Supplementary material

IRT for Voting Advice Applications: A multi-dimensional test that is adaptive and interpretable

1 The use of a model based MAP prior

For the adaptive test described in this paper, we recommend the MAP estimate of the ability parameter. If we let $\theta_{j,k}$ denote the ability of respondent j in dimension k, then $\theta_{j,\ell}$, $\ell \neq k$ is the ability of respondent j in a dimension other than k. As an alternative to using a standard normal prior, we suggest a prior that takes advantage of the provisional estimates of $\theta_{j,\ell}$ when we estimate $\theta_{j,k}$. Thus, with regards to $\theta_{j,k}$ we treat $\hat{\theta}_{j,\ell}$ as auxiliary information about the respondent. The prior mean for the MAP estimate of $\theta_{j,k}$ we denote $\widetilde{\theta}_{j,k}$.

For the MAP estimate of test taker j, we use the prior distribution $N(\mu_{j,k} = \widetilde{\theta}_{j,k}, \sigma^2 = 1)$, where $\widetilde{\theta}_{j,k} = g(\hat{\theta}_{j,\ell})$, i.e. $\widetilde{\theta}_{j,k}$ is some function of $\hat{\theta}_{j,\ell}$.

The value for the prior mean $\widetilde{\theta}_{j,k}$ should not be confused with $\hat{\theta}_{j,k}$, which is the MAP estimate that takes both the prior distribution and the response data in scale k into account.

$g(\cdot)$ can be any function that predicts $\theta_{j,k}$ from the ability estimates in the remaining dimensions. In the results section of this paper, we use a simple linear regression for the two-dimensional case. Predicting $\theta_{j,2}$ from the provisional $\hat{\theta}_{j,1}$ then gives

$$\widetilde{\theta}_{j,2} = \beta_0 + \beta_1 \hat{\theta}_{j,1} \tag{1}$$

This is equivalent to

$$\widetilde{\theta}_{j,2} = \overline{\hat{\theta}}^*_{.,2} + r^*_{\hat{\theta}_{.,1},\hat{\theta}_{.,2}} \cdot \frac{s^*_{\hat{\theta}_{.,2}}}{s^*_{\hat{\theta}_{.,1}}} \left(\hat{\theta}_{j,1} - \overline{\hat{\theta}}^*_{.,1} \right), \tag{2}$$

where $r^*_{\hat{\theta}_{.,1},\hat{\theta}_{.,2}}$ is the correlation between the estimates $\hat{\theta}^*_{.,1}$ and $\hat{\theta}^*_{.,2}$, $s^*_{\hat{\theta}_{.,k}}$ is the sample standard deviation of $\hat{\theta}^*_{.,k}$ and $\overline{\hat{\theta}}^*_{.,k}$ is the mean of the ability estimates on scale k. Estimates marked with * are estimates from previous data.

In implementation, we could in a first phase use a standard normal prior to estimate the ability of early participants who complete the full test. In preparation of a second phase, we can then fit a model $g(\cdot)$ from these early estimates. The model will allow for better priors, and hence better estimates for subsequent users. If we want to, we can keep updating the model $g(\cdot)$ throughout the lifetime of the test.

With a large enough sample of previous test takers, the model-based prior mean can substantially improve the ability estimates in the early stages of the test. Equation 2 shows that a weak correlation coefficient makes the method less useful, but it is also quite harmless to use the method on weakly correlated data. With $r^*_{\hat{\theta}_{\cdot,1},\hat{\theta}_{\cdot,2}} = 0$ we get a prior mean close to the average ability. However, if we only have a small number of previously completed tests it may be advisable to instead use a standard normal prior.

The subject of augmented scores has been previously explored, often in the context of tests where the total score is the sum of subscores. An example could be a test in mathematics, divided into subtests that address algebra, geometry, etc. These tests raise the question whether the subscores are reliable on their own to measure an ability, and whether augmentation strategies can make them more reliable. When a subscore is based on only a few items, one proposal is to estimate the corresponding ability by using a weighted average of that subscore and the mean subscore of all participants. This is known as Kelley's formula. Other similar approaches take advantage of the full test score as auxiliary data, which may sometimes improve the estimates of the subscores (Haberman, 2008). Multivariate Bayesian approaches have been used to simultaneously estimate more than one subscore (de la Torre and Patz, 2005).

1.1 The effects of a model based MAP prior

To observe the effect of the model-based MAP prior, we simulated two-dimensional response data, and estimated $\hat{\theta}_{j,1}$ and $\hat{\theta}_{j,2}$. The approximate correlation between the estimates was 0.442, which is comparable to the correlation in the empirical VAA dataset.

Table 1 shows the RMSE of $\hat{\theta}_{j,2}$, i.e. the RMSE of ability of respondent j in dimension 2. The first row in the table shows the scenario where the respondent has answered 2 items in dimension 2. The second row shows a scenario where the respondent has answered 5 items in dimension 2, etc.

In the first results column we used a standard normal prior. In the second and third results columns we used model based priors. In the second column, we have the scenario that 5 items in dimension 1 have been answered. In the third column, we have the scenario that 10 items in dimension 1 have been answered. With more items answered in dimension 1, the estimate for dimension 1 becomes more accurate, and therefore the prior which we use to estimate the ability in dimension 2 becomes better.

In Table 1 we see that the model based priors improve the accuracy of the ability estimate in dimension 2. This is especially true when few items have been answered in dimension 2. When 10 or more items have been answered in dimension 2, the model based prior no longer has a substantial effect.

Table 1 The RMSE for $\hat{\theta}_{.,2}$ using MAP with a standard normal prior and a model-based prior respectively

	RMSE		
#items $\theta_{.,2}$	Std normal prior	Prior 5 items	Prior 10 items
2	0.79	0.74	0.72
5	0.67	0.64	0.63
10	0.54	0.53	0.52

Thus, model based priors can help produce more accurate ability estimates when only few items in a VAA have been answered.

Comparisons between score augmentation through the MAP prior and alternative score augmentation approaches could be a future topic to study. Such comparisons may show whether the easy-to-implement augmentation approach used here results in an accuracy similar to that of more complex methods.

In a fixed-length test that includes a large number of mandatory items, a custom prior mean appears to give only a modest gain in accuracy. In that situation, the test designer may be better off using a standard normal MAP prior. Nevertheless, the results indicate that model-based priors are useful in adaptive tests that users can conclude early.

2 Calculate item usefulness

After each response, we want to find the most useful item to present next. Our proposed method ranks the items from an item pool in order of usefulness. For each remaining item $i \in \{1, 2, ..., n_I\}$ that belongs to a scale $k \in \{1, 2, ..., K\}$, we can calculate the item's usefulness as

$$\xi_{i,k} = \eta_{i,k} \cdot \delta_k \tag{3}$$

where

$$\delta_k = \left(\frac{1}{K}(1 - \Phi) + \psi_k \Phi \right), \quad \Phi = \Phi\left(\frac{I_{\min} - \mu}{h} \right) \tag{4}$$

The parameter $\eta_{i,k}$ is the estimated reduction of the standard deviation in dimension k resulting from selecting item i in scale k. The parameter δ_k gives the importance of scale k.

The parameter δ_k represents the importance of scale k relative to the other scales that we use in our low-dimensional representation. In the early phase of the test we may be primarily interested in increasing the accuracy of the estimates, and less interested in the party positions or candidate positions in the neighborhood of the provisional ability estimate. Therefore we may want to consider each scale to be of equal importance at the start the test. Further into the test when the accuracy is somewhat better, we may want to give more weight to the scale importance scores.

The value of Φ specifies how much weight we put on the scale importance score. Φ is the cumulative standard normal distribution. μ is the cumulative Fisher information considered sufficient to transition to a phase where the choice of item is

affected by party positions in the near neighborhood of $\hat{\theta}_j$. I_{min} is $\min_{k\in\{1,\dots,K\}}(I_k)$, i.e. the summed Fisher information of the given responses in the scale that has least such information.

The parameter ψ_k is the importance of the scale to which the item belongs when $\Phi = 1$, and $\frac{1}{K}$ represents the importance of the scale in the case where all K scales have equal importance. Thus, when I_{min} is sufficiently large, we transition from the scale importance $\frac{1}{K}$ to the scale importance ψ_k.

The parameter h controls whether the transition from equal weights to context-aware weights is abrupt or gradual. Higher values of h give more gradual transitions.

An algorithm that continuously chooses the item that gives the greatest value of $\xi_{i,k}$ can be said to greedily minimize a weighted sum of the standard deviations over all scales, where the weights are δ_k.

It follows from Equations (3) and (4) that

$$\mu \to \infty \implies \xi_{i,k} \propto \eta_{i,k} \tag{5}$$

in which case we can disregard the dimensional weights. This special case gives a simplified algorithm where we rank the items in the item pool in descending order of $\eta_{i,k}$.

2.1 Variance reduction

In the univariate case, the maximum likelihood estimate $\hat{\theta}_{(ML)}$ has a normal distribution asymptotically with the true parameter value as its mean and a variance that equals the inverse of the sum of the Fisher information of the answered items (Reckase, 2009).

$$\hat{\theta}_{j,k(ML)} \sim N\left(\theta_{j,k}, \frac{1}{\sum_{i\in U_k} I_{i,k}}\right), \tag{6}$$

where $I_{i,k}$ is the Fisher information for item i in dimension k, and U_k is the set of indices of the already answered items in dimension k. By selecting the question with the highest Fisher information given the current location $\hat{\theta}_{j,k}$, we minimize the variance conditioned on the estimated ability. We are not guaranteed to minimize the actual variance, since this depends on the true unknown value of the latent parameter $\theta_{j,k}$.

A uniscale setting only requires that we continuously choose the item with the highest Fisher information conditioned on our current estimate of θ_j. In a setting with multiple scales we are also required to choose which scale that we want our next item to measure.

The overall best item will always be the best item in its own dimension based on the maximum information criterion. That gives us K candidate items that could be the best next item. We here suggest that an item is better if it gives a greater reduction in the standard deviation. For any of the scales, that reduction from a candidate item c can be calculated as

$$\eta_{c,k} = \frac{1}{\sqrt{\sum_{i \in U_k} I_{i,k}}} - \frac{1}{\sqrt{\sum_{i \in U_k} I_{i,k} + I_{c,k}}}, \tag{7}$$

where $I_{c,k}$ is the information of item c in dimension k.

This selection criterion favors items in scales where the current variance is high.

2.2 Scale importance

The scale importance ψ_k measures how important scale k is compared to the other scales. A more important scale is here defined as a scale that has a greater influence on the ranking of the parties in the neighbourhood of the current ability estimate.

The Euclidean distance from $\hat{\theta}_j$ to the position of party s is

$$D_s = \sqrt{\sum_{k=1}^{K} (\hat{\theta}_{j,k} - R_{s,k})^2}, \tag{8}$$

where $R_{s,k}$ is the position of party s in dimension k.

The derivative of the Euclidean distance w.r.t $\hat{\theta}_{j,k}$ tells us how the Euclidean distance between $\hat{\theta}_j$ and the party position changes with $\hat{\theta}_{j,k}$.

$$\frac{\partial D_s}{\partial \hat{\theta}_{j,k}} = \frac{\hat{\theta}_{j,k} - R_{s,k}}{\sqrt{\sum_{\ell=1}^{K} (\hat{\theta}_{j,\ell} - R_{s,\ell})^2}} \tag{9}$$

Equation (9) gives the relationship between a change in $\hat{\theta}_{j,k}$ and a change in D_s at the current location of $\hat{\theta}_j$, and is therefore more valid for small movements of $\hat{\theta}_{j,k}$ than for large movements. This makes the scale importance more suitable to consider toward the later part of the procedure when the changes in $\hat{\theta}_j$ are fairly small.

The expression

$$\frac{\partial |D_s - D_t|}{\partial \hat{\theta}_{j,k}} = \left| \frac{\partial D_s}{\partial \hat{\theta}_{j,k}} - \frac{\partial D_t}{\partial \hat{\theta}_{j,k}} \right| \tag{10}$$

indicates how much a movement of $\hat{\theta}_{j,k}$ at its current location changes the difference between the distances D_s and D_t. The extent to which this difference is influenced by a movement in $\hat{\theta}_{j,k}$ is a measure of how significant such movement is for the ranking of parties s and t.

The ranking of pairs of parties positioned closer to the current estimate $\hat{\theta}_j$ is considered more important than the ranking of pairs that are further away. We calculate the distance from $\hat{\theta}_j$ to a *pair* of parties s, t as

$$D_{s,t}^2 = \sum_{v \in \{s,t\}} \sum_{\ell=1}^{K} (\hat{\theta}_{j,\ell} - R_{v,\ell})^2 \tag{11}$$

which is the sum of the squared Euclidean distances to from $\hat{\theta}_j$ to the party positions. We use the inverse of this sum as a weight for a pair of parties as shown in Equation (12).

The use of the squared Euclidean distance puts high emphasis on parties in the close neighborhood of the current estimate of θ_j.

We calculate the importance of dimension k as

$$\psi_k = \frac{\sum_{\{s,t\}\in\Omega} \frac{1}{D_{s,t}^2} \cdot \frac{\partial |D_s - D_t|}{\partial \hat{\theta}_{j,k}}}{\sum_{\ell=1}^{K} \left(\sum_{\{s,t\}\in\Omega} \frac{1}{D_{s,t}^2} \cdot \frac{\partial |D_s - D_t|}{\partial \hat{\theta}_{j,\ell}} \right)} \tag{12}$$

where Ω is the set of all pairs of parties. If n_R is the number of parties in the VAA then we have $n_R(n_R-1)/2$ pairs of parties. Note that

$$\sum_{\ell=1}^{K} \psi_\ell = 1 \tag{13}$$

2.3 Numerical example

We will here look at a numerical example where we have two scales, and the respondent has already answered a set of items for each scale. As a result we have a provisional estimate of θ_j, and Fisher information for both scales. We also have 3 reference points, which can represent 3 parties or candidates that the respondent is matched against.

We let the Fisher information be 7 for scale 1 and 5 for scale 2. The smallest Fisher information is thus $I_{min} = 5$. The provisional ability estimate is $\hat{\theta}_j = [1 \;\; 0.3]$, and we set $\mu = 4, h = 1$. The scenario is illustrated in Figure 1.

We have 5 items to consider. To select the next item, we need to calculate the reduction of the standard deviation $\eta_{i,k}$ for every remaining item, and the scale importance ψ_k for the two scales.

2.3.1 Numerical calculation of the variance reduction

In Table 2 we calculate $\eta_{c,k}$ for the five items under consideration. The values in the table are conditioned on the most recent ability estimate. For items on the same scale, higher Fisher information implies a larger reduction in the standard deviation. However, for two items on different scales this is not necessarily the case. Here we can see that the information is higher in item 2 than in item 4, but since the information is also higher in the set of answered items from scale 1, the reduction in standard deviation from selecting item 2 is smaller.

2.3.2 Numerical calculation of the scale importance

Table 3 shows the relationship between the provisional ability estimate and the three reference points, which may represent political parties or candidates. Table 4 shows

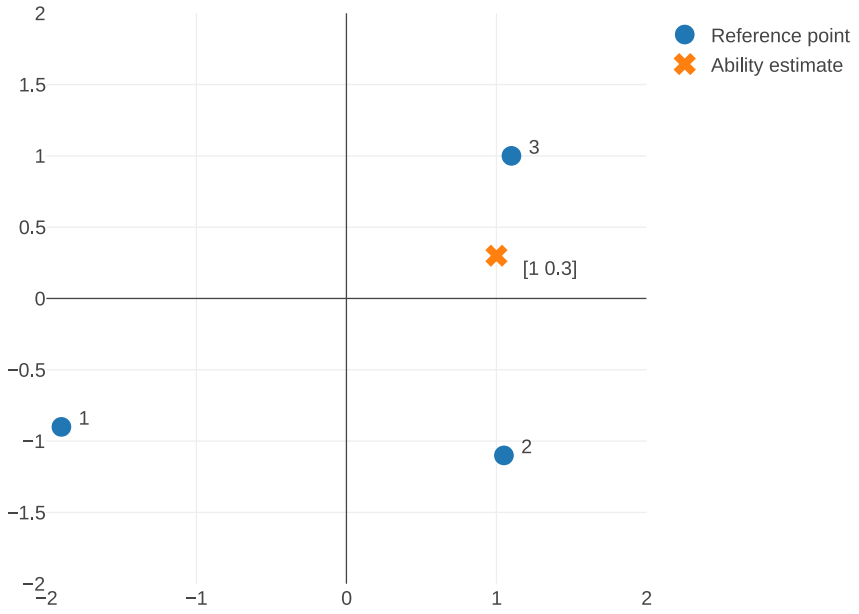

Fig. 1 In this scenario we have a provisional estimate of our ability on two scales. We also have three reference points, and we are interested in how close the true ability is to each of these reference points.

Table 2 Items under consideration in our numerical example

item	scale	$I_{c,k}$	$\sum_{i \in U_k} I_i$	$\frac{1}{\sqrt{\sum_{i \in U} I_{i,k}}}$	$\frac{1}{\sqrt{\sum_{i \in U} I_{i,k} + I_{c,k}}}$	$\eta_{c,k}$
1	1	2.90	7.00	0.38	0.32	0.06
2	1	1.70	7.00	0.38	0.34	0.04
3	1	1.10	7.00	0.38	0.35	0.03
4	2	1.50	5.00	0.45	0.39	0.05
5	2	1.20	5.00	0.45	0.40	0.05

the impact of a change in the ability estimate with regards to pairs of reference points. By summing each of the last two columns of Table 4 and normalizing the result we get the scale importance for our two scales.

We find that $\psi_1 \approx 0.19, \psi_2 \approx 0.81$. The reason why scale 2 is attributed a greater importance can be found in Table 4, or intuitively in Figure 1. Whereas movements in the first dimension influence how close we are to point 1 compared to the other points, this is of less interest since we are far from point 1. The ranking of point 2 vs point 3, which is mostly determined by dimension 2, is considered more interesting since we are closer to this pair of points, measured as the sum of squared Euclidean distances.

Table 3 Properties of the reference points in relation to $\hat{\theta}_j$

reference point	$R_{s,1}$	$R_{s,2}$	D_s	$\frac{\partial D_s}{\partial \hat{\theta}_{j,1}}$	$\frac{\partial D_s}{\partial \hat{\theta}_{j,2}}$
1	-1.90	-0.90	3.14	0.92	0.38
2	1.05	-1.10	1.40	-0.04	1.00
3	1.10	1.00	0.71	-0.14	-0.99

Table 4 Properties of pairs of reference points in relation to $\hat{\theta}_j$

	R_s	R_t	$\frac{\partial \lvert D_s - D_t \rvert}{\partial \hat{\theta}_{j,1}}$	$\frac{\partial \lvert D_s - D_t \rvert}{\partial \hat{\theta}_{j,2}}$	$D^2_{s,t}$	$\frac{1}{D^2_{s,t}}$	$\frac{1}{D^2_{s,t}} \frac{\partial \lvert D_s - D_t \rvert}{\partial \hat{\theta}_{j,1}}$	$\frac{1}{D^2_{s,t}} \frac{\partial \lvert D_s - D_t \rvert}{\partial \hat{\theta}_{j,2}}$
Pair1	1	2	0.96	0.62	11.81	0.08	0.08	0.05
Pair2	1	3	1.07	1.37	10.35	0.10	0.10	0.13
Pair3	2	3	0.11	1.99	2.46	0.41	0.04	0.81
Sum							0.23	0.99
Sum, Normalized							0.19	0.81

2.3.3 The usefulness of the items

Now that we know how much an item reduces the standard deviation and the importance of the two scales, we can use Equation (3) to calculate the usefulness of each item.

Table 5 Calculations of the usefulness of each item under consideration

item	scale	$\eta_{i,k}$	ψ_k	Φ	$\eta_{k,i} \cdot \left(\frac{1}{K}(1-\Phi) + \psi_k \Phi \right)$
1	1	0.06	0.19	0.84	0.014
2	1	0.04	0.19	0.84	0.009
3	1	0.03	0.19	0.84	0.006
4	2	0.05	0.81	0.84	0.042
5	2	0.05	0.81	0.84	0.035

From Table 5 we find that item 4 is the most useful item. This is the case even though item 1 contributes to a greater reduction in standard deviation. Therefore we select item 4 as our next item.

In a simplified version of the algorithm where we do not account for scale importance, we would instead have selected item 1.

3 Example application, results

Table 7 shows the results from the VAA data. Table 6 shows the results from the simulated data. The tables show the same information as the plots included in the article.

Table 6 The proportion of respondents matched to their best reference point after 1, 2, ..., 60 items

	2 reference points		5 reference points		10 reference points		15 reference points	
#items	static	adaptive	static	adaptive	static	adaptive	static	adaptive
1	0.69	0.66	0.48	0.47	0.39	0.41	0.22	0.31
2	0.73	0.81	0.52	0.59	0.41	0.48	0.28	0.31
3	0.78	0.85	0.54	0.64	0.44	0.57	0.30	0.45
4	0.82	0.88	0.56	0.67	0.45	0.54	0.30	0.47
5	0.82	0.89	0.60	0.71	0.47	0.61	0.35	0.49
10	0.88	0.91	0.69	0.78	0.57	0.69	0.44	0.60
20	0.89	0.93	0.75	0.82	0.68	0.75	0.57	0.71
40	0.94	0.94	0.84	0.86	0.76	0.79	0.73	0.76
60	0.94	0.94	0.85	0.86	0.80	0.79	0.76	0.75

Table 7 The proportion of users in the test dataset for whom the best party is identified after n question items

	Proportion correct best party	
#items	Adaptive method	Static IRT
1	0.42	0.40
2	0.57	0.44
3	0.64	0.58
4	0.71	0.58
5	0.76	0.59
6	0.79	0.66
7	0.81	0.71
8	0.85	0.72
9	0.90	0.83
10	0.92	0.86
11	0.94	0.87
12	1.00	1.00

References

de la Torre, J. and Patz, R. J. (2005). Making the Most of What We Have: A Practical Application of Multidimensional Item Response Theory in Test Scoring. *Journal of Educational and Behavioral Statistics*, 30(3):295–311.

Haberman, S. J. (2008). When Can Subscores Have Value? *Journal of Educational and Behavioral Statistics*, 33(2):204–229.

Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer, New York, NY.