New multi-parameter item response models

April 16, 2008

Prepared by:
Murray Aitkin and Irit Aitkin
School of Behavioural Science
University of Melbourne

Prepared for:

US Department of Education
Office of Educational Research and Improvement
National Center for Education Statistics

This project was an activity of the NAEP Education Statistics Services Institute.

Contents

1 Aim of the project

The aim of this project was to assess the identifiability of the 3PL model for guessing on NAEP items, and to examine alternative models for guessing which are ability-based and are more readily identifiable on NAEP-scale data sets.

2 Summary

The study established that the 3PL model could be identified using Latent Gold for all items on the 30-item Knowledge and Skills subscale of the 1986 NAEP Math data (age 9/Grade 3), and that 15 of the 30 items had non-zero guessing parameters. These items corresponded closely with the 3PL items identified in the original analysis.

The 3PL model gave a substantial improvement in fit over the 2PL model. The reporting group estimates for the 3PL model were not directly comparable with those from the 2PL model because of the compression of the logit scale implied by the 3PL model, but were nearly proportional by a scale factor.

Polynomial models – quadratic and cubic generalizations of the linear logit model – were also examined. The quadratic model – the 3QL – was more easily fitted than the 3PL and had no identification difficulty. It was not equivalent to the 3PL model, but on the NAEP Math data gave a similar improvement to the 3PL model over the 2PL model. A further model generalization – the 3PQL model – including both guessing and quadratic regression could also be identified on the NAEP data, though it nearly reached the limit of identifiability.

A further class of models was considered, based on a two-component normal mixture distribution of ability. This has a different implication from the polynomial models or the 3PL model – that students in the two components may respond differently to the items. Most models in this class gave similar fit and results to the 3PL/3QL models, but those with different difficulty *and* discrimination parameters in each component gave a greatly improved fit to the data.

For the sub-class of *guessing models*, with zero discrimination parameters in one (the guessing) component, and a full 2PL model in the other component, the fit of the model was substantially better than that of the 3PL model. A further extension of the model allowed the modeling of the probability of being in the guessing component; this further increased the improvement in fit. Membership in the guessing component was positively related to black and Hispanic ethnicity, and to attendance at South-East regional schoools, and negatively related to white ethnicity and to attendance at high metropolitan and urban fringe schools.

The general two component mixture with different and non-zero discrimination parameters in each component improved substantially over the two-component guessing model. This model showed that a smaller subset of the population responded consistently differently from the larger subset, finding the items generally much harder. The additional modeling of the smaller component membership probability further increased the improvement in fit.

Combinations of categories with high probabilities of being in the smaller component were blacks, Hispanics and American Indian students in Southeast and West region schools in low metropolitan areas, and those with high probabilities of being in the larger component were white students in the Northeast and Central region schools in high metropolitan and urban fringe areas whose parents were college graduates. Girls had a marginally higher probability of being in the larger component.

The reporting group differences in this model were changed substantially from those in the 2PL and other models. They were generally substantially decreased relative to those in the 2PL model, by as much as 4 SEs for the Black-White and 3SEs for the Hispanic-White differences. *None* of the size and type of community estimates, or the parents' education estimates, were now significantly different from zero: these variables do *not* affect, or contribute to, the variation in item responses, but they *do* contribute to the identification of the latent component membership which identifies how difficult the students find the items.

Further investigation of factors at the student, school or class level which could identify this latent component structure would be valuable both educationally and politically; this will be carried out on the 2005 NAEP math data in a Secondary Analysis project.

For the comparison of complex models on sparse item data, the likelihood ratio test comparison of models may not follow its asymptotic χ^2 distribution, and a fully Bayesian model comparison approach is needed for reliable inferences.

3 The 2PL model

The 2PL model for the probability p_{ij} of student i with ability θ_i answering correctly ($y_{ij} = 1$) a binary item j is given by

$$\begin{aligned} p_{ij} = \Pr[y_{ij} = 1 \mid \theta_i] &= \exp(\phi_{ij})/[1 + \exp(\phi_{ij})] \\ \phi_{ij} &= a_j(\theta_i - b_j), \\ &= \alpha_j + \beta_j\theta_i, \\ \theta_i &\sim N(0, \sigma^2), \end{aligned}$$

where b_j is the difficulty of item j and a_j is its discrimination, and the alternative parameters α_j and β_j are given by $\beta_j = a_j$, $\alpha_j = -a_j b_j$. For

model identification one constraint is needed on either σ or one of the slope parameters a_j. A non-zero ability mean for θ is not identifiable: it is aliased or confounded with one of the intercept terms α_j. We give above the equivalent simple linear regression form of the model as well, as this facilitates the notation for the extensions we describe.

Maximum likelihood analysis in this model has been well-established since the EM algorithm approach of Bock and Aitkin (1981). Much large-scale analysis for binary response items is based on this model, and on the 3PL model for items on which guessing is expected.

Detection of departures from the 2PL model is an important issue in the use of the model; in particular, ambiguous or badly-written items may lead to a non-monotone item characteristic curve (ICC) which could affect the estimation of both regression model parameters and individual abilities. We give examples of both these effects in real data.

Various measures of item fit and item residuals have been used for this purpose, but as noted by van der Linden and Hambleton (1997 p.16), "Well-established statistical tests [for the 2PL and 3PL models] do not exist, and even if they did, questions about the utility of statistical tests in assessing model fit can be raised, especially with large samples."

We address these difficulties by extensions of the 2PL model. We consider first the 3PL model.

4 Extended models

4.1 The 3PL model

The 3PL model for the probability p_{ij} of student i with ability θ_i answering correctly item j is given by

$$\begin{aligned} p_{ij} &= c_j + (1 - c_j)\phi_{ij}, \\ \theta_i &\sim N(0, \sigma^2), \end{aligned}$$

where c_j is the *guessing* parameter for item j – the probability of answering item j correctly for students who are guessing independently of ability. This model is a form of *two-component mixture* which is more difficult to estimate by maximum likelihood as it has *two* kinds of latent structure.

4.2 The 3QL model

We consider a 3-parameter model which is the quadratic extension of the 2PL model, with a quadratic regression of logit response probability on ability. We call this model the 3QL – the 3-parameter Quadratic Logit model. The parallel 3QP probit model can be defined similarly, but we

restrict consideration to the logit version:

$$\begin{aligned} \Pr[y_{ij} = 1 \mid \theta_i] &= \exp(\psi_{ij})/[1 + \exp(\psi_{ij})] \\ \psi_{ij} &= \alpha_j + \beta_j\theta_i + \gamma_j\theta_i^2, \\ &= \phi_{ij} + \gamma_j\theta_i^2, \\ \theta_i &\sim N(0, \sigma^2). \end{aligned}$$

Here γ_j is the *curvature* parameter for item j. This 3QL model requires an identifiability constraint on one of the quadratic coefficients (conveniently $\gamma_1 = 0$) if σ is not constrained.

Polynomial logit item response models were considered briefly by McDonald (1989), who predicted identification difficulties with these models, by analogy with the well-known difficulties with the 3PL model. We comment on this point below.

The 3QL model can have a wide range of ICC forms, as is clear from the behaviour of the quadratic. With a small positive quadratic coefficient, the model is close to the 2PL model, and may be similar to the 3PL in its ICC shape. For a larger positive quadratic coefficient it may decrease and then increase, while for a large negative quadratic coefficient it may increase to a maximum and then decrease, suggesting an ambiguous item for high ability students.

Strong curvature would generally indicate a badly written item, since the intended design of good items is that the success probability increases monotonically with latent ability. Consequently the model may be regarded as providing a test for item consonance with the 2PL model. As such it provides much more information than merely a single global test statistic of item fit: each item can be assessed through its own estimated parameters, rather than from residuals from the 2PL model.

If the items do fit the 2PL model, this will be indicated by a non-significant difference in deviances ($-2 \log L_{max}$) between the 2PL and 3QL models, and this will provide a strong goodness of fit test for the 2PL model (as for the corresponding deviance comparison for the Rasch model compared with the 2PL model). In large samples the deviance difference will be distributed as χ^2_ν with ν the difference between the numbers of identifiable parameters in the two models.

4.2.1 Maximum likelihood model fitting

Maximum likelihood estimation in the 3QL model parallels closely that for the 2PL model described by Bock and Aitkin (1981). We first extend the 2PL model with explanatory variables $\mathbf{x}_i$:

$$\begin{aligned} \operatorname{logit} p_{ij} &= \alpha_j + \beta_j\theta_i + \boldsymbol{\beta}'\mathbf{x}_i \\ \theta_i &\sim N(0, \sigma^2); \end{aligned}$$

the likelihood for responses y_{ij} is

$$L(\boldsymbol{\beta}, \{\alpha_j, \beta_j\}) = \prod_i \int \prod_j [p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}]\phi(\theta_i/\sigma)/\sigma \mathrm{d}\theta_i,$$

where $\phi(x)$ is the normal density function. The maximization of the likelihood is achieved through numerical integration over the θ_i: the approximate log-likelihood is (Bock and Aitkin 1981, Aitkin 1999)

$$\log l \doteq \sum_i \sum_k \pi_k \prod_j [p_{ijk}^{y_{ij}}(1-p_{ijk})^{1-y_{ij}}]$$

where

$$\operatorname{logit} p_{ijk} = \alpha_j + \beta_j z_k + \boldsymbol{\beta}' \mathbf{x}_i$$

where z_k is the Gaussian quadrature masspoint with mass π_k.

For the 3QL model we have

$$\operatorname{logit} p_{ijk} = \alpha_j + \beta_j z_k + \gamma_j z_k^2 + \boldsymbol{\beta}' \mathbf{x}_i$$

with the same marginal distribution for θ_i.

Since z_k is *observed* in the discrete computation, *the Gaussian quadrature approximation procedure is unchanged by the quadratic extension*, apart from the additional estimation of the γ_j curvature parameters for each item. The quadratic extension requires only a very minor change in the M step of the EM algorithm, with an additional set of score equations for these parameters. The E step is unchanged, except for the inclusion of the estimated curvature parameters in the probability model and the maximized likelihood; the weight calculation needed in this step is unchanged.

This is an important point: as for the 2PL model, each 3QL item provides information about each student's ability, in contrast to the 3PL model in which guessed items do *not* provide information about student ability, and the fact of guessing on an item has to be assessed from the inconsistency of responses on this item relative to responses on other items on which there is *no* guessing – which has to be *assumed*.

The identifiability issue raised by McDonald (1989) is circumvented by using the EM algorithm: even if the model is underidentified, the algorithm will converge to a point on the (generally multidimensional) ridge in the parameter space on which the unidentifiable parameters are related. The resulting parameter estimates are not unique, since any other point on the ridge would give the same maximized likelihood; however the inversion of the information matrix reveals the singularities in the model and the unidentifiable parameters.

We give examples below; models were fitted using either Gllamm or Latent Gold.

4.3 The 4CL model

The quadratic regression can be extended to higher-order polynomials, to the extent that the parameters are identifiable. This is a simple matter of defining the appropriate ability power variable and extending the item regression on ability to include it. We extend the quadratic model with a cubic term, and call the resulting model the 4CL model.

4.4 The composite 3PQL model

This model allows *both* guessing *and* quadratic regression:

$$\begin{aligned} p_{ij} &= c_j + (1 - c_j)\exp(\psi_{ij})/[1 + \exp(\psi_{ij})], \\ \psi_{ij} &= \alpha_j + \beta_j\theta_i + \gamma_j\theta_i^2, \\ \theta_i &\sim N(0, \sigma^2). \end{aligned}$$

This model may look far too complicated to fit, but it is identifiable on the NAEP data example.

4.5 A generalized 4-parameter guessing model

$$\begin{aligned} p_{ij} &= c_j \cdot g_j + (1 - c_j)\exp(\phi_{ij})/[1 + \exp(\phi_{ij})], \\ \theta_i &\sim N(0, \sigma^2), \end{aligned}$$

where c_j is the proportion of students who guess on item j, and g_j is the probability of guessing item j correctly.

Comparison of this model with the 3PL model shows that the latter's single c_j parameter is split into two: the *proportion of guessers on item* j, c_j, and the *probability of a correct guess on item* j, g_j. It follows immediately that in the 3PL model, $g_j = 1$ for all j, that is, *those guessing on item* j *all guess correctly*! This is an unexpected and unreasonable feature of the 3PL model.

This consideration leads us to further extended models, in which we recognise explicitly *heterogeneity in the ability population.*

4.6 Mixture of logits model

This is a *population heterogeneity* model: the population is a *mixture* of two components. In the first component which contains a proportion δ of the population, the probability of a correct answer on item j is given by the 2PL model 1 with parameters α_{1j} and β_{1j}, while in the second component, containing the proportion $(1 - \delta)$ of the population, the probability of a correct answer on item j is given by the 2PL model 2 with parameters α_{1j} and β_{1j}. The probability of a correct answer on item j is then

$$p_{ij} = \delta \cdot \exp(\phi_{1ij})/[1 + \exp(\phi_{1ij})] + (1 - \delta) \cdot \exp(\phi_{2ij})/[1 + \exp(\phi_{2ij})]$$

where

$$\begin{aligned}\phi_{1ij} &= \alpha_{1j} + \beta_{1j}\theta_i, \\ \phi_{2ij} &= \alpha_{2j} + \beta_{2j}\theta_i,\end{aligned}$$

and the parameters in the two components are in general unrelated. It might be thought that the ability distribution in the two components can be correspondingly general, with different means and variances. However because of the same issue in the 2PL model, in this model the component variances are each aliased with one of the 2PL component discrimination parameters, and the component means are aliased with one of the 2PL component intercepts. So the ability distributions cannot be generalized to be different without constraining other model parameters. We therefore treat the ability distributions as identical in the reported analyses.

4.6.1 Maximum likelihood model fitting

The mixture of logits model can be fitted straightforwardly by nested EM algorithms, since the component logit models can themselves be fitted by EM, and any finite mixture can be fitted by an EM algorithm. We do not give details.

4.7 Mixture guessing model

This is a special case of the generalized 4-parameter guessing model above (and of the mixture of logits model), with c_j replaced by δ, constant over items. In this model, students in the first component guess at random, with probability g_j of guessing correctly on item j. In the second component, the probability of a correct answer is given by the 2PL model.

The probability of a correct answer by student i on item j is then

$$\begin{aligned}p_{ij} &= \delta \cdot g_j + (1-\delta) \cdot \exp(\phi_{ij})/[1+\exp(\phi_{ij})], \\ \phi_{ij} &= \alpha_j + \beta_j\theta_i.\end{aligned}$$

This model can be considered a representation of *engagement*: those in the first component are not *engaged* in the test. Since guessing by definition does not provide information about student ability, those unengaged students in the first component do not contribute to the estimation of their own abilities, and therefore to group differences in ability. These group differences are estimated only from those in the second component. There may therefore be considerable change in reporting group estimates in this model relative to the single-component 2PL model, if the proportion in the guessing component is at all appreciable.

To the extent that these models are identifiable, they provide a rich class of alternative models to the 3PL. With explanatory variables in the model

at the student level, these models can be made even richer by allowing the explanatory variables to affect *the probability of membership in the components*, as well as the probability of correct responses on the items. This is achieved by replacing the c_j parameter by

$$c_{ij} = \exp(\lambda' \mathbf{x}_i)/[1 + \exp(\lambda' \mathbf{x}_i)],$$

which has a logistic regression of the probability of component membership on the explanatory variables. This requires only a slight extension of the EM algorithm for the finite mixture of logit models. We do not give details.

We illustrate these models on a number of examples.

5 Examples: the LSAT6 and LSAT7 data sets

Bock and Aitkin illustrated the EM algorithm for the two-parameter probit (2PP) analysis with two small data sets of 1000 students from the Law School Aptitude Test. These scales each had 5 items; scale 6 was well fitted by the 2PP model but scale 7 was not, pointing to either multidimensionality of these items or a mis-specified item response model. Both data sets are given in Table 12.

5.1 LSAT6

For this data set, the deviances ($-2 \log L_{max}$) for the Rasch and 2PL models are 4933.87 and 4933.30, a deviance change of 0.57 on 4 df. For the 3QL model the deviance is 4930.70, a reduction of 2.60 on 4 df compared to the 2PL model. It is clear that there is no evidence of curvature over all items, and that the Rasch model is a very good fit. The G^2 goodness of fit index, comparing the model deviance with the deviance for the "saturated" multinomial model reproducing the observed data (Maydeu-Olivares and McArdle 2005 p. 90) is 22.04 on 25 df for the Rasch model, 21.47 on 21 df for the 2PL model, and 18.87 on 17 df for the 3QL model.

Parameter estimates for the Rasch, 2PL and 3QL models are shown in Table 1.

5.2 LSAT7

For this data set, the deviances for the Rasch and 2PL models are 5329.82 and 5317.54, a deviance change of 12.28 on 4 df, which has a χ^2_4 p-value of 0.0154. The 2PL is a better fit. However the goodness-of-fit G^2s for the Rasch and 2PL models are 44.24 on 25 df ($p = 0.0102$) and 31.94 on 21 df ($p = 0.0594$). The 2PL does not fit very well.

The 3PL model has a deviance of 5316.17, a reduction of only 1.37 on 4 df from the 2PL model. Only item 2 has a non-zero guessing parameter. The 3PL model is clearly not supported by the data.

Aitkin, Francis and Hinde (2005 p.535) noted that adding a common quadratic term to the 2PL decreased the deviance by only 0.12, so that this *common* term was not needed. For the full 3QL model the deviance is 5304.04, a reduction of 13.50 on 4 df ($p = 0.0091$) compared to the 2PL model. The G^2 for this model is 18.44 on 17 df, a good fit. Parameter estimates for the 2PL and 3QL models are shown in Table 2.

Fitting the 4CL model, we obtain a deviance of 5294.05, a reduction of 9.99 on 4 df ($p = 0.0406$) compared with the 3QL model. This is just significant at the 5% level. The goodness-of-fit G^2 is now 8.45 on 13 df ($p = 0.8131$); the high p-value suggests over-fitting.

The 3QL model is not the only generalization of the 2PL which could provide a model check: both multi-dimensional ability and latent class (mixture) models could achieve this. We discuss these in detail below for the NAEP data, but present the results here for comparison with the other models. A two-dimensional 2PL model for the LSAT7 example has a deviance of 5307.03, a reduction from the 2PL of 10.51 with 5 df, while a two-latent-class 2PL mixture model with different intercepts but the same slopes in each class has a deviance of 5300.93, a reduction of 16.60 on 5 df. These compare with the 3QL reduction of 13.50 on 4 df.

Parameter estimates for the 2PL mixture model are shown in Table 2. The second latent class is very small, with only 2.5% of the 1000 students; these people found the items (except for item 2) very much harder than those in the first latent class.

6 Model comparisons

The 2PL (solid curve), 3QL (dotted curve) and 4CL (dot-dashed curve) ICCs for each LSAT7 item are shown in Figures 11–15, over the range -3 to 3 for θ.

There are striking departures from linearity, with both positive and negative curvature. We interpret only the quadratic model; the cubic model appears to be overfitted and gives only a marginally significant improvement in deviance.

Items 1, 4 and 5 all show downward curvature; item 1 has a maximum response probability of 0.92 at $\theta = 0.61$, item 4 a maximum of 0.74 at $\theta = 0.53$ and item 5 a maximum of 0.9 at $\theta = 0.81$. Item 2 shows a very high response probability for low θ, falling rapidly to less than 0.3 at $\theta = -1.2$ and then increasing to 1. The cubic model does not change the lower-tail behaviour. Item 3 shows close agreement of the three models for large θ, but the quadratic model suggests a 3PL model with guessing parameter about 0.2, as the 3QL ICC flattens out near this value at $\theta = -3$.

How do we interpret these results? First, since the 3QL model is a much better fit than the 2PL, the inference about individual ability from the 2PL

needs to be checked. Figure 16 graphs the posterior mean (EAP) of the ability variable for each response pattern under the 3QL model against the corresponding posterior mean under the 2PL model.

The correlation is high (0.8155) but there are some notable discrepancies; because three items have negative curvature, high total scores do not necessarily indicate high ability.

Second, for the general case in which the model contains explanatory variables, the sensitivity of their estimates and standard errors should also be checked.

Finally, the items themselves may need to be re-written to achieve a monotone ICC which can be better represented by the simpler 2PL model (or the Rasch model) in future administrations of the items.

A further major issue is whether there are other models which also fit the data which do not have the non-monotone property of the 3QL. For the LSAT7 data the two-component mixture of logits model fits even better than the 3QL (with one extra parameter), and has a quite different interpretation: it is not the items which are badly written, but the student population which is heterogeneous.

7 The effect of mis-specification of the 3PL model

One major question in this study was whether the 3QL model could serve as a more easily computed alternative to the 3PL without its identifiability difficulties. To assess this we carried out a simulation study of the effect of mis-specification in the analysis of item data from a 3PL model. We generated 332 samples with 10 binary item responses for each of 1000 subjects, from five items with 2PL models and five with 3PL models. The item parameters are given in Table 3.

In addition to the items, the logit linear predictor had a regression model with four explanatory variables which we labelled ethnic group (4 levels), sex, poverty (2 levels) and homework (3 levels). The generating parameters for this model are given in Table 4; these were intended to be more extreme representations of NAEP data, to highlight estimation differences.

Results for one random sample from this model are also shown in Table 4. The incorrect 2PL model for all items had a deviance of 11,444.03. Fitting the 3QL model gave a deviance of 11,423.54, an improvement of 20.49 for the additional 10 parameters. Fitting the 3PL model gave a deviance of 11,258.40, an improvement of 165.14 compared to the 3QL, with the same number of parameters.

Regression parameter estimates for all three models are shown in Table 4 with the true values. All 2PL parameter estimates appear to be considerably biased downwards in magnitude, as would be expected from the range restriction in the 3PL model. This effect appears less severe for the 3QL

model parameters. The 3PL parameter estimates had smaller biases, in the opposite direction to the other two.

The estimated 3QL ICCs are shown in Figures 1–10 (dashed lines) over the range -3 to 3 for θ, with the true 2PL or 3PL ICCs (solid lines). The 3QL ICCs for the 3PL items are similar to those for the 3PL model for most items (6, 8, 9, 10), while item 4 has a marked negative curvature. For the 2PL items most 3QL ICCs (1, 2, 5, 7) are in general agreement with the 2PL ICCs, while item 3 shows marked negative curvature.

Of the 332 simulations, 21 failed to give an output file, apparently because of the instability of the current beta version of Latent Gold running under wine on the Linux cluster. In the 311 successful simulations we fitted the 2PL, 3QL and 3PL models, and two mixed 2PL models with normal mixture distributions for ability, the first with different means but common variance, the second with different means and variances. The deviance improvement on the 2PL model, averaged over the 311 samples, was 24.70 for the 3QL model, 193.54 for the 3PL model, 3.14 for the equal-variance mixture model and 4.34 for the unequal variance mixture model. (The deviance change for the 3QL model was in many samples significant at the 5% level of $\chi^2_{10}(18.3)$, indicating *some* failure of the 2PL model.) Estimation warnings occurred frequently for the 3PL model and occasionally for the 3QL model, indicating that convergence had not been achieved, because the number of EM and Gauss-Newton iterations specified was insufficient. For the 3PL model this resulted from the estimated guessing parameters approaching zero (five of them *were* zero); as these are estimated on the logit scale, they tried to approach $-\infty$, with minutely increasing likelihood. For the 3QL model this did not indicate any parameter estimation difficulty, only the need for more iterations. Iterating to convergence in both these models would have the effect of increasing their likelihoods slightly, and increasing their improvement over the 2PL model.

Parameter estimates are shown in Table 5. The bias of the 3PL estimates had the opposite sign to the biases of the other methods, reflecting the compression of the probability scale in the 3PL models generating the data. The 3PL estimates had the smallest (absolute) biases except for the largest ethnic2 parameter, where the 3QL estimate had the smallest bias.

The 3QL estimates had consistently smaller biases than the 2PL estimates. The biases of the 2PL mixture model estimates were very similar to those of the 2PL estimates, not surprising as these models barely improved on the likelihood for the 2PL model.

Surprisingly (since the 3PL model was correct), the 3PL estimates always had the largest across-sample variability (measured by sdb), followed by the 3QL estimates; the 2PL estimates had the smallest variability, and the mixture model sdbs were just slightly larger. As a consequence, the MSE of the 3QL model was generally the smallest, though it was equalled by the 3PL for sex and ethnic 4, and equalled or slightly bettered by the 2PL for

the small homework estimates.

The "se" column of Table 5 gives the average across samples of the estimated standard error given by the Latent Gold package. In several samples the mixture models failed to give an identifiable mixture and the information matrix in the regression variables, latent class means and variances was singular. This led to infinite standard errors for some parameter estimates, which are set to 1000 by the package, and resulted in a very large mean(se). This does not lead to any difficulty in actual analysis as it is immediately clear that the mixture model is unidentifiable and over-complex for the data. However the average of one or more values of 1000 with other values like 0.06 indicates this unidentifiability in a number of samples – the item responses *do* come from a mixture, but at the item level, not the population level.

For the 2PL, 3PL and 3QL which did not have this difficulty, the average se underestimated the actual variability across samples, slightly for the 2PL and 3PL, and more seriously for the 3QL.

8 NAEP data analysis

We applied these models to the analysis of a large NAEP survey. The data are the 30 items of the Numbers and Operations – Knowledge and Skills subscale of the 1986 NAEP Age 9/Grade 3 math test. Estimated item parameters for these items can be found in the ETS Technical Report for this survey and are reproduced in Table 3; some items were fitted in the original ETS analysis by the 3PL model, while most were fitted by the 2PL model. These parameters were estimated in a "null" model with no explanatory variables, and so need not correspond to the estimates we obtain below.

There were 10,463 children who attempted at least one item from this subscale. The sample design used a two-stage clustered and stratified sample of PSUs, schools within PSUs, and students within schools, with stratification by student and school ethnicity, and oversampling of lower frequency ethnic groups. A regression model was fitted to the item responses, additive on the logit scale to the 2PL model. The 20-parameter regression model included the main effects of the NAEP reporting group variables sex, ethnicity, region, size and type of community (STOC), and parents education (PARED).

In the analyses described below, all models included the ethnicity stratifying variable and the school random effect, so the different sampling fractions among ethnic groups and across school ethnicity do not require reweighting (Pfefferman 1993).

A four-level analysis of the sample data using Gllamm (Skrondal and Rabe-Hesketh 2004) showed that the PSU level had a very small variance component; the school variance component however was large. The detailed analysis reported below used a three-level model ignoring the slight PSU

clustering, and was carried out in Latent Gold 4.5.

In using Latent Gold, at the termination of iterations the models had sometimes not converged by the convergence criterion (10^{-5} on the successive log-likelihood differences). So some small increases in the maximized log-likelihoods may occur for the most complex models; this does not change the major conclusions from the analyses.

8.1 Models fitted

We adopt a consistent form for presentation for the models, since there is a large number of parameters in all models. The *reporting group* estimates are shown for all the polynomial models in Table 7, with standard errors and the maximized log-likelihood. These are the important model parameters; substantial variations in these estimates across models would show the need for a careful choice of model for reporting these estimates.

8.1.1 The 2PL model

This is the basic model against which we compare all others, though many of the items in the original analysis required the 3PL model (Table 6). The model has a maximized log-likelihood of $-39,930.05$.

8.1.2 The 3PL model

This model fitted the 3PL for all items. The notorious identification difficulties of this model led us to expect unidentifiability. However Latent Gold 4.5 *was* able to identify the model, with a maximized log-likelihood of $-39,848.49$, an improvement of 81.56 (an equivalent χ^2 of 163.12) for the additional 30 parameters: there is no question of the inadequacy of the 2PL model.

The guessing parameters are shown in Table 9. The Latent Gold analysis reported 13 unidentifiable parameters; we interpreted these as guessing parameters which were all approaching zero. We refitted the model with these c_j parameters fixed at 0, together with two other items with large logit values (≥ 5); values less than 5 were retained. The maximized log-likelihood for this model was unchanged by the 15 constrained guessing parameters, and converged much faster.

It is of interest that the original NAEP analysis identified 12 3PL items: 4, 6, 9, 10, 15, 19, 20, 24, 27, 28, 29, 30, whereas the full 3PL analysis here identified 15 3PL items, 10 of the original 12 plus five others: items 1, 2, 3, 7, and 11. Items 4 and 19 are not 3PL items in the current analysis.

For the reporting group estimates, those from the 3PL are almost all larger than those from the 2PL, as are their standard errors and the variance components. The reason is clear: the compressed logit scale for the 3PL items, and the large number of them, mean that effects on this scale must

be larger than on the full 2PL logit scale to reproduce the data fit. The two sets of parameter estimates are not directly proportional, but they are very similar in their relations within each set. The improved fit of the 3PL model does not appear to change much the relative differences in the reporting group categories.

8.1.3 The 3QL model

Latent Gold had no difficulty in estimating this model. The maximized log-likelihood was $-39,845.75$. This is slightly higher than for the 3PL model, although there are more parameters (29 curvature parameters versus 15 guessing parameters). The curvature parameters are shown in Table 9.

Only 7 of the curvature parameters exceed twice their standard errors, for items 2, 12, 13, 14, 21, 22 and 23. Item 2 was identified as a 3PL item in the previous analysis. All the curvatures are negative except for item 12: the negative curvature is the opposite of 3PL curvature.

The reporting group estimates and standard errors show the *opposite* effect from those for the 3PL: they are generally *smaller*, and again show similar proportionality to the 2PL estimates. However this cannot be due to the different variance component estimates, as these are very similar to those for the 2PL model.

8.1.4 The 3PQL model

This very large model (140 parameters) imposed the same constraints on zero guessing parameters as in the 3PL model. Surprisingly, it required fewer EM iterations than the 3PL model. The maximized log-likelihood was $-39,777.87$, an improvement of 67.88 on the 3QL model for the additional 15 parameters. Guessing and curvature parameters are shown in Table 9, and reporting group estimates in Table 7. The guessing parameters are very similar to those for the 3PL, and the curvature parameters to those for the 3QL, though their standard errors have increased so much that only four items (13, 21, 22, 23) now have estimates more than twice their standard errors. The reporting group estimates and standard errors are generally slightly smaller than those for the 3PL model, reflecting the "shrinkage" of the 3QL model over the 2PL model.

8.2 Mixed ability models

The following models all represent the ability distribution as a *two-component mixture of normals*, with different regressions (slopes and intercepts) of items on ability in the two components. They vary in the extent to which the reporting group parameters vary over the two components, and whether membership in the two components is itself modeled. We also examined three-component mixtures, but found so many unidentifiable parameters

that we concluded that the 1986 data cannot support this level of latent structure.

In the two-component mixtures we found that it was necessary to constrain the student ability variance parameter to be 1 in all models, and to free the discrimination parameters. It proved impossible to identify a second variance parameter for the two groups, and so all models fix the ability variance parameter to be 1 in both groups. This also gave much faster ML estimation.

We label the models by the parameters which are different in each component – the other parameters are understood to be the same. The code is 2x for 2-component mixture, and

- I – intercept
- D – discrimination
- R – reporting group
- C – component membership
- G – guessing model

We consider first guessing models, with the discrimination parameters set to zero in one component. Three models were examined, the first with common reporting group regressions and no modeling of component membership, the second with different reporting group regressions and no modeling of component membership, and the third with a common reporting group regression and modeling of component membership.

8.2.1 Common regression, component membership not modeled 2xG

This model has a maximized log-likelihood of $-39,777.88$ with 110 parameters, a large increase of 152.17 for the 31 extra parameters in the mixture and the guessing component. The proportion in the guessing component is estimated to be 0.219, with 95% confidence interval (0.184, 0.259). Reporting group parameter estimates are given in Table 8, and estimated guessing parameters, on both the logit and probability scale, are in Table 11. The reporting group estimates are little changed (less than 0.5 SEs) from those for the 2PL model. The guessing parameter estimates are much greater, for nearly all items, than the random guessing values $1/k$, with k the number of response categories. The among-school variance component, 0.143 (.016), is close to that for the 2PL model.

8.2.2 Component-specific regressions, component membership not modeled 2xRG

This model has a maximized log-likelihood of $-39,743.12$ with 130 parameters, an increase of 34.76 relative to the 2xG for the 20 extra parameters. The deviance change of 69.52 is highly significant relative to the asymptotic χ^2_{20} distribution. Parameter estimates are not shown for this model, for reasons given below.

8.2.3 Common regression, component membership modeled 2xCG

This model has a maximized log-likelihood of $-39,711.77$ with 130 parameters, an increase of 66.11 relative to the 2xG model for the 20 extra parameters. This is substantially larger (by 31.35) than the improvement of the 2xRG model, with the same number of parameters. Reporting group and logit probability model parameter estimates are given in Table 8. The reporting group estimates are little changed from those for the 2PL model. The largest change is less than 1 SE, and most are much smaller. The school variance component is very similar 0.132 (.015).

The component membership parameters are for the probability of being in the non-guessing component. Combinations of categories with high probabilities of being in the guessing component are Blacks and Hispanic students in Southeast region schools, and those with high probabilities of being in the non-guessing component are white students in high metropolitan and urban fringe schools.

8.2.4 Item difficulty and item discrimination only model 2xID

This model has a maximized log-likelihood of $-39,649.41$ with 142 parameters, an increase of 280.64 for the 60 additional parameters compared to the 2PL model. The reporting group estimates are given in Table 9 and the item parameter estimates are given in Table 12. The among-school variance component was 0.130 (.016), very similar to that for the 2PL model.

8.2.5 Reporting group model 2xIDR

This model has a maximized log-likelihood of $-39,623.38$ with 162 parameters, an increase of 26.03 for the 20 additional parameters compared to the 2xID model. The deviance change of 52.06 is significant compared with χ^2_{20}. Parameter estimates are not shown for this model, for reasons given below.

8.2.6 Component membership model 2xIDC

This model has a maximized log-likelihood of $-39,547.41$, also with 162 parameters. This is an increase of 102.00 for the 20 additional parameters

relative to the 2xID model. The item parameters are give in Table 15 and the reporting group and logit probability model estimates in Table 12. The among-school variance component was 0.121 (.015), very similar to that for the 2PL model. The reporting group parameter estimates are generally substantially decreased relative to those in the 2PL model, by as much as 4 SEs for the Black-White and 3SEs for the Hispanic-White differences. *None* of the size and type of community estimates, or the parents' education estimates, is now significantly different from zero. The modeling of membership in the two components has a substantial effect on the reporting group differences. The SE region is now 2 SEs above the reference NE region.

The discrimination parameters in Table 13 are all positive apart from three non-significant negative ones. It is difficult from the table to see the nature of the differences; the ICCs are shown for each component in Figures 17-46 (component 1 solid curve, component 2 dashed curve). It is very clear that the items were found much easier, and many less discriminating, in component 2 than in component 1. So component 1 (containing 45.6% of the population) appears a lower-achieving group than component 2 (55.4%).

Items 27 and 28 have non-significant negative discriminations in the first component – it is clear that this corresponds to random guessing in this component; the items are hard, and the more able group in the second component also found them hard.

Item 1 has a non-significant negative discrimination in the second component; the figure shows that in this component the success probability is essentially 1: the item is extremely easy for those in this component. The wording of the item questions is given in Table 15.

The nature of the second component membership can be seen from the logistic model estimates in Table 9. The parameter estimates in this model are log-oddds values for the probability of being in the second component. By exponentiating, they are converted to odds ratios for component 1 to component 2, relative to the odds ratio for the first category of the variable. These odds ratios for each variable are given in the last column of Table 9.

Combinations of categories with high probabilities of being in the first component are Blacks, Hispanics and American Indian students in Southeast and West region schools in low metropolitan areas, and those with high probabilities of being in the second component are white students in the Northeast and Central region schools in high metropolitan and urban fringe areas whose parents are college graduates. Girls have a marginally higher probability of being in the second component.

8.2.7 Reporting group and component membership model 2xIDRC

This model has a maximized log-likelihood of $-39,532.44$, with 182 parameters, an increase of 14.97 over the component membership model 2xIDC for the 20 additional parameters. The deviance change of 29.94 is not sig-

nificant for χ^2_{20}. Given that the component membership is modeled, the additional interaction modeling of separate reporting group effects in the two components is unnecessary. We do not further discuss this model, or the reporting group model 2xIDR. We conclude that the model 2xIDC provides a sufficient representation of the relationship between test outcomes, latent class membership and reporting group variables.

9 Multidimensional models

We noted in the section on the LSAT7 test above that a multidimensional ability model is another alternative to the 2PL model. We examined the two-dimensional ability model for the 30 scale items: though the items are intended and designed to be unidimensional, this may not be the case. The improvement found in fitting the mixed distribution model may be due to a different departure from the 2PL model.

In fitting the two-dimensional 2PL model, we constrained the ability variance to be 1 on both dimensions, and constrained the correlation between the factors to be zero. This does not imply any loss of generality, because the two-dimensional model, like all multi-factor models, is invariant (in terms of its maximized likelihood) under arbitrary rotations, and in particular under the orthogonal rotation which makes the rotated factors uncorrelated. Since the rotated ability factor loadings have no simple interpretation, we do not present them in the tables, but give just the reporting group estimates and standard errors in Table 8. They are very close to those for the 2PL model, with slightly greater standard errors reflecting the larger number of parameters estimated. The maximized log-likelihood for the two-dimensional model is $-39,685.69$ with 110 parameters, a very substantial improvement over the 2PL model, but a smaller improvement than that of the 2xID model ($-39,649.41$ with 142 parameters). These models are not nested, but the improvement of the 2xID model is 36.28 greater with its 32 extra parameters; this would be highly significant if the models were nested. The component modeling 2xIDC extension of the 2xID model gives a further improvement of 102.00 in maximized log-likelihood over the 2xID model, with 20 extra parameters.

10 Discussion

We began this study with the search for a more easily identified ability-based model for guessing as an alternative to the 3PL. However our experience with the Latent Gold implementation of the EM algorithm showed that the identification difficulties of this model are less than we had believed: even for the very sparse NAEP data we were able to identify not only the 3PL model, but also the more complex 3PQL model with both guessing and

quadratic regression. Our first attempt at an alternative model, the 3QL, is very easily fitted: as for the 2PL model, every item contributes to estimation of the student ability, as there is no mixture guessing structure.

These generalized item response models – 3QL, 3PQL – can represent a wide variety of ICCs. What is not clear is whether real items actually have non-monotone ICCs, or whether the fitted non-linearity is purely a feature of forcing the model on data which show only random improvement, as occurs with any unnecessarily complex model.

We rely at present on the asymptotic distribution of the likelihood ratio test to assess the real need for more complex item response functions, so it would be helpful to know how these asymptotic properties deteriorate with increasingly parametrized models, and to have Bayesian methods for model comparison which work better in heavily parametrized models.

The simulation study showed that when the true model is the 3PL, the 3QL fails to reproduce it (for the sets of parameter values considered), so the 3QL is not a *replacement for the 3PL*, but is a *diagnostic for the failure of the 2PL*: for several of the examples in the report the 2PL model failed to represent adequately the response functions. In the simulation study, the 3QL model, though incorrect, still gave the best regression parameter estimates (in terms of MSE), despite its biases. The 3PL model also showed biases, combined with less precision in the estimates: this model is inherently difficult, and its parameter estimates are not as well behaved as those of the 2PL model. This may be a consequence of its unnatural structure, as described above.

For the NAEP data, the 3PL and 3QL models gave similar fits and reporting group estimates and standard errors; the differences are interpretable as a scale change caused by the logit scale compression (for the 3PL and 3PQL models).

The mixture models *without* different difficulty *and* discrimination parameters were similar to the 3PL and 3QL models in fit and reporting group estimates. For the guessing models with zero discrimination parameters in one component, the fit of the model was substantially better than that of the 3PL model, with maximized log-likelihoods of $-39,848.49$ for the 3PL and $-39,777.88$, with almost the same number of parameters, for the 2xG guessing model without modeling of the guessing component probability. The additional modeling of this probability increased the maximized log-likelihood to $-39,711.77$ for the 2xGC model with 20 additional parameters. Membership in the guessing component was positively related to black and Hispanic ethnicity, and to attendance at South-East regional schoools, and negatively related to white ethnicity and to attendance at high metropolitan and urban fringe schools. So it appears that modeling guessing through the 3PL model, with guessing a function of the item only independently across items, gives a poorer representation of the test results than the guessing model with a subgroup of the student population who guess on all items.

However for the models with sets of *both* difficulties *and* discriminations in the two components, a *very large improvement* in fit occurs: the maximized log-likelihood for the 2xID model was $-39,649.40$, an increase of 128.48 over the 2xG model for the 30 additional discrimination parameters, and the modeling of the component membership probability increased this further for the 2xIDC model to $-39,547.41$, an even larger improvement of 164.36 over the 2xGC model.

If these maximized likelihood comparisons can be relied on to compare the models, the existence of a mixture of response types in the population is very clear, and the ICCs for the two components show a markedly greater difficulty of the items in one component (with 46% of the population), while in the other component (with 54%) the items are relatively easy, and much less discriminating.

The effect of modeling the component membership probabilities is to separate the effects of the reporting group variables into two categories: those affecting the latent component membership, and those affecting both the latent component membership *and* the correct response probability on the items.

The sex, region, size and type of community, and parents education variables *do not affect the probability of a correct response on the items* (though the Central region has a marginally lower level of correct responses than the other regions). Their effect is on the *latent group membership*: students in Southeast and West region schools in low metropolitan areas have higher probabilities of being in the first component, and students in the Northeast and Central region schools in high metropolitan and urban fringe areas whose parents are college graduates have higher probabilities of being in the second component. Girls have a marginally higher probability of being in the second component. The component grouping affects how difficult the students find the items: those in the first component find the items much more difficult than those in the second component.

The ethnic origin variables affect *both* latent component membership *and* the correct response probabilities on the items, and in the same way: all non-white ethnic groups have both higher probabilities than whites of being in the second component, and lower probabilities of a correct response on the items. The ethnic group differences on the item responses are smaller than for the 2PL model, but this is counter-weighted by the higher probability of being in the component which finds the items much harder.

The two-factor 2PL model did not nearly reach the fit of the mixed ability models: we conclude that heterogeneity of ability in the student population, rather than guessing or multidimensionality of the Knowledge and Skills subscale, provides the best representation of the data among the models we examined.

The implications of these findings deserve careful study. It would be intriguing to investigate whether the latent component can be identified by

manifest variables at the student, class and school level – is this an issue of the class teacher, syllabus, type of school attended by the student, or something else? We will be investigating the 2005 NAEP math data to see whether the models identified here appear in a similar form.

11 References

Aitkin, M. (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117-128.

Aitkin, M. and Aitkin, I. (2005) *Multi-level model analysis of the Knowledge and Skills scale of the NAEP 1986 math data (final report).* NCES report.

Aitkin, M., Francis, B. and Hinde, J. (2005) *Statistical Modelling in GLIM4.* Oxford University Press.

Bock, R.D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika* **46**, 443-459.

Garcia-Perez, M. (1999) Fitting logistic IRT models: small wonder. *The Spanish Journal of Psychology* **2**, 74-94.

Hutchinson, T.P. (1991) *Ability, partial information and guessing: statistical modelling applied to multiple-choice tests.* Rumsby Scientific Publishing, Rundle Mall, South Australia.

Lord, F.M. and Novick, M.R. (1968) *Statistical Theories of Mental Test Scores.* Addison-Wesley, Reading Mass.

McDonald, R.P. (1989) Future directions in IRT. *Int. J. Ed. Res.* **13**, 205-220.

Maydeu-Olivares, A. and McArdle, J.J. (2005) *Contemporary Psychometrics: A Festschrift for Roderick P. McDonald.* Lawrence Erlbaum Associates, Mahwah NJ.

Mislevy, R.J. (1985) Estimation of latent group effects. *Journal of the American Statistical Association* **80**, 993-997.

Pfefferman, D. (1993) The role of sampling weights when modelling survey data. *Int. Statist. Rev.*, **61**, 317-37.

San Martin, E., del Pino, G. and De Boeck, P. (2006) IRT models for ability-based guessing. *Applied Psychological Measurement* **30**, 183-203.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Relations Models.* Chapman and Hall/CRC, Boca Raton FL.

van der Linden, W. and Hambleton, R.K. (1997) *Handbook of Modern Item Response Theory.* Springer, New York.

12 Tables

Table 1: Parameter estimates, LSAT 6

	Rasch		2PL		3QL		
Item j	α_j	β_j	α_j	β_j	α_j	β_j	γ_j
1	2.730	0.755	2.773	0.826	3.334	0.505	-0.490
2	1.000	0.755	0.990	0.703	0.884	0.885	0.176
3	0.240	0.755	0.249	0.891	0.449	0.733	-0.260
4	1.306	0.755	1.285	0.699	1.158	0.883	0.214
5	2.099	0.755	2.053	0.657	1.832	1.269	0.564
dev	4933.87		4933.30		4930.70		

Table 2: Parameter estimates, LSAT 7

	2PL		3QL			2PLmix		
Item j	α_j	β_j	α_j	β_j	γ_j	α_{1j}	α_{2j}	β_j
1	1.856	0.988	2.305	0.643	-0.524	1.899	-3.008	0.750
2	0.808	1.081	0.693	2.597	1.000	0.915	1.236	1.506
3	1.805	1.708	1.647	1.868	0.296	1.798	-1.241	1.518
4	0.486	0.765	0.950	0.580	-0.542	0.521	-1.724	0.637
5	1.855	0.736	2.106	0.488	-0.301	1.927	-5.542	0.453
dev	5317.54		5304.04			5300.93		

Table 3: Item parameters, simulation

j	α_j	β_j	c_j
1	-0.217	1.738	0
2	-0.359	1.202	0
3	0.540	0.841	0
4	0.789	1.090	0.238
5	0.735	0.855	0
6	0.339	1.150	0.208
7	0.985	1.162	0
8	0.858	0.894	0.280
9	-1.643	0.898	0.352
10	-1.159	0.620	0.225

Table 4: Regression parameter estimates, simulation

Variable	True	2PL	3QL	3PL	(SE)
Sex 2	-0.472	-0.426	-0.406	-0.577	0.093
Ethnic 2	-2.359	-1.600	-1.733	-2.643	0.168
Ethnic 3	-1.877	-1.274	-1.432	-2.166	0.158
Ethnic 4	0.944	0.458	0.480	0.773	0.201
Poverty	-0.800	-0.522	-0.639	-0.931	0.143
Homework 2	0.100	-0.017	-0.025	0.036	0.107
Homework 3	0.300	0.158	0.187	0.281	0.130

Table 5: Biases and MSEs, simulated data

method	parameter	true	mean	bias	MSE	sdb	se
2PL	Sex	-0.472	-0.324	0.148	0.026	0.067	0.067
3QL			-0.343	0.128	0.022	0.076	0.067
3PL			-0.571	-0.099	0.022	0.108	0.106
2xE			-0.323	0.149	0.027	0.068	9.906
2xU			-0.323	0.149	0.027	0.068	6.630
2PL	Ethnic2	-2.359	-1.527	0.831	0.705	0.118	0.113
3QL			-1.640	0.719	0.555	0.196	0.117
3PL			-3.123	-0.764	0.667	0.290	0.278
2xE			-1.513	0.846	0.730	0.124	3.396
2xU			-1.513	0.846	0.731	0.126	9.950
2PL	Ethnic 3	-1.887	-1.257	0.630	0.410	0.115	0.109
3QL			-1.342	0.545	0.324	0.164	0.112
3PL			-2.409	-0.522	0.328	0.236	0.218
2xE			-1.243	0.644	0.428	0.115	3.391
2xU			-1.244	0.643	0.427	0.117	13.223
2PL	Ethnic 4	0.944	0.716	-0.228	0.079	0.165	0.152
3QL			0.750	-0.193	0.073	0.188	0.153
3PL			1.090	0.146	0.073	0.226	0.220
2xE			0.718	-0.226	0.079	0.168	6.711
2xU			0.719	-0.225	0.079	0.169	13.266
2PL	Poverty	-0.800	-0.518	0.282	0.090	0.103	0.101
3QL			-0.556	0.244	0.074	0.120	0.100
3PL			-1.026	-0.226	0.081	0.172	0.173
2xE			-0.515	0.285	0.092	0.103	6.658
2xU			-0.515	0.285	0.092	0.104	9.936
2PL	Homework2	0.100	0.068	-0.032	0.007	0.077	0.078
3QL			0.074	-0.026	0.008	0.084	0.077
3PL			0.119	0.019	0.015	0.119	0.120
2xE			0.067	-0.033	0.007	0.078	16.471
2xU			0.067	-0.033	0.007	0.078	13.191
2PL	Homework3	0.300	0.205	-0.095	0.021	0.108	0.098
3QL			0.218	-0.082	0.021	0.120	0.098
3PL			0.362	0.062	0.029	0.159	0.150
2xE			0.204	-0.096	0.021	0.110	19.768
2xU			0.204	-0.096	0.021	0.110	13.212

Table 6: Item parameters (NAEP)

item	a	SE	b	SE	c	SE
1	0.503	0.019	-3.780	0.143	0	0
2	0.769	0.017	-2.066	0.049	0	0
3	0.841	0.018	-0.642	0.019	0	0
4	1.090	0.045	-0.724	0.044	0.238	0.015
5	0.855	0.023	-0.860	0.032	0	0
6	1.150	0.065	-0.295	0.040	0.208	0.013
7	1.162	0.022	-0.848	0.024	0	0
8	1.738	0.125	0.125	0.018	0	0
9	0.894	0.032	-0.960	0.044	0.280	0.014
10	0.898	0.047	-0.716	0.050	0.352	0.015
11	0.886	0.020	-0.900	0.028	0	0
12	1.288	0.021	-1.101	0.025	0	0
13	1.300	0.025	-0.445	0.017	0	0
14	1.234	0.023	-0.554	0.018	0	0
15	0.620	0.037	-0.256	0.032	0.225	0.013
16	0.942	0.022	-1.273	0.039	0	0
17	1.202	0.059	0.299	0.034	0	0
18	0.865	0.023	-0.047	0.014	0	0
19	1.058	0.038	-1.152	0.053	0.198	0.020
20	1.101	0.053	-0.817	0.055	0.257	0.018
21	0.899	0.014	-1.871	0.034	0	0
22	0.893	0.014	-1.839	0.033	0	0
23	1.017	0.016	-1.042	0.021	0	0
24	1.185	0.027	-1.074	0.034	0.232	0.012
25	1.096	0.025	-0.376	0.020	0	0
26	0.998	0.024	-0.484	0.021	0	0
27	1.766	0.296	1.115	0.248	0.197	0.006
28	1.149	0.034	0.365	0.021	0.164	0.006
29	0.955	0.044	-0.544	0.040	0.247	0.013
30	0.974	0.051	-0.454	0.042	0.243	0.013

Table 7

Reporting group estimates and SEs - guessing/quadratic models

	2PL	3PL	3QL	3PQL
male	0			
femal	.012 (.028)	.022 (.034)	.007 (.027)	.003 (.033)
white	0			
black	-.667 (.047)	-.837 (.059)	-.651 (.046)	-.812 (.059)
hispa	-.460 (.043)	-.609 (.053)	-.427 (.043)	-.562 (.053)
as/pa	-.203 (.117)	-.223 (.139)	-.161 (.114)	-.162 (.136)
amind	-.471 (.093)	-.602 (.119)	-.419 (.089)	-.542 (.113)
other	-.200 (.752)	-.035 (.820)	.075 (.661)	.072 (.821)
NE	0			
SE	-.020 (.077)	-.073 (.095)	.022 (.079)	-.022 (.093)
Cent	-.172 (.074)	-.211 (.092)	-.166 (.075)	-.196 (.089)
West	-.182 (.069)	-.217 (.085)	-.146 (.076)	-.183 (.084)
extru	0			
lomet	-.201 (.113)	-.348 (.138)	-.156 (.141)	-.343 (.140)
himet	.497 (.116)	.586 (.145)	.423 (.110)	.524 (.135)
manct	.150 (.106)	.162 (.135)	.144 (.111)	.143 (.136)
urbfr	.158 (.112)	.171 (.151)	.142 (.113)	.140 (.152)
medct	.092 (.097)	.115 (.122)	.074 (.096)	.092 (.121)
smplc	- .019 (.095)	-.029 (.118)	-.029 (.097)	-.030 (.119)
nfnhs	0			
finhs	-.179 (.206)	-.186 (.274)	-.247 (.188)	-.333 (.223)
smcol	.045 (.200)	.057 (.267)	-.064 (.181)	-.127 (.216)
colgr	.398 (.205)	.524 (.273)	.271 (.187)	.308 (.223)
DK	.382 (.198)	.489 (.266)	.265 (.179)	.287 (.214)
nores	.027 (.197)	.057 (.265)	-.101 (.178)	-.151 (.213)
s^2_sch	.139 (.017)	.233 (.028)	.137 (.018)	.224 (.026)
s^2	1.682 (.365)	2.225 (.608)	1.654 (.354)	2.243 (.650)
log Lmax	-39,930.05	-39,848.49	-39,845.75	-39,777.87

Table 8

Reporting group estimates and SEs - 2PL, guessing and two-factor models

	2PL	2xG	2xGC	2xGC(mem.ship)	two-factor
intercept				-1.087 (.540)	
male	0	0	0		
femal	.012 (.028)	.038 (.028)	.022 (.028)	.086 (.104)	.033 (.030)
white	0				
black	-.667 (.047)	-.674 (.046)	-.617 (.048)	-.856 (.168)	-.692 (.049)
hispa	-.460 (.043)	-.453 (.043)	-.415 (.044)	-.593 (.150)	-.470 (.046)
as/pa	-.203 (.117)	-.123 (.118)	-.069 (.119)	-.622 (.407)	-.197 (.126)
amind	-.471 (.093)	-.519 (.094)	-.480 (.095)	-.698 (.386)	-.489 (.100)
other	-.200 (.752)	-.133 (.633)	.015 (.596)	-.268 (2.42)	-.287 (.809)
NE	0				
SE	-.020 (.077)	.055 (.077)	.116 (.079)	-.349 (.163)	-.025 (.078)
Cent	-.172 (.074)	-.146 (.082)	-.165 (.079)	.096 (.160)	-.191 (.076)
West	-.182 (.069)	-.098 (.078)	-.180 (.079)	-.099 (.144)	-.196 (.070)
extru	0				
lomet	-.201 (.113)	-.230 (.132)	-.191 (.116)	-.542 (.300)	-.221 (.118)
himet	.497 (.116)	.428 (.122)	.384 (.118)	.811 (.248)	.501 (.118)
manct	.150 (.106)	.153 (.113)	.152 (.101)	.110 (.251)	.143 (.112)
urbfr	.158 (.112)	.168 (.128)	.134 (.115)	.569 (.247)	.156 (.115)
medct	.092 (.097)	.078 (.107)	.075 (.097)	.367 (.231)	.078 (.104)
smplc	- .019 (.095)	-.007 (.106)	-.016 (.095)	.243 (.230)	-.034 (.101)
nfnhs	0				
finhs	-.179 (.206)	-.177 (.184)	-.235 (.184)	-.318 (.574)	-.185 (.203)
smcol	.045 (.200)	.068 (.176)	-.009 (.176)	-.069 (.514)	.054 (.195)
colgr	.398 (.205)	.376 (.182)	.259 (.182)	.497 (.528)	.421 (.202)
DK	.382 (.198)	.396 (.174)	.298 (.173)	.348 (.500)	.403 (.193)
nores	.027 (.197)	.020 (.173)	-.074 (.172)	.127 (.500)	.033 (.192)
s^2_sch	.139 (.017)	.143 (.016)	.132 (.015)		.137 (.018)
s^2	1.682 (.365)	1.0 -	1.0 -		1.0
					1.0
log Lmax	-39,930.05	-39,777.88	-39,711.77		-39,685.69

Table 9

Reporting group estimates and SEs - 2PL and mixed models

	2PL	2xID	2xIDC	2xIDC(mem.ship)	odds ratio
male	0	0	0		
femal	.012 (.028)	.034 (.029)	.008 (.035)	.168 (.089)	1.18
white	0				
black	-.667 (.047)	-.675 (.048)	-.458 (.056)	-.935 (.122)	0.39
hispa	-.460 (.043)	-.451 (.044)	-.313 (.051)	-.610 (.117)	0.54
as/pa	-.203 (.117)	-.149 (.122)	-.073 (.139)	-.300 (.312)	0.74
amind	-.471 (.093)	-.480 (.097)	-.307 (.111)	-.720 (.274)	0.49
other	-.200 (.752)	-.256 (.763)	-.186 (.937)	-.016 (1.82)	0.98
NE	0				
SE	-.020 (.077)	.029 (.074)	.116 (.079)	-.395 (.138)	0.67
Cent	-.172 (.074)	-.177 (.076)	-.165 (.079)	-.054 (.146)	0.95
West	-.182 (.069)	-.135 (.068)	-.078 (.073)	-.299 (.129)	0.74
extru	0				
lomet	-.201 (.113)	-.227 (.112)	-.112 (.119)	-.493 (.217)	0.61
himet	.497 (.116)	.361 (.122)	.222 (.126)	.742 (.208)	2.10
manct	.150 (.106)	.119 (.105)	.091 (.114)	.127 (.203)	1.14
urbfr	.158 (.112)	.109 (.110)	-.016 (.119)	.554 (.216)	1.74
medct	.092 (.097)	.012 (.100)	-.041 (.108)	.291 (.192)	1.34
smplc	- .019 (.095)	-.078 (.098)	-.100 (.105)	.134 (.185)	1.14
nfnhs	0				
finhs	-.179 (.206)	-.244 (.191)	-.315 (.214)	.202 (.430)	1.22
smcol	.045 (.200)	-.031 (.184)	-.093 (.203)	.265 (.394)	1.30
colgr	.398 (.205)	.347 (.190)	.163 (.210)	.764 (.413)	2.15
DK	.382 (.198)	.315 (.181)	.181 (.200)	.644 (.383)	1.90
nores	.027 (.197)	-.060 (.181)	-.178 (.199)	.543 (.382)	1.72
s^2_sch	.139 (.017)	.130 (.016)	.121 (.015)		
s^2	1.682 (.365)	1.0 -	1.0 -		
log Lmax	-39,930.05	-39,649.40	-39,547.41		

Table 10

	Guessing parameters				Curvatures		3PQL			
	3PL		constrained 3PL		3QL		guessing		curvature	
item	c*	SE	c*	SE	gamma	SE	c*	SE	gamma	SE
1	0.150	(0.898)	0.150	(0.898)	0	-	-0.019	(0.805)	0	-
2	3.057	(3.369)	3.058	(3.370)	-0.257	(0.102)	2.240	(1.734)	-0.198	(0.113)
3	3.096	(0.817)	3.096	(0.817)	-0.184	(0.110)	2.988	(0.757)	-0.158	(0.155)
4	30.005	(1000)	100	-	-0.070	(0.082)	100	-	-0.153	(0.083)
5	14.118	(478)	100	-	-0.108	(0.076)	100	-	-0.155	(0.078)
6	3.283	(0.625)	3.283	(0.625)	0.174	(0.088)	3.416	(0.765)	0.065	(0.102)
7	2.880	(0.761)	2.880	(0.761)	-0.028	(0.103)	3.020	(0.924)	0.027	(0.113)
8	67.174	(1000)	100	-	-0.066	(0.283)	100	-	0.095	(0.236)
9	0.563	(0.159)	0.563	(0.159)	-0.048	(0.073)	0.522	(0.159)	0.018	(0.166)
10	2.221	(0.495)	2.221	(0.495)	-0.130	(0.071)	2.396	(0.609)	-0.109	(0.078)
11	1.464	(0.214)	1.464	(0.214)	-0.179	(0.090)	1.463	(0.218)	-0.151	(0.134)
12	43.718	(1000)	100	-	0.469	(0.177)	100	-	0.300	(0.266)
13	242.982	(1000)	100	-	-1.047	(0.174)	100	-	-0.655	(0.276)
14	112.153	(1000)	100	-	-1.126	(0.196)	100	-	-0.624	(0.410)
15	1.482	(0.210)	1.482	(0.210)	-0.068	(0.066)	1.472	(0.217)	-0.235	(0.167)
16	21.735	(1000)	100	-	-0.010	(0.082)	100	-	-0.030	(0.094)
17	119.377	(1000)	100	-	0.098	(0.114)	100	-	0.001	(0.141)
18	52.183	(1000)	100	-	0.090	(0.090)	100	-	0.072	(0.101)
19	19.732	(1000)	100	-	0.033	(0.164)	100	-	0.012	(0.162)
20	2.731	(1.176)	2.731	(1.176)	0.167	(0.152)	2.761	(1.266)	0.064	(0.182)
21	49.881	(1000)	100	-	-1.006	(0.105)	100	-	-1.049	(0.099)
22	69.430	(1000)	100	-	-1.149	(0.133)	100	-	-1.175	(0.129)
23	59.832	(1000)	100	-	-0.490	(0.095)	100	-	-0.490	(0.090)
24	4.303	(5.002)	4.303	(5.001)	0.155	(0.095)	3.099	(1.436)	0.161	(0.104)
25	124.321	(1000)	100	-	-0.008	(0.077)	100	-	-0.026	(0.075)
26	17.078	(1000)	100	-	0.109	(0.100)	100	-	0.123	(0.101)
27	2.115	(0.155)	2.115	(0.155)	0.096	(0.058)	2.195	(0.177)	0.010	(0.154)
28	1.924	(0.181)	1.924	(0.181)	0.130	(0.100)	1.933	(0.196)	-0.058	(0.219)
29	1.062	(0.221)	1.062	(0.221)	0.128	(0.084)	1.174	(0.255)	0.096	(0.122)
30	1.013	(0.165)	1.013	(0.165)	0.174	(0.104)	1.058	(0.176)	0.170	(0.212)

c*_j = logit(1 - c_j)

Table 11

Guessing parameters and SEs, 2xG model

item	2xG	p
1	22.14 (1000)	1.000
2	2.078 (0.384)	0.889
3	-0.291 (0.341)	0.428
4	0.741 (0.289)	0.677
5	0.763 (0.305)	0.682
6	0.096 (0.283)	0.524
7	-0.459 (0.312)	0.387
8	-4.675 (1.050)	0.009
9	0.688 (0.246)	0.666
10	0.403 (0.244)	0.599
11	0.238 (0.243)	0.559
12	39.00 (1000)	1.000
13	1.151 (0.268)	0.760
14	1.543 (0.293)	0.824
15	-0.203 (0.242)	0.449
16	0.603 (0.257)	0.646
17	-2.625 (0.414)	0.068
18	-0.849 (0.263)	0.300
19	1.617 (0.379)	0.834
20	0.773 (0.349)	0.684
21	0.939 (0.287)	0.719
22	0.940 (0.287)	0.719
23	0.608 (0.274)	0.648
24	2.014 (0.380)	0.882
25	-0.385 (0.257)	0.405
26	-0.113 (0.272)	0.472
27	-2.180 (0.352)	0.102
28	-1.302 (0.370)	0.214
29	0.465 (0.300)	0.614
30	0.078 (0.327)	0.520

% c*_j = logit(1 - c_j)

Table 12

Intercept and slope parameters for the 2xID mixture model

item	comp1 int (SE)	comp2 int (SE)	comp1 slope (SE)	comp2 slope (SE)
1	2.740 (0.296)	4.563 (0.538)	1.128 (0.214)	-0.674 (0.474)
2	0.815 (0.237)	2.748 (0.256)	0.960 (0.157)	0.099 (0.197)
3	-1.611 (0.264)	0.641 (0.228)	0.815 (0.226)	0.407 (0.131)
4	-0.107 (0.230)	0.627 (0.225)	1.048 (0.173)	0.968 (0.153)
5	-0.329 (0.234)	0.439 (0.220)	1.032 (0.184)	0.658 (0.126)
6	-0.982 (0.233)	0.041 (0.228)	0.567 (0.160)	1.066 (0.161)
7	-0.819 (0.238)	0.845 (0.230)	0.594 (0.166)	0.751 (0.159)
8	-4.278 (0.516)	-2.548 (0.312)	0.211 (0.596)	1.081 (0.248)
9	0.468 (0.219)	1.272 (0.221)	0.861 (0.129)	0.892 (0.128)
10	-0.189 (0.220)	0.898 (0.216)	0.856 (0.126)	0.652 (0.105)
11	-0.429 (0.221)	0.721 (0.217)	1.017 (0.148)	0.771 (0.113)
12	-1.173 (0.249)	2.713 (0.251)	0.641 (0.164)	0.627 (0.215)
13	-7.750 (2.652)	0.749 (0.223)	2.898 (1.247)	0.758 (0.138)
14	-3.189 (0.377)	1.012 (0.224)	0.834 (0.346)	0.776 (0.143)
15	-0.296 (0.219)	0.140 (0.214)	0.773 (0.123)	0.463 (0.094)
16	0.256 (0.229)	1.386 (0.235)	1.460 (0.217)	1.135 (0.164)
17	-3.494 (0.442)	-1.779 (0.249)	2.102 (0.363)	1.263 (0.181)
18	-1.382 (0.238)	-0.489 (0.225)	1.047 (0.173)	1.085 (0.146)
19	0.105 (0.263)	1.694 (0.252)	1.316 (0.354)	0.765 (0.230)
20	-0.322 (0.268)	1.266 (0.254)	0.895 (0.256)	0.907 (0.227)
21	2.673 (0.377)	2.447 (0.290)	3.525 (0.409)	1.917 (0.276)
22	2.926 (0.446)	2.429 (0.280)	4.399 (0.899)	1.951 (0.241)
23	-0.400 (0.243)	1.139 (0.233)	1.688 (0.330)	1.140 (0.176)
24	0.238 (0.230)	1.949 (0.275)	0.746 (0.147)	1.110 (0.205)
25	1.173 (0.265)	-0.058 (0.223)	0.979 (0.259)	0.691 (0.139)
26	-1.228 (0.266)	0.164 (0.232)	1.007 (0.286)	0.835 (0.158)
27	-1.710 (0.257)	-1.585 (0.242)	-0.267 (0.156)	0.387 (0.160)
28	-1.161 (0.275)	-0.941 (0.254)	-0.117 (0.231)	0.701 (0.211)
29	0.136 (0.238)	0.815 (0.226)	0.351 (0.168)	0.555 (0.154)
30	-0.348 (0.246)	0.589 (0.235)	0.406 (0.206)	0.820 (0.193)

Table 13

Intercept and slope parameters for the 2xIDC mixture model

item	comp1 int (SE)	comp2 int (SE)	comp1 slope (SE)	comp2 slope (SE)
1	3.250 (0.345)	3.931 (0.364)	1.446 (0.236)	-0.191 (0.368)
2	1.328 (0.261)	2.426 (0.255)	1.482 (0.187)	0.361 (0.202)
3	-0.964 (0.278)	0.279 (0.237)	1.662 (0.250)	0.633 (0.143)
4	-0.612 (0.243)	1.251 (0.245)	0.405 (0.115)	0.811 (0.183)
5	-0.778 (0.253)	0.921 (0.237)	0.597 (0.134)	0.375 (0.135)
6	-1.161 (0.252)	0.349 (0.240)	0.316 (0.149)	0.985 (0.165)
7	-0.187 (0.248)	0.482 (0.242)	1.231 (0.176)	1.231 (0.176)
8	-3.672 (0.443)	-3.292 (0.436)	1.368 (0.342)	1.670 (0.337)
9	0.489 (0.232)	1.376 (0.236)	0.827 (0.123)	0.856 (0.132)
10	-0.114 (0.231)	0.999 (0.231)	0.810 (0.119)	0.678 (0.109)
11	-0.319 (0.233)	0.792 (0.232)	1.039 (0.152)	0.799 (0.119)
12	-0.944 (0.252)	2.877 (0.263)	0.683 (0.161)	0.505 (0.255)
13	-9.012 (7.070)	0.954 (0.237)	3.185 (2.659)	0.691 (0.153)
14	-3.168 (0.483)	1.167 (0.238)	1.059 (0.458)	0.627 (0.140)
15	-0.287 (0.232)	0.254 (0.229)	0.765 (0.121)	0.404 (0.092)
16	0.310 (0.242)	1.504 (0.250)	1.418 (0.217)	1.134 (0.172)
17	-3.111 (0.399)	-1.618 (0.261)	1.668 (0.332)	1.274 (0.188)
18	-1.373 (0.251)	-0.311 (0.239)	0.930 (0.174)	1.083 (0.153)
19	-0.024 (0.276)	1.917 (0.265)	1.554 (0.401)	0.605 (0.219)
20	-0.217 (0.264)	1.369 (0.268)	0.889 (0.221)	0.972 (0.236)
21	2.456 (0.332)	2.615 (0.313)	3.454 (0.375)	1.910 (0.254)
22	2.424 (0.332)	2.603 (0.305)	3.705 (0.513)	1.890 (0.234)
23	-0.363 (0.252)	1.255 (0.245)	1.787 (0.290)	1.031 (0.170)
24	0.197 (0.240)	2.276 (0.296)	0.705 (0.120)	1.053 (0.225)
25	-1.220 (0.265)	0.116 (0.237)	0.971 (0.197)	0.617 (0.123)
26	-1.203 (0.259)	0.319 (0.245)	0.944 (0.207)	0.757 (0.143)
27	-1.604 (0.260)	-1.567 (0.258)	-0.282 (0.142)	0.531 (0.176)
28	-1.147 (0.268)	-0.810 (0.255)	-0.122 (0.200)	0.675 (0.192)
29	0.283 (0.240)	0.817 (0.240)	0.284 (0.135)	0.591 (0.144)
30	-0.208 (0.248)	0.622 (0.243)	0.477 (0.165)	0.821 (0.181)

13 Figures

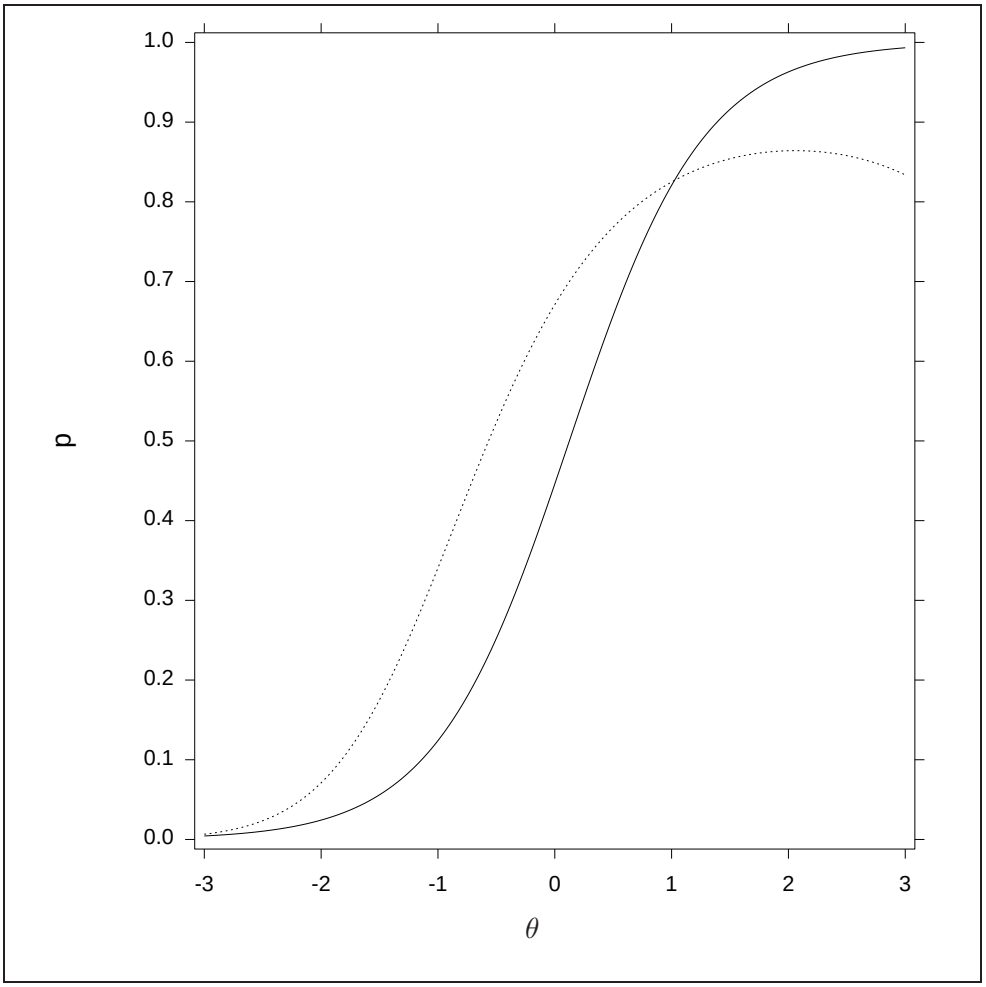

Figure 1: Item 1 ICCs, simulated data

Table 14: Data sets, LSAT 6 and 7

Item	1	2	3	4	5	LSAT6	LSAT7
	0	0	0	0	0	3	12
	0	0	0	0	1	6	19
	0	0	0	1	0	2	1
	0	0	0	1	1	11	7
	0	0	1	0	0	1	3
	0	0	1	0	1	1	19
	0	0	1	1	0	3	3
	0	0	1	1	1	4	17
	0	1	0	0	0	1	10
	0	1	0	0	1	8	5
	0	1	0	1	0	0	3
	0	1	0	1	1	16	7
	0	1	1	0	0	0	7
	0	1	1	0	1	3	23
	0	1	1	1	0	2	8
	0	1	1	1	1	15	28
	1	0	0	0	0	10	7
	1	0	0	0	1	29	39
	1	0	0	1	0	14	11
	1	0	0	1	1	81	34
	1	0	1	0	0	3	14
	1	0	1	0	1	28	51
	1	0	1	1	0	15	15
	1	0	1	1	1	80	90
	1	1	0	0	0	16	6
	1	1	0	0	1	56	25
	1	1	0	1	0	21	7
	1	1	0	1	1	173	35
	1	1	1	0	0	11	18
	1	1	1	0	1	61	136
	1	1	1	1	0	28	32
	1	1	1	1	1	298	308

Table 15: NAEP items

Report item	NAEP Block	NAEP item	
1	M1	4	35 + 42 = (77)
2	M1	5	55 + 37 = (92)
3	M1	6	59 + 46 + 82 + 68 = (255)
4	M1	11	? represents nine tens (90)
5	M1	13	Number 10 more than 95 (105)
6	M1	15	The digit in thousands place in 45,372 (5)
7	M1	16	Product of 21 and 3 (63)
8	M1	17	Product of 314 and 12 (3768)
9	M2	3	Which is greater: 2573, 2537, 2735 or (2753)
10	M2	6	(7 > 5), 7 = 5 or 7 < 5
11	M2	8	7 + 24 + 9 = (40)
12	M2	9	64 - 27 = (37)
13	M2	10	604 - 207 = (397)
14	M2	11	231 - 189 = (42)
15	M2	12	Number of birds in picture [< 100, (100-1000), > 1000, > 10,000]
16	M2	21	15 / 5 = (3)
17	M2	22	52 / 4 = (13)
18	M2	23	29 - (13) = 16
19	M3	15	One dollar and 86 cents means [$.186, ($1.86), $10.86, $18.60, $186.00]
20	M3	17	The digit in the tens place in 3058 (5)
21	M4	8	39 -26 (13)
22	M4	9	79 - 45 (34)
23	M4	10	65 - 7 (58)
24	M4	11	If 10 in each bag, 150 marbles in [10, (15), 25, 140, 150, 160] bags
25	M4	17	Three-fourths is (3/4)
26	M4	20	If N*13 = 13, N = (1)
27	M4	22	4.32 is [forty-three and 2/10, four hundred 32, (four and 32/100), forty-three hundred]
28	M6	21	152 - 59 - 93 [four possibilities]
29	M7	15	Which picture shows 3/4 shaded [four possibilities]
30	M7	23	82 - 39 is closest to [80 - 30, (80 - 40), 90 - 30, 90 - 40]

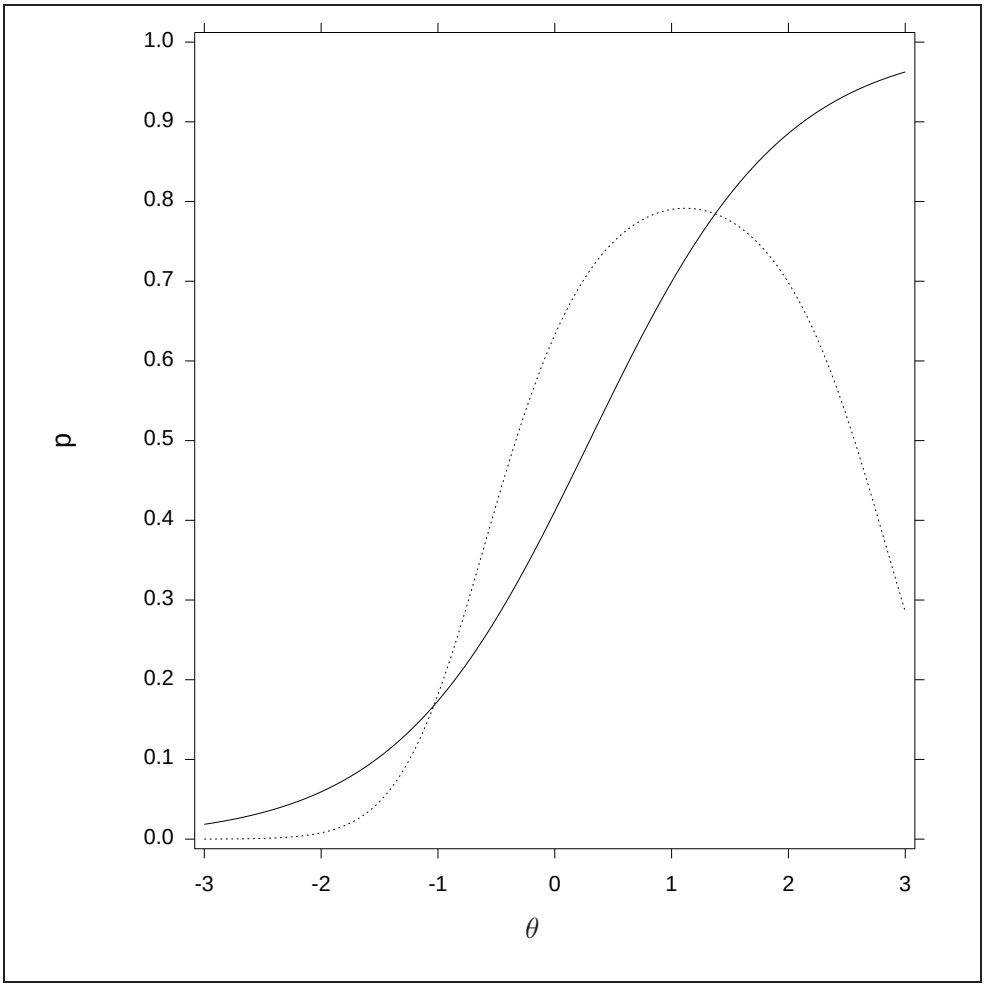

Figure 2: Item 2 ICCs, simulated data

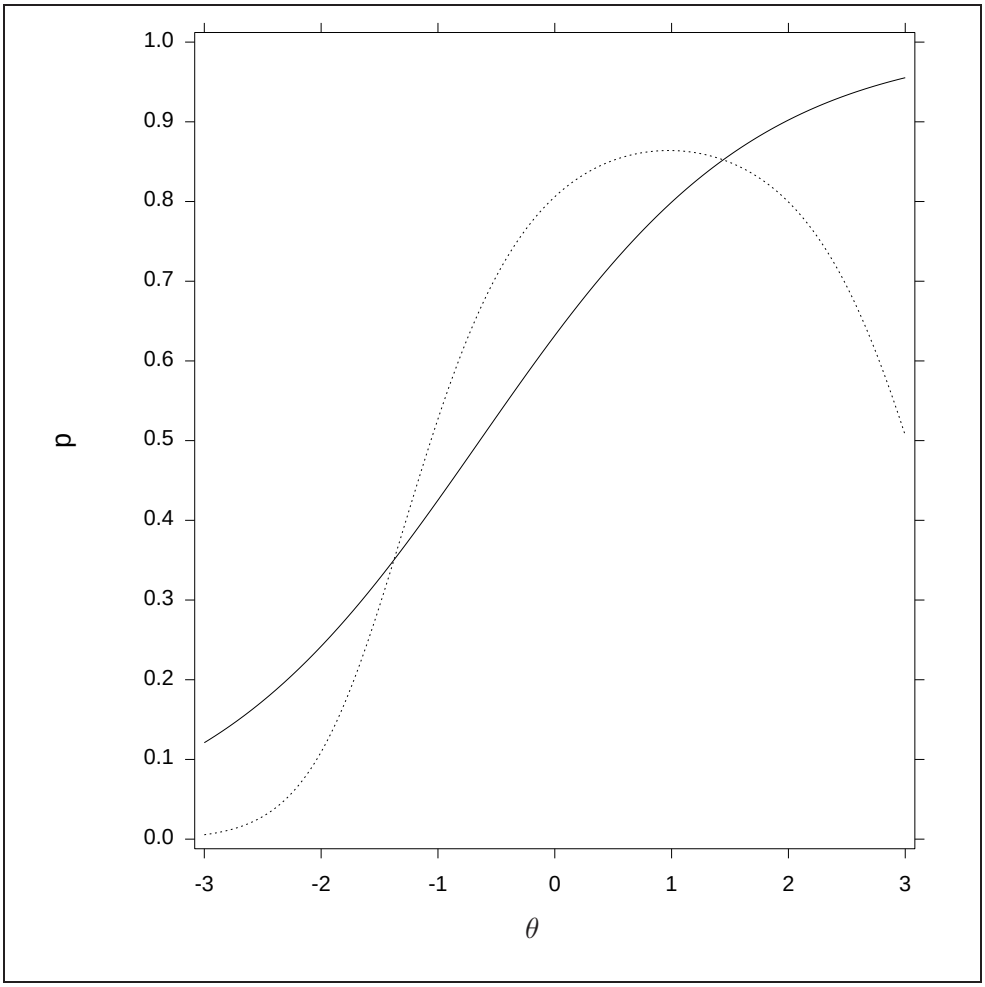

Figure 3: Item 3 ICCs, simulated data

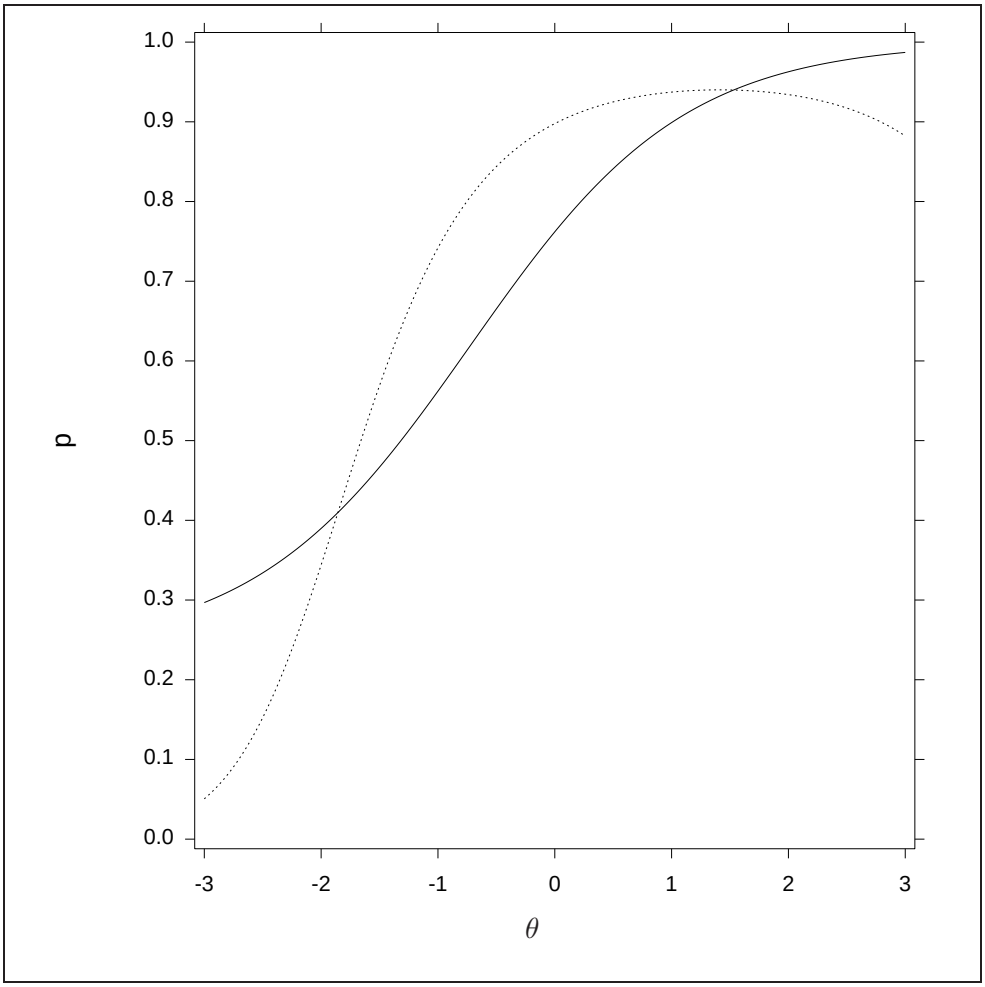

Figure 4: Item 4 ICCs, simulated data

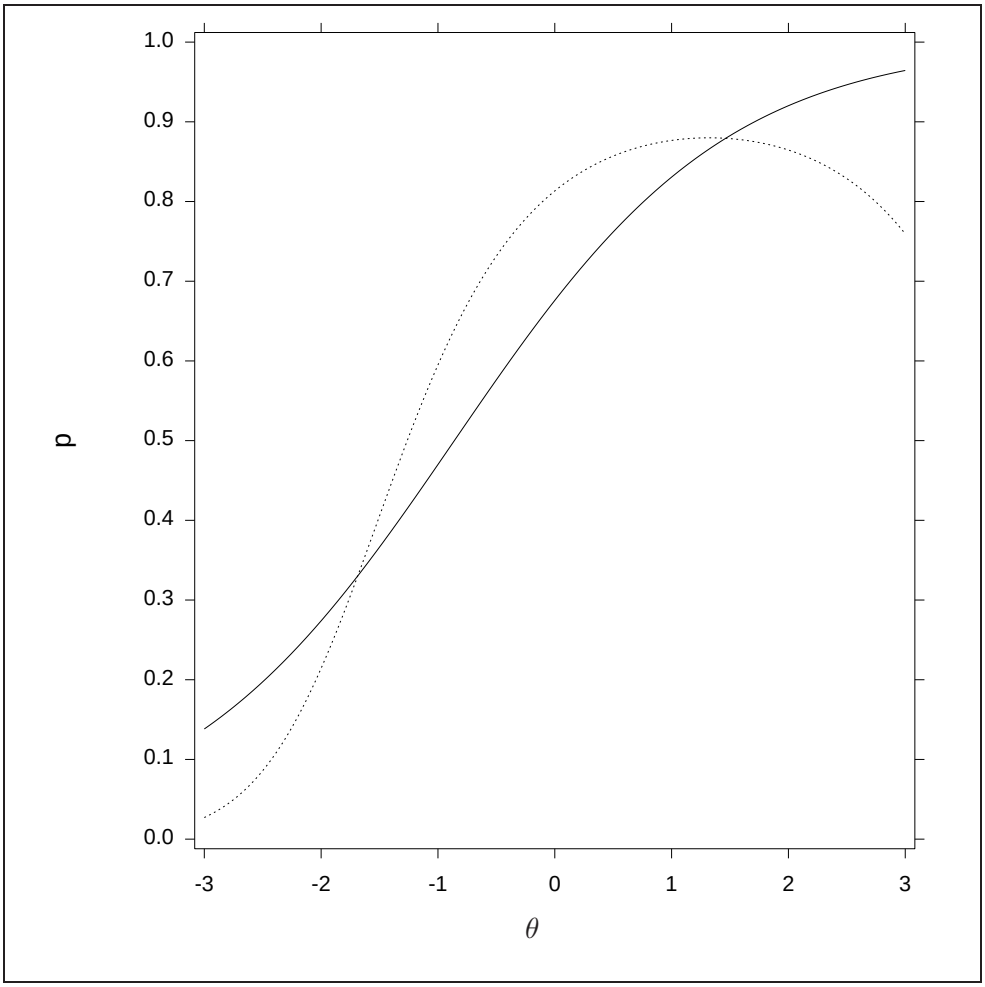

Figure 5: Item 5 ICCs, simulated data

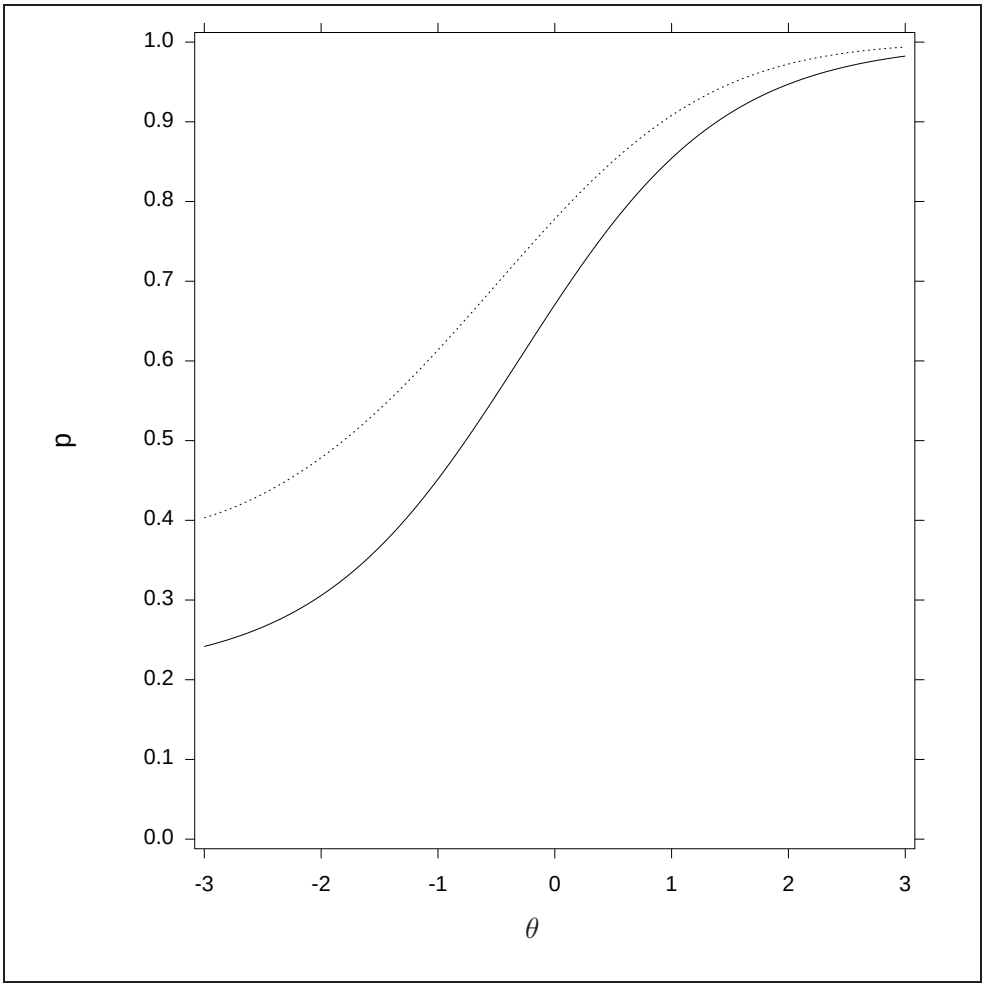

Figure 6: Item 6 ICCs, simulated data

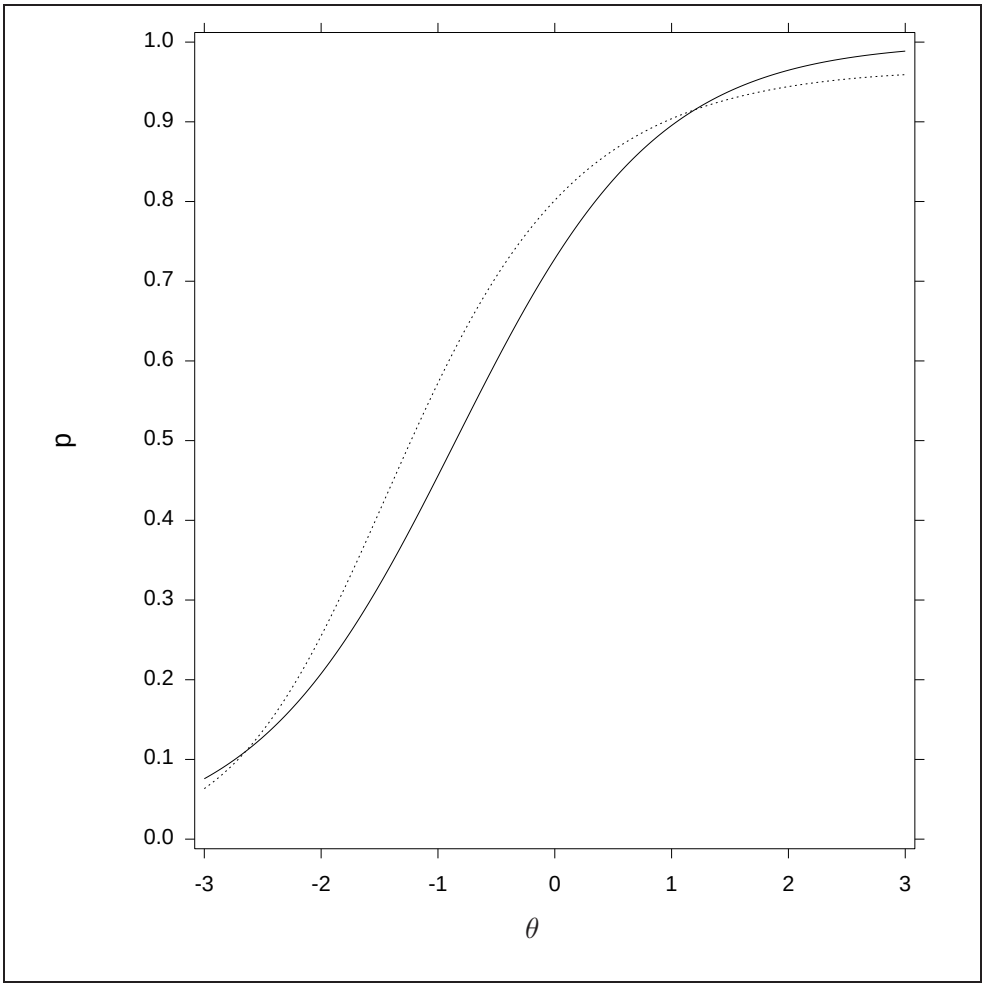

Figure 7: Item 7 ICCs, simulated data

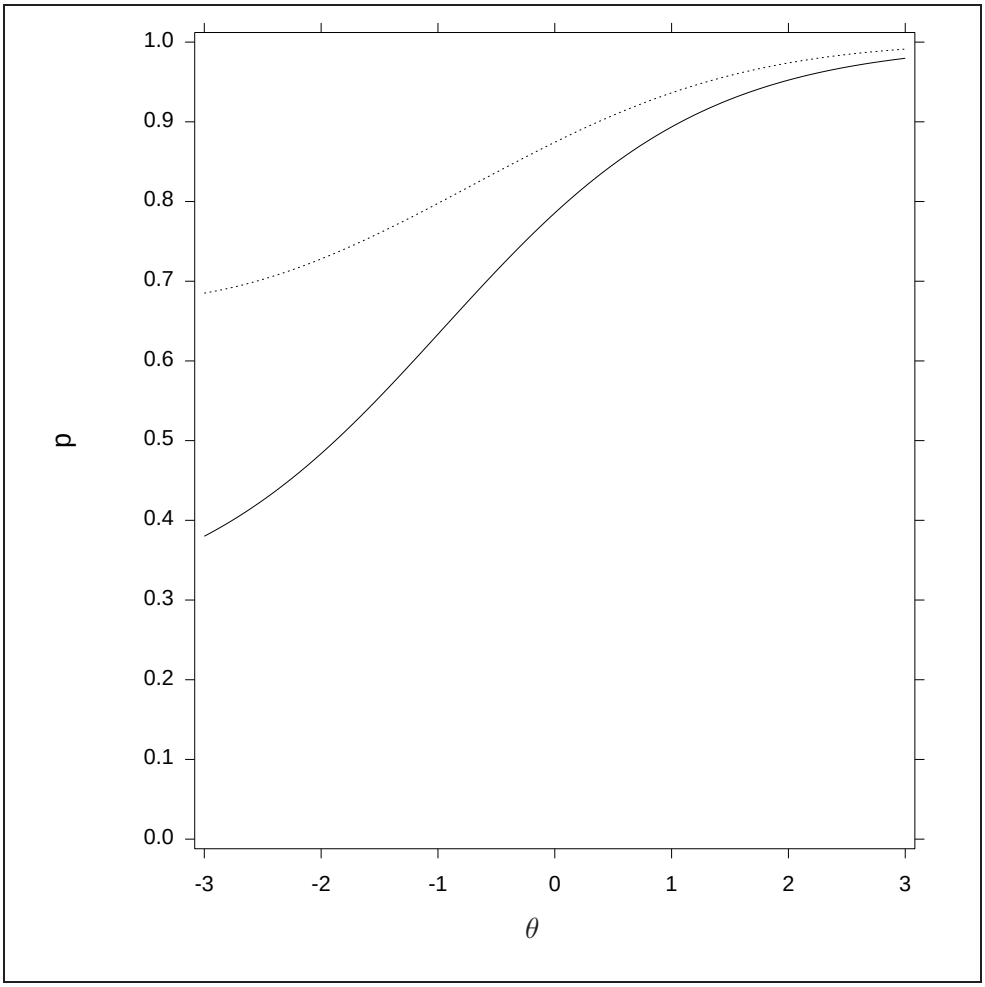

Figure 8: Item 8 ICCs, simulated data

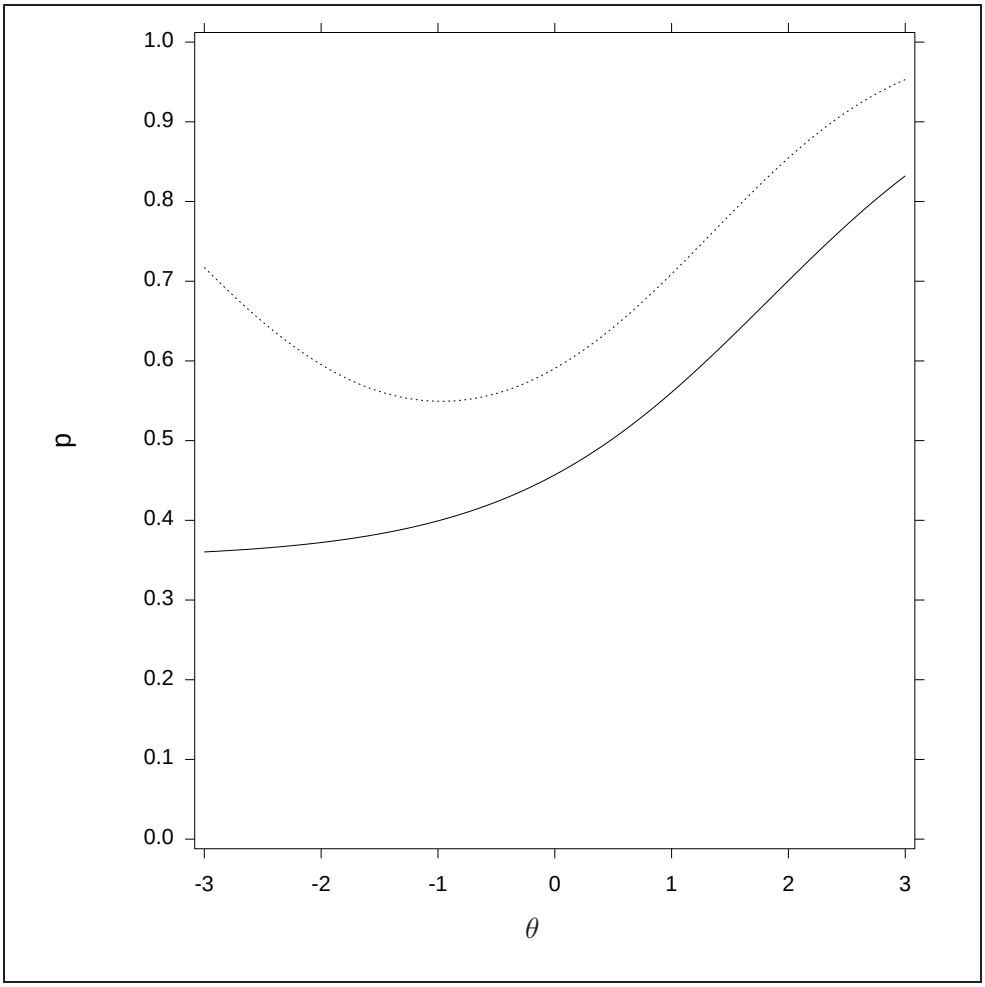

Figure 9: Item 9 ICCs, simulated data

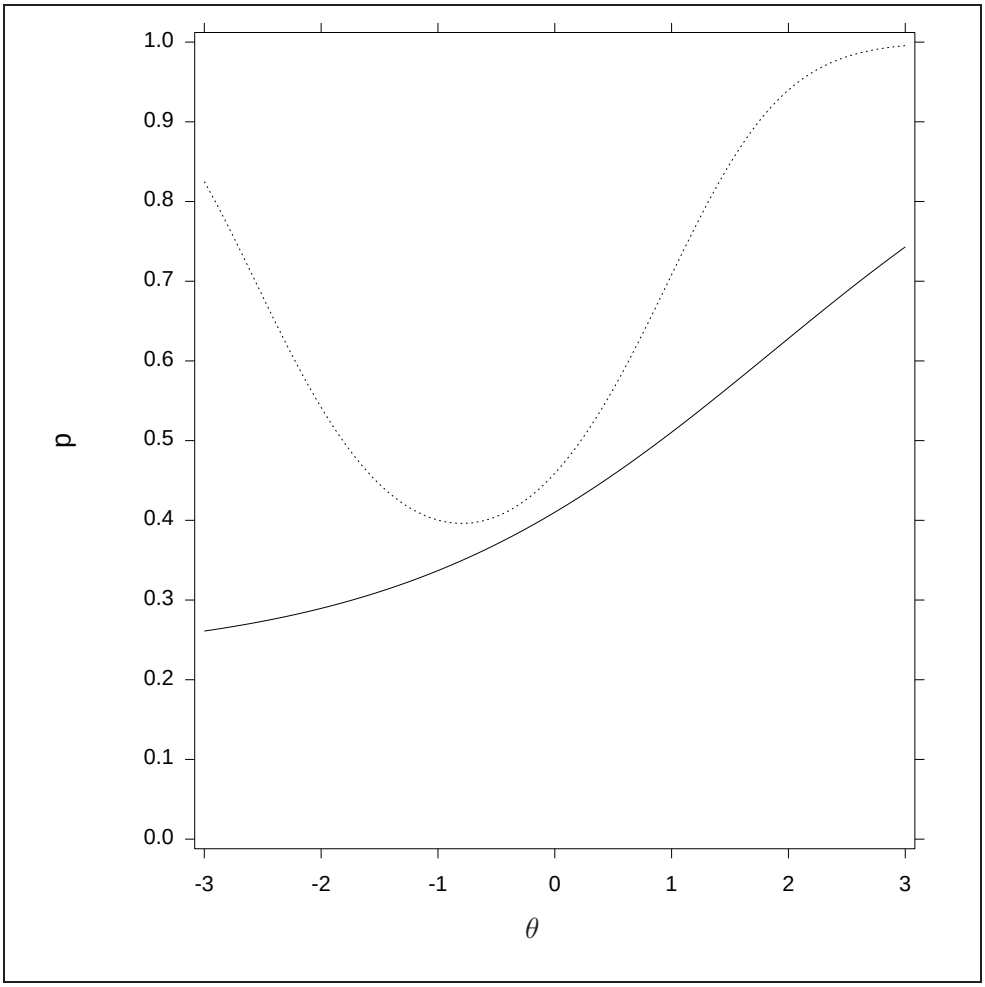

Figure 10: Item 10 ICCs, simulated data

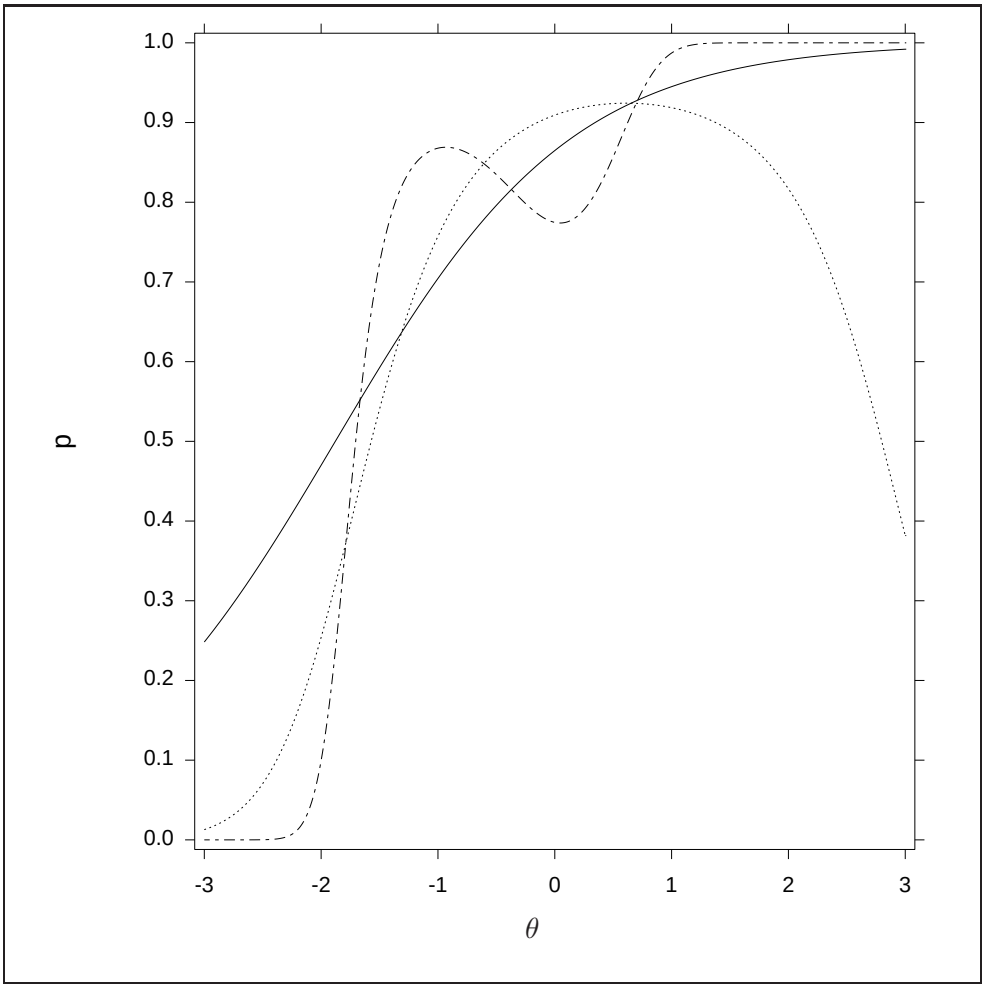

Figure 11: LSAT7, item 1 ICCs

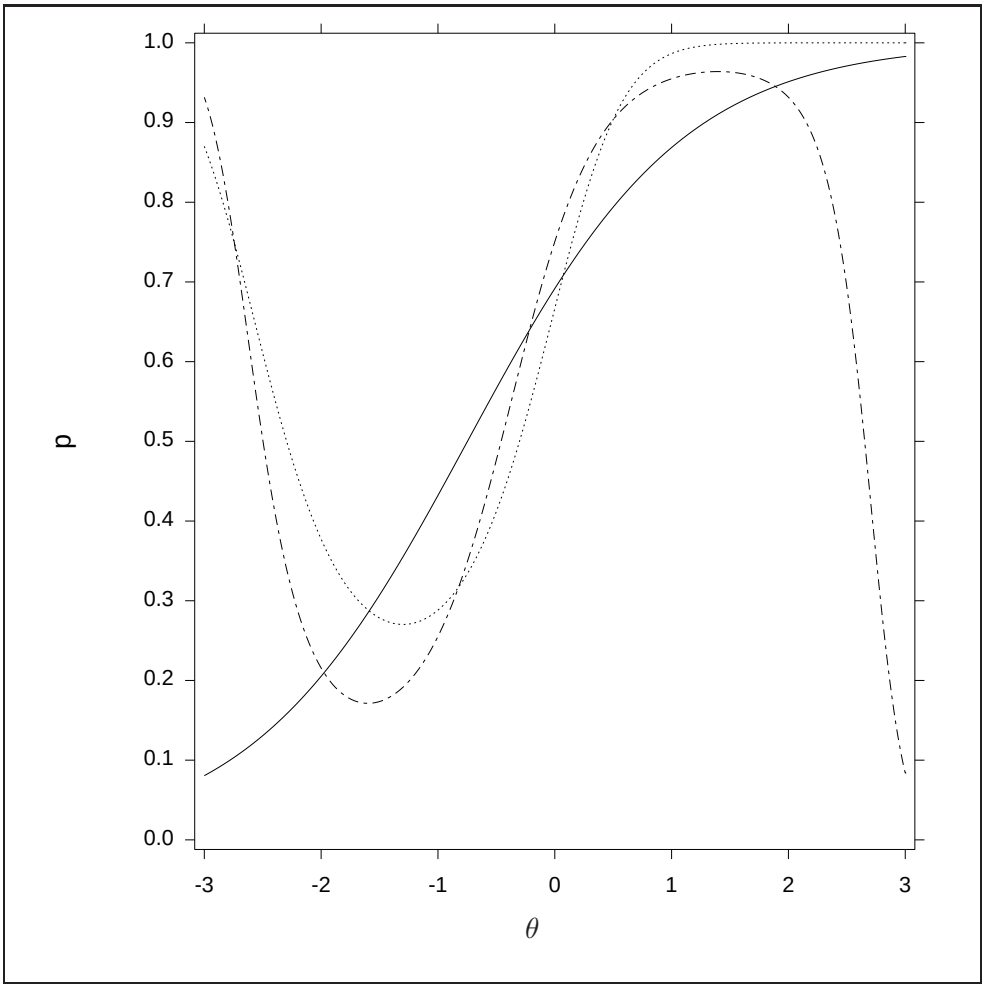

Figure 12: LSAT7, item 2 ICCs

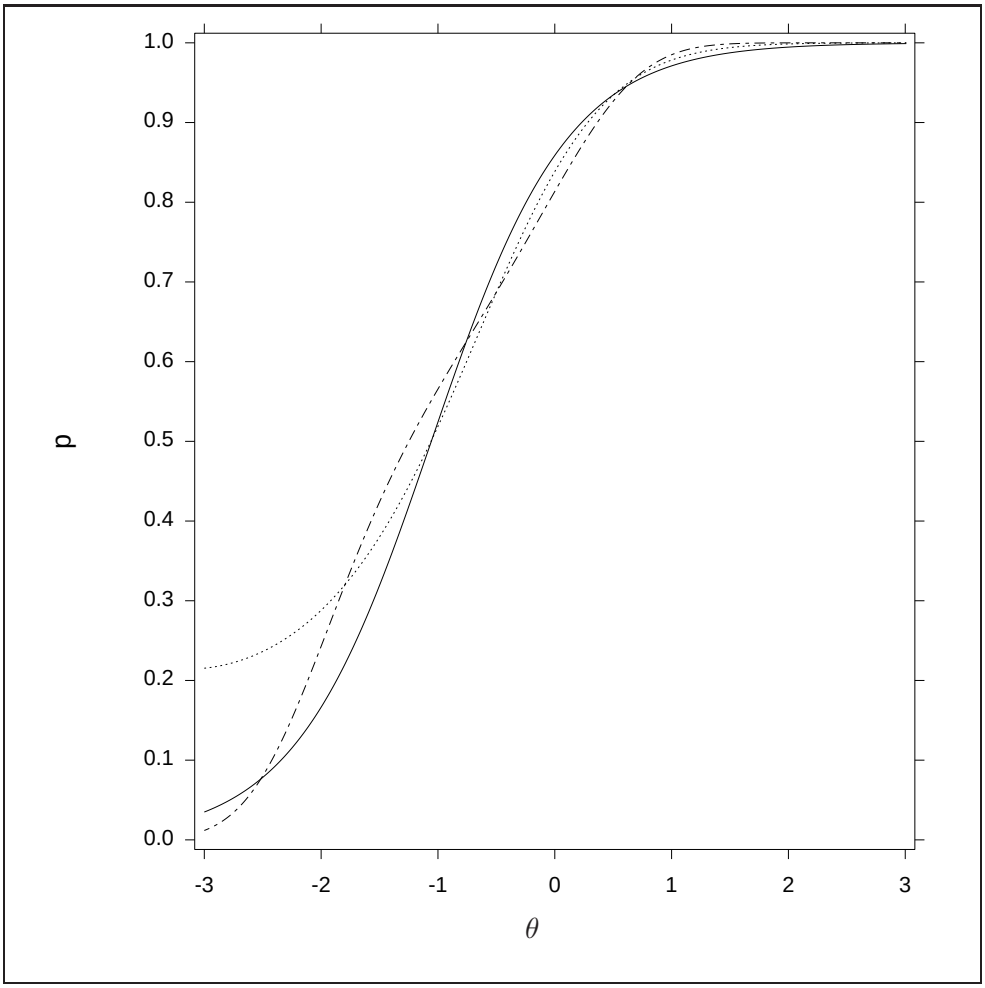

Figure 13: LSAT7, item 3 ICCs

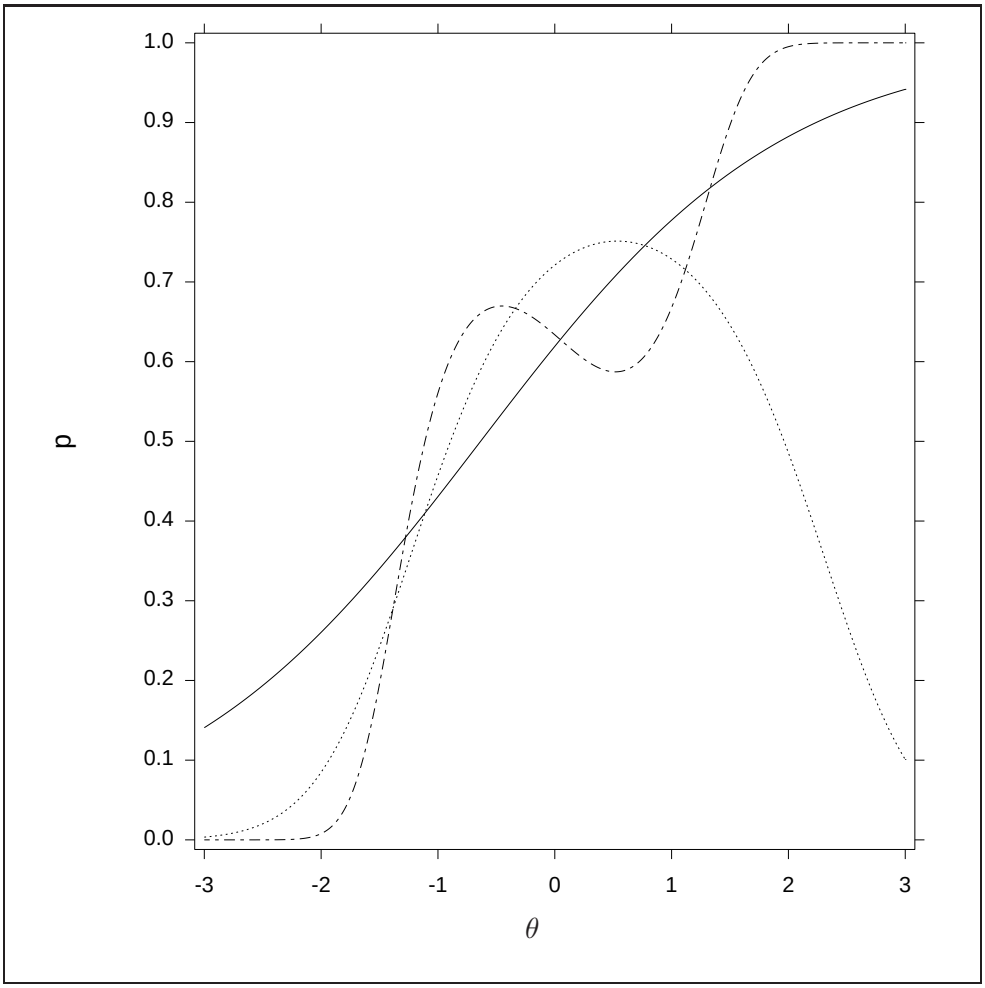

Figure 14: LSAT7, item 4 ICCs

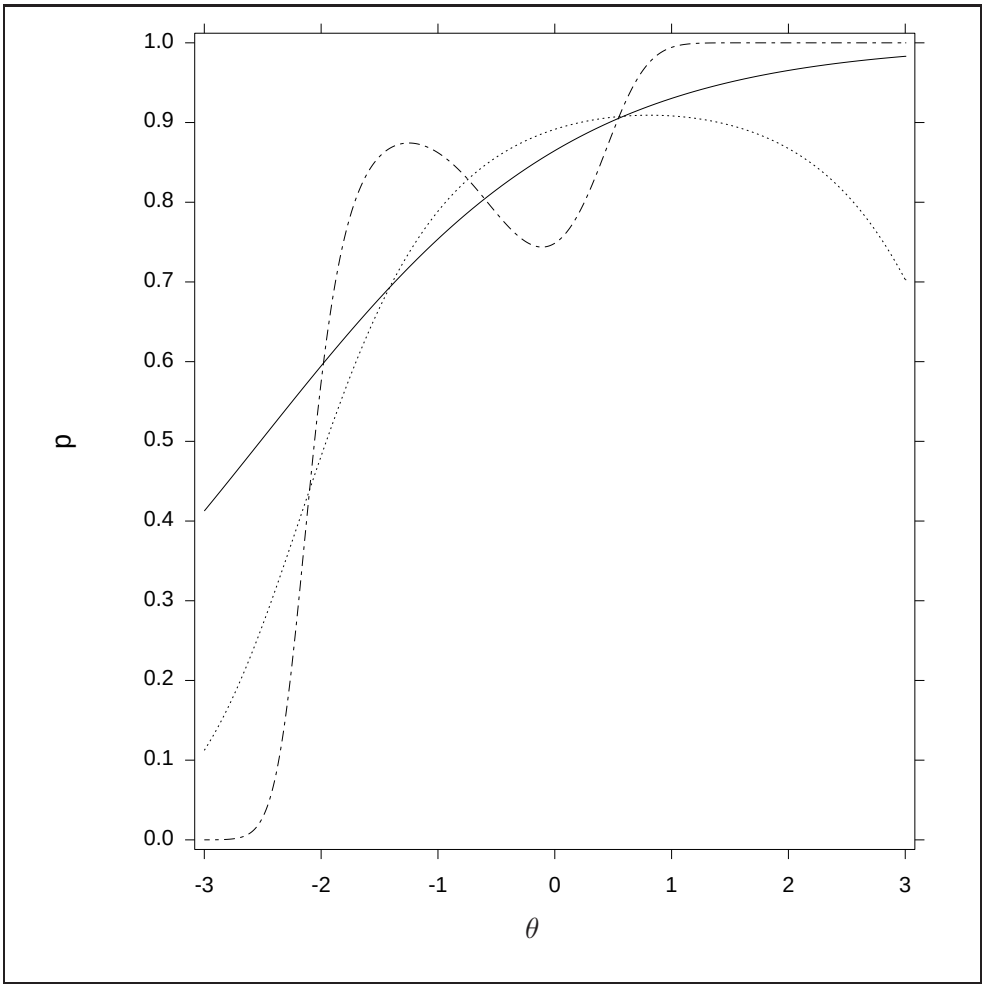

Figure 15: LSAT7, item 5 ICCs

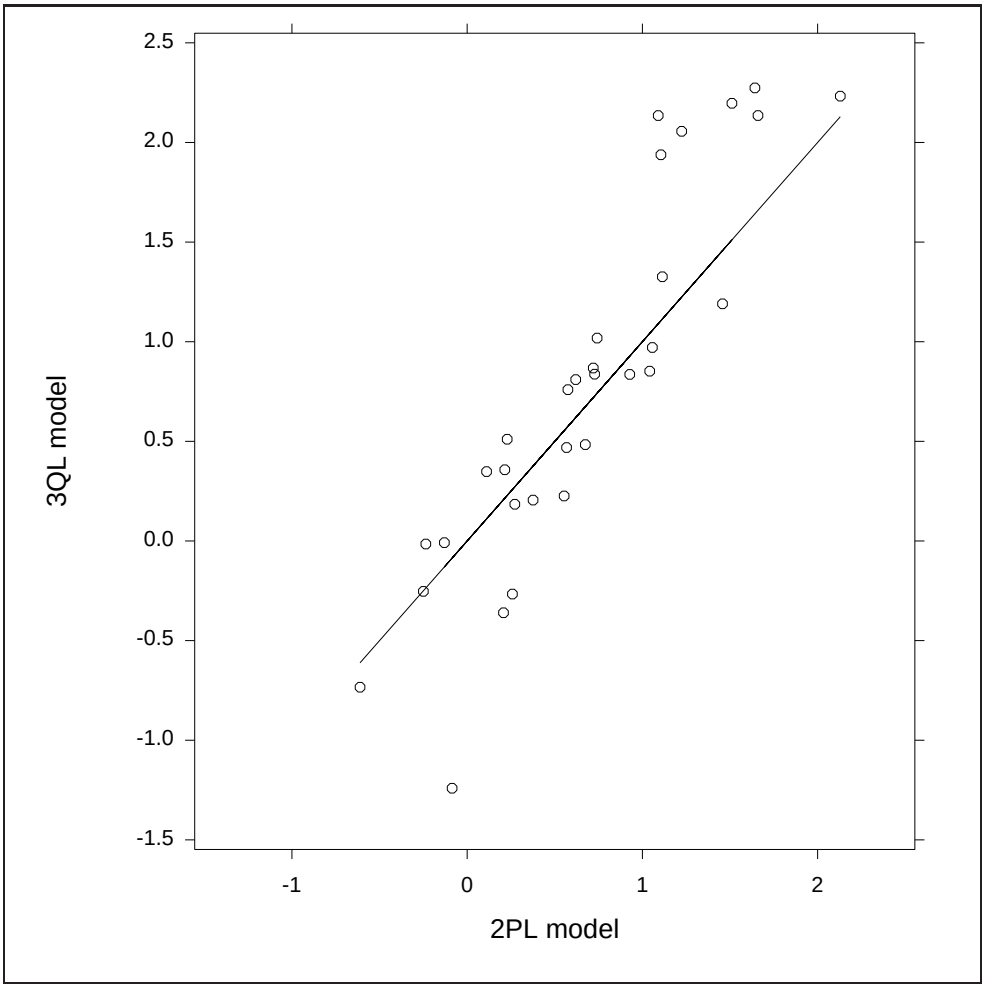

Figure 16: Posterior means, 3QL vs 2PL

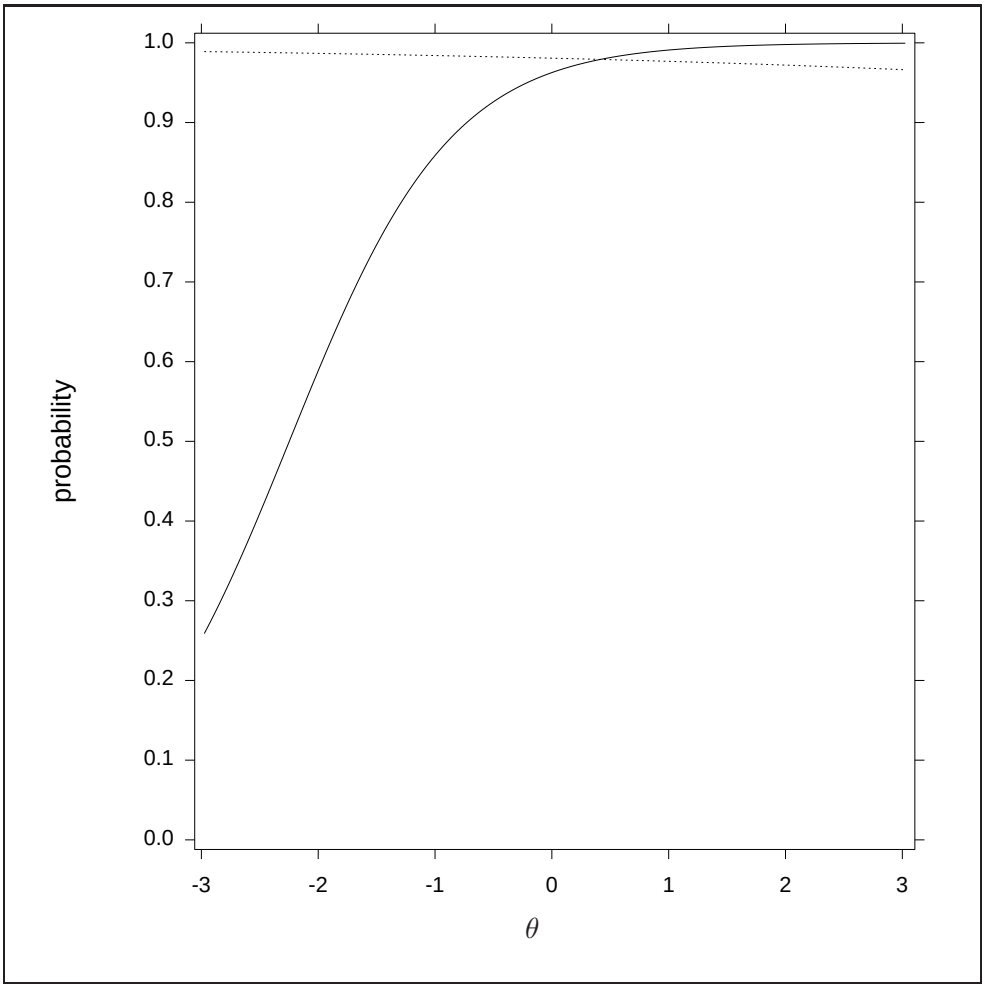

Figure 17: NAEP, item 1 ICCs

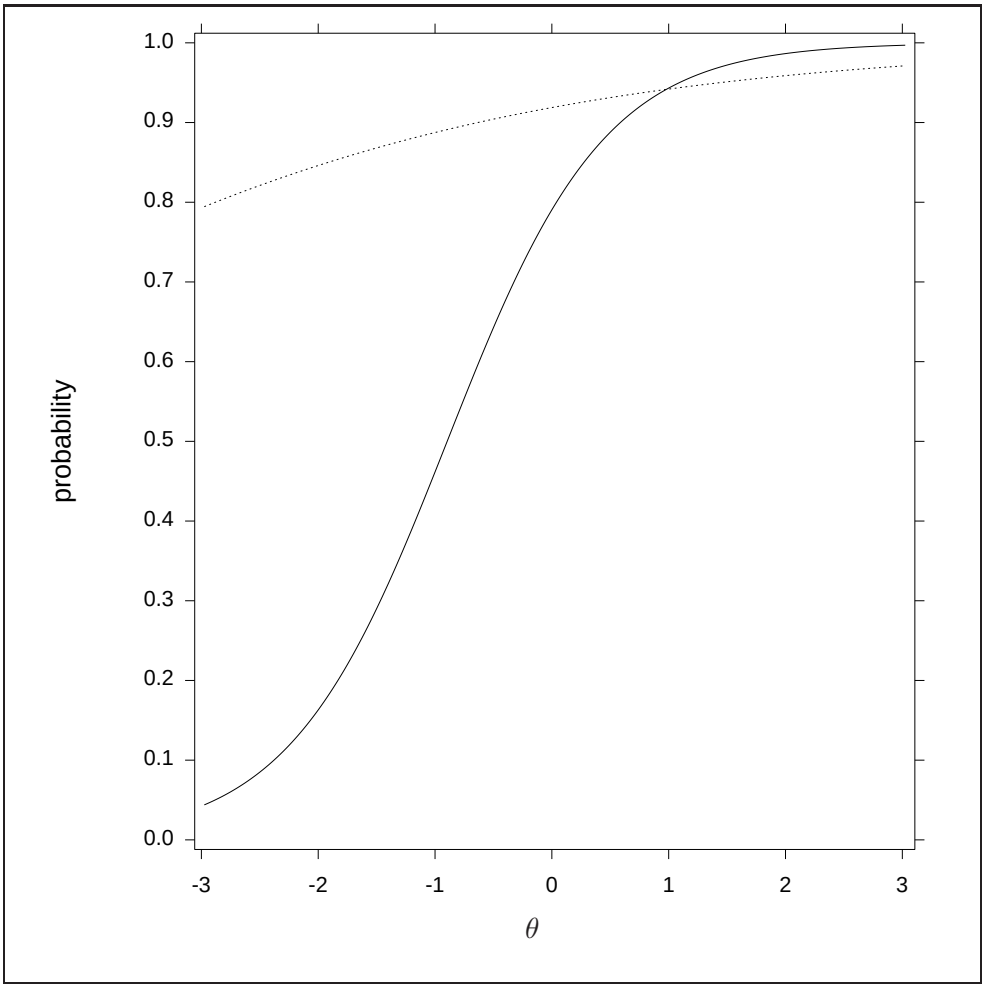

Figure 18: NAEP, item 2 ICCs

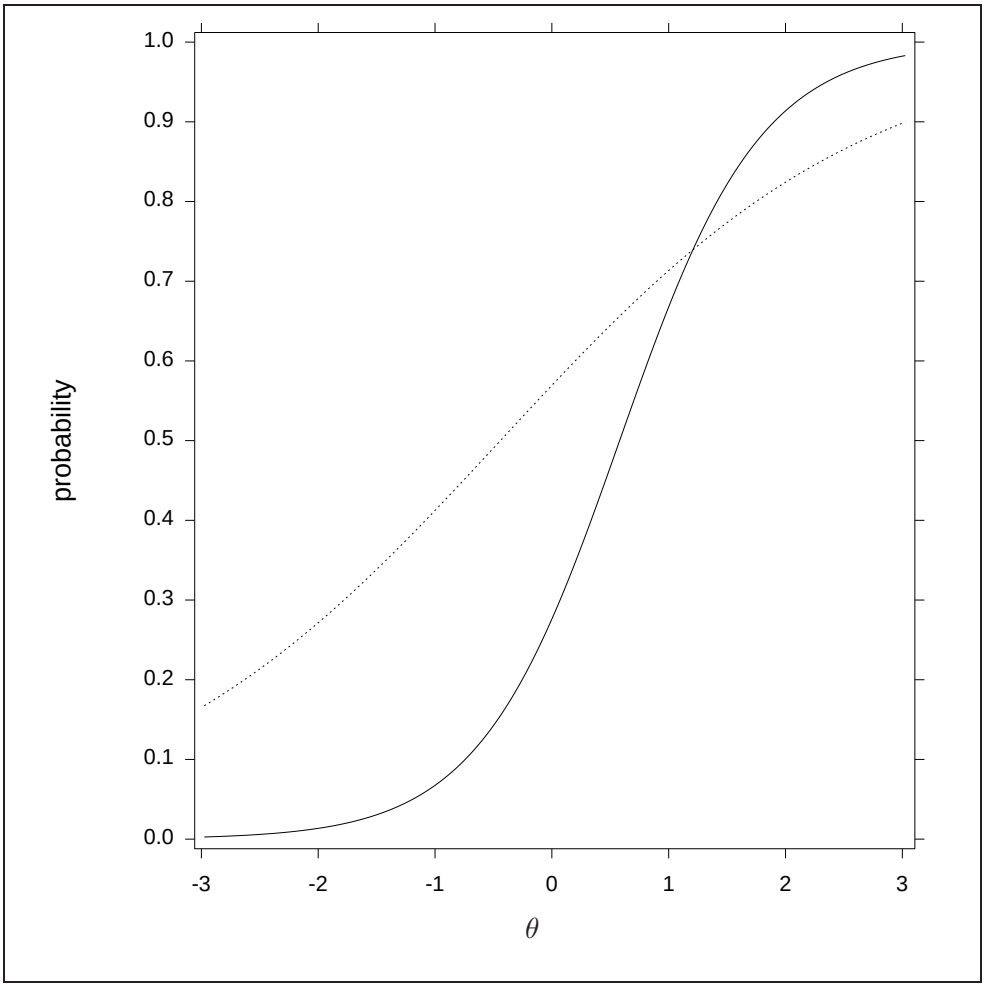

Figure 19: NAEP, item 3 ICCs

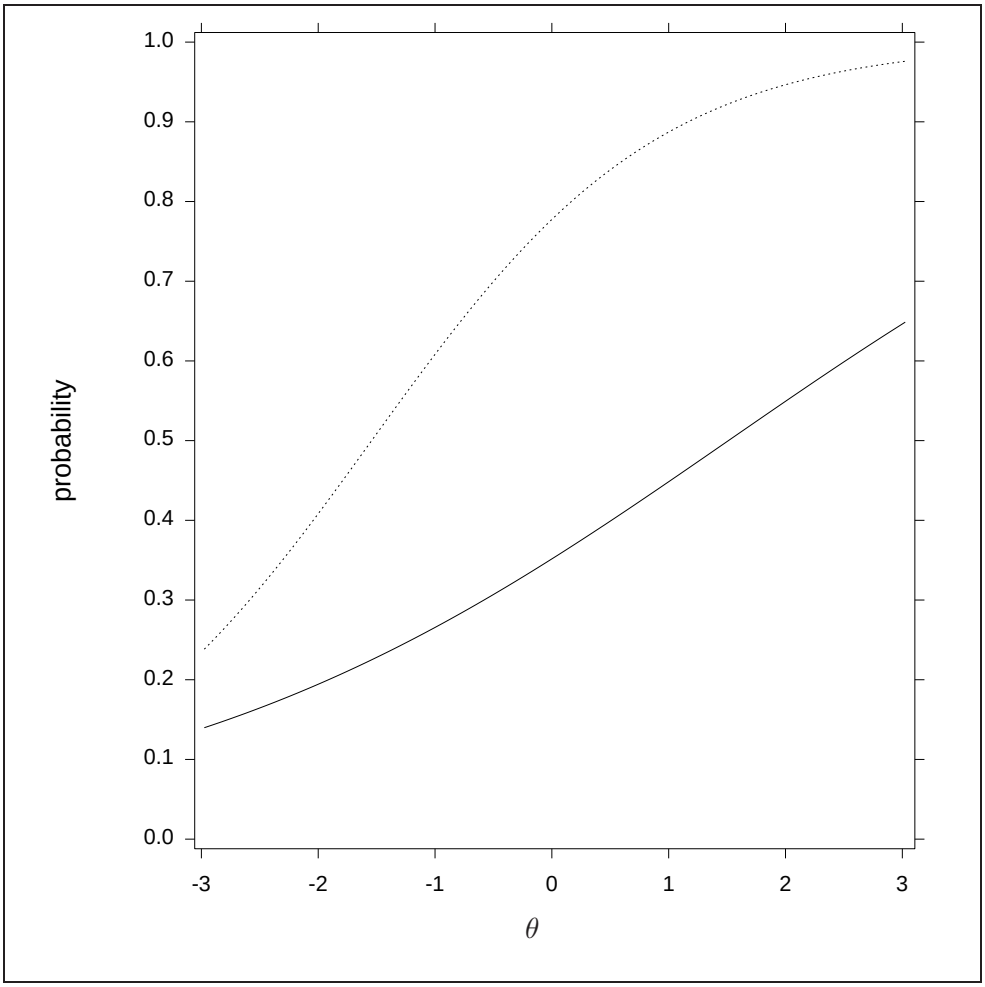

Figure 20: NAEP, item 4 ICCs

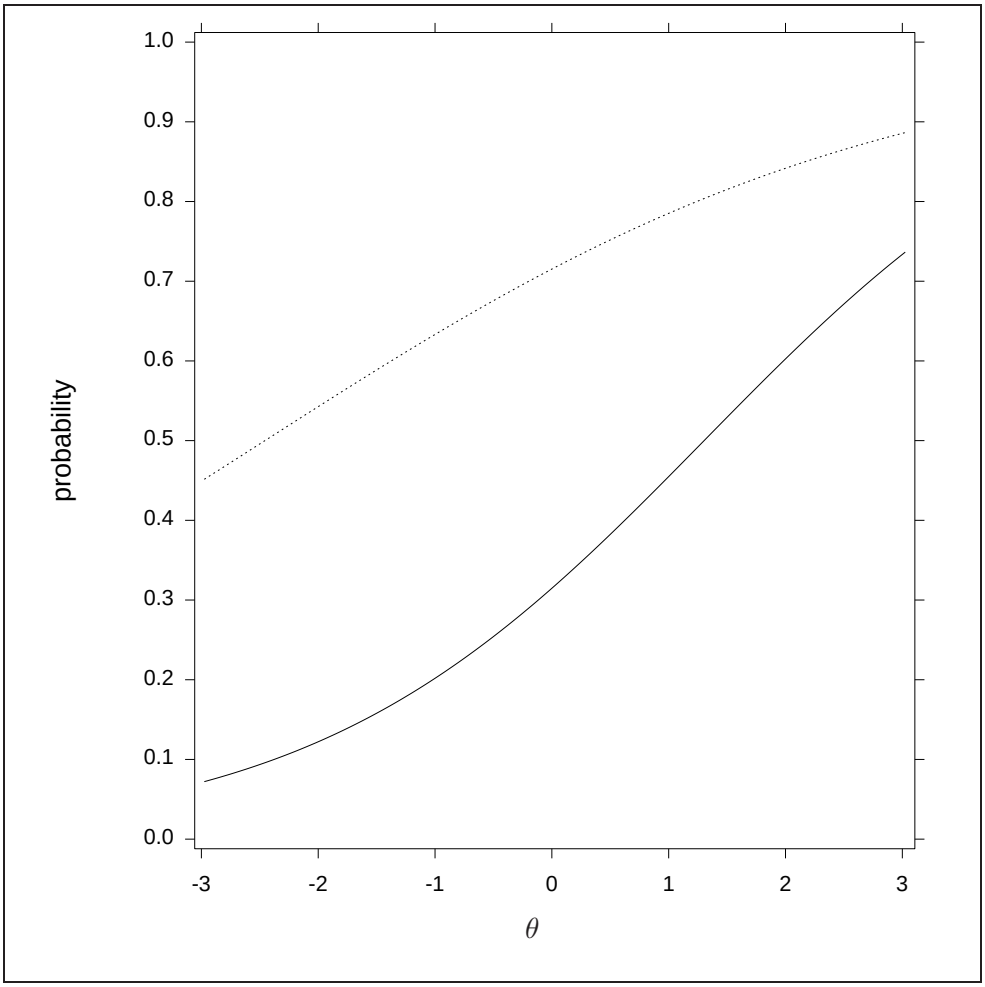

Figure 21: NAEP, item 5 ICCs

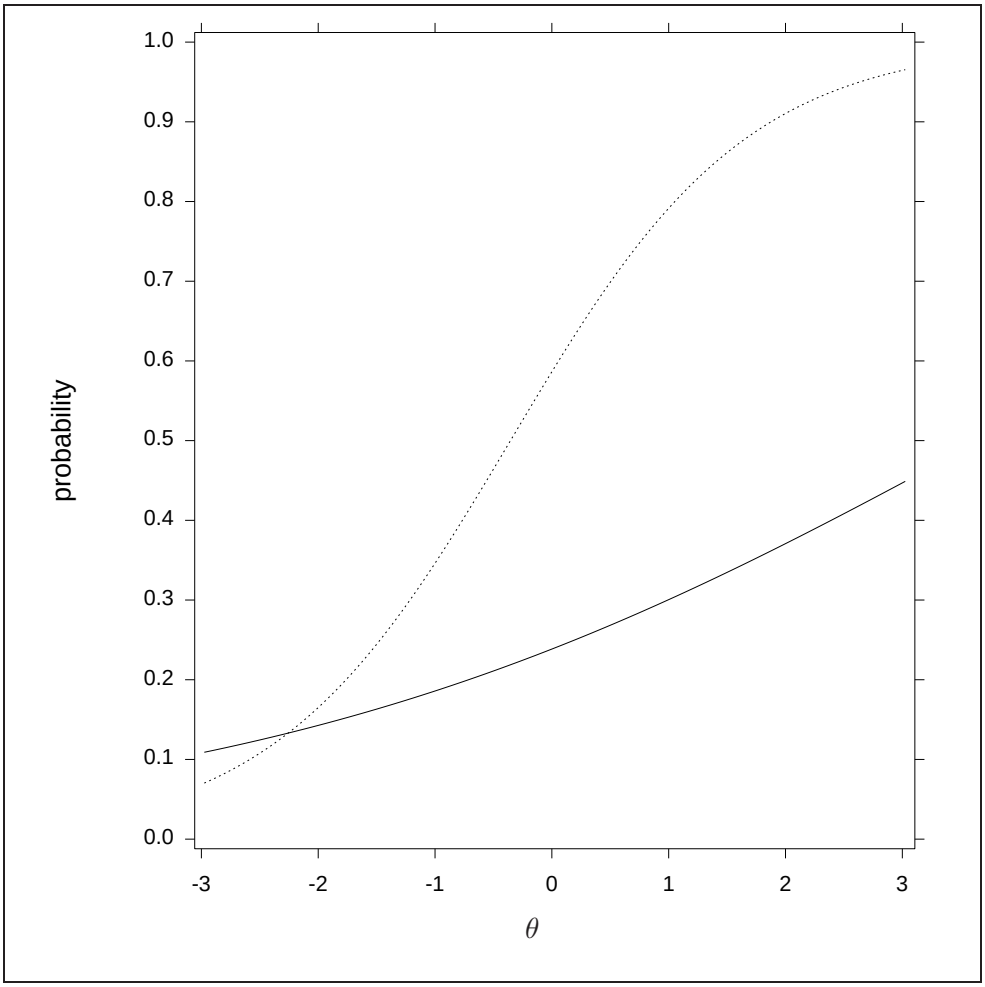

Figure 22: NAEP, item 6 ICCs

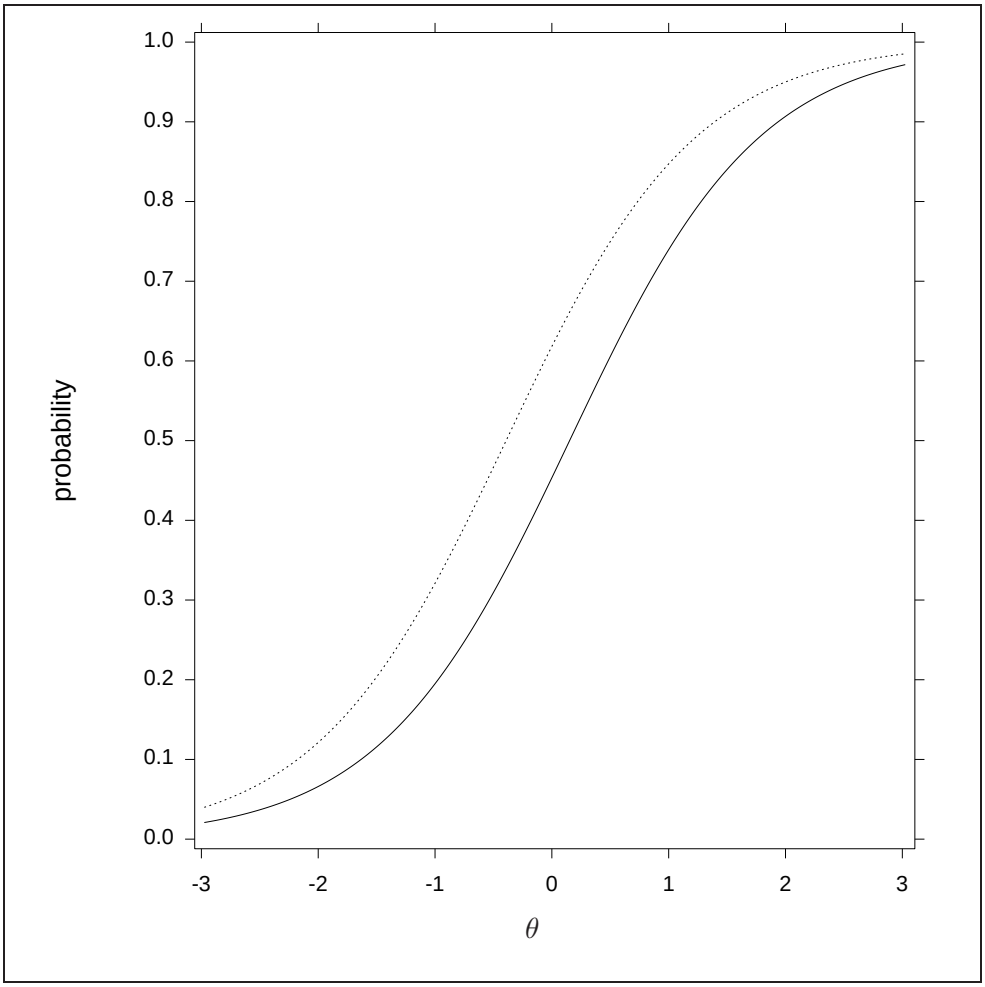

Figure 23: NAEP, item 7 ICCs

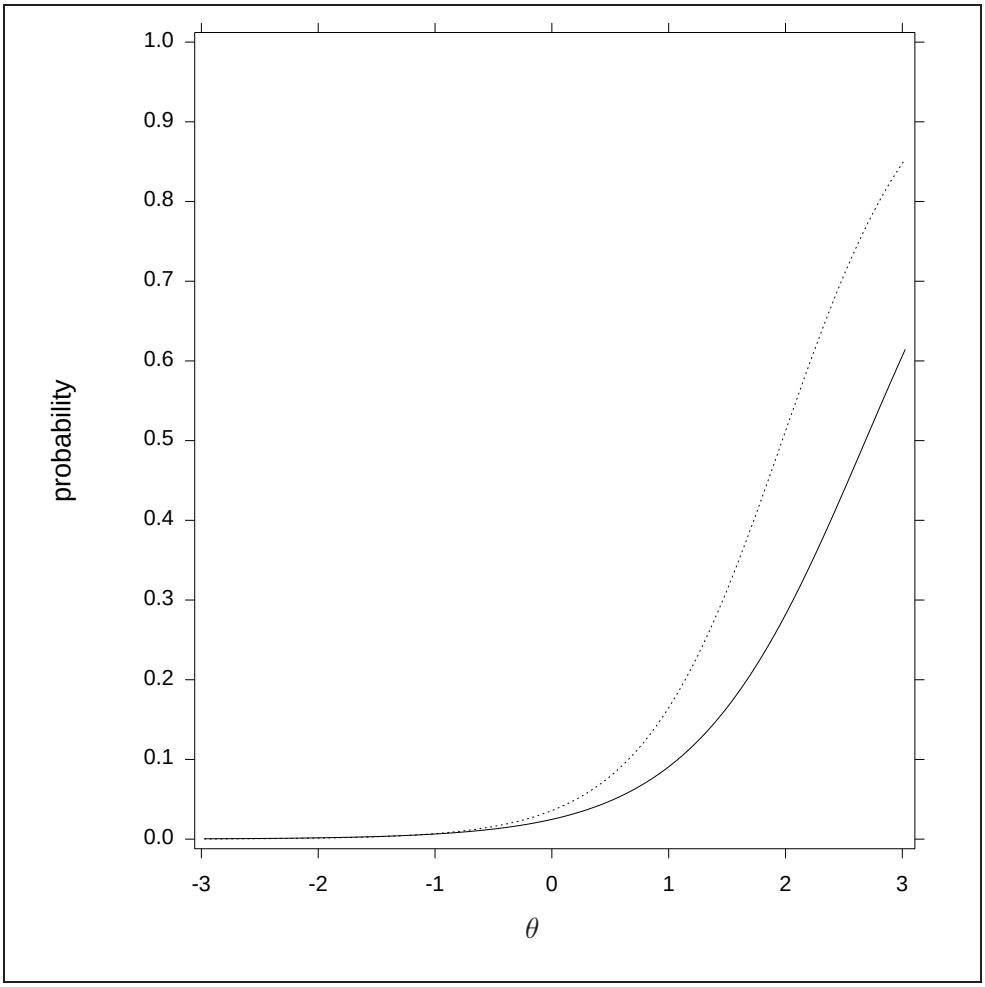

Figure 24: NAEP, item 8 ICCs

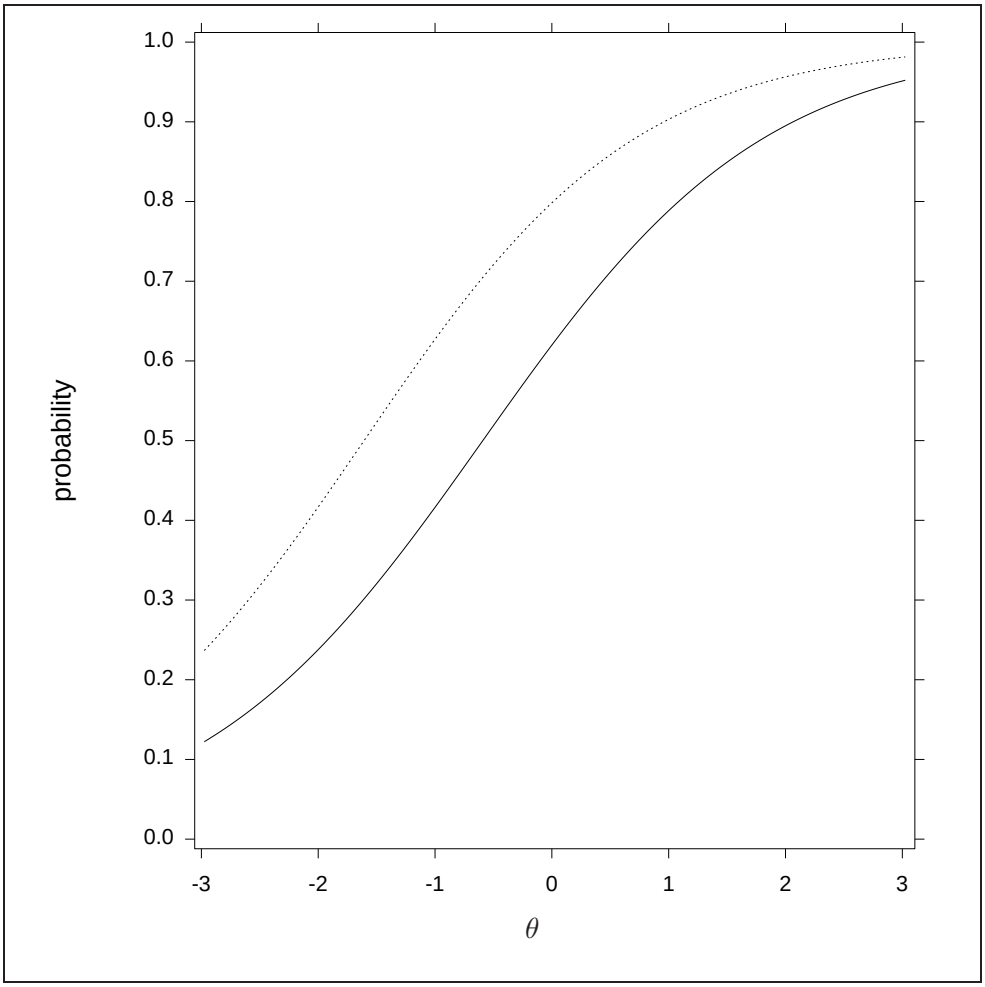

Figure 25: NAEP, item 9 ICCs

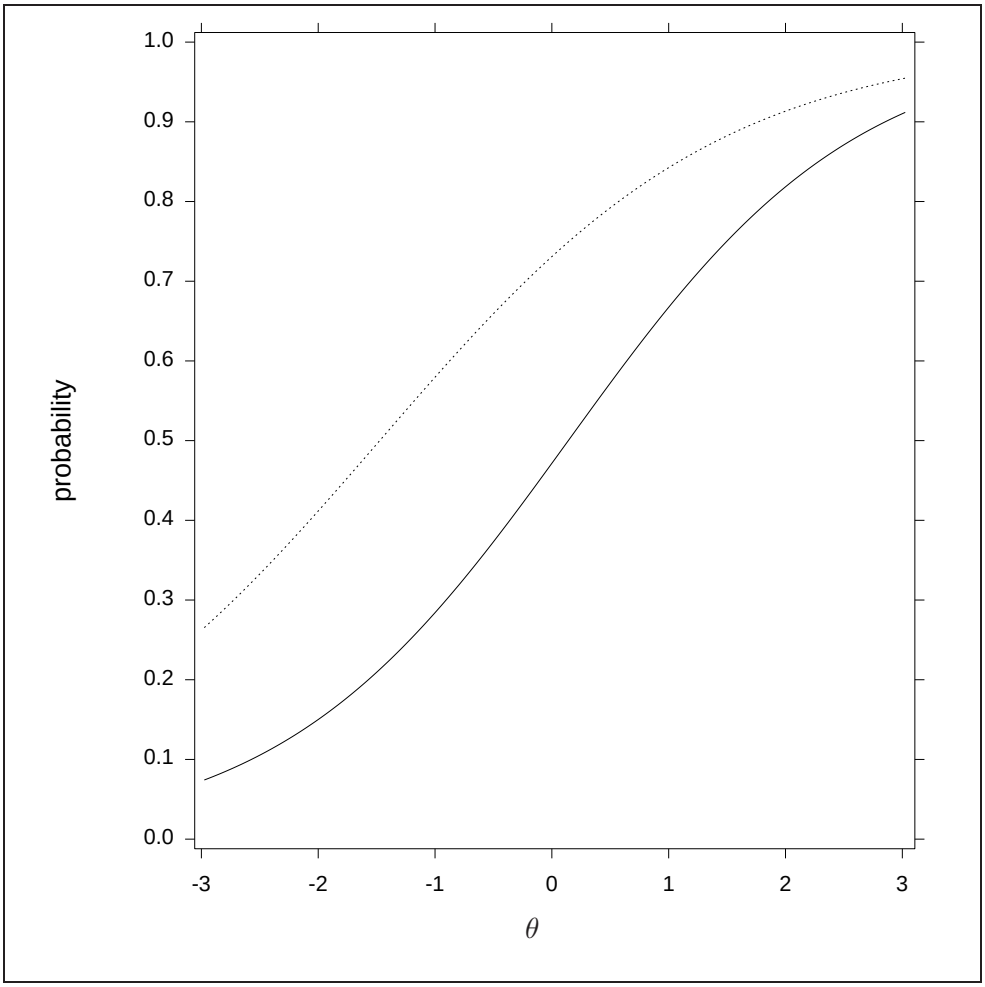

Figure 26: NAEP, item 10 ICCs

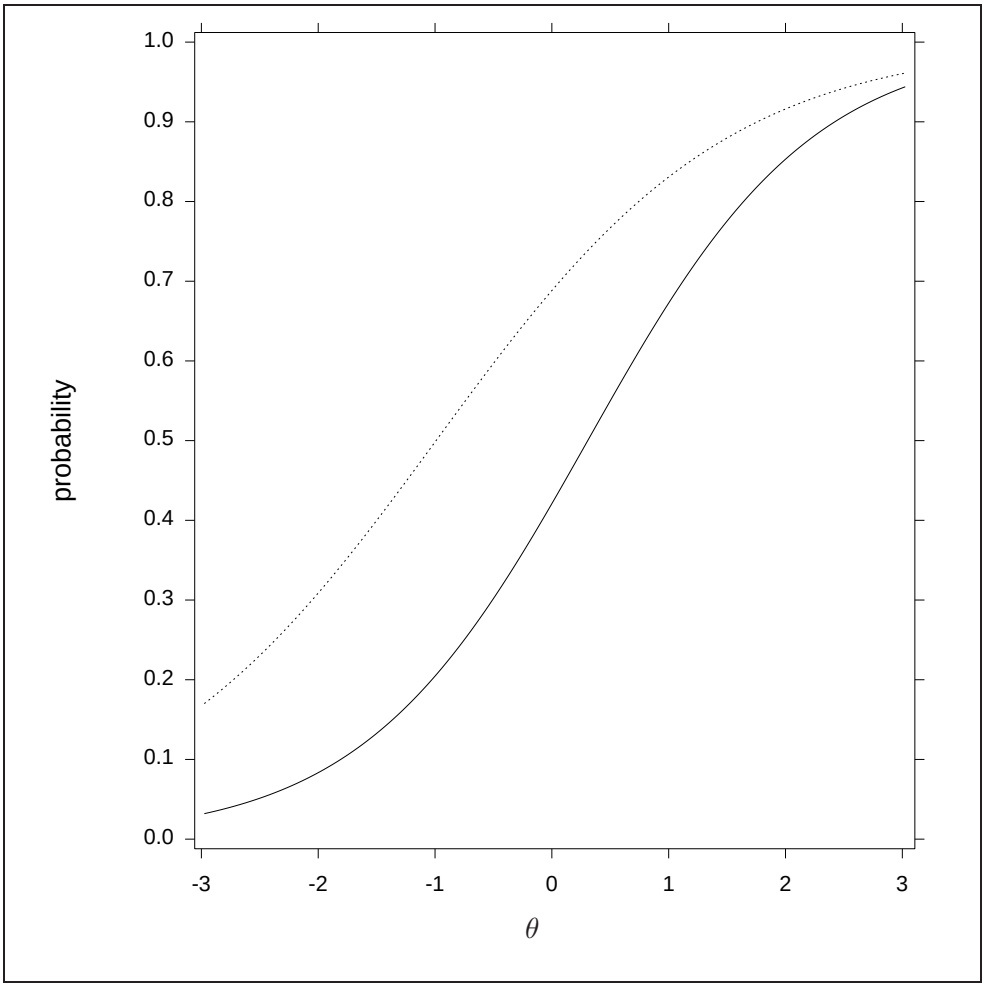

Figure 27: NAEP, item 11 ICCs

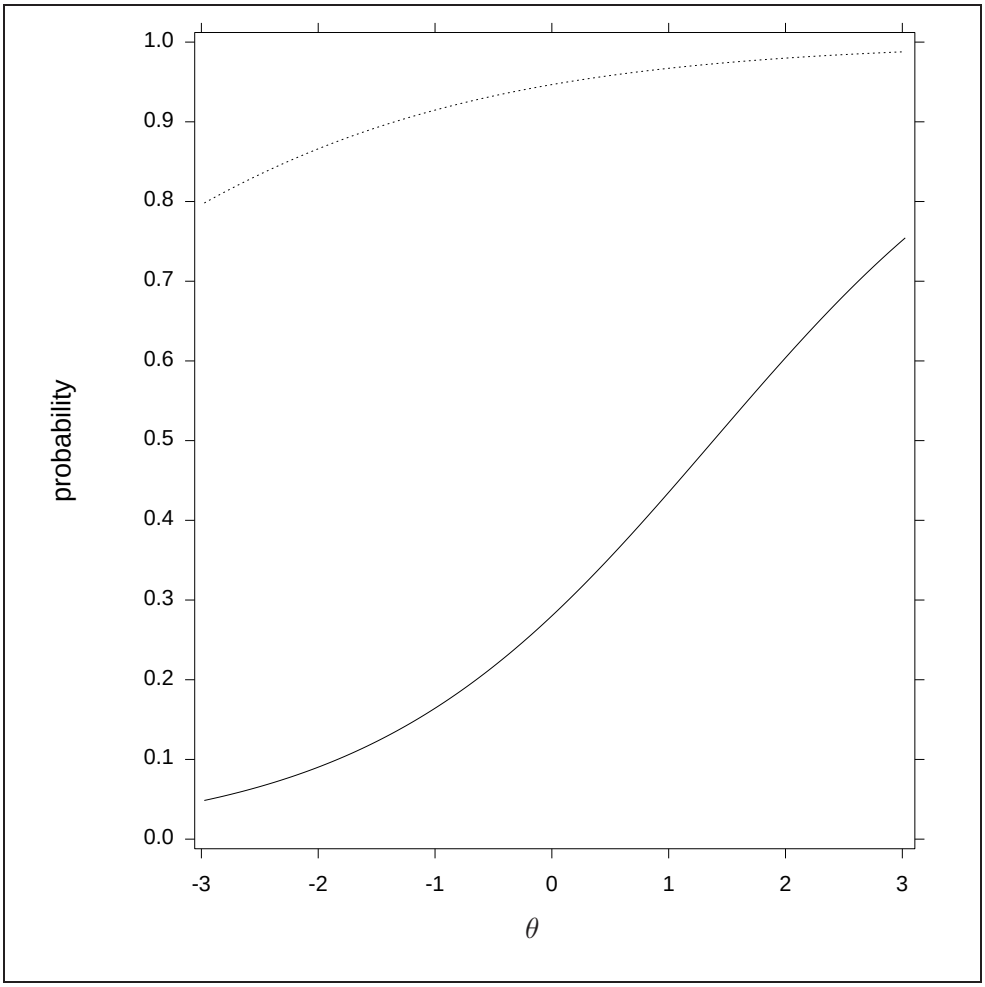

Figure 28: NAEP, item 12 ICCs

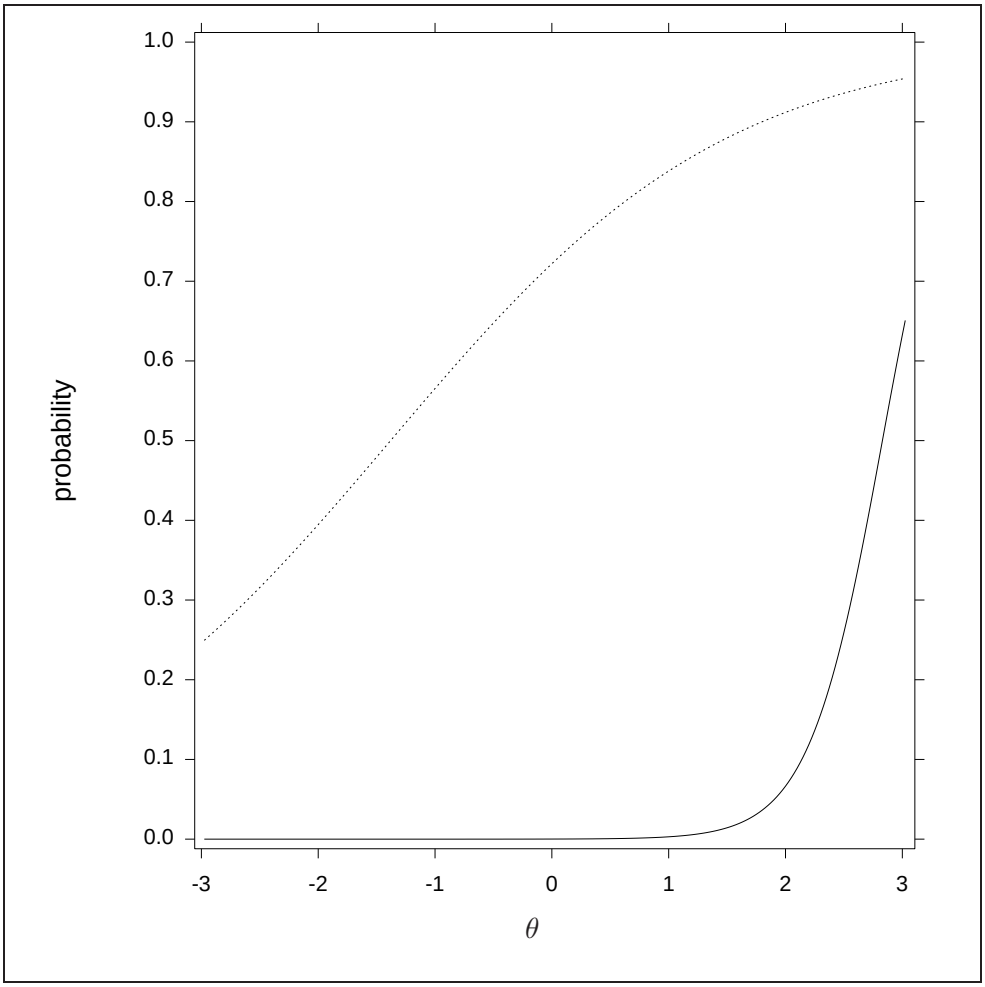

Figure 29: NAEP, item 13 ICCs

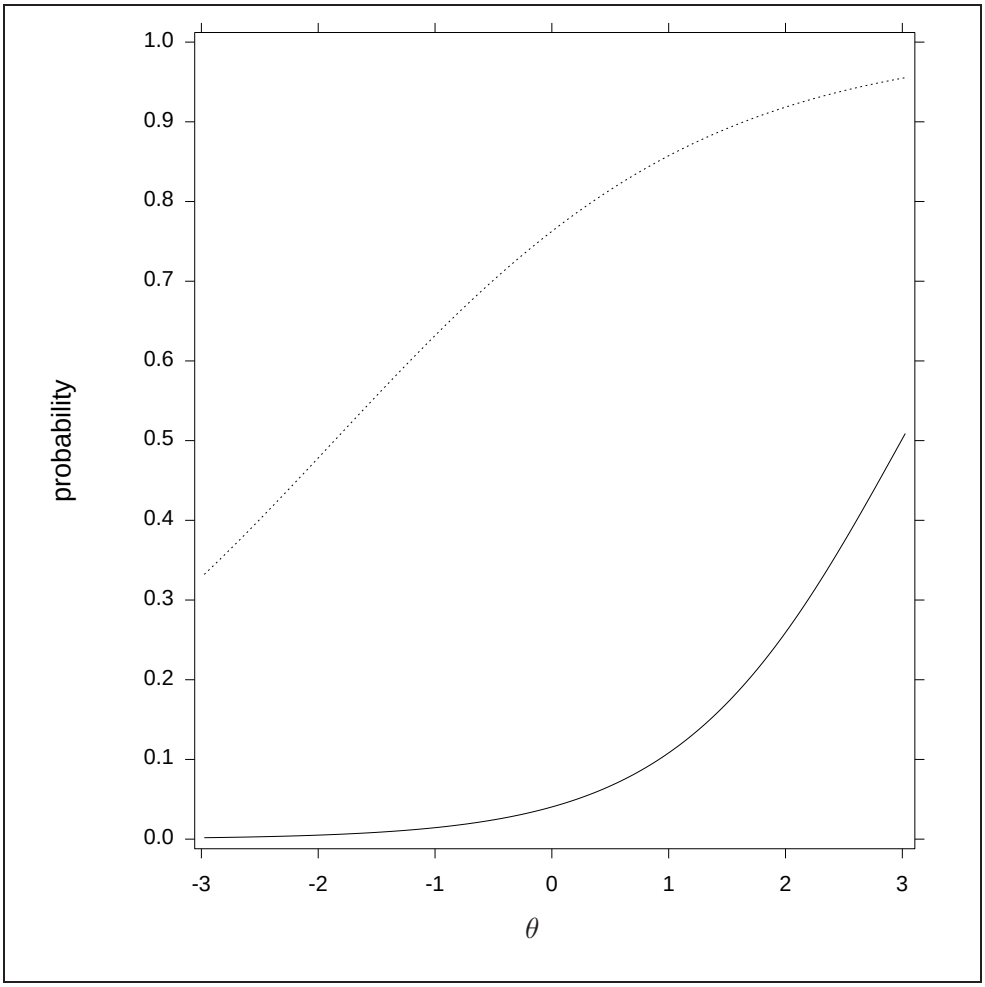

Figure 30: NAEP, item 14 ICCs

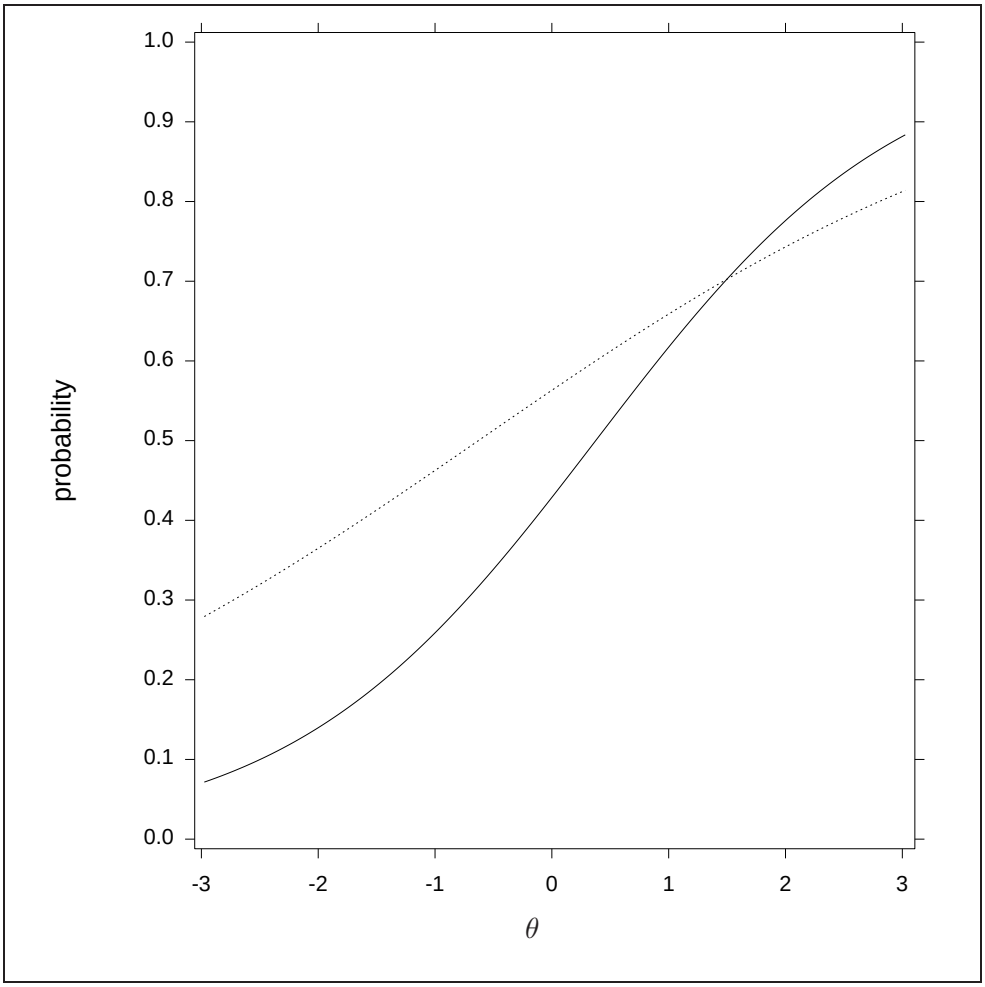

Figure 31: NAEP, item 15 ICCs

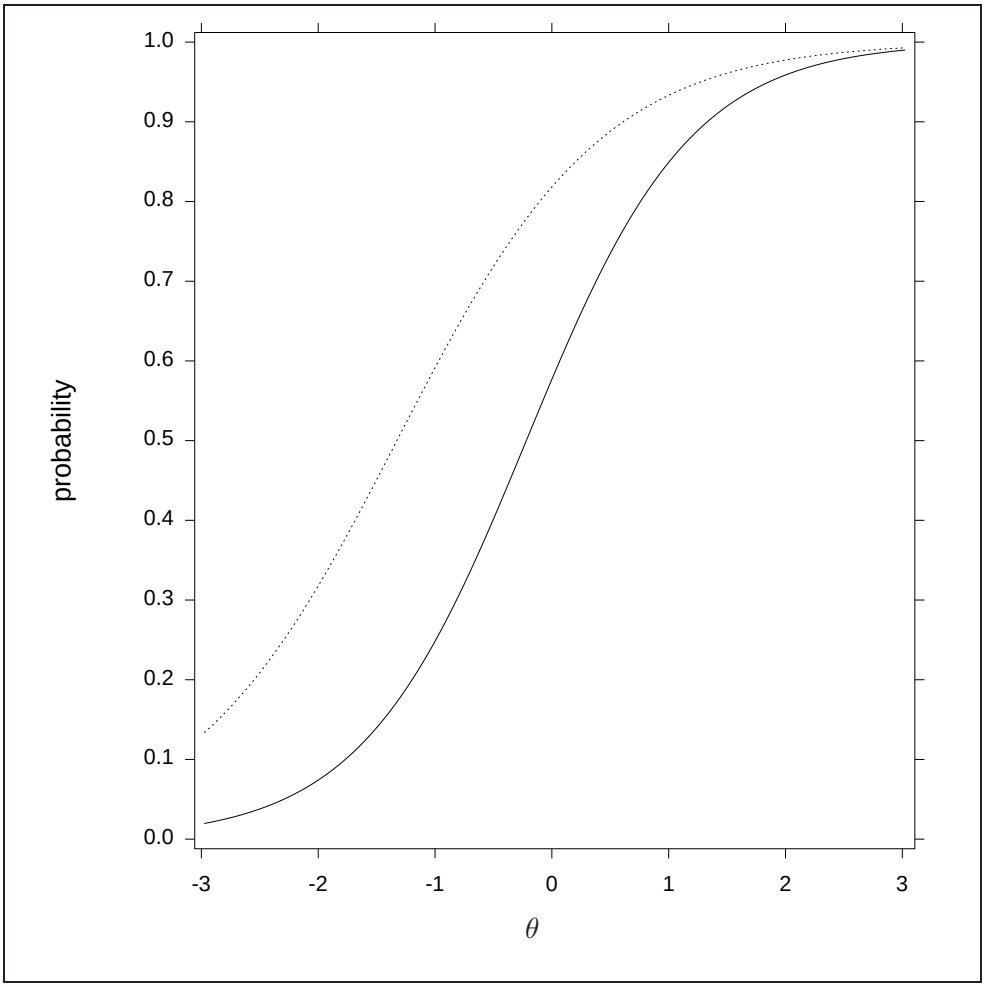

Figure 32: NAEP, item 16 ICCs

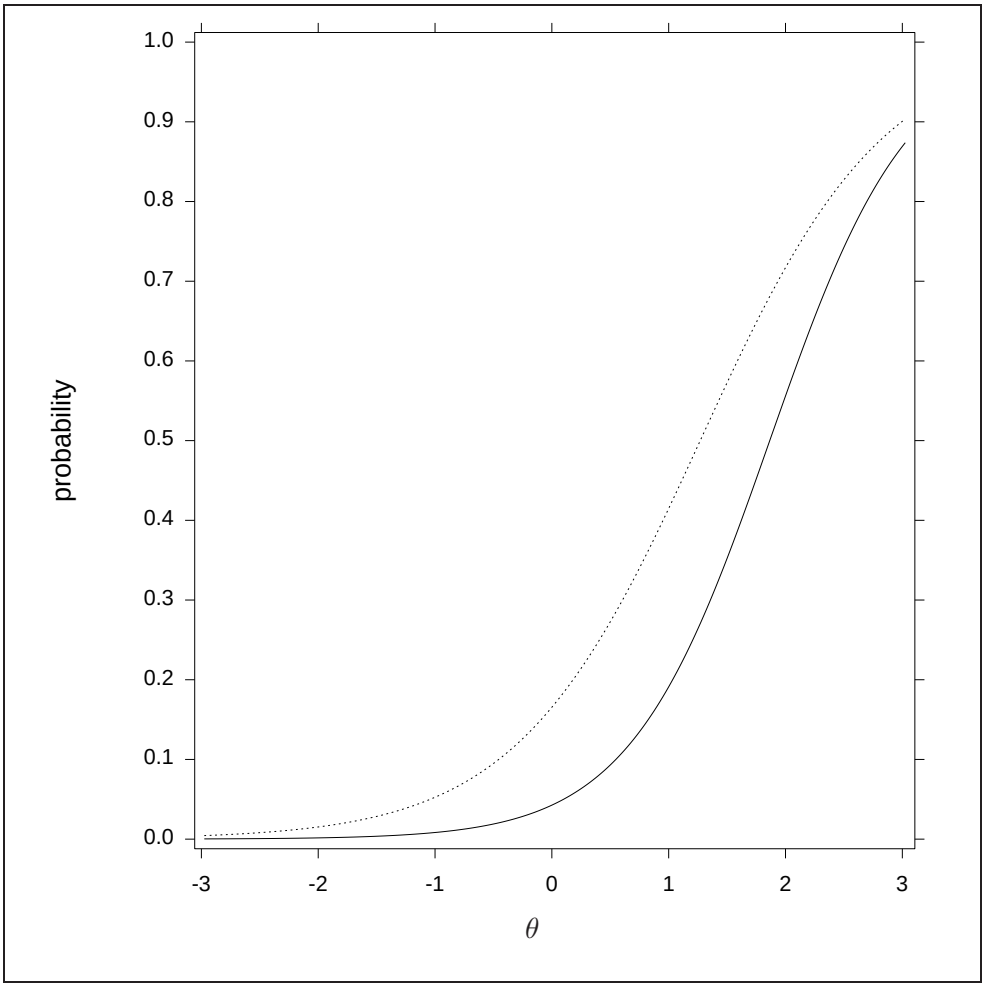

Figure 33: NAEP, item 17 ICCs

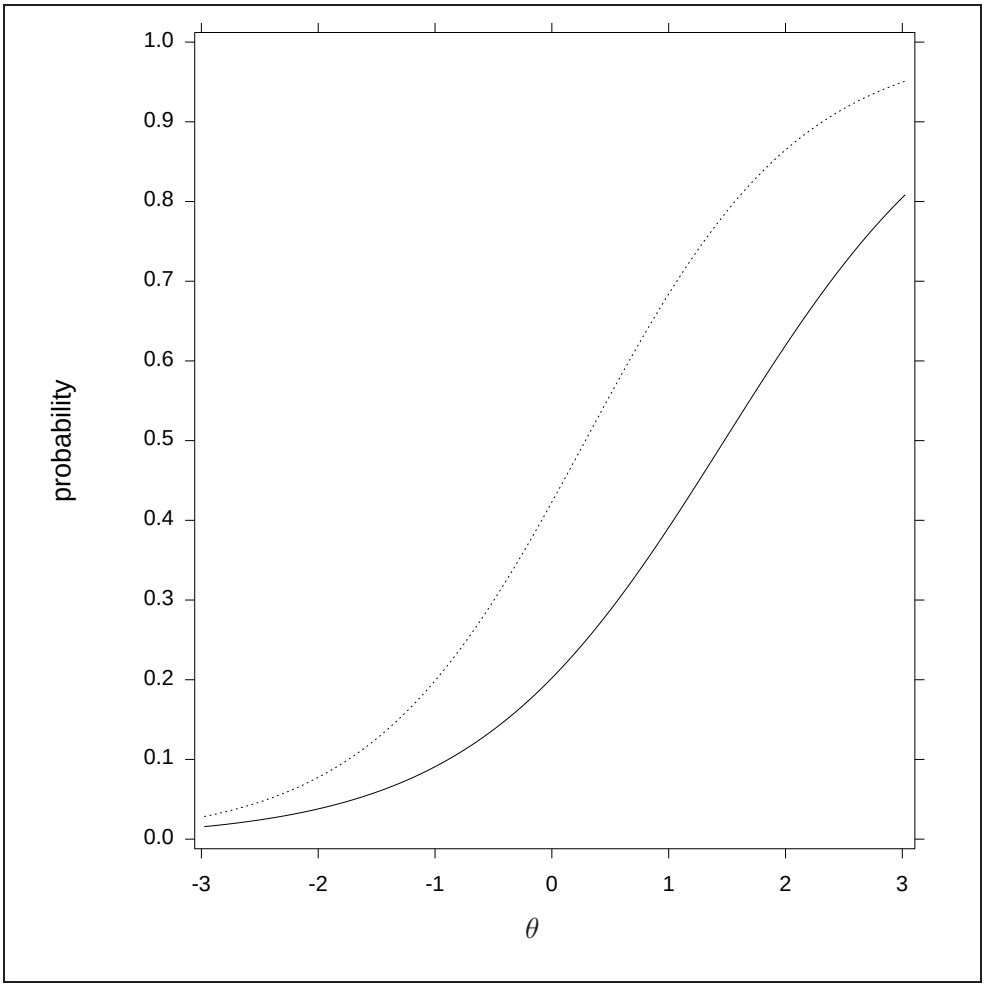

Figure 34: NAEP, item 18 ICCs

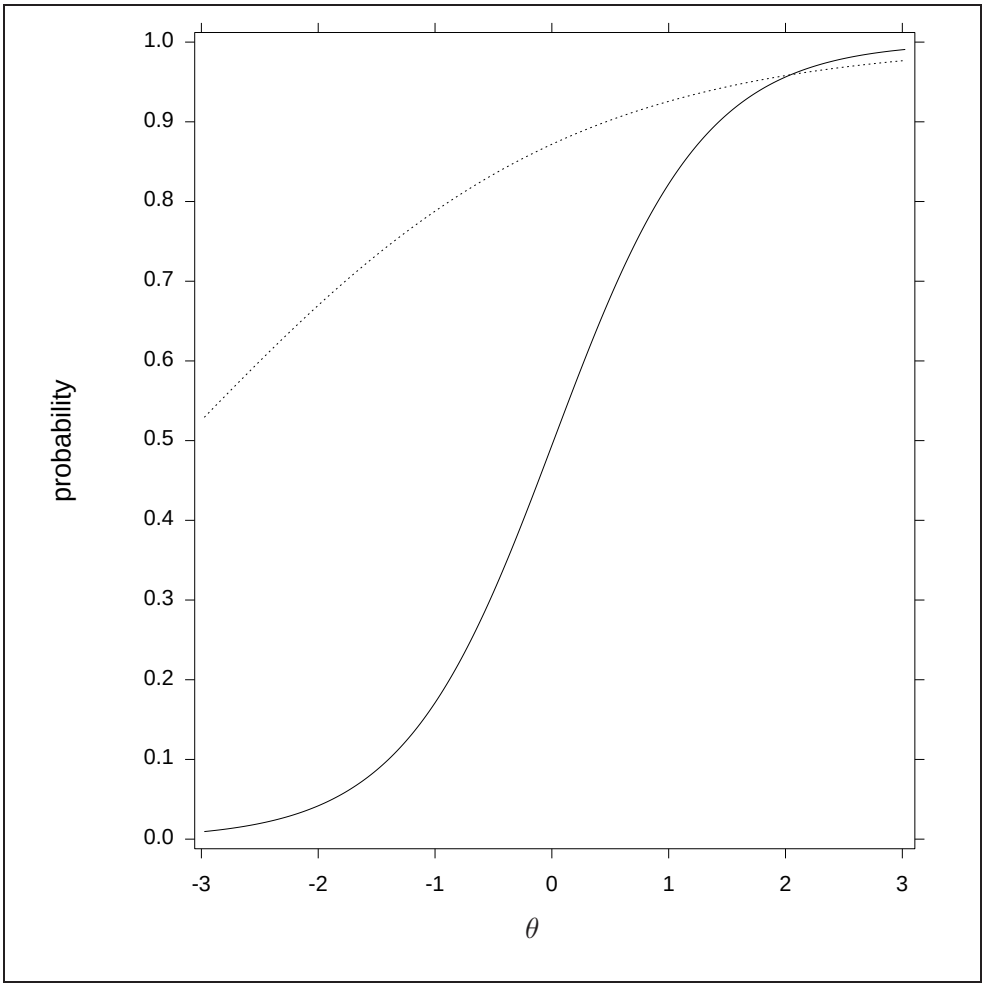

Figure 35: NAEP, item 19 ICCs

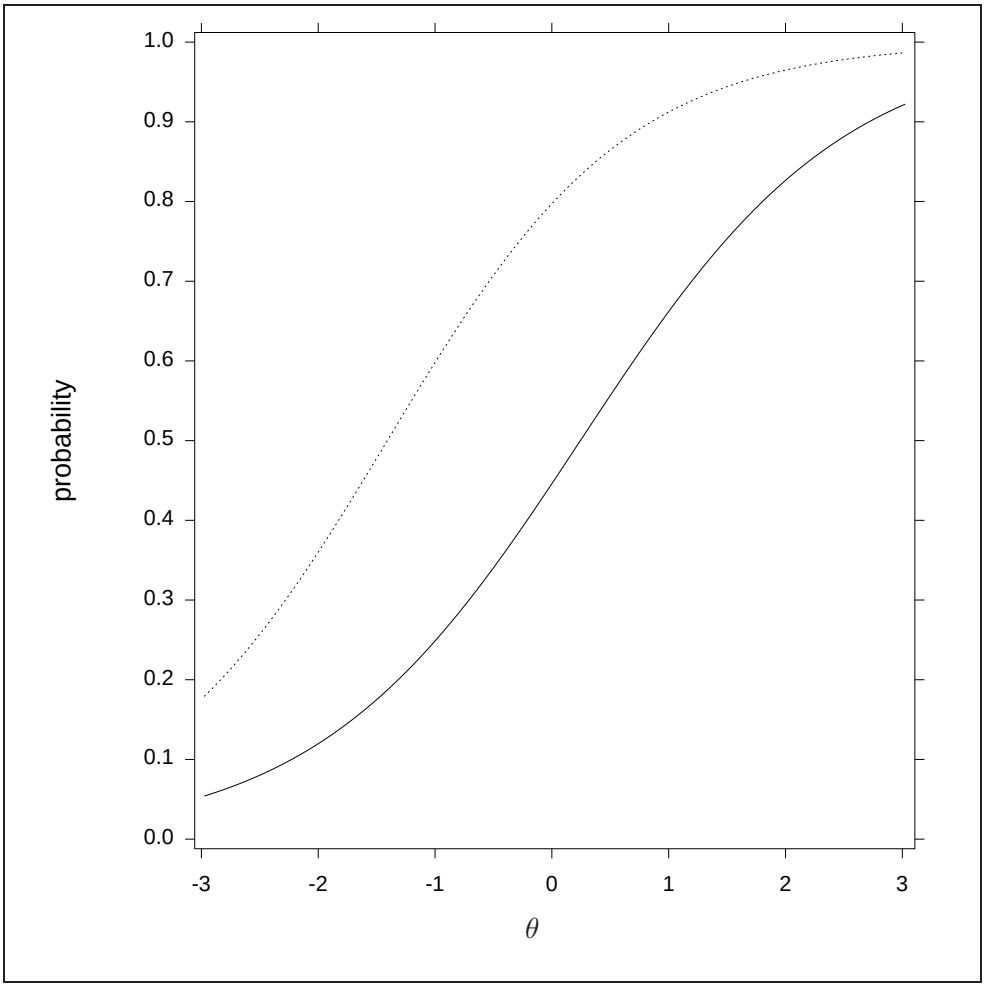

Figure 36: NAEP, item 20 ICCs

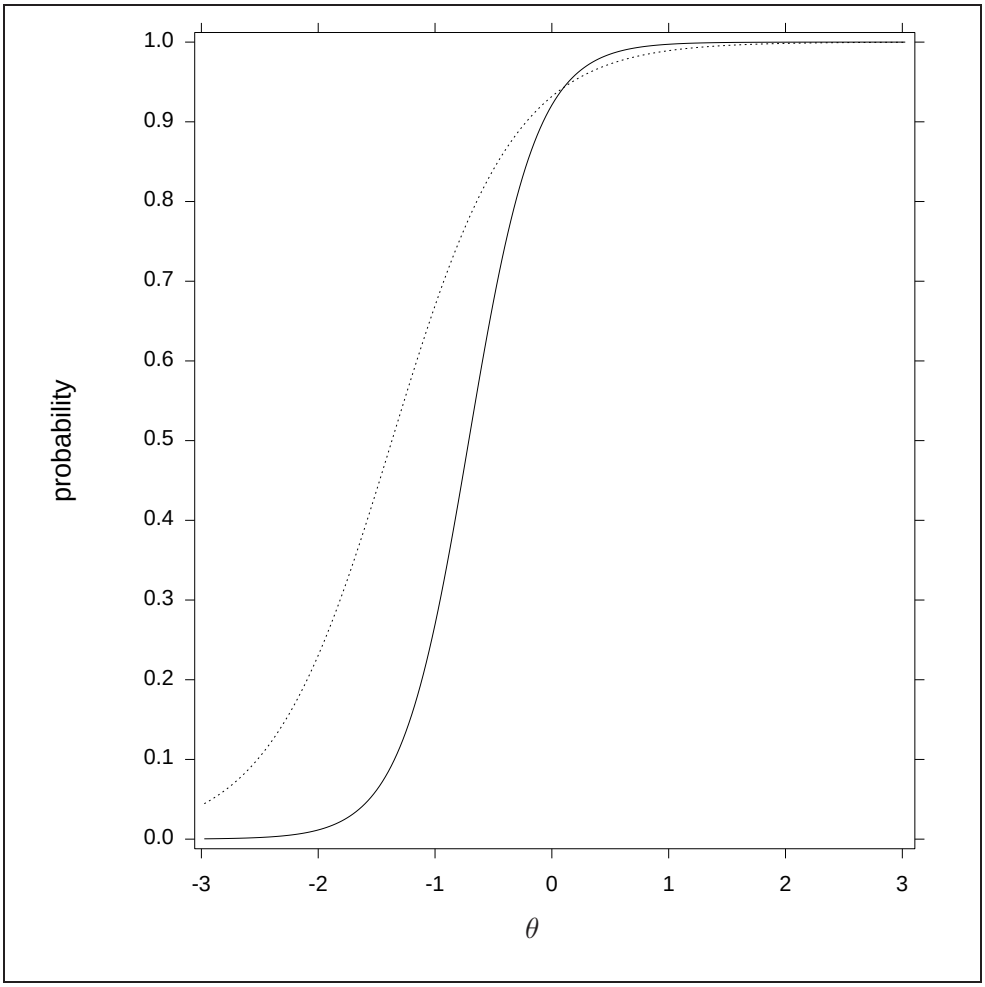

Figure 37: NAEP, item 21 ICCs

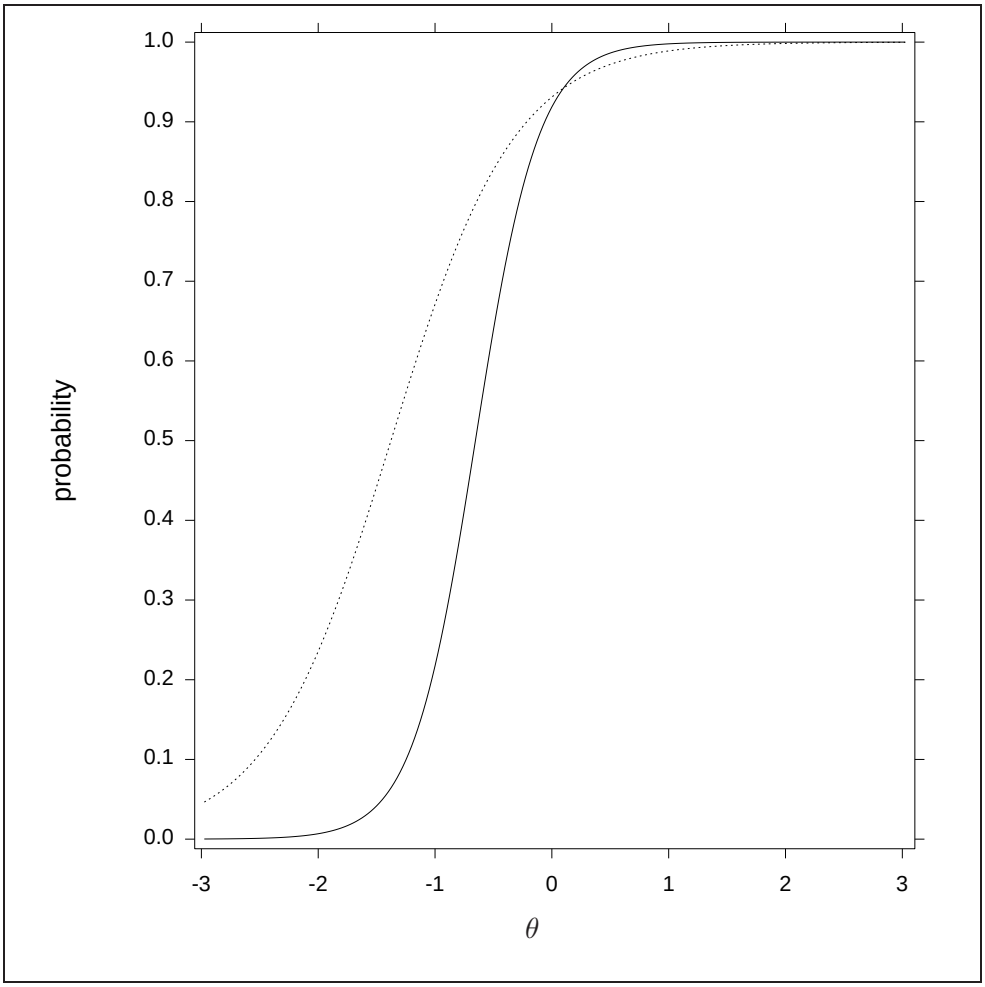

Figure 38: NAEP, item 22 ICCs

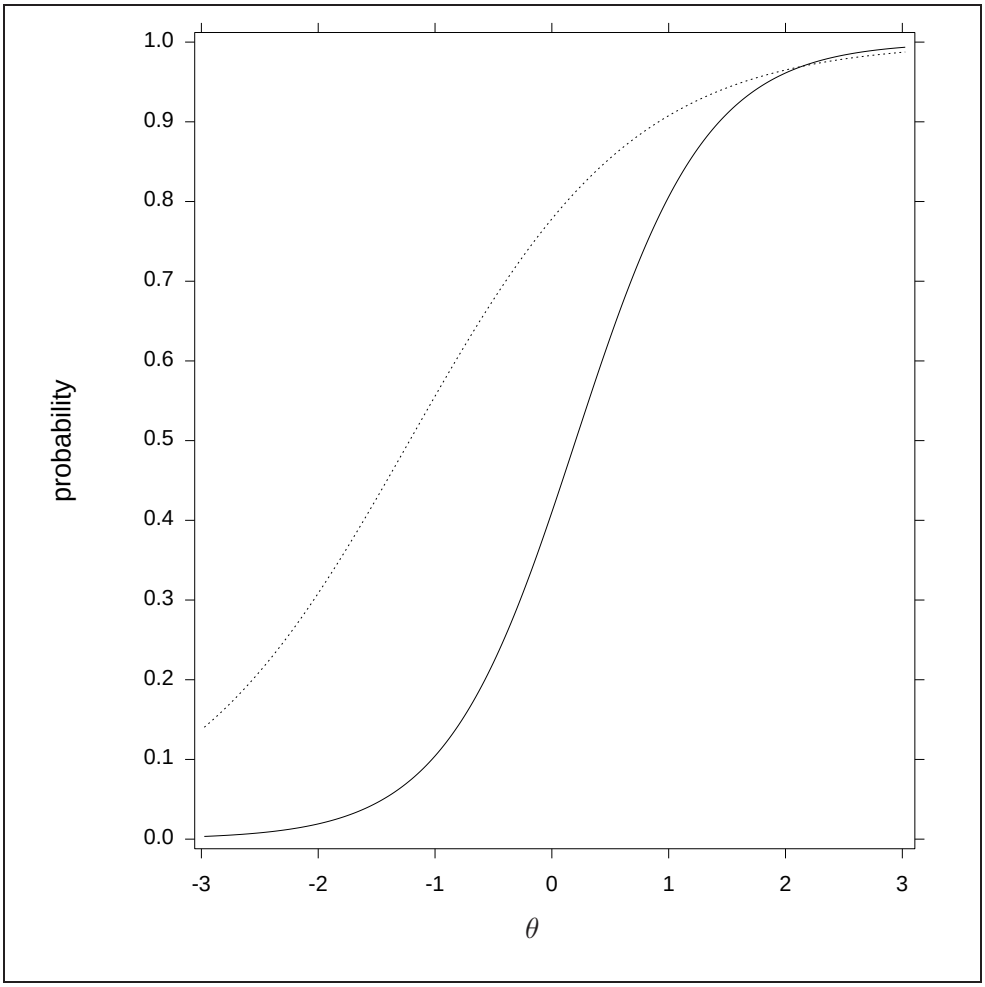

Figure 39: NAEP, item 23 ICCs

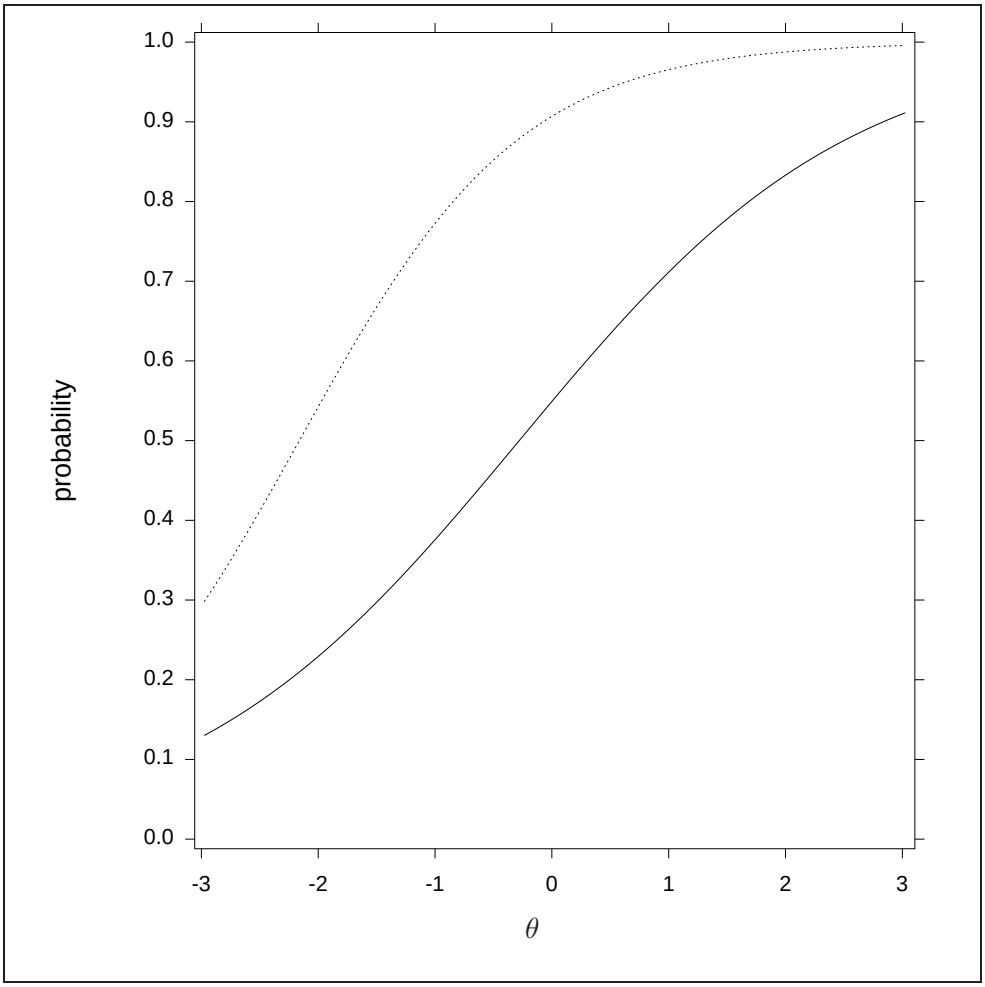

Figure 40: NAEP, item 24 ICCs

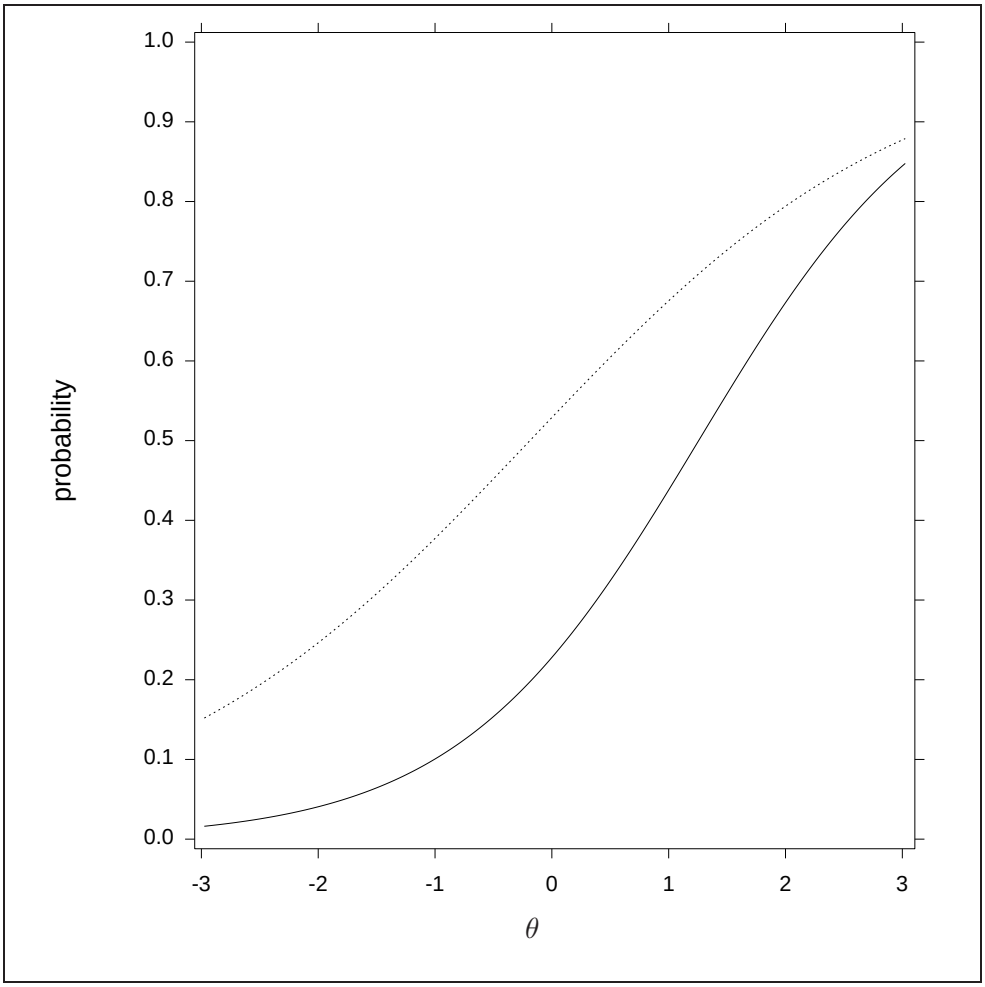

Figure 41: NAEP, item 25 ICCs

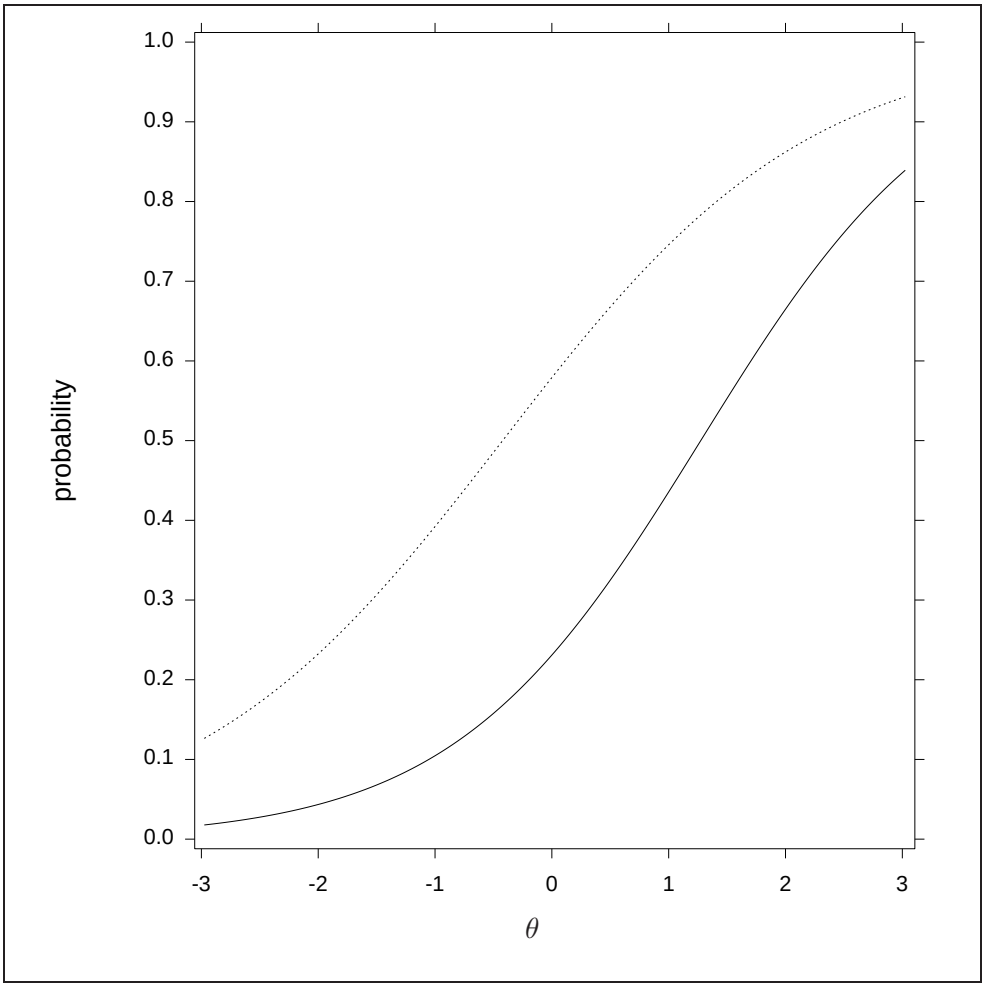

Figure 42: NAEP, item 26 ICCs

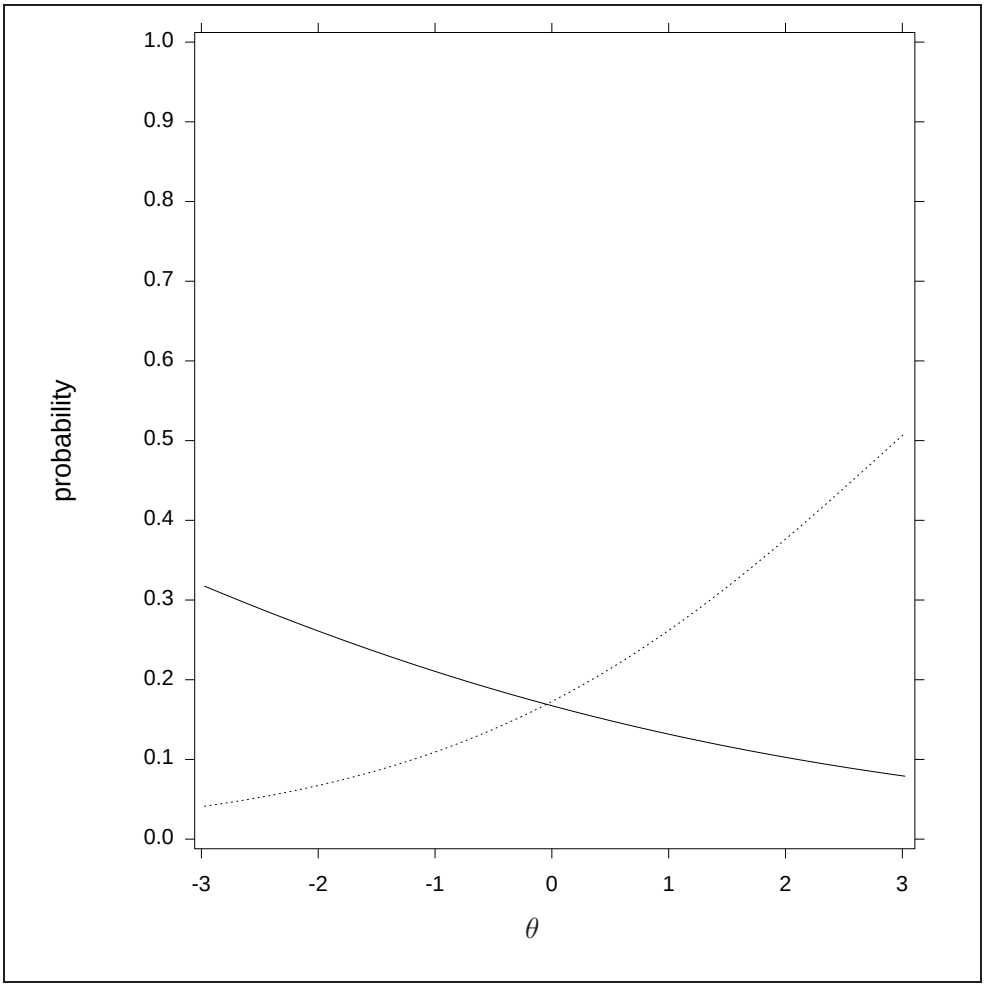

Figure 43: NAEP, item 27 ICCs

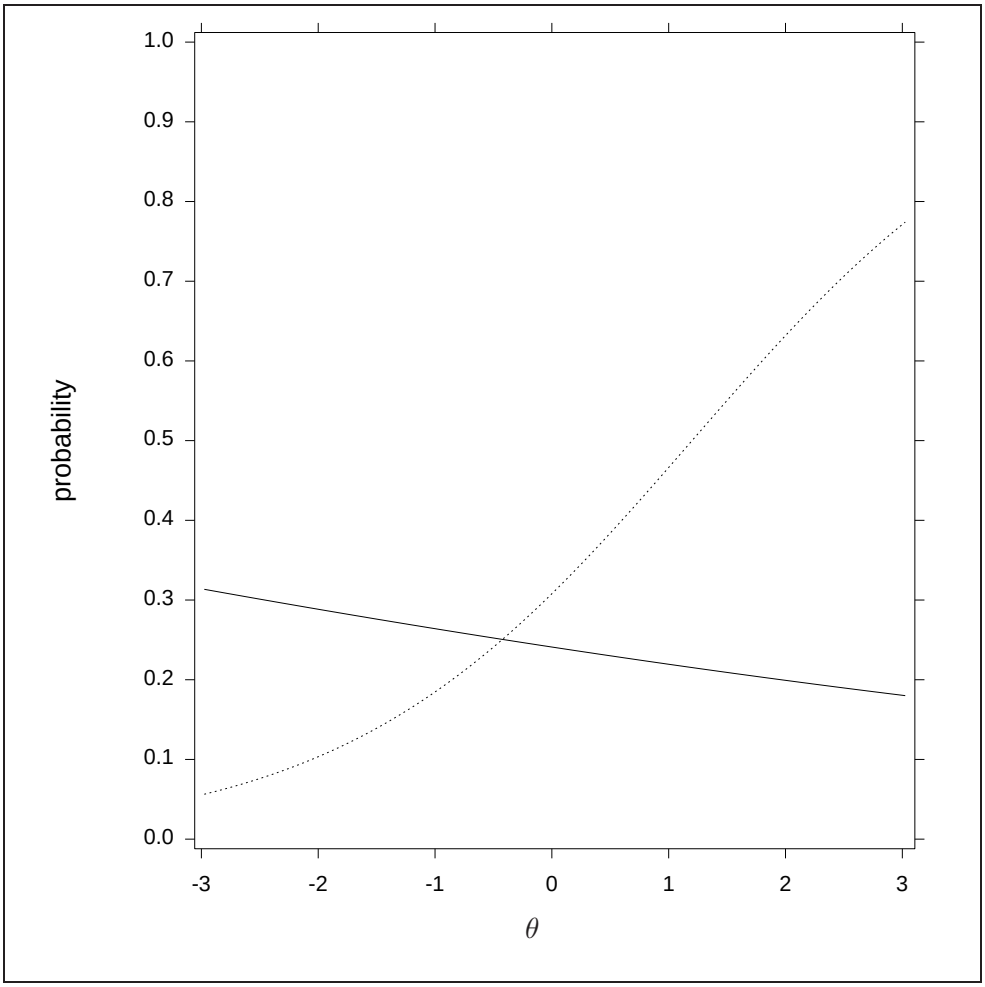

Figure 44: NAEP, item 28 ICCs

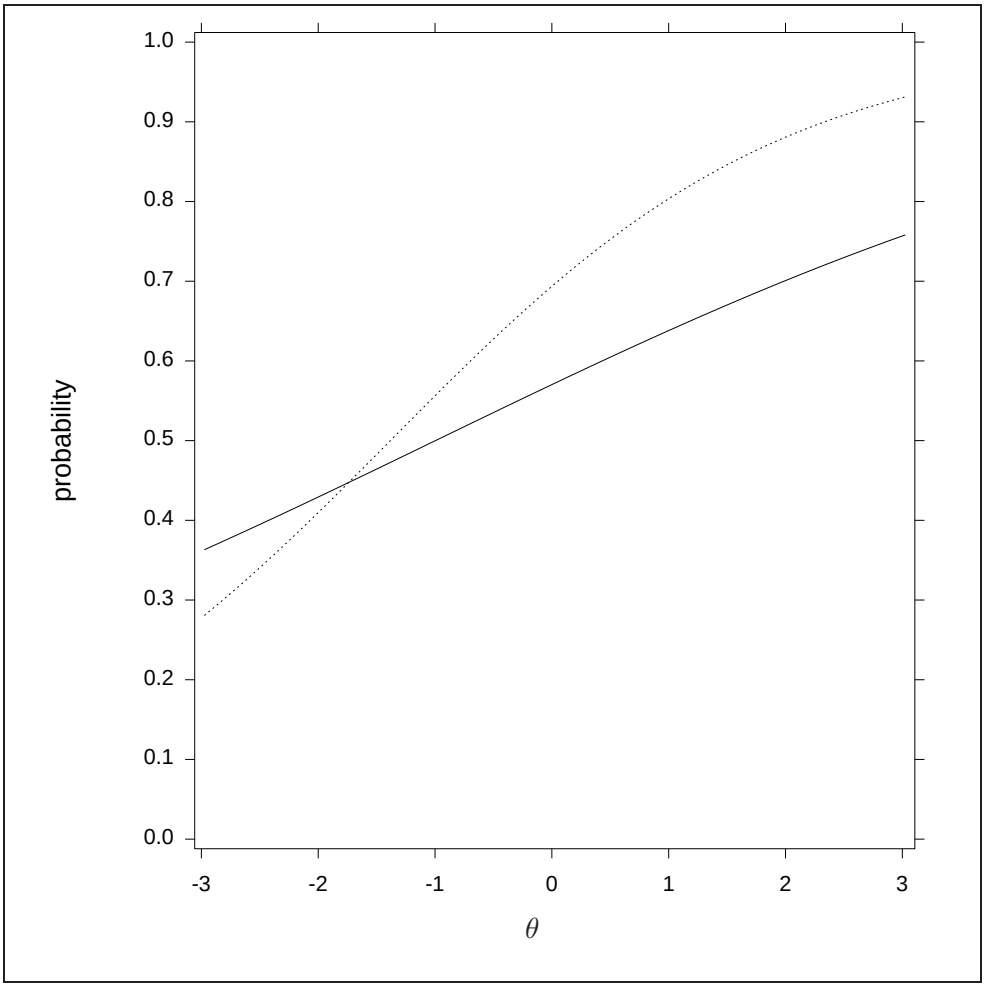

Figure 45: NAEP, item 29 ICCs

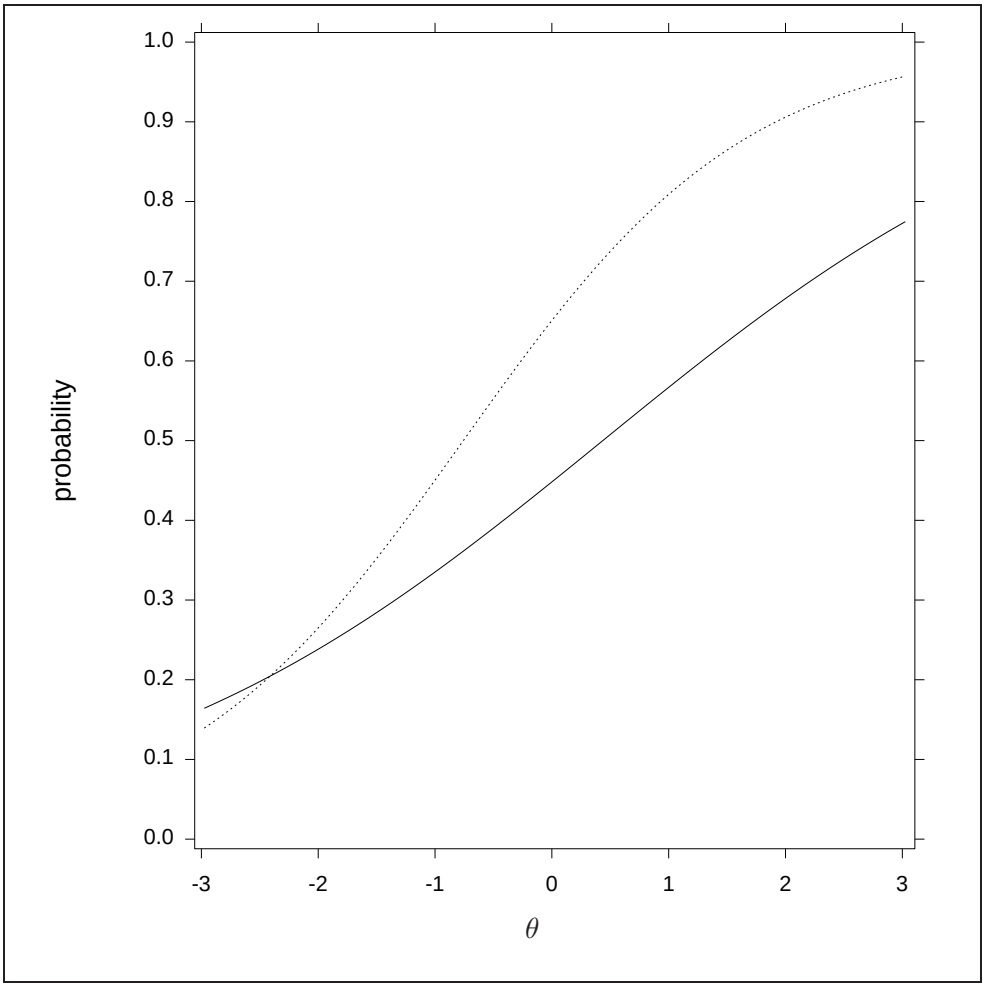

Figure 46: NAEP, item 30 ICCs