SHORT NOTE

Addressing model uncertainty in item response theory person scores through model averaging

Jason D. Rights[1] · Sonya K. Sterba[1] · Sun-Joo Cho[1] · Kristopher J. Preacher[1]

Abstract Item banks are often created in large-scale research and testing settings in the social sciences to predict individuals' latent trait scores. A common procedure is to fit multiple candidate item response theory (IRT) models to a calibration sample and select a single best-fitting IRT model. The parameter estimates from this model are then used to obtain trait scores for subsequent respondents. However, this model selection procedure ignores *model uncertainty* stemming from the fact that the model ranking in the calibration phase is subject to sampling variability. Consequently, the standard errors of trait scores obtained from subsequent respondents do not reflect such uncertainty. Ignoring such sources of uncertainty contributes to the current replication crisis in the social sciences. In this article, we propose and demonstrate an alternative procedure to account for model uncertainty in this context—*model averaging* of IRT trait scores and their standard errors. We outline the general procedure step-by-step and provide software to aid researchers in implementation, both for large-scale research settings with item banks and for smaller research settings involving IRT scoring. We then demonstrate the procedure with a simulated item-banking illustration, comparing model selection and model averaging within sample in terms of predictive coverage. We conclude by discussing ways that model averaging and IRT scoring can be used and investigated in future research.

Keywords Item response theory · Model uncertainty · Model averaging · Item banks

---

✉ Jason D. Rights
     jason.d.rights@vanderbilt.edu

[1] Quantitative Methods Program, Department of Psychology and Human Development, Vanderbilt University, Peabody #552, 230 Appleton Place, Nashville, TN 37203, USA

⚛ Springer

1 Introduction

Item banks are often created in large-scale research and testing settings in the social sciences to predict individuals' latent trait scores (e.g., mathematics ability scores or health expectancy scores), often in conjunction with computer adaptive testing (CAT; Edelen et al. 2014; Meijer and Nering 1999). A common procedure is to fit multiple candidate item response theory (IRT) models to a *calibration* sample and select a single IRT model by some criterion, such as Bayesian Information Criterion (BIC; Schwarz 1978; see Cohen and Cho 2017 for reviews). From this best-fitting model, item parameter estimates for each item in the bank are obtained. For subsequent respondents (which can be thought of as composing a *validation* sample), item parameter estimates are treated as known and used to generate person scores and their standard errors.

Recent advances in item banking have involved accounting for one source of uncertainty—sampling variability in item parameter estimates from the calibration phase (e.g., Sinharay et al. 2003). However, another source of uncertainty has been ignored—*model uncertainty*. This uncertainty stems from the fact that the model ranking in the calibration phase is subject to sampling variability (even the "best-fitting" model can vary across samples, Lubke et al. 2017; Preacher and Merkle 2012; Sterba and Rights 2017). Conventional standard errors of person scores obtained from the validation phase do not reflect model uncertainty. Furthermore, ignoring model uncertainty in such a manner is a contributing factor to the current replication crisis in the social sciences (Lubke and Campbell 2016).

An increasingly employed approach to account for model uncertainty is *model averaging* (MA; Burnham and Anderson 2002). In MA, predictions in the validation sample are computed using a weighted average of quantities taken from all models considered during calibration, with the most plausible models given the most weight. In a variety of contexts from linear regression to structural equation modeling, when there is greater model uncertainty, MA has yielded predicted outcomes ($\hat{y}$'s) with better predictive coverage than model selection (MS) (Hoeting et al. 1999; Kaplan 2016; Kaplan and Lee 2016). Under little model uncertainty, MS and MA perform similarly.

MA has not been used to compute predicted IRT person scores ($\hat{\theta}$'s) in an item-banking context. Our purpose in this short note is to (a) introduce and describe MA for IRT person score predictions, where it has not before been applied, (b) provide new software for its implementation, and (c) show the predictive coverage of MA and MS person scores using a simulated item-banking illustration. Predictive coverage is particularly important for person scores in CAT, because a common stopping criterion is the width of the person score prediction interval (Bjorner et al. 2007). When such an interval is less accurate, the assessment may end at an inappropriate time.

Before continuing, we wish to make note of several caveats. First, we will focus on averaging person scores specifically in item-banking contexts. This allows us to demonstrate predictive coverage in a validation sample, similar to demonstrations done with other types of model averaging (Hoeting et al. 1999;

Kaplan 2016; Kaplan and Lee 2016). However, as we explain further in the discussion section, the general averaging procedure we will describe here can also be used in smaller research settings without item banks or validation samples. Second, we note that large research companies often might not select IRT models and their items based on fit indices (and instead for political or philosophical reasons, for instance). Selecting a model in such a way still in essence ignores the potential uncertainty of scores had a different model (that might be similarly plausible) been chosen, but nonetheless, the specific procedures discussed in the current paper will pertain to the use of fit indices in the applications of IRT models (e.g., Anderson 1973; Baker and Kim 2004; Bock and Aitkin 1981; de Ayala 2009). Finally, among the goals of the current paper we just mentioned, we do not include a thorough investigation of the conditions wherein MA and MS approaches with IRT person scores perform most differently, nor the extent to which MA would perform better than MS. Such an investigation is outside the scope of this short note, but is an important future direction.

2 Model averaging of person scores

We focus on frequentist MA (Hjort and Claeskens 2003), though extensions to Bayesian MA are possible (Hoeting et al. 1999). Suppose that the researcher has K competing models. In item-banking applications, K will typically consist of a small number of commonly applied IRT models, though our approach could be implemented with a larger set of models. In the calibration sample, we fit these K candidate models and obtain model-specific item parameter estimates and selection criterion values (e.g., BIC). In the validation sample, we obtain model-specific person scores from all K models by fixing the item parameter estimates to those obtained in the calibration sample for that model. To ensure scale comparability of person scores across models, we rescale scores within model to have the same mean (here, 0) and variance (here, 1) across models:

$$\hat{\theta}_{kj}^r = \frac{\hat{\theta}_{kj} - E_j(\hat{\theta}_{kj})}{\sqrt{\text{var}_j(\hat{\theta}_{kj})}},\tag{1}$$

where $\hat{\theta}_{kj}^r$ is the person score for model k ($k=1,\ldots,K$) and person j ($j=1,\ldots,J$). An r

superscript indicates "rescaled." We rescale person-score standard errors proportionally:

$$\text{SE}(\hat{\theta}_{kj}^r) = \frac{\text{SE}(\hat{\theta}_{kj})}{\sqrt{\text{var}_j(\hat{\theta}_{kj})}}.\tag{2}$$

Next, we compute *model-specific weights* (measures of plausibility ranging from 0 to 1). We use BIC weights (Burnham and Anderson 2002), though other weighting approaches could be employed. The K BIC values from the calibration sample are used to compute K model-specific weights. The kth model's weight is

$$w_k = \frac{\exp\left(-\frac{1}{2}\mathrm{BIC}_k\right)}{\sum_{k=1}^{K} \exp\left(-\frac{1}{2}\mathrm{BIC}_k\right)}. \tag{3}$$

The K weights sum to 1.[1] Using the K sets of person scores from the validation sample and K weights from the calibration sample, we calculate model-averaged person scores as

$$\hat{\theta}_{\cdot j}^r = \sum_{k=1}^{K} w_k \hat{\theta}_{kj}^r \tag{4}$$

and their standard errors as

$$\mathrm{SE}(\hat{\theta}_{\cdot j}^r) = \sum_{k=1}^{K} w_k \sqrt{\mathrm{SE}(\hat{\theta}_{kj}^r)^2 + (\hat{\theta}_{kj}^r - \hat{\theta}_{\cdot j}^r)^2}. \tag{5}$$

In contrast, when using MS, we would simply select the best-fitting model (e.g., the one with the lowest BIC) and use the person score and standard error from this model alone. This is equivalent to applying our MA approach and giving the best-fitting model a weight of 1 and all others a weight of 0. Thus, when a single model is heavily favored over all other models, such that the MA weight (Eq. 3) for this model is essentially 1, MA and MS will yield the same results. When no single model is heavily favored, the weights will be distributed across the models and MA and MS results will differ (as we demonstrate in our upcoming simulated example). Note that, whether using MA or MS, the number of items is the same.

3 Software implementation

To aid researchers in implementing MA for person scores, we developed an R function, *modelavgIRT*, that reads in and rescales $J \times K$ $\hat{\theta}_{kj}$ and $\mathrm{SE}(\hat{\theta}_{kj})$ from the validation sample and K information criterion values from the calibration sample and outputs J $\hat{\theta}_{\cdot j}$ and $\mathrm{SE}(\hat{\theta}_{\cdot j})$. This function is provided in the appendix. In cases wherein there is no validation sample (e.g., small research settings involving IRT scoring),

---

[1] When computing Eq. 3, software might round all K values summed in the denominator to 0, yielding an undefined solution. The software we provide in the Appendix accounts for this by using an equivalent mathematical reformulation that is not susceptible to this issue.

researchers can use the same procedure by simply inputting the $J \times K$ $\hat{\theta}_{kj}$ and $SE(\hat{\theta}_{kj})$ for their entire sample. Note that, although R functions exist for MA in linear regression, they are inapplicable here because they are not designed to average person-specific quantities across IRT models.

4 Illustration

We next provide an illustrative comparison of person score predictive coverage using MS and MA in a simulated item-banking context. We use simulated data because we are assessing coverage regarding latent ability levels, which are unknown with empirical data. The generating model is a three-parameter logistic (3-PL) bifactor model with a small guessing parameter and two weak secondary dimensions with ten items each. More specifically, in the generating model, the probability of item response "1" is given by

$$P(y_{ji} = 1|\theta_j, \theta_{jd}) = c_i + \frac{1 - c_i}{1 + \exp[-(\alpha_i\theta_j + \alpha_{id}\theta_{jd} - \beta_i)]}. \tag{6}$$

Here, y_{ji} is the item response (0 or 1) for person j and item i. θ_j is the person score for person j (primary dimension) that is generated from a standard normal distribution and it is the person score of substantive interest in our illustration. θ_{jd} is the person score for person j (secondary dimension). Each item loads onto one of the two secondary dimensions, $d = 1$ or 2. The first ten items load on $d = 1$ and next ten load on $d = 2$. The person scores for the two secondary dimensions are not of substantive interest in our illustration. β_i is the item location for item i. It is generated from a standard normal distribution. α_i is the (primary dimension) item discrimination for item i. It is generated from a log-normal distribution with $\mu = 0.08$ and $\sigma = 0.3$. α_{id} is the (secondary dimension) item discrimination for item i and the secondary dimension d. It is generated as 0.378 for all items, which induces an explained common variance (ECV, Reise et al. 2013) for the primary dimension equal to 0.90, implying very weak secondary dimensions. Finally, c_i is the pseudo-guessing parameter (lower asymptote) for item i. It is generated as 0.1 (for all items).

As typical of practice, we suppose that substantive interest lies in person scores for the primary dimension only (Reise 2012). To reflect realistic practice, none of our $K = 4$ fitted models are completely correct (Preacher and Merkle 2012), though all are commonly fit in practice: 1-PL ($k = 1$), 2-PL ($k = 2$), 3-PL ($k = 3$), and 2-PL-bifactor with two secondary dimensions ($k = 4$). We use 1000-person calibration and validation samples. In the calibration sample, K models were fit using *mirt* (Chalmers 2012); model weights were $w_1 = 0.441$, $w_2 = 0.559$, $w_3 < 0.001$, and $w_4 < 0.001$ for 1-PL, 2-PL, 3-PL, and 2-PL-bifactor with two secondary dimensions, respectively. In the validation sample, we obtained model-specific person scores (for the primary dimension) through expected a posteriori (EAP) scoring by fixing item

parameter estimates to those obtained in calibration (Bock and Aitkin 1981). MA person scores and standard errors in the validation sample were obtained using earlier-described procedures. MS person scores and standard errors in the validation sample were obtained from the best-fitting 2-PL (having the lowest BIC in the calibration sample) and were similarly rescaled, as shown in Eqs. 1 and 2.

Person-specific 90% prediction intervals in the validation sample are computed as $\hat{\theta}^r_{\cdot j} \pm 1.645 \times \text{SE}(\hat{\theta}^r_{\cdot j})$ for MA and $\hat{\theta}^r_{k'j} \pm 1.645 \times \text{SE}(\hat{\theta}^r_{k'j})$ for MS, where k' denotes the best-fitting model. Here, predictive coverage is defined as the percentage of persons in the *single* validation sample whose true ability, θ^r_j, falls within their prediction interval. However, *nominal* coverage is defined on average across *repeated* samples, so coverage in a single validation sample would not be expected to match the nominal rate exactly, even if on average across samples coverage is nominal. Hence, here we are most interested in the within-sample comparison of coverage from MA and MS. Predictive coverage was 85.6% for MS and 87.2% for MA. Lower coverage for MS over MA can be due to the former ignoring model uncertainty (Kaplan 2016). Mirroring previous research in other modeling contexts, MA's coverage was modestly closer to nominal. With even greater model uncertainty, benefits of MA over MS can be more pronounced.

5 Discussion

In this paper, we demonstrated how person scores computed with item banks can account for model uncertainty using MA and we provided user-friendly software for implementation. Employing model-averaged person scores with model-averaged standard errors in health and educational applications involving item banks can help prevent model uncertainty from contributing to the current replication crisis in these fields (Lubke and Campbell 2016).

Because of our restricted focus in this short note, there are several limitations that serve as future directions. First, we focused on assessing model-level fit via information criteria, but in practice, researchers might also assess item-level fit (Ames and Penfield 2015); for instance, one might remove items with low discrimination. In this context, one can still use either MS or MA, but the process would need to be adapted to account for the removal of items. Second, mirroring work done in other MA contexts, we focused on assessing predictive coverage for MS vs. MA, noting that a predominant issue with MS is that standard errors do not reflect model uncertainty and thus can be too small, yielding coverage that is too low. Nonetheless, in practice, researchers would also be interested in the point estimates of the scores themselves. In our example, though MA provided better coverage than MS, the point estimates of the person scores were highly correlated (0.99). Future work can determine whether or not MA and MS point estimates of scores are typically similar. We suspect, for instance, that in CAT settings wherein person score interval length is a stopping criterion, MS scores and MA scores would likely be more dissimilar, given that, with MS, the assessment would end too soon (i.e., before the desired level of precision). More broadly speaking, as a final future direction, future

work should determine (a) the conditions wherein MA and MS would be most dis-similar in terms of either interval and/or point prediction and (b) the extent to which MA performs better than MS under varying degrees of model uncertainty. Here, we presented an illustrative simulation with a single set of generating conditions; these can be expanded in the future to include a variety of such conditions.

Compliance with ethical standards

Conflict of interest   On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix: *modelavgIRT* R function

modelavgIRT R function description

This function reads in person scores (e.g., EAP scores) and their standard errors from the validation sample, and information criteria values (BIC, AIC) from the calibration sample from each of a set of candidate IRT models and outputs model-averaged person scores and standard errors (see manuscript equations 4 and 5).

modelavgIRT R function input

personscores	A data set consisting of person scores obtained from each candidate model in the validation sample, with rows denoting person and columns denoting model
personSEs	A data set consisting of person score standard errors obtained from each candidate model in the validation sample, with rows denoting person and columns denoting model
selectionindex	List of information criteria values (BIC, AIC) for each model, in the order of the columns of personscores and personSEs
rescale	Logical; if set to TRUE (default), prior to averaging, each models' person scores will be rescaled to have mean of 0 and a variance of 1 and standard errors will be rescaled proportionally

modelavgIRT R function Code

```
modelavgIRT <- function(personscores,personSEs,selectionindex,rescale=TRUE) {
    ##rescale personscores to have mean 0 and var 1
    #rescale personSEs proportionally
    if(rescale==TRUE){
    for(i in seq(ncol(personscores))){
    personscores[,i]        <-      (personscores[,i]      -      mean(personscores[,i]))/
sd(personscores[,i])
    personSEs[,i] <- personSEs[,i]/sd(personscores[,i])
```

```
}
}
##compute weights
weights <- c(rep(NA,length(selectionindex)))
for(i in seq(length(selectionindex))){
weights[i] <- sum(exp(-
.5*selectionindex[1:length(selectionindex)]+.5*selectionindex[i]))^(-1)
}
##compute averaged person scores
avg.personscore <- matrix(NA,nrow(personscores),1)
for(i in seq(nrow(personscores))){
avg.personscore[i,] <- sum(weights*personscores[i,])
}
##compute averaged person SEs
avg.personSE <- matrix(NA,nrow(personSEs),1)
for(i in seq(nrow(personSEs))){
avg.personSE[i,] <- sum(weights*sqrt(personSEs[i,]^2+(personscores[i,]-
avg.personscore[i,])^2))
}
output <- list(weights,avg.personscore,avg.personSE)
names(output) <- c("weights","Average person score","Average person SE")
return(output)
}
```

References

Ames AJ, Penfield RD (2015) An NCME instructional module on item-fit statistics for item response theory models. Educ Meas Issues Pract 34:39–48

Anderson EB (1973) A goodness of fit test for the Rasch model. Psychometrika 38:123–140

Baker FB, Kim S-H (2004) Item response theory: parameter estimation techniques, 2nd edn. Marcel Dekker, New York

Bjorner JB, Chang CH, Thissen D, Reeve BB (2007) Developing tailored instruments: item banking and computerized adaptive assessment. Qual Life Res 16:95–108

Bock RD, Aitkin M (1981) Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 46:443–459

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach. Springer, Berlin

Chalmers RP (2012) Mirt: a multidimensional item response theory package for the R environment. J Stat Softw 48:1–29

Cohen AS, Cho S-J (2017) Information criteria. In: van der Linden WJ, Hambleton RK (eds) Handbook of item response theory, models, statistical tools, and applications. CRC, Boca Raton

de Ayala RJ (2009) The theory and practice of item response theory. Guilford Publishing, New York

Edelen MO, Tucker JS, Shadel WG, Stucky BD, Cerully J, Zhen L, Hansen M, Cai L (2014) Development of the PROMIS® health expectancies of smoking item banks. Nicotine Tob Res 16:S222–S230

Hjort NL, Claeskens G (2003) Frequentist model average estimators. J Am Stat Assoc 98:879–899

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14:382–401

Kaplan D (2016) On the utility of Bayesian model averaging for improving prediction in the social and behavioral sciences. Society of multivariate behavioral research meeting, Richmond

Kaplan D, Lee C (2016) Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. Struct Equ Model 23:343–353

Lubke G, Campbell I (2016) Inference based on the best-fitting model can contribute to the replication crisis: assessing model selection uncertainty using a bootstrap approach. Struct Equ Model 23:479–490

Lubke G, Campbell I, McArtor D, Miller P, Luningham J, van den Berg S (2017) Assessing model selection uncertainty using a bootstrap approach: an update. Struct Equ Model 24:230–245

Meijer RR, Nering ML (1999) Computerized adaptive testing: overview and introduction. Appl Psychol Meas 23:187–194

Preacher KJ, Merkle EC (2012) The problem of model selection uncertainty in structural equation modeling. Psychol Methods 17:1–14

Reise SP (2012) The rediscovery of bifactor measurement models. Multivar Behav Res 47:667–696

Reise SP, Bonifay WE, Haviland MG (2013) Scoring and modeling psychological measures in the presence of multidimensionality. J Pers Assess 95:129–140

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Sinharay S, Johnson MS, Williamson DM (2003) Calibrating item families and summarizing the results using family expected response functions. J Educ Behav Stat 28:295–313

Sterba SK, Rights JD (2017) Effects of parceling on model selection: parcel-allocation variability in model ranking. Psychol Methods 22:47–68