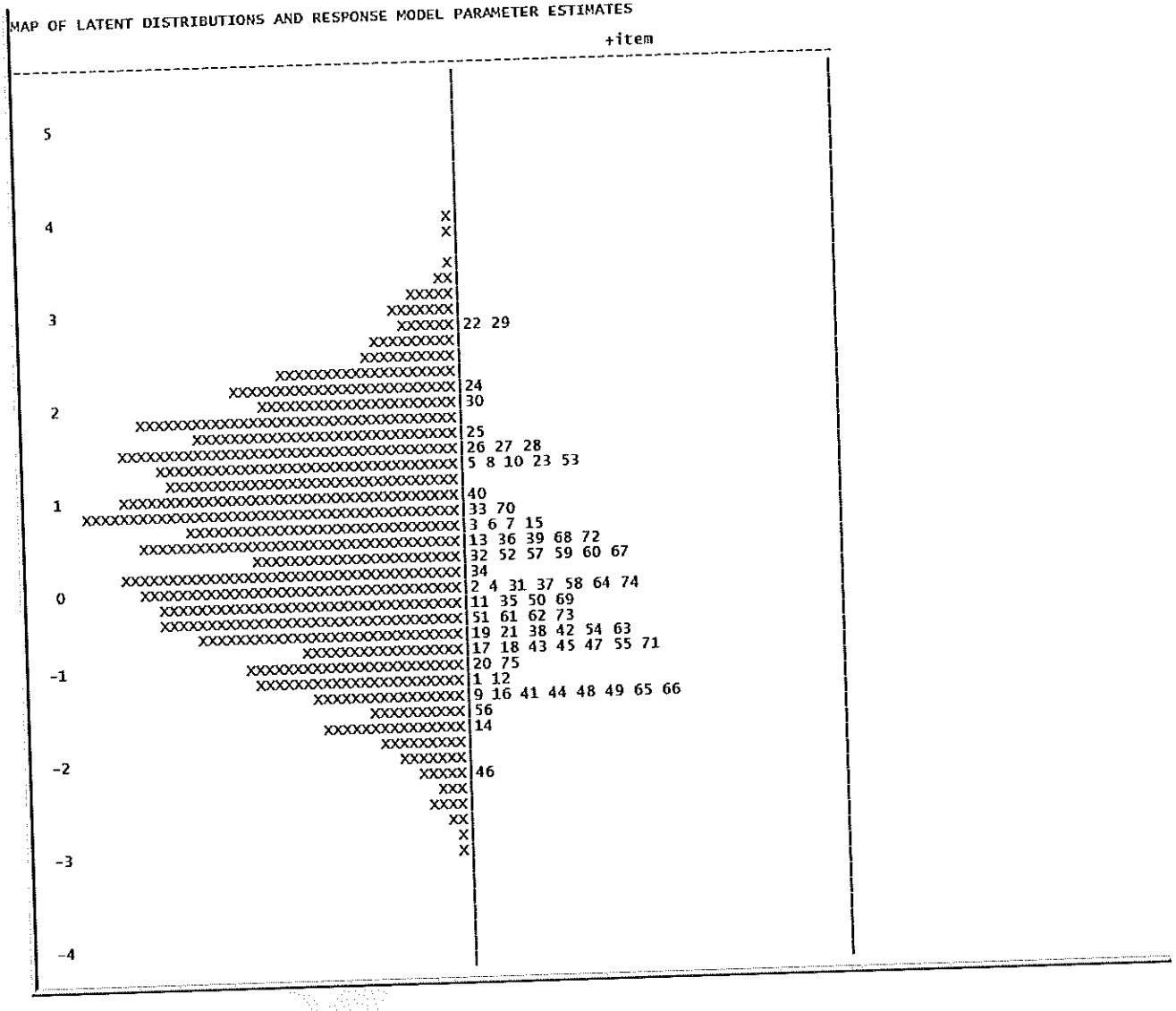
Appendix 3 – Item Map for P1 Literacy



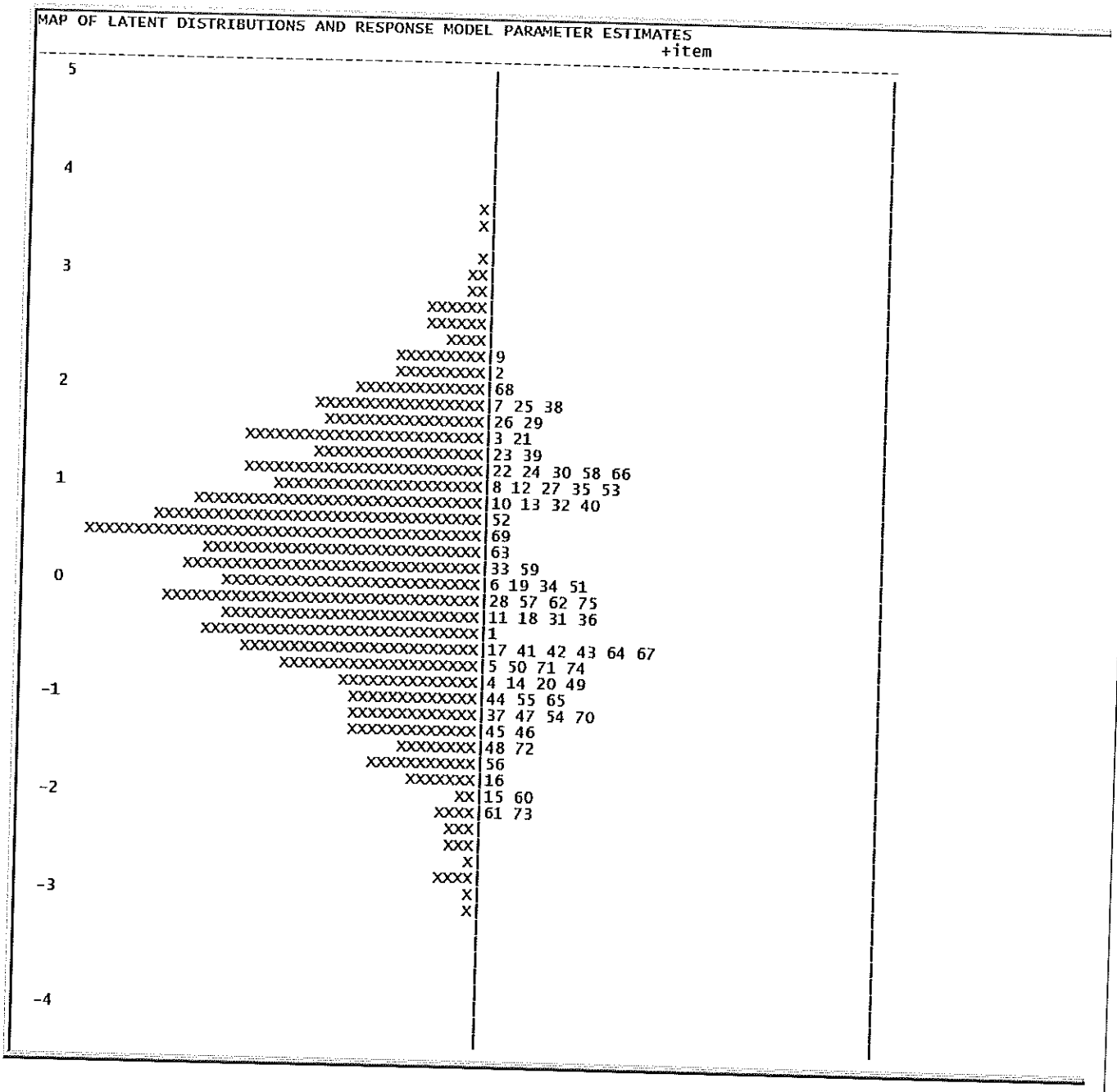
Appendix 4 – Item Map for P4 Numeracy



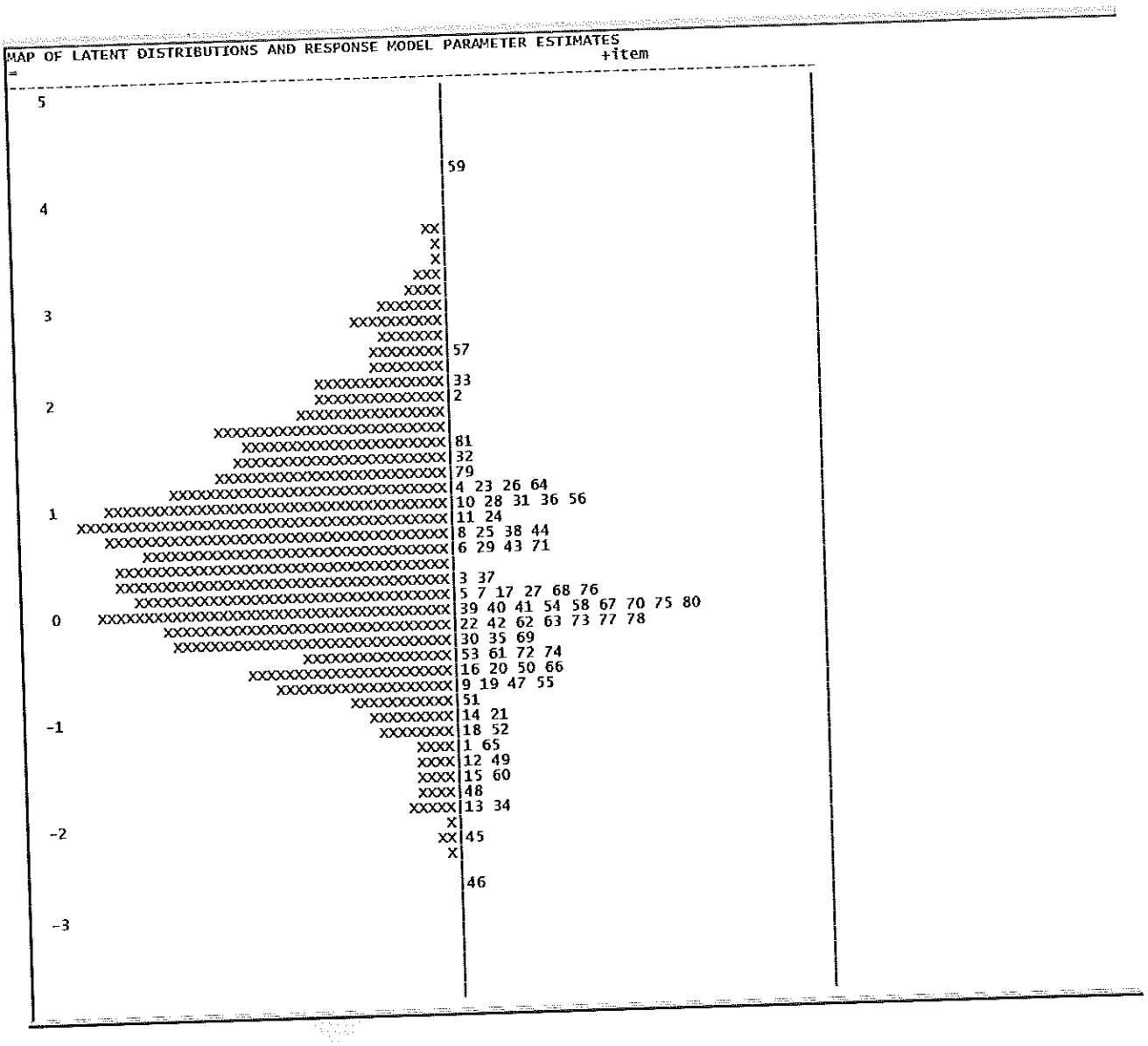
Appendix 5 – Item Map for P4 Reading



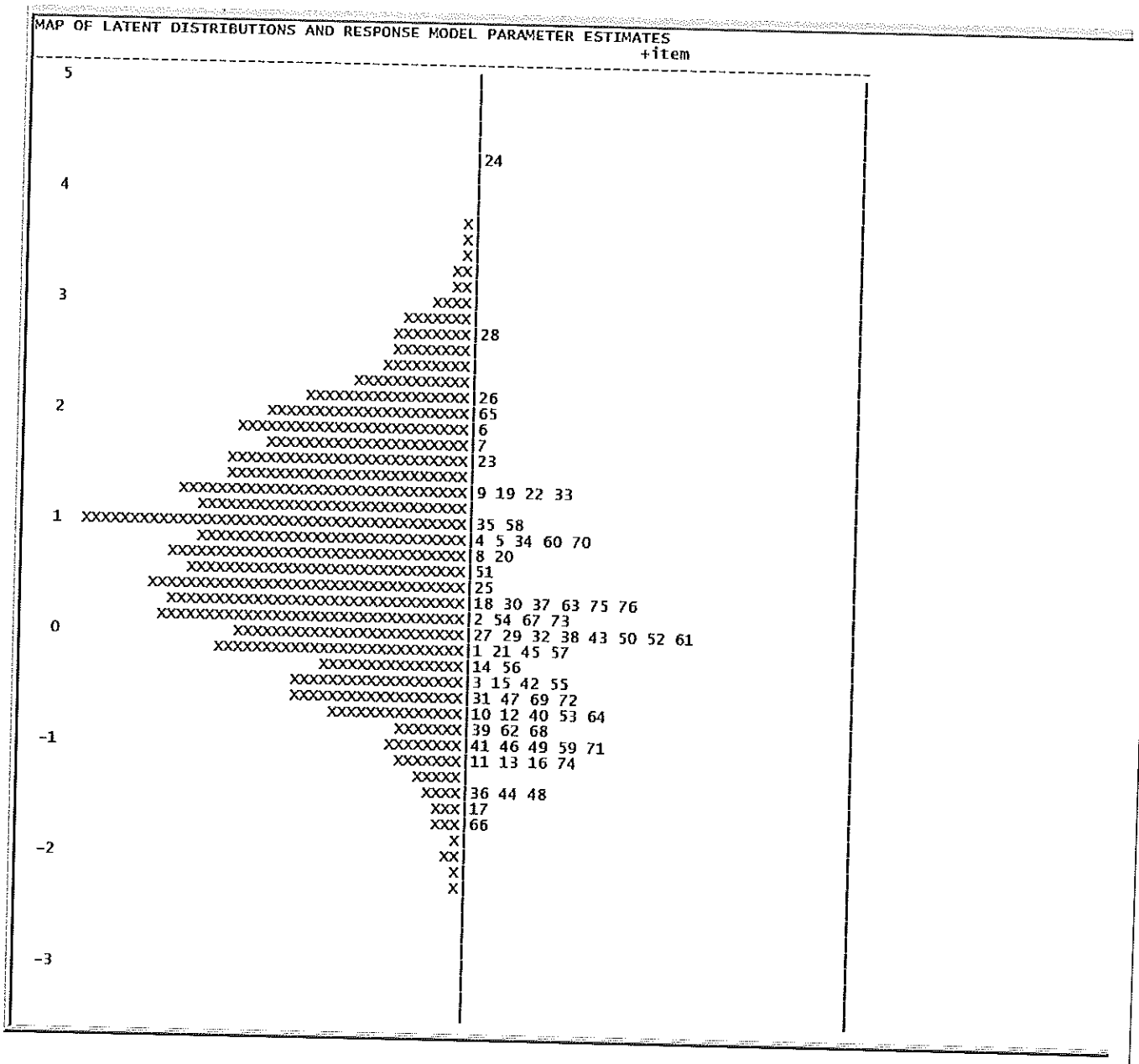
Appendix 6 – Item Map for P4 Writing



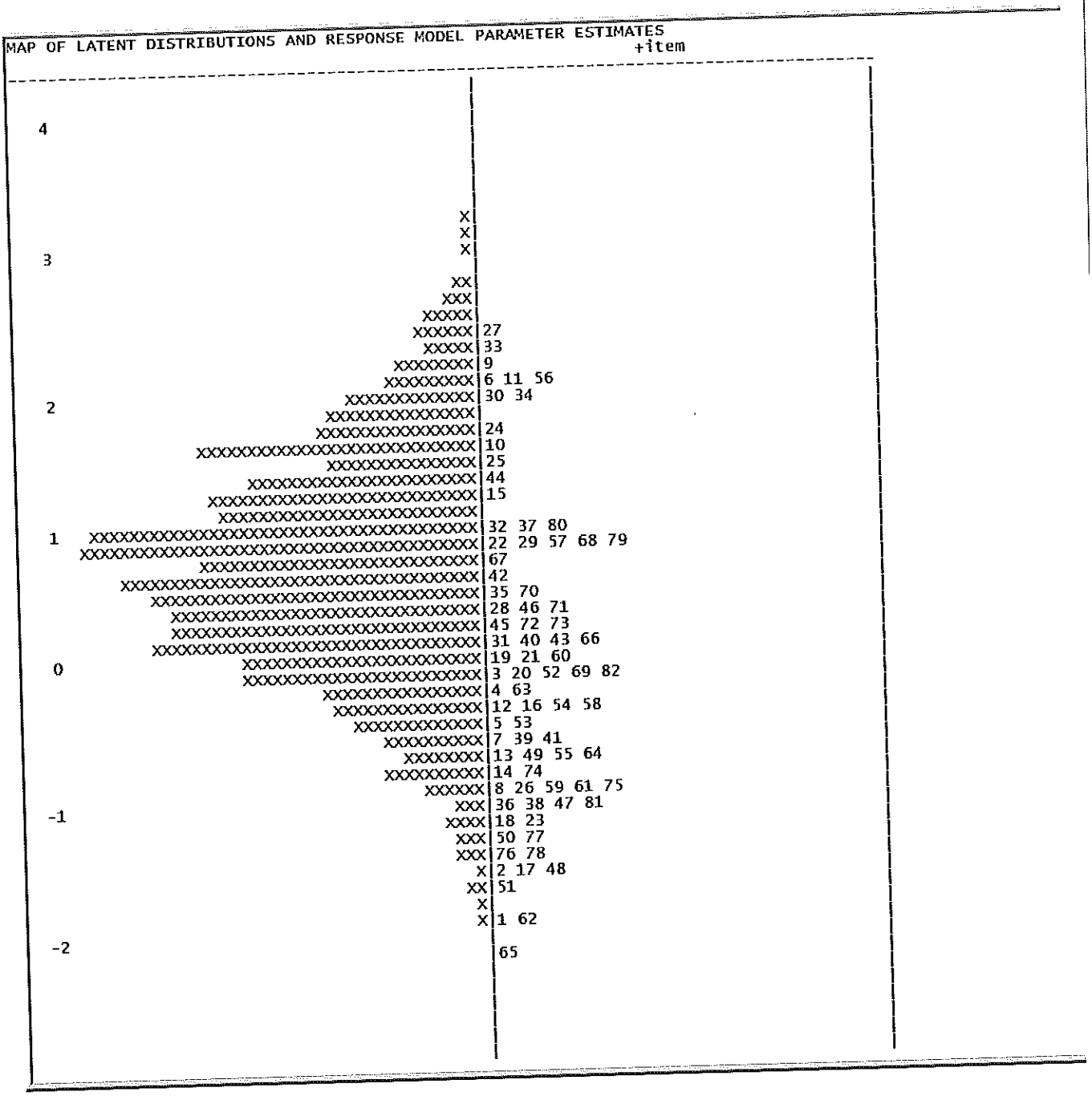
Appendix 7 – Item Map for P7 Numeracy



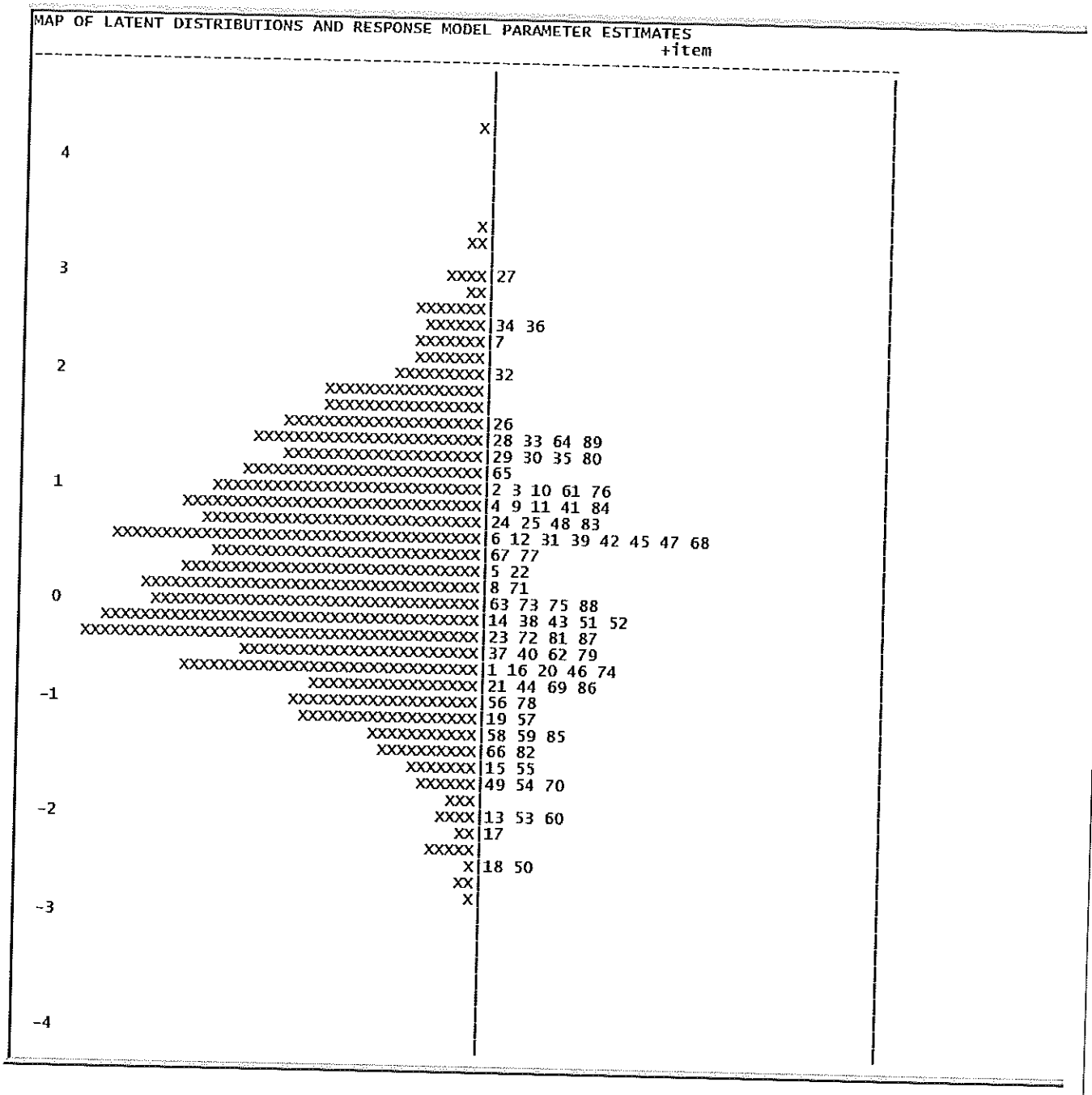
Appendix 8 – Item Map for P7 Reading



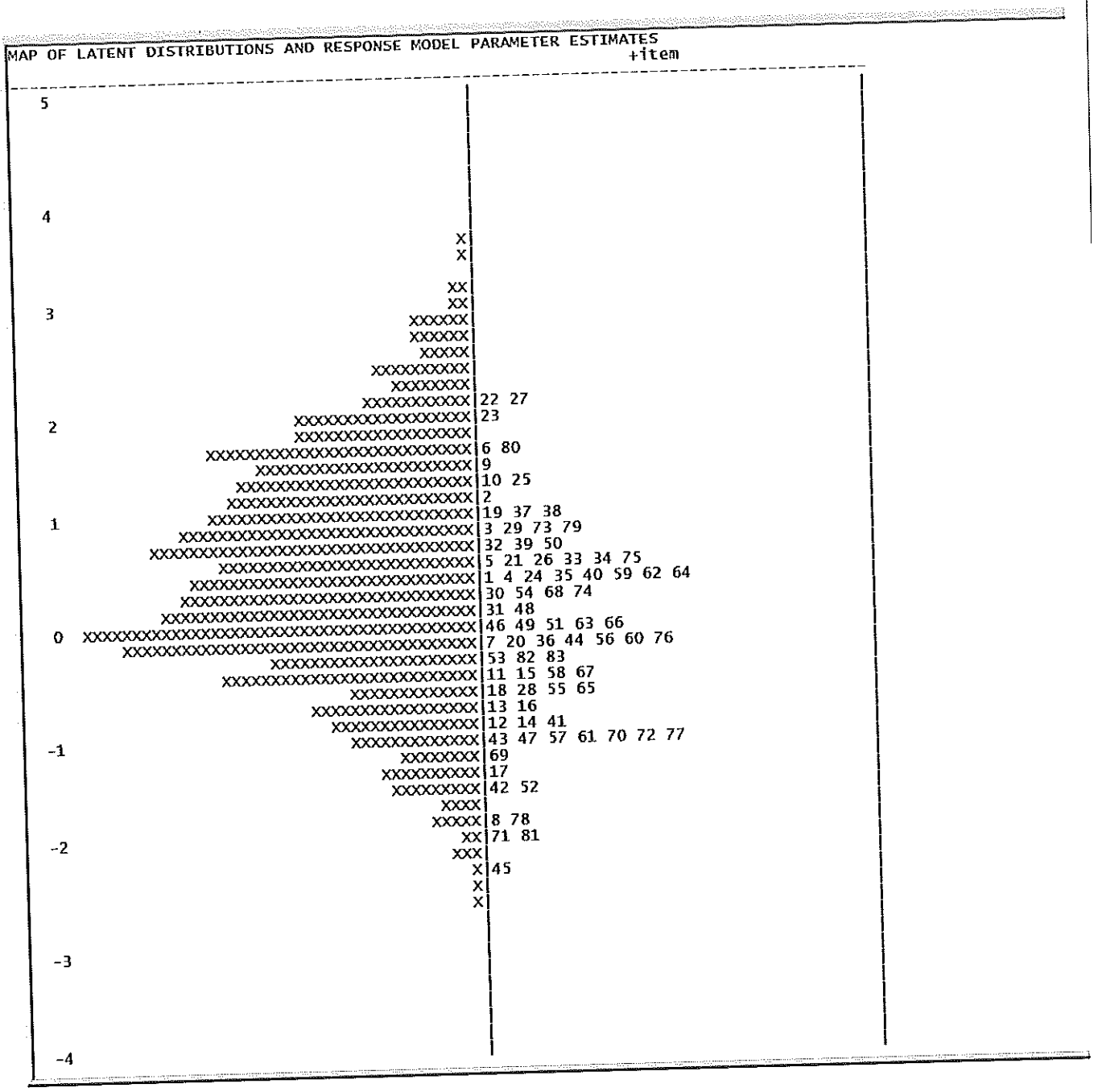
Appendix 9 – Item Map for P7 Writing



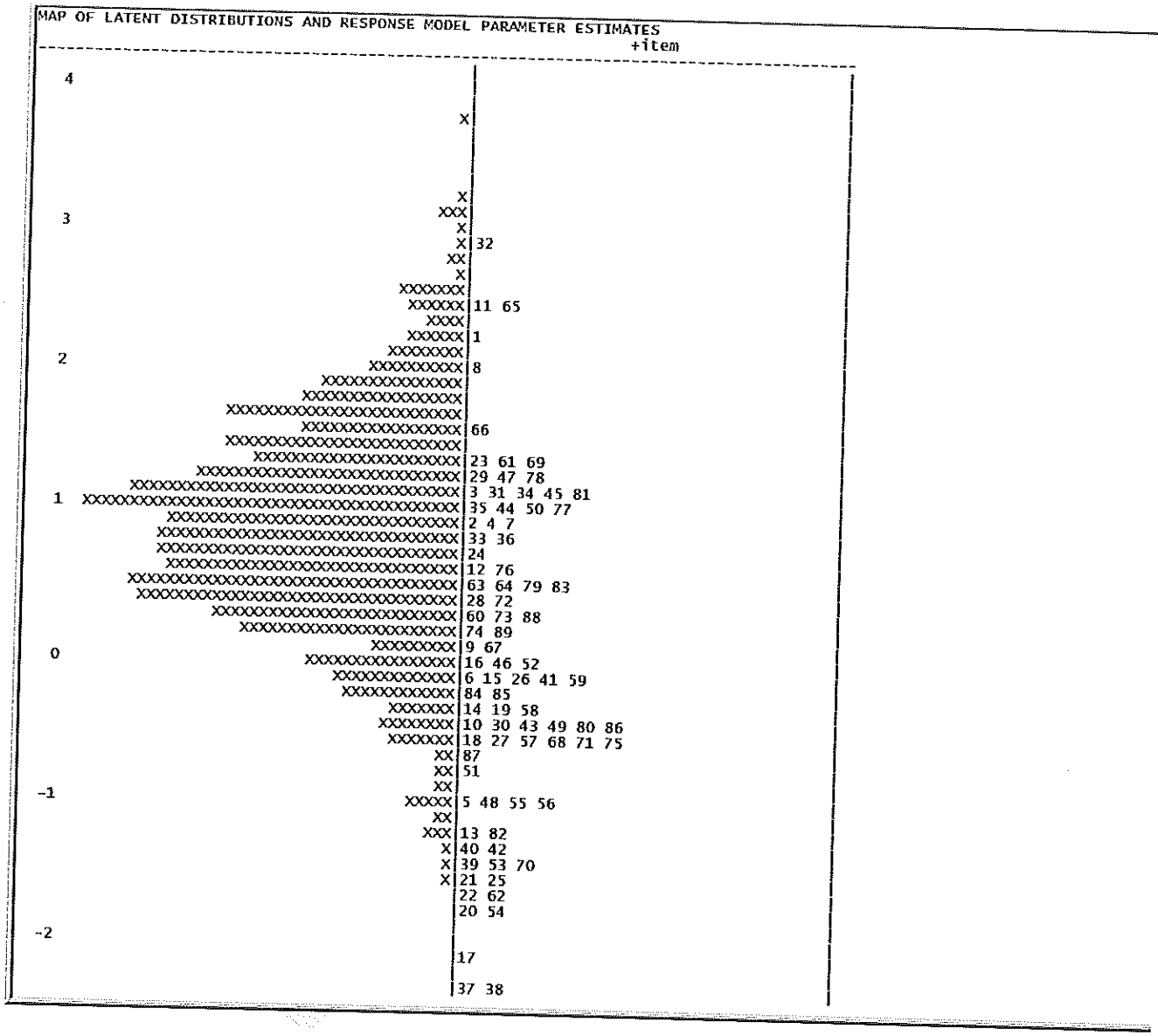
Appendix 10 – Item Map for S3 Numeracy



Appendix 11 – Item Map for S3 Reading



Appendix 12 – Item Map for S3 Writing



Redacted - 1 page - exemption applies

Redacted - 1 page - exemption applies

Redacted - 1 page - exemption applies

Redacted - 1 page - exemption applies

Redacted - 1 page - exemption applies

Redacted - 1 page - exemption applies

Redacted - 1 page - exemption applies

Redacted - 1 page - exemption applies

Redacted - 1 page - exemption applies

Redacted - 1 page - exemption applies

Redacted - 1 page - exemption applies