Investigation of the ability distribution in the NAEP 1986 Math survey

June 17, 2006

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 Federal Statistics Program

Contents

1 Aim of the project

The aim of this project was to assess the information about the student ability distribution in the Knowledge and Skills subscale of the 1986 NAEP Math survey for Grade 3/Age 9 students, through model fitting with both parametric and nonparametric models.

2 Summary

Several parametric distributions (including the normal distribution) were fitted to the student ability distribution in three-level analyses of the Knowledge and Skills subscale. The heavily skewed log cubic distribution fitted best amongst these, and substantially better than the normal.

Fully nonparametric estimation with the three-level model was not possible,[1] and only two-level nonparametric analysis was possible. This confounded student ability with the school random effect; the resulting composite distribution was left-skewed with a heavy left tail relative to the normal distribution.

Changes in reporting group parameter estimates between the normal distribution and the best-fitting log cubic distribution were small, the largest being 0.5 of an SE. Most changes were between 0.1 and 0.3 SEs. Changes in SEs were very small. Parameter estimates and SEs in the two-level nonparametric analysis were similarly stable.

We conclude that, as for the 2PL model analysis of simulated binary item data, Gaussian quadrature appears so far to work well on real NAEP test data, if the interest is in the reporting group parameter estimates and SEs. However, if interest is in the percentiles of the ability distribution the nonparametric estimate is unsuitable, and reliance on parametric ability distributions appears essential.

3 Background

Current analysis of the NAEP math survey requires a distribution of ability in the fitting of the large conditioning model. The model-based ML approach also requires a distributional assumption to evaluate the likelihood. The usual assumption is a *normal* distribution, which is convenient as it allows the use of Gaussian quadrature to compute the likelihood. This assumption raises two important questions: is the assumption correct? – and if not, does it matter?

An earlier report under Study 1.3.301.2: *Identification of Ability Distributions in IRT models for NAEP items* showed that in simulation studies in

[1] A software bug in Gllamm, so far not corrected, allowed only the highest level distribution to have more than two estimated quadrature points.

the 2-level 2PL model, reporting group estimates and standard errors were very robust to different shapes of the true ability distribution.

So for a unidimensional ability underlying all items, the true ability distribution was *essentially irrelevant* to the estimation of upper-level (reporting group) effects in the 2PL model, and thus the normal distribution was *a convenient computational assumption*, not a restrictive model. A further consequence was that there is no need to use the more computationally intensive Bock-Aitkin semi-nonparametric method currently used to estimate the ability distribution; Gaussian quadrature methods assuming a normal ability distribution could be used quite generally to estimate reporting group differences over different true ability populations for the 2PL model.

4 The ability distribution

However, NCES publishes *percentiles* of the ability distribution (on the NAEP reporting scale) by reporting groups, and so the dependence of percentiles on variations in the ability distribution is also an important question.

Current methods use the percentiles of the distribution of *plausible values* by reporting group. The latter are generated from a normal distribution of ability across all students, though there is evidence from the semi-nonparametric estimation that this distribution is not normal.

A difficulty with this approach is its circularity: percentiles are computed from plausible values which are themselves generated from a normal ability distribution which is not clearly appropriate. In a fully model-based framework percentiles depend *strongly* on the form of the ability distribution, which therefore needs to be investigated. There are two possible approaches:

- fit parametric distributions more general than the normal to find the most appropriate;
- *estimate* the distribution semi- or fully non-parametrically.

For the first approach, we considered the following distribution families:

- The extreme value and reversed extreme value, with fixed left and right skew;
- the log cubic distribution of Holland and Thayer (2000), with adjustable skewness left or right determined by a skewness parameter;
- the symmetric t distribution, with adjustable tail heaviness determined by the degrees of freedom parameter.

Since these distributions, like the normal, do not give an analytic likelihood function, they require numerical integration analogous to Gaussian quadrature, but with masses determined by the specific densities. For direct

comparison with Gaussian quadrature, the continuous densities were converted to 8-point discrete distributions on the Gaussian quadrature masspoints z_k, $k = 1, \ldots, 8$ for the numerical integrations. For each candidate distribution density $f(z)$, the masses f_k at the z_k were calculated by

$$f_k = f(z_k) / \sum_{k=1}^{8} f(z_k).$$

For the second approach, we used fully nonparametric estimation, fitting up to 10 masspoints and estimating both the locations of the masspoints and their probability masses.

4.1 Data analysis for the 1986 NAEP Math survey

There were 21,287 Grade 3/Age 9 students in the 1986 Math survey, but only about half of these had responses on any of the items on the Knowledge and Skills scale, so the "full" data set for this scale has 10,463 students clustered in 440 schools. The schools are themselves clustered in 94 PSUs, but the PSU level is ignored in this analysis as the variance component at this level is very small.

The number of students per school varied from 5 to about 45, with an average of 24, and there was an average of 7 items answered per student. We used a minimal set of reporting variables: sex, race (6 levels), region (4), size and type of community (stoc, 7) and parents education level (pared, 6), to give us some feel for the results. We used a main effect model with 20 dummy variables for these categorical variables.

We ran a 3-level analysis using Gllamm in Stata. To increase the speed of convergence for the non-normal analyses we used as starting values the parameter estimates from the normal (Gaussian quadrature) 3-level analysis. A full discussion of the normal analysis is give in the report *Multi-level model analysis of the Knowledge and Skills scale of the NAEP 1986 math data.*

4.2 Parametric distributions

For each candidate distribution, the three-level main effect regression model was fitted by maximum likelihood, and the maximized log-likelihoods are given in the table below.[2] For the t and log cubic families, the log-likelihood at the MLE of the additional parameter can be compared with that for the normal distribution by the usual large-sample likelihood ratio test: the difference in $-2 \log L_{max}$ is compared to the χ^2_1 percentage points. For the t distribution the difference (at 2 or 3 df) is 32.90, and for the log cubic (at $\hat{\gamma} = -0.075$) it is 41.68. The 1% point of χ^2_1 is 6.64: the above values are

[2]We ran the log cubic analysis over a finer and wider grid but report here only the relevant range.

far beyond any critical values from the χ^2_1 distribution. There is no question that these distributions are more appropriate than the normal for the ability distribution.

Comparing the t family to the log cubic family is more complex because these families are not nested, but the log-likelihood improvement of 4.39 for the best log cubic strongly suggests skew in the distribution; this is supported by the extreme value distribution result, with its very heavy left tail, and is validated by the nonparametric analysis which follows.

Table 1: Log-likelihoods for parametric models

normal	-39,930.05
extreme value	-39,918.44
t_1	-39,916.36
t_2	-39,913.60
t_3	-39,913.60
t_4	-39,914.54
log cubic(+.02)	-39,937.52
log cubic(+.01)	-39,933.89
log cubic(0.00)	-39,930.05
log cubic(-.01)	-39,927.61
log cubic(-.02)	-39,924.70
log cubic(-.03)	-39,921.72
log cubic(-.04)	-39,918.53
log cubic(-.05)	-39,915.10
log cubic(-.06)	-39,911.70
log cubic(-.07)	-39,909.30
log cubic(-.075)	-39,909.21
log cubic(-.08)	-39,910.55

Of these distributions, the heavily skewed log cubic (with the ML estimate of $\gamma = -0.075$) fitted the data best, but *all* the alternative distributions (with negative skew or heavier tails) fitted much better than the normal. The somewhat better fit of the log cubic than the t family points to heavier tails on the left than the right: a longer tail of low-ability than high-ability candidates.[3]

4.3 Nonparametric distributions

Nonparametric estimation (discrete) of the ability distribution may be achieved in two ways:

[3]Another possible reason for the skewed distribution is that the 3PL model was not used; accounting for guessing might change the form of the ability distribution.

- *semi*-nonparametrically, by estimating the probability ordinates π_k of the distribution on a fixed grid of ability values z_k, typically an equally-spaced set of 20-40 values covering a range like -5 to +5;
- *fully* nonparametrically, by estimating *both* the probability ordinates π_k *and* their locations z_k.

There is a finite limit to the resolution – the number of estimable parameters – of the discrete estimate (Aitkin 1996, 1999), which for binary response data on all respondents is of the order of half the number of scale items (15 here). Since a long-tailed distribution could be expected, we preferred to estimate both locations and ordinates fully nonparametrically rather than to fix the locations and estimate only the ordinates.

However, in trying to use Gllamm to estimate a reasonable number of masspoint locations and ordinates at the student level we encountered a program bug, which did not allow more than a two points to be estimated at the lower (student ability) level. This bug does not affect the upper level estimation, but our aim was to estimate the student abilty distribution, not the school ability distribution. We consulted Sophia Rabe-Hesketh about the bug and at the date of writing, no patch for it has been given.

To obtain a nonparametric estimate with more than two points, we were therefore restricted to the two-level model, ignoring the school level. The consequence of this constraint is that the school random effects are *confounded* with the student ability random effects: the distribution estimated nonparametrically is a *composite* of the student ability and school random effect distributions. The model for student i with ability z_i in school k with random effect η_k, answering item j with probability p_{ijk} of a correct answer,

$$\begin{aligned} \operatorname{logit} p_{ijk} \mid z_i, \eta_k &= \alpha_j + \beta_j z_i \\ z_i \mid \eta_k &\sim N(\boldsymbol{\gamma}'\mathbf{x}_i + \eta_k, \sigma^2) \\ \eta_k &\sim N(0, \sigma^2_{sch}) \end{aligned}$$

becomes, if the school level is suppressed,

$$\begin{aligned} \operatorname{logit} p_{ij} \mid z_i &= \alpha_j + \beta_j z_i \\ z_i &\sim N(\boldsymbol{\gamma}'\mathbf{x}_i, \sigma^2 + \sigma^2_{sch}), \end{aligned}$$

so the student-level variance is inflated by the school level variance; also the correlation structure of responses of students within the same school is lost. A further consequence is that the three-level parametric models fitted above are not comparable to the two-level model below.

To assess the effect of increasing the number of masspoints, we fitted 4-, 6-, 8- and 10-point ability distributions by nonparametric maximum likelihood. More than seven points (with 13 distribution parameters: 7

masspoints and 6 probability masses) could not be identified – any additional points gave the same maximized log-likelihood and degenerated to the 7-point estimate.

4.3.1 Masspoint estimates

The sets of estimated masspoints and masses are given in the table below, rescaled to have mean zero and variance 1, together with the Gaussian quadrature masspoints and masses, and the maximized log-likelihoods ℓ.

K	4		6		8		10		GQ(8pt)	
k	z_k	pi_k	z_k	pi_k	z_k	pi_k	z_k	pi_k	z_k	pi_k
1							-3.897	.012		
2					-3.767	.013	-1.937	.066	-4.146	.0001
3			-2.469	.058	-1.919	.065	-0.850	.170	-2.803	.010
4	-2.075	.102	-1.010	.178	-0.859	.187	-0.843	.019	-1.637	.117
5	-0.561	.361	-0.237	.292	-0.174	.275	-0.168	.272	-0.539	.373
6	0.542	.440	0.566	.330	*0.580	.326	0.580	.331	0.539	.373
7	1.810	.097	1.142	.111	1.133	.103	1.137	.083	1.637	.117
8			2.523	.031	2.465	.031	1.240	.018	2.803	.010
9							1.317	.0001	4.146	.0001
10							2.490	.029		
l	-40,067.95		-40,049.00		-40,046.98		-40,046.98		-40,077.26	

* Two points coincident here

The maximized log-likelihood increased with increasing number of masspoints up to 7 points, when it remained unchanged with more points. The 8-point distribution has a degeneracy at 0.580, where two points are coincident. Increasing the number of masspoints to 10 gives three near-degeneracies, with the points at -0.850 and -0.843, and those at 1.137 and 1.240, almost coincident, and the point at 1.317 having almost no mass: no more than 7 points can be identified. Combining the two points near -0.85 and 1.2, and eliminating the point at 1.317 gives a distribution very similar to the 7-point estimate.

The likelihood is extremely flat in the estimated masspoints and masses, and the Gauss-Newton algorithm is extremely slow to converge for all the nonparametric estimates, showing the difficulty of identifying the ability distribution for these data. It had not converged for any of them after 30 iterations, though the parameter estimates, standard errors and log-likelihoods were then unchanged to 3 decimal places.

For reasons given below the estimated masses and masspoints are shown without standard errors, though these are computed by the package. The estimated 7-point distribution is in no sense precise; however it is clearly both left-skewed and heavy left-tailed.

This distribution is unfortunately not comparable with the parametric distributions, as mentioned above, but we show for reference the 8-point parametric and 7-point nonparametric cdfs on the same (probability) scale in Figure 1, with all distributions scaled to have mean zero and variance 1.

The two-level nonparametric estimate is in red, the best three-level parametric model (the log cubic) is in dark blue, the normal in green, the t_4 in dark green and the extreme value in blue. The heavy left, and light right, tail of the nonparametric estimate are clearly visible.

4.4 Inference about percentiles

The maximized log-likelihoods for the continuous models clearly identified the left-skewed log cubic distribution as the best-fitting among those that we examined. Percentiles of these non-normal continuous distributions can be obtained directly from the analytical form of the density, or by numerical computation of the cdf. The report *Identification of Ability Distributions in IRT models for NAEP items* gave a table of moments and major percentiles for the log cubic family with various degress of skewness, computed directly from the density.

Inference about the percentiles of these distributions, in terms of estimates and standard errors, follows by standard delta-method theory from the estimates and covariance matrix of the model parameters.

However the nonparametric estimate does not lend itself to such inference, because the estimated distribution is discrete, with discrete jumps in the cdf. Thus percentiles are not generally available except by assumptions of linearity or other simple forms, or smoothing, between the estimated masspoints. Even if this assumption is made, the standard errors and covariances of the masspoints and their probability masses make very complicated the calculation of a standard error for an estimated percentile.

Thus the estimation of ability distribution percentiles requires a parametric model, as we found in the simulations. However, the nonparametric estimate is very useful in identifying the type of skew and tail behaviour of the ability distribution, and so can point to a suitable family of continuous distributions which can be used with some confidence for parametric model inference about the percentiles.

5 Effect of the different ability distributions on reporting group estimates

A natural question is the extent to which the form of the ability distribution affects the reporting group estimates. We show in the table below the maximum likelihood estimates and standard errors of the reporting group effects for the 3-level model, for the normal and the other distributions using 8-point quadrature. The model fit improves across the table (increasing maximized log-likelihood).

Changes in parameter estimates were small, the largest (between GQ and the best-fitting log cubic) being 0.5 of an SE for `race4`. Most changes were between 0.1 and 0.3 SEs. Changes in SEs were very small.

ML estimates for 8pt 3-level normal and ev, t_2, and log cubic($\gamma = -0.075$) distributions

	GQ	SE	ev	t2	log cubic
sex2:	.012(.028)		.007(.028)	.008(.028)	.005(.027)
race2:	-.667(.047)		-.672(.045)	-.673(.046)	-.668(.046)
race3:	-.460(.043)		-.449(.042)	-.458(.043)	-.447(.043)
race4:	-.203(.117)		-.108(.115)	-.211(.123)	-.135(.118)
race5:	-.471(.093)		-.479(.091)	-.485(.093)	-.474(.092)
race6:	-.200(.752)		-.190(.586)	-.202(.695)	-.167(.625)
regi2:	-.020(.077)		-.029(.073)	.023(.074)	-.007(.076)
regi3:	-.172(.074)		-.209(.073)	-.167(.072)	-.196(.074)
regi4:	-.182(.069)		-.169(.066)	-.149(.067)	-.158(.069)
stoc2:	-.201(.113)		-.227(.109)	-.175(.111)	-.202(.111)
stoc3:	.497(.116)		.505(.111)	.456(.114)	.498(.113)
stoc4:	.150(.106)		.124(.109)	.118(.108)	.116(.109)
stoc5:	.158(.112)		.107(.111)	.166(.110)	.118(.111)
stoc6:	.092(.097)		.090(.102)	.077(.101)	.075(.102)
stoc7:	-.019(.095)		-.023(.098)	-.040(.097)	-.042(.098)
pred2:	-.179(.206)		-.162(.187)	-.133(.198)	-.167(.181)
pred3:	.045(.200)		.037(.181)	.098(.192)	.039(.174)
pred4:	.398(.205)		.401(.186)	.445(.197)	.399(.180)
pred5:	.382(.198)		.372(.179)	.420(.190)	.371(.171)
pred6:	.027(.197)		.017(.178)	.068(.189)	.016(.171)
log L:	-39,930.05		-39,918.44	-39,913.60	-39,909.21

As described above, because of the Gllamm bug we are unable to assess this question for the three-level model and could examine it only for the two-level model. We show in the table below the maximum likelihood estimates of the reporting group effects for Gaussian quadrature and for the 4-, 6-, 8- and 10-point nonparametric estimates. The model fit improves with more points up to 8 (of which only 7 are distinct), then remains constant.

ML estimates for 2-level normal and nonparametric ability distributions

	GQ(8pt) SE	NP (4pt)	(6pt)	(8pt)	(10pt)
sex2:	.015(.028)	.010(.028)	.015(.028)	.012(.028)	.013(.029)
race2:	-.792(.041)	-.776(.041)	-.782(.041)	-.786(.041)	-.786(.042)
race3:	-.514(.041)	-.502(.040)	-.496(.041)	-.499(.041)	-.499(.041)
race4:	-.240(.114)	-.180(.113)	-.170(.125)	-.188(.120)	-.187(.122)
race5:	-.588(.092)	-.582(.091)	-.569(.093)	-.573(.093)	-.574(.095)
race6:	-.380(.855)	-.336(1.14)	-.057(.725)	-.096(.711)	-.106(.713)
regi2:	.003(.046)	.014(.046)	.014(.045)	.016(.046)	.015(.046)
regi3:	-.219(.046)	-.199(.046)	-.205(.046)	-.208(.046)	-.208(.047)
regi4:	-.193(.042)	-.175(.041)	-.170(.042)	-.174(.042)	-.174(.045)
stoc2:	-.126(.075)	-.103(.074)	-.139(.074)	-.131(.074)	-.129(.079)
stoc3:	.475(.074)	.460(.074)	.454(.072)	.455(.072)	.456(.079)
stoc4:	.133(.072)	.152(.072)	.106(.072)	.117(.072)	.119(.080)
stoc5:	.121(.074)	.134(.073)	.107(.072)	.111(.072)	.113(.078)
stoc6:	.031(.067)	.031(.067)	.028(.066)	.034(.066)	.036(.070)
stoc7:	-.055(.066)	-.042(.065)	-.062(.064)	-.056(.064)	-.054(.070)
pred2:	-.205(.152)	-.250(.150)	-.226(.153)	-.235(.153)	-.233(.158)
pred3:	.061(.141)	.011(.140)	.018(.143)	.013(.142)	.016(.143)
pred4:	.460(.149)	.405(.147)	.430(.150)	.432(.150)	.435(.150)
pred5:	.453(.138)	.382(.136)	.416(.140)	.412(.139)	.414(.140)
pred6:	.047(.137)	-.011(.135)	.008(.139)	.008(.138)	.011(.139)
log L:	-40,077.26	-40,067.95	-40,049.00	-40.046.92	-40046.98

Parameter estimates were again very stable, the largest change (between GQ and the 8-point NPML estimate) being 0.5 SE for `race4` and `region4`. Most changes were between 0.1 and 0.3 SEs. Standard errors were very

similar for all the analyses, with a slight increase as the number of estimated masspoints increased.

We were unable to determine whether, in a full three-level nonparametric analysis, the student ability distribution would show the same structure as in the two-level analysis, or whether the school random effect distribution induces heavy left-tail behaviour, or both.

6 Conclusion

It is clear that for this data set the parameter estimates from the Gaussian quadrature analysis using the normal model for ability were very little affected by the actual distribution, which was left-skewed with a heavy left tail. This result for the scale we examined was consistent with the earlier simulation studies we reported with the 2-level model (*Identification of Ability Distributions in IRT models for NAEP items*). It supports our conclusion that Gaussian quadrature appears to be sufficient for robust inference about reporting group parameters in the 2PL model analysis of binary item data, and it appears so far to work well on real NAEP test data also.

The nonparametric estimation provides standard errors for the locations and ordinates, but the estimates and standard errors are difficult to use for inference about percentiles, though they help identify a suitable parametric family for the ability distribution. The families of parametric models examined provided a "best-fitting" one in the log cubic family which was clearly superior to the normal distribution, and inference about its percentiles would be straightforward using standard theory and the covariance matrix of the parameter estimates.

7 References

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing 6*, 251-262.
Aitkin, M. (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics 55*, 117-128.
Holland, P.W. Thayer, D.T. (2000) Univariate and bivariate loglinear models for discrete test score distributions. *J. Educational and Behavioral Statistics*, **25**, 133-183.

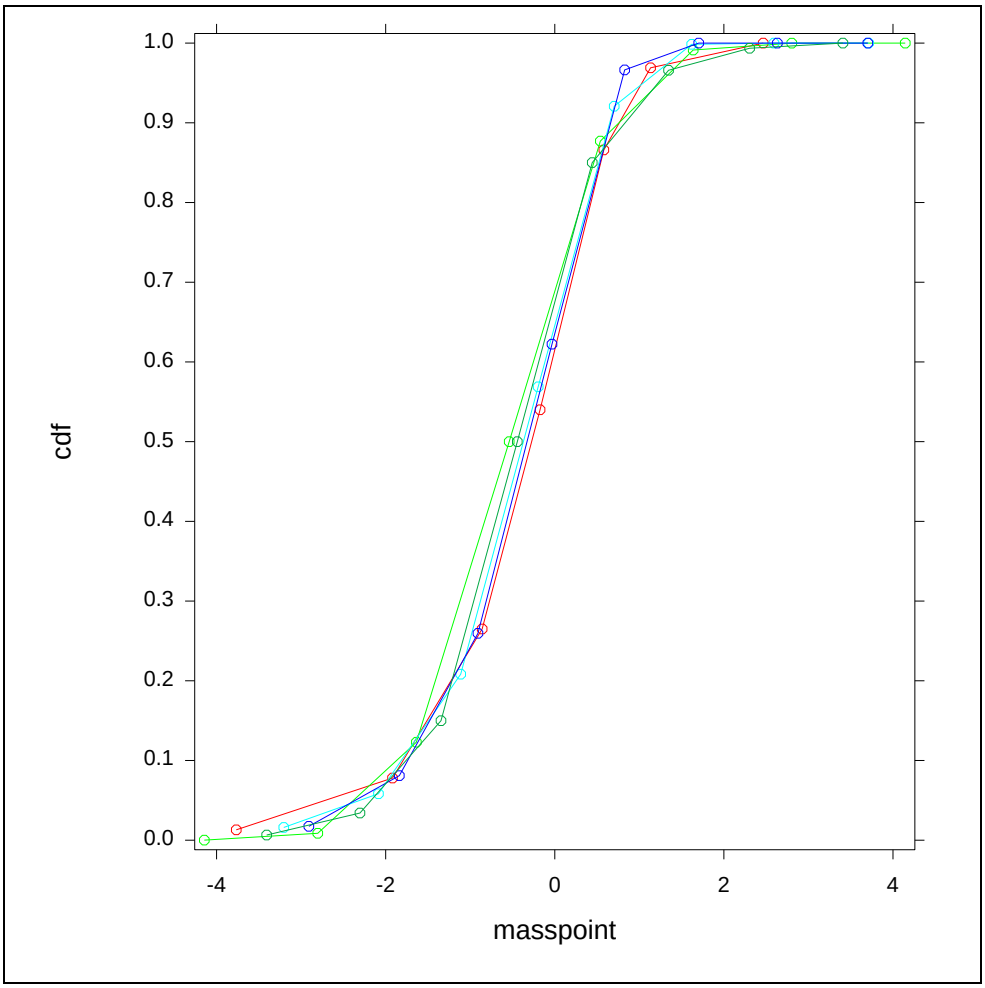

Figure 1: Parametric cdfs and 8-point NPML ability estimates