The Impact of Rapid Guessing on Model Fit and Factor-Analytic Reliability: An Exploratory Analysis

Alfonso J. Martinez[a] & Joseph A. Rios[b]

[a]University of Iowa

[b]University of Minnesota

Author's Note

Alfonso J. Martinez: https://orcid.org/0000-0002-5337-9654

Joseph A. Rios: https://orcid.org/0000-0002-1004-9946

Author contribution statement: The first author identified the applied datasets, conducted the analysis, created all figures and tables, and wrote the article. The second author conceived of the present study, identified the applied datasets, and contributed to editorial review. All authors conducted critical revisions of the article throughout the review process and approved of the final version to be published.

Correspondence concerning this article should be sent to Alfonso J. Martinez, University of Iowa.

Email: alfonso-martinez@uiowa.edu

Abstract

Rapid guessing (RG) is form of non-effortful responding whereby examinees provide a response in a timeframe that is incommensurate with the amount of time needed to thoroughly engage with the item. Previous research has found that RG leads to biased parameter estimates, biased ability estimates, and measurement non-invariance if not properly accounted for. The consequences of RG on model fit and factor-analytic reliability (MF&R), however, are not well understood. To address this gap in the literature, the present study explores the effects of rapid guessing on model fit and reliability across a corpus of 20 diverse low-stakes assessments. Three RG scoring approaches (Naïve, Penalized, Effort-moderated with imputation) were compared across four model fit indices (CFI, TLI, RMSEA, SRMR) and two reliability coefficients (McDonald's omega and Cronbach's alpha). We found evidence that model fit is influenced by choice of scoring procedure, with the effort-moderated with imputation method producing better model fit than naïve and penalized scoring procedures when RG rates were greater. RG was also found to differentially impact reliability indices, however no systematic trends across scoring method and RG rates emerged. Implications of the research, recommendations for practitioners, and a discussion of future directions are discussed.

The use of low-stakes assessments (LSAs) is common practice in educational settings for providing evaluations of examinee performance through formative and summative assessments and are routinely used for accountability purposes (Wise & DeMars, 2005). Despite their ubiquity in education, a defining feature of LSAs is that they are not generally used to make inferences or decisions at the individual-level, but rather at an aggregate-level (e.g., performance of a country as a whole; Akyol, Krishna, & Wang, 2021). As a result, LSAs possess minimal consequences to examinees; this feature challenges the assumption that examinees expend maximum effort throughout the course of a testing event (Wise, 2017). Indeed, previous research has found that LSAs are likely to induce a construct-irrelevant behavior known as non-effortful responding (NER; Wise & Gao, 2017).

NER is a behavior characterized by random responding without consideration for the item content (Rios, 2022; Wise & DeMars, 2005). According to the expectancy-value theory of achievement motivation (e.g., Pintrich & Schunk, 2002), the likelihood an examinee engages in NER is dictated by two principal factors: (1) the examinee's *perception* of the importance and usefulness of the presented item (i.e., task value; TV), and (2) the recognition that the examinee does not possess the requisite background needed to successfully solve a given problem (i.e., perceived probability of success; PPS). Figure 1 provides a conceptual diagram of this mechanism. As the figure illustrates, TV and PPS influence the effort an examinee is willing to put towards a particular task; if the effort put towards the task is less than the effort required by the task, the likelihood of NER increases.

One form of NER that is of particular interest to practitioners and applied researchers is rapid guessing (RG; Wise, 2017). RG is a behavior that occurs when examinees provide a response in a timeframe that is incommensurate with the amount of time needed to meaningfully engage with a given item. Previous research on RG has found that inferences from assessments can be substantially biased if RG is not accounted for in the analysis phase (Wise, 2017). For instance, RG has been shown to bias item parameter and person (ability) estimates (Rios et al., 2017), biases equating and scaling analyses (Deng & Rios, 2022), and distorts psychometric item properties (e.g., item information; van Barnevald, 2007). Importantly, RG has been found to be a prevalent behavior in LSAs, with a recent a meta-analysis reporting that up to 28% of examinees engage in RG to some degree at least once throughout the course of a testing event (Rios, Deng, & Ihlenfeldt, 2022). These findings highlight the importance of identifying RG and correcting its adverse effects; however, the impact of RG on model fit and factor-analytic reliability (MF&R) are not well understood. Because MF&R provide information about the measurement and psychometric quality of an assessment, it is important to investigate the extent to which RG affects these metrics. Before describing our research questions in more detail, we briefly review a popular RG identification procedure based on response times.

Identifying Rapid Guessing with Response Times

Before one can account for the presence of RG in a psychometric analysis, one must first identify responses that are believed to be the result of RG. Several procedures for identifying RG have been proposed in the literature, including mixture models and response time threshold approaches. Most RG identification procedures rely on the ability to obtain response time information at the item level. Recently, however, RG identification methods that do not rely on response time information have been developed (Rios, 2022) and these have been found to be especially useful in contexts in which response times may not be available (e.g., paper-and-pencil administrations). In this research, we focus on the response time thresholds method due to its ease of implementation, ubiquity in operational settings (Rios & Deng, 2021), and unobtrusiveness. Detailed overviews of the mixture modeling framework can be found in Rios, Abulela, and Deng (2022).

The RT thresholds approach is implemented as follows: first a user-specified value $t \in (0, 1)$ is chosen to represent the threshold. Response times that fall below $100t\%$ of item i's mean response time ($\overline{RT_\imath}$) are classified as RG. For instance, if $t = 0.30$ and $\overline{RT}_i = 10$ seconds, then response times less than 3 seconds are indicative of RG. After identification of RG, all responses that were classified as RG are subjected to scoring procedures that attempt to correct or adjust the RG responses. In practice, it is common to specify $t = 0.30$ so that response times that fall under 30% of the response time mean are classified as RG. This value has been found to perform relatively well across a variety of settings (Rios & Deng, 2021), however, users are free to choose any value of t that they believe is appropriate for their needs.

Scoring Procedures for Correcting Responses Flagged as RG

Identification of RG from response time information is the first step in mitigating the potentially deleterious effects of RG. The next step is to apply a procedure that corrects the responses that were classified as RG. Note that responses classified as solution behavior (i.e., non-rapid guessing responses) are not subject to the corrective methods discussed in this section (i.e., SB responses are left unaltered). We restrict our attention to four commonly used corrective methods: naïve scoring, penalized scoring, effort-moderated scoring (EM), and EM scoring with imputation (EM-I) (see Rios, Abulela, and Deng (2022) for a comprehensive review).

Naïve Scoring. Naïve scoring is the simplest corrective procedure as it does not modify responses classified as RG in any way. Hence, an analysis that utilizes the naïve scoring approach simply utilizes the original response patterns for all further inferences and completely ignores the presence of RG. Naïve scoring is usually applied if response time information from log file data is unavailable or if the investigator has prior evidence that RG is not an influential factor (e.g., the students were highly motivated to perform well on the assessment). In this work, the naïve scoring approach serves as a baseline method for comparative purposes with the scoring approaches discussed next.

Penalized Scoring. The penalized scoring approach was developed by Wright (2016) and rescores responses flagged as RG as incorrect regardless of the value of the original response. The rationale behind this procedure is based on the notion that an examinee who engages in RG is likely to expend minimal cognitive effort throughout the assessment. As a result, the examinee will likely possess a lower overall score. An advantage of this approach is that it adjusts for the inflation in scores due to the random guessing strategies examinees engage in. An assumption of penalized scoring is that examinees engage in RG on a given item because the item difficulty is greater relative to their ability (Rios, Abulela, & Deng, 2022). In other words, the penalized scoring approach assumes that, had the examinee not engaged in RG, said examinee would have likely gotten the item wrong anyways because they do not possess the ability needed to successful solve the problem. It follows from this assumption that penalized scoring assumes that RG is related to ability such that examinees who are low in ability have a higher likelihood of engaging in RG. A growing body of research suggests that RG is idiosyncratic and evenly distributed across the ability scale with low ability examinees engaging in RG at the same rate as their average and high ability counterparts (e.g., Wise & DeMars, 2006), hence, the assumption that RG manifests at a higher rate for low ability examinees is likely to be untenable.

Effort-Moderated Scoring. The effort-moderated (EM) scoring method was developed by Wise and DeMars (2006) and is based on the premise that RG responses are psychometrically uninformative and do not provide meaningful information with respect to the latent construct of interest (e.g., ability). As a result, the contribution of responses classified as RG are downweighed and all RG responses are treated as missing data. Simulation-based research has found that EM scoring is capable of mitigating bias in both item and ability parameter estimates (Rios & Soland, 2021). An assumption of EM scoring is that RG occurs idiosyncratically, is independent of ability, and is equally distributed across the ability scale, i.e., examinees of all abilities are equally likely to engage in RG on a given item. Note that this assumption is equivalent to the missing completely at random assumption from the missing data literature (Enders, 2022).

EM Scoring with Imputation. The EM-I scoring approach was developed by Hauser and Kingsbury (2009) as a modification to EM scoring to address computational issues associated from with the original EM method. In particular, EM scoring tends to produce response matrices that suffer from model non-convergence and inflated standard errors as the rates of RG increase. This results in low power and potentially unreliable or unstable parameter estimates. The EM-I method introduces an imputation step into the scoring procedure based on the concept of plausible values. This procedure produces several imputed responses which are then aggregated to produce the final imputed response. This approach is advantageous as it 'fills-in-the-blanks' and results in a complete data matrix which can improve computational stability and parameter estimation. A detailed overview of the EM-I procedure is given in the Appendix.

The Present Study

The deleterious effects of RG on important psychometric quantities are well-documented in the literature; less widely known, however, are the effects RG has on model fit and reliability (MF&R). These are important aspects to consider as both provide essential information for evaluating the quality of the data and can be used as a basis for empirically evaluating test validity claims (Stanley & Edwards, 2016). Moreover, model fit information can be used to help practitioners identify items that are negatively affected by the presence of RG (e.g., items which examinees do not respond well to). Additionally, it is well-known that the scoring approaches discussed above can mitigate the effects of RG, however it is not known how MF&R indices are influenced by choice of scoring procedure, if at all. Such information is especially helpful for providing practitioners with insights that may help guide the choice of scoring approach in operational settings.

The present study fills this gap in the literature by exploring the effects of RG on MF&R across a large corpus of LSAs. The following research questions were addressed in this study:

1. Is there evidence of differential MF&R when RG is/is not accounted for?
2. If RG is found to affect MF&R, for what proportion of rapid guessers do we observe such differences?

The first research question seeks to examine how MF&R is influenced in the presence of RG via the scoring methods discussed previously. Given the assumptions underlying each scoring and the resulting treatment to RG responses, it is of interest to evaluate the extent to which scoring approach yield *meaningful* differences in MF&R. If the use of one approach results in systematically higher/lower MF&R, such information can be used to guide selection of a scoring approach in practical applications. The second research question is concerned with evaluating how the rates of RG influence MF&R. If MF&R are largely unaffected by rates of RG, then this provides evidence that MF&R is robust to differences in RG rates. However, if there is a significant discrepancy in MF&R as a function of RG rates, then it is imperative to investigate how large RG rates must be before MF&R results become unreliable.

Method

We investigate our RQs across a diverse set of 20 operational assessments (see Table 1). The assessment corpus was deliberately chosen to be diverse to explore general trends and not limit generalizability of the results to a specific type of assessment. The assessments in the corpus are diverse in several respects, including content domain, testing population, language, test length, and sample size. For instance, test length ranged from 8 to 50 items (all dichotomously scored) and sample size ranged from 256 to 14,593 examinees. A comprehensive overview of the assessment corpus including a detailed summary of the assessment selection process can be found in Martinez & Rios (under review).

Identifying RG with Sequential Thresholds

Although the response time thresholds approach is one of the most popular RG identification procedures in applied and operational research, a major criticism of this approach is that there are no theoretical justifications for the threshold an investigator specifies. Several operational testing programs utilize an NT30 criterion (e.g., Wise, 2017), whereby responses that fall below 30% of the response time means for a given item are flagged as RG. Such a criterion is generally chosen based practical considerations, such as the need to balance false positive and false negative rates (Rios & Deng, 2021), however, in practice, the true false positive and false negative rates are unknown, and it is very likely that any threshold value chosen by the investigator will under- or over-estimate these rates to some degree. To account for uncertainty that arises from the choice of a given threshold, we propose and implement the sequential thresholds procedure.

The sequential thresholds procedure is an extension of the original threshold approach that allows an investigator to evaluate how inferences are affected across various threshold values. The idea behind the sequential thresholds approach is to select a range of threshold values between two extremes, say l and u. Then, K evenly spaced values between l and u serve as thresholds. All psychometric analyses are conducted at each threshold value to give K sets of results. The results can then be compared across threshold range to evaluate the robustness of the results. This approach takes into consideration the possibility that RG status will likely be under- or over-classified for any given threshold value and provides a sensible way for evaluating the extent to which inferences remain stable (or not) as a function of threshold value.

For the present analysis, values between 0.05 and 0.35 in increments of 0.01 served as thresholds (31 total) and response times that were less than $100x\%$ of the item's mean response time were classified as RG ($x = 0.05, \dots, 0.35$). After identification of RG responses, a modified response matrix was generated whereby RG responses were recoded according to the three scoring approaches described earlier: naïve (NS), penalized (PS), and effort-moderated scoring with imputation (EM-I). For each assessment in the corpus, the sequential-thresholds procedure generates 63 unique threshold-generated datasets (one for NS, 31 for PN, and 31 for EM-I), resulting in a total of 1,260 unique threshold-generated datasets across the entire corpus.

Model Fit and Reliability

MF&R was investigated within the structural equation modeling (SEM) framework. The SEM framework has a rich history in psychometrics and several well-known model fit and reliability indices can be readily evaluated within this framework (Shi, Maydeu-Olivares, & Rosseel, 2019). In this analysis we investigate the following model fit indices: comparative fix index (CFI), the Tucker-Lewis fit index, root mean square error of approximation (RMSEA), and the square root mean residual (SRMR). Briefly, the CFI and TLI indices are incremental fit indices that provide information about 'badness of fit' of a hypothesized (fitted) model relative to a 'worst-fitting' baseline model (see Bentler (1990) and Bentler & Bonett (1980) for more details). RMSEA is an absolute fit index that captures ‘how far’ the hypothesized (fitted) model is from a hypothetical ‘perfect’ model (Xia & Wang, 2019), and SRMR provides a measure of distance between model-implied and observed covariance matrices (Ximénez et al., 2022). Comprehensive overviews of these model fit indices can be found in Xia & Wang (2019).

Reliability was assessed via two estimators: coefficient omega and coefficient alpha (McNeish, 2018). The latter is the classical reliability estimator and the most commonly used measure of internal consistency of an assessment; the former is a factor-analytic reliability index that takes in to account the factor structure of the data (McNeish, 2018). When the assumption of tau-equivalence does not hold (e.g., factor loadings are not statistically equivalent), coefficient omega has been found to perform better than coefficient alpha and the two estimators are equivalent when the assumption of tau-equivalence is tenable (i.e., coefficient alpha is a special case of coefficient omega).

The R package *lavaan* was used to fit each threshold-generated dataset and estimate all models (Rosseel, 2012). A one-factor item-factor analytic model was specified for each threshold-generated dataset. All models were estimated using the diagonally weighted least-squares estimator. After model fitting, the model fit indices were extracted and stored for analysis. Coefficient alpha was estimated using the R package *psych* (Revelle, 2015) and coefficient omega was estimated using the Green & Yang (2009) procedure as implemented in the R package *semTools* (Jorgensen et al., 2022).

Results

Rates of RG

Figure 2 displays the proportion of examinee-by-item interactions that were classified as RG at each level of the thresholds for all assessments in the corpus. Proportions ranged from

0.007 at the first threshold value to 0.061 at the final threshold value ($M = 0.034, SD = 0.016$). In general, the proportion of RG classifications increased as the threshold value increased. This result makes sense given that higher threshold values are more liberal in their classifications and lower threshold values are more conservative as only the most extreme response times are classified as RG.

Figure 3 displays the proportion of examinees who engaged in RG on at least one item on the respective assessment. Values ranged from 0 to 0.64 and, in general, there was an upward trend in proportions as a function of threshold value ($M = 0.196, SD = 0.145$). A similar trend was observed for the proportion of examinees who engaged in RG on over 50% of items on each respective assessment, with values ranging from 0 to 0.107 ($M = 0.02, SD = 0.02$) (see Figure 4), however the proportions themselves were generally much smaller in magnitude. This suggests that it is common for examinees to engage in RG at least once throughout the testing event but relatively uncommon for examinees to continuously engage in RG throughout the majority of the testing event. Indeed, the proportion of examinees who engaged in RG on all items was very small, with values ranging from 0 to 0.011 ($M = 0.001, SD = 0.002$) (see Figure 5).

Model Fit

Results from the model fit indices (CFI, TLI, RMSEA, SRMR) are displayed in Figures 6 through 9, respectively. Examining the CFI fit indices first, over 80% of all 1,260 threshold-generated datasets exhibited excellent model fit according to the criterion CFI ≥ 0.95 (Hu & Bentler, 1999). In general, CFI values exhibited the tendency to increase as a function of threshold value when evaluated under EM-I scoring but decreased as a function of threshold value when evaluated under penalized scoring. In other words, when EM-I scoring was used, better model fit is observed – according to the CFI index – as the RG threshold becomes more liberal but worse model fit was observed if penalized scoring was utilized.

For most assessments, the difference in CFI for the different scoring procedures at a given threshold value was not large (e.g., differences at the third decimal place) and larger differences occurred at around the midpoint of the threshold range (e.g., a threshold of 0.20). Furthermore, for several assessments, the CFI index under EM-I scoring remained relatively stable relative to naïve scoring whereas the indices were much more variable across threshold values under penalized scoring.

With respect to the TLI fit indices, a similar trend to that of the CFI index was observed. More specifically, there was a tendency for EM-I to display better model fit relative to naïve scoring as the threshold became more liberal. Moreover, penalized scoring displayed a tendency to display worse model fit relative to naïve scoring at more liberal threshold values, however in many instance the difference between the two scoring approaches was not meaningfully significant (e.g., a difference of up to 0.004 in TLI between EM-I and penalized scoring at a given threshold value).

As was the case with the CFI index, EM-I scoring was more stable throughout the threshold range, with notable exceptions being the values from the PIAAC literacy and Socioemotional Competency assessments whereby TLI values were much larger and smaller relative to the TLI values from naïve scoring, respectively. In general, over 77% of all threshold-generated datasets exhibited excellent model fit according to the Hu & Bentler (1999) criterion TLI ≥ 0.95, with majority of the datasets that exhibited excellent fit being those at more liberal threshold values and when scored with EM-I.

In terms of RMSEA, over 93% of threshold-generated datasets exhibited excellent model fit according to the Hu & Bentler (1999) criterion of RMSEA $\leq$ 0.06. A general trend emerged whereby penalized scoring exhibited worse model fit as the threshold value increased and EM-I scoring exhibited better model fit as the threshold value became more liberal. However, we note that, in most instances, the difference in RMSEA was not very large with the discrepancy in model fit between the two scoring approaches often occurring in the third decimal digit. This suggests that RMSEA is performs relatively the same for both scoring methods, however slightly better fit can be obtained if EM-I scoring is utilized. Interestingly, the RMSEA index displayed high variability under EM-I (e.g., lower stability across the threshold range), suggesting that this fit index is more sensitive to threshold value.

Finally, just under 74% of threshold-generated datasets exhibited excellent model fit according to the Hu & Bentler (1999) criterion of SRMR $\leq$ 0.08. Similar to the results from the RMSEA index, a general trend emerged whereby more liberal threshold values resulted in better model fit under EM-I scoring but worse model fit under penalized scoring. We note, however, that this trend was not universal to across the corpus, as there were instances in which penalized scoring exhibited better model fit at more liberal threshold values than EM-I (e.g., see Higher Education English 2 assessment in Figure 9), however, differences in SRMR for these instances were generally small.

Reliability

Results from the reliability analysis can be found in Figure 10. Across the entire corpus, the trends with respect to the two reliability estimators were nearly identical. Moreover, the omega coefficient displayed systematically higher reliability at every threshold value regardless of the scoring approach (e.g., omega reliability was higher when estimated under penalized scoring than alpha reliability when estimated under penalized scoring, etc.). Hence, here we report results from the omega reliability, but the alpha reliability estimates are included in Figure 10.

Interestingly, results from the reliability analysis were much more variable than the results from the model fit analyses. For instance, for the Amsterdam chess dataset, omega reliability from EM-I scoring was nearly identical to omega reliability from naive scoring for all threshold values, but omega reliability from penalized scoring systematically decreased after a threshold value of 0.15. By contrast, for the eTIMMS science assessment, omega reliability under EM-I scoring was systematically lower than omega reliability under penalized scoring at nearly all threshold values but remained consistent with omega reliability under naive scoring. Interestingly, for the PIAAC literacy assessment, omega reliability under naive scoring was systematically higher than omega reliability from *both* penalized and EM-I scoring, and, in addition, omega reliability from penalized scoring outperformed omega reliability from EM-I scoring as well. Hence, it appears that reliability coefficients are highly sensitive to scoring approach and are assessment-dependent in the sense that one scoring approach (e.g., penalized) may produce a higher reliability for one assessment than an alternative scoring approach (e.g., EM-I) but the opposite may be true for another assessment. This suggests that reliability should be investigated on a case-by-case basis.

Discussion

Rapid guessing is a form of non-effortful responding whereby examinees provide responses to an item in a timeframe that is incommensurate with the amount of time needed to meaningfully engage with said item. The deleterious effects of rapid guessing on the psychometric properties of a test such as quality of calibration estimates and ability estimates are well documented; however, the effects of rapid guessing on model fit and reliability are not well known. These factors are of practical importance as model fit and reliability both provide information about the structure of the test after accounting for rapid guessing and can be used for validity purposes. Via an exploratory analysis using a large corpus of operational datasets, the present study investigated the impact of RG on model fit and reliability.

We examined model fit within the structural equation modeling framework and reliability with classical and factor-analytic reliability estimators. Results from the analysis indicated that, in general, better model fit could be achieved with more liberal threshold values, provided that EM-I scoring is used. If, however, penalized scoring is used, then it would be preferable to use more conservative threshold values there was a tendency for model fit for decrease otherwise.

The reason for EM-I exhibiting better model fit at more liberal threshold values is not entirely clear, however a potential factor that could have contributed to this could be from the imputation procedure. As mentioned in the Appendix, the EM-I procedure utilizes the notation of plausible values to fill-in-the-blanks for responses classified as RG to construct a complete data matrix (Hauser & Kingsbury. 2009). This procedure utilizes an estimated ability score and all available examinee information to impute the most likely response for a given examinee-by-item interaction (provided said interaction was classified as RG). It is possible that this procedure results in a higher quality response matrix compared to the penalized scoring method which simply rescores all RG responses as incorrect. At higher threshold values, the rescoring of RG responses as incorrect may create a bias in the results which leads to a decrease in model fit. It would be beneficial to explore this question in depth via simulation-based research.

We found empirical evidence that reliability estimators are sensitive to the context of the assessment and the scoring procedure used. Evidence of this comes from our finding that one scoring approach may perform better than another scoring approach for one assessment but the opposite may be true for a different assessment. Hence, in contrast to the model fit analyses in which there was a general tendency for EM-I scoring to perform better at liberal threshold values, there were no such trends for the reliability analyses. These results corroborate the literature review by Deng & Rios (2022) who found that coefficient alpha did not perform consistently in applied datasets. We recommend practitioners and researchers investigate reliability on a case-by-case basis as no general trends were found in this study.

A strength of the present study lies in the use of the sequential thresholds approach. This approach was motivated by the observation that operational testing programs generally choose one threshold value for all psychometric inferences. As noted by several researchers (e.g., Martinez & Rios, under review), the choice of a single threshold is problematic in the sense that the true RG classification rates are unknown, and the use of a single threshold will likely under- or over-classify the RG states of many examinee-by-item interactions. The sequential thresholds approach attempts to examine the effect of RG more holistically by considering multiple threshold values to see how inferences change across said thresholds. Thus, the sequential thresholds method can be thought of as a method for evaluating the robustness of the results. As can be seen from the Figures below, there is evidence that model fit and reliability indices are influenced to some degree as a function of threshold value. We recommend that practitioners consider utilizing the sequential thresholds method with a few pre-selected threshold values to gauge if and how inferences are affected as a result of the selected values.

We note the following limitations of the present research. First, a major limitation of the sequential thresholds method is that analyses need to be repeated several times. In operational settings where time is of the essence, running an analysis for multiple threshold values may not be feasible. However, this approach may be particularly useful during the developmental stages of an assessment as it would allow an operational testing program to examine several candidate threshold values and a single one can be picked after extensive preliminary analyses.

An additional limitation of the study was that the analysis was exploratory in nature and differences with respect to the characteristics of the assessment were not considered. Although our decision to incorporate diverse assessments was intentional to prevent limiting the generalizations of our findings, it is possible that characteristics of the assessment such as test length and sample size played a role in our analysis. For instance, the sample size of the Amsterdam chess test was relatively small (256 examinees) whereas the Understanding America datasets contained responses from several thousand examinees. A systematic simulation study where factors such as test length and sample size are kept fixed could provide a more nuanced picture of the role of RG on model fit and reliability.

References

Akyol, P., Krishna, K., & Wang, J. (2021). Taking PISA seriously: How accurate are low-stakes exams? *Journal of Labor Research*, *42*, 184-243.

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107, 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Deng, J., & Rios, J. A. (2022). Investigating the Effect of Differential Rapid Guessing on Population Invariance in Equating. *Applied Psychological Measurement*, *46*(7), 589-604.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Hauser, C., & Kingsbury, G. G. (2009, April 14–16). *Individual score validity in a modest-stakes adaptive educational testing setting* [Paper presentation]. The annual meeting of the National Council on Measurement in Education, San Diego, CA.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). semTools: Useful tools for structural equation modeling. R package version 0.5-6. Retrieved from https://CRAN.R-project.org/package=semTools

Martinez & Rios (under review). How tenable are modeling assumptions around rapid guessing behavior? Results from a large corpus of low-stakes assessments.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological methods*, *23*(3), 412.

Pintrich, P. R., & Schunk, D. H. (2002). Motivation in education: Theory, research, and applications (2nd ed.). Upper Saddle, NJ: Merrill Prentice-Hall.

Revelle, W. (2023). psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 2.3.3. Retrieved from https://CRAN.R-project.org/package=psych.

Rios, J. A. (2022). A comparison of robust likelihood estimators to mitigate bias from rapid guessing. *Applied Psychological Measurement*, *46*(3), 236-249.

Rios, J. A., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-scale Assessments in Education*, *9*(1), 1-25.

Rios, J. A., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-scale Assessments in Education*, *9*(1), 1-25.

Rios, J. A., & Deng, J. (2022). Quantifying the Distorting Effect of Rapid Guessing on Estimates of Coefficient Alpha. *Applied Psychological Measurement*, *46*(1), 40-52.

Rios, J. A., & Soland, J. (2021). Investigating the impact of noneffortful responses on individual-level scores: Can the Effort-Moderated IRT model serve as a solution?. *Applied Psychological Measurement*, *45*(6), 391-406.

Rios, J. A., Abulela, A. A, & Deng, J. (2022). A Review of Rapid Guessing Scoring Approaches in Low-Stakes Multiple-Choice Assessments.

Rios, J. A., Deng, J., & Ihlenfeldt, S. D. (2022). To What Degree Does Rapid Guessing Distort Aggregated Test Scores? A Meta-analytic Investigation. *Educational Assessment*, *27*(4), 356-373.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not?. *International Journal of Testing*, *17*(1), 74-104.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1-36.

Shi, D., Maydeu-Olivares, A., & Rosseel, Y. (2020). Assessing fit in ordinal factor analysis models: SRMR vs. RMSEA. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 1-15.

Stanley, L. M., & Edwards, M. C. (2016). Reliability and model fit. *Educational and Psychological Measurement*, *76*(6), 976-985.

van Barneveld, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement*, *31*(1), 31-46.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52-61.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational assessment*, *10*(1), 1-17.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343-354.

Wright, D. B. (2016). Treating all rapid responses as errors (TARRE) improves estimates of ability (slightly). Psychological Test and Assessment Modeling, 58(1), 15–31.

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior research methods*, *51*, 409-428.

Ximénez, C., Maydeu-Olivares, A., Shi, D., & Revuelta, J. (2022). Assessing cutoff values of SEM fit indices: Advantages of the unbiased SRMR index and its cutoff criterion based on communality. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 368-380.

Appendix I Effort-moderated Scoring with Imputation

The multiple imputation EM scoring approach was implemented to address issues arising from the introduction of missing values into the response matrix. Most SEM software implement limited information procedures (e.g., diagonally weighted least squares) for model estimation. As a result, any cases with missing data (i.e., those arising from EM scoring) will be dropped from the analysis. Failure to account for this might artificially impact model fit and reliability depending on the magnitude of respondents who engaged in RG behavior on at least one item (see Figure 2). To compare NS, PS, and EM scoring procedures more fairly, we used a multiple imputation (MI) for datasets generated under EM scoring to impute RG responses (Hauser & Kingsbury, 2009).

For a given assessment and threshold, the dataset generated from application of the EM scoring procedure (i.e., after recoding responses identified as RG as missing) was modeled using full-information maximum likelihood under a 2-parameter logistic model. The estimated item and ability parameters were then used to calculate the probability of a correct response (p). The resulting value was compared to a random number (v) that was generated from a $U(0,1)$ distribution. If $p < v$, then an incorrect response (i.e., a value of 0) was imputed in place of the missing response; otherwise, a correct response (i.e., a value of 1) was imputed.

A total of $M = 20$ iterations were specified, and the average of the iterations was computed. This procedure has no effect on observed values (i.e., observed correct responses average out to 1 and observed incorrect responses average out to 0, as expected). For imputed responses, the average will fall between 0 and 1, depending on the number of 1s that were imputed across the 20 iterations (i.e., a value of 0.25 indicates that 5 out of the 20 imputed values were 1s). This value was then recoded to a 0 if it was less than 0.50 or recoded to a 1 if it was greater than or equal to 0.50.

Hence, the resulting matrix no longer contains missing responses and was used in the estimation of the model fit and reliability indices described in the main text.

Table 1. Summary of Assessment Corpus

Assessment Name	Test Language	Test Length	Sample Size
PISA 2018 Mathematics	English	14	2,706
PISA 2018 Science	Spanish	30	1,315
PISA 2018 Global Competency	Spanish	28	783
PISA 2018 Financial Literacy	Indonesian	25	644
PISA 2018 Reading	English	27	801
PISA 2015 Mathematics	Multiple	22	5,158
UA Numeracy	English	8	6,856
UA Cognitive Numeracy	English	15	8,629
UA Picture Vocabulary	English	15	11,355
UA Verbal Analogies	English	15	11,112
PIAAC 2016 PSTRE	Dutch	14	691
PIAAC 2016 Numeracy	Spanish	20	753
PIAAC 2016 Literacy	French	20	637
eTIMSS 2019 Grade 8 Math	English	25	1,864
eTIMSS 2019 Grade 8 Science	English	16	4,565
Amsterdam Chess Test	English, Dutch	40	256
College Entrance ELA (US)	English	36	826
College Entrance ELA (Canada)	English	26	14,593
Communication Competence	English	40	1,664
Sustainability Principles	English	50	1,001

Note: PISA = Program for International Student Assessment, UA = Understanding America, PIAAC = Program for the International Assessment of Adult Competencies, PSTRE = Problem Solving in Technology - Rich Environments, eTIMMS = Trends in International Mathematics and Science Study, ELA = English language arts. Complete details about corpus can be found in Martinez & Rios (under review).

Figure 1. Conceptual diagram of factors that influence non-effortful responding.

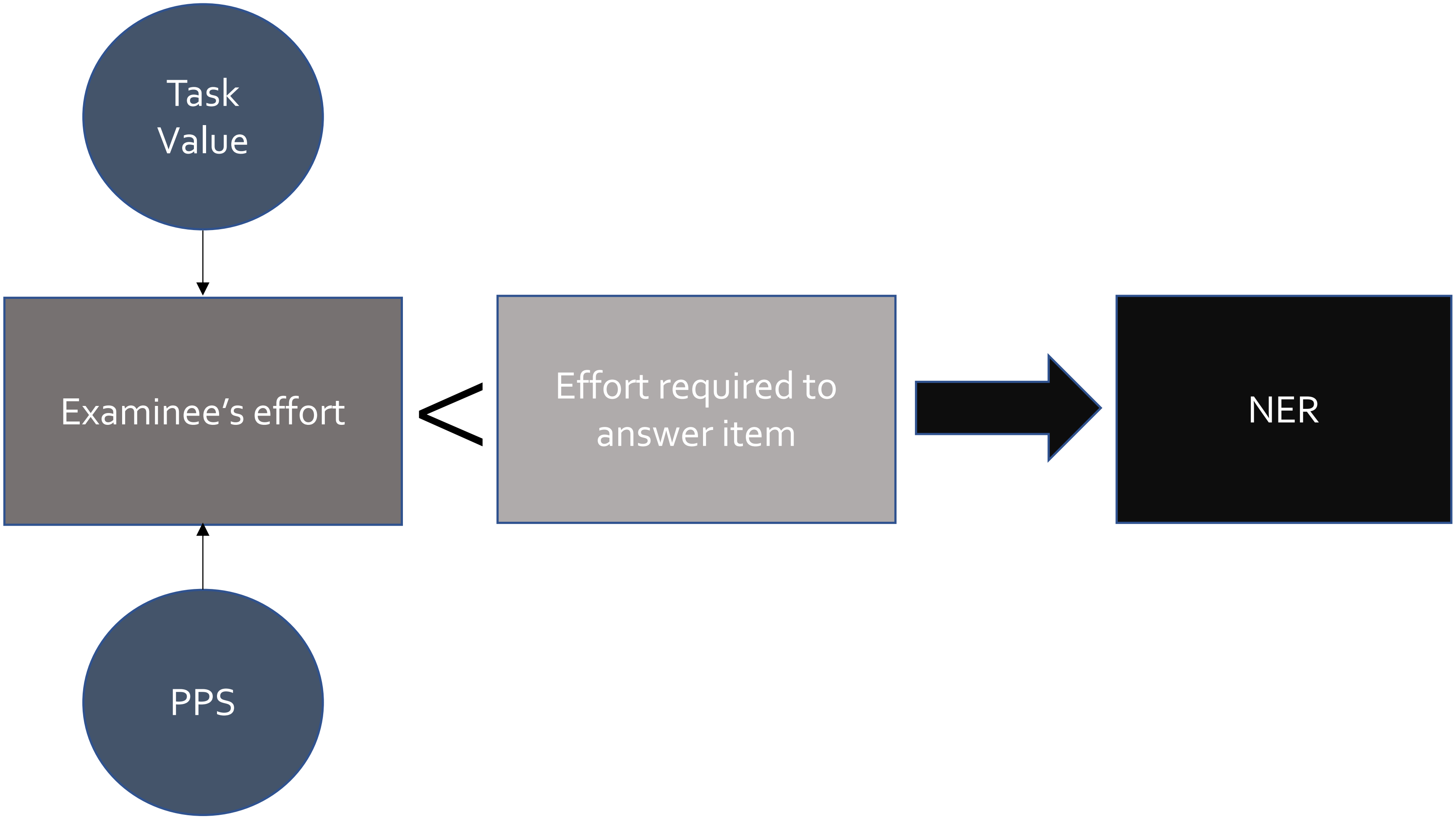


Figure 2. Proportion of examinee-by-item interactions classified as rapid guessing.

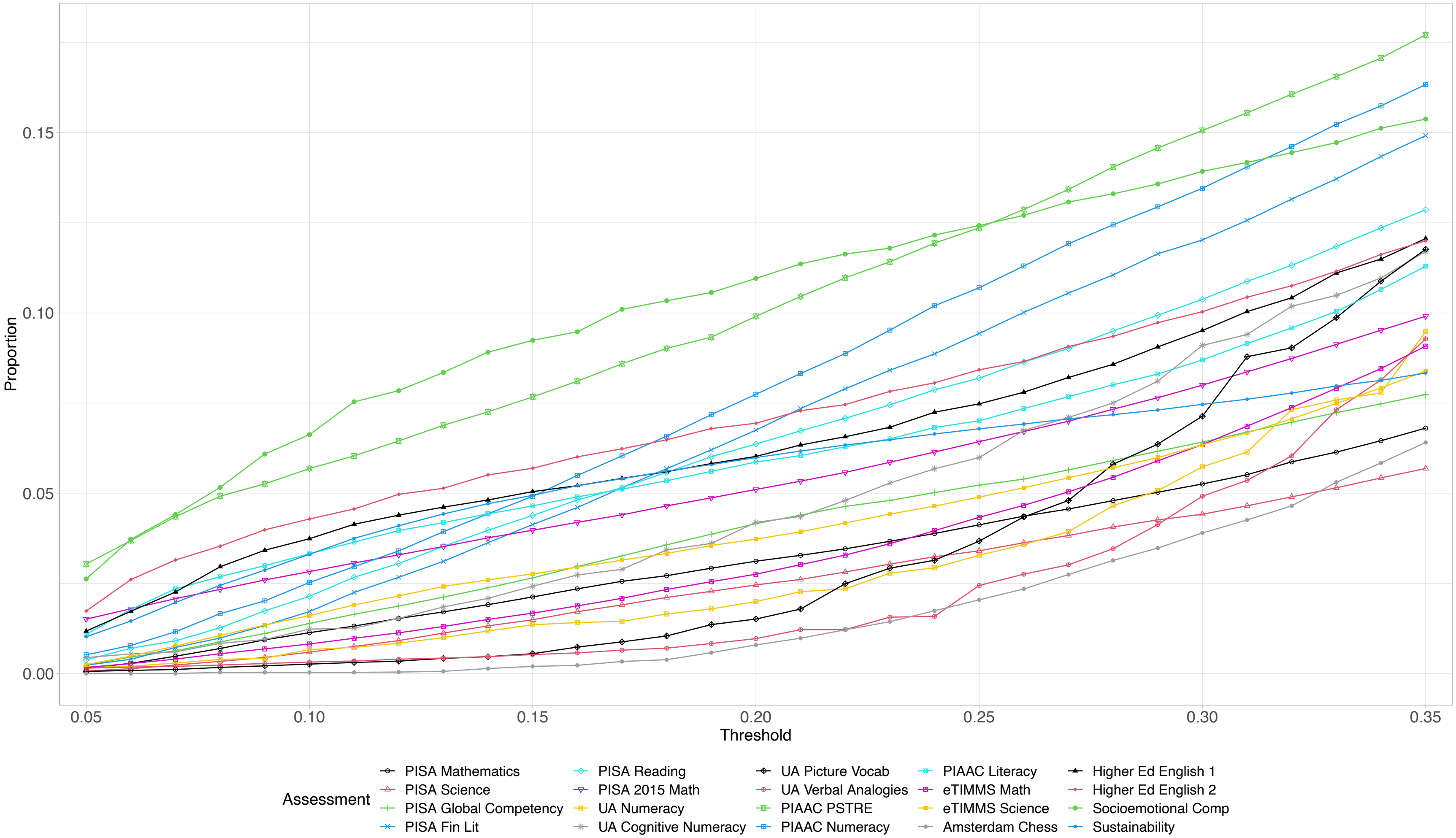


Figure 3. Proportion of examinees who engaged in rapid guessing behaviors on at least one item.

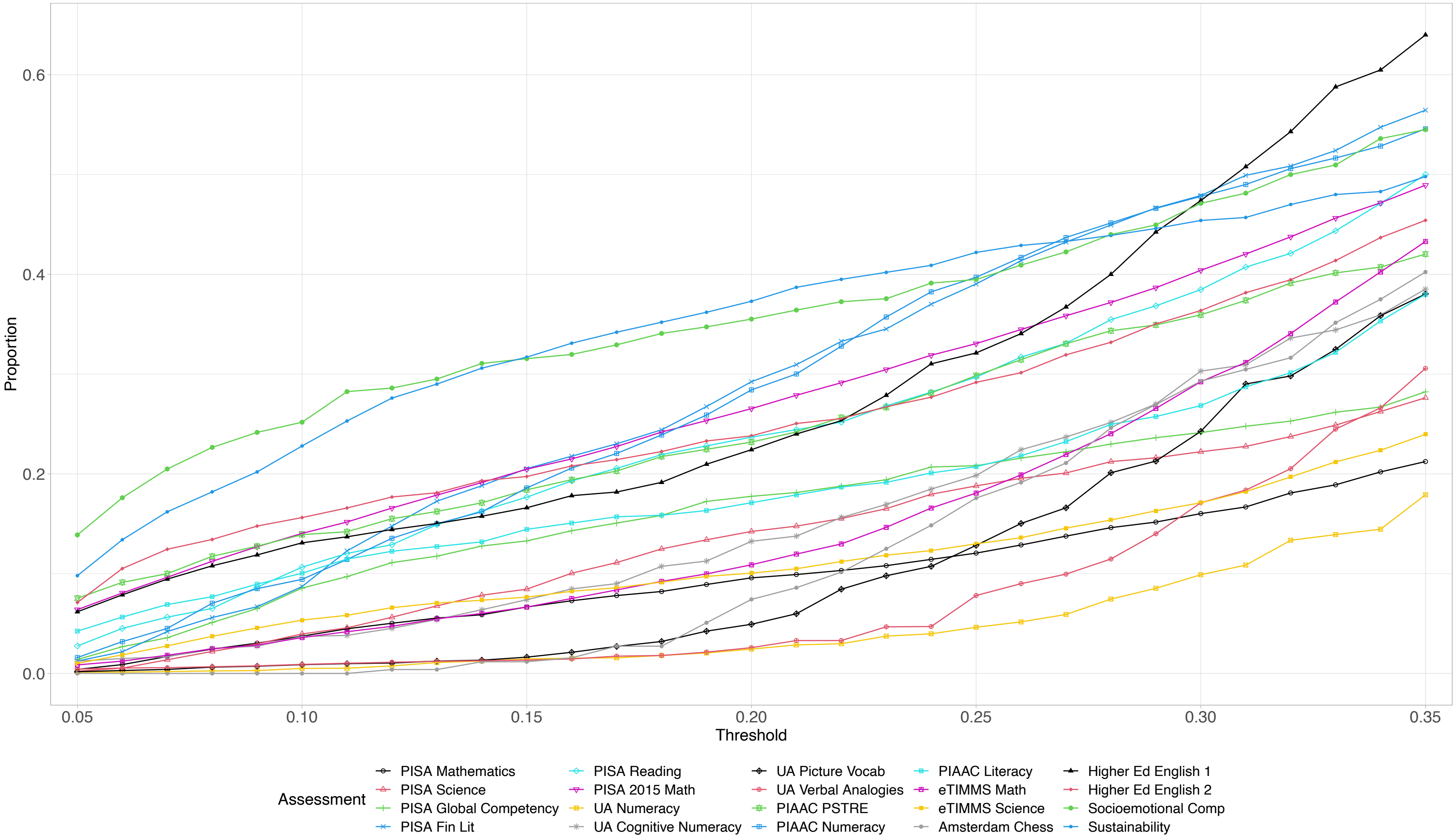


Figure 4. Proportion of examinees who engaged in rapid guessing behaviors on over 50% of items.

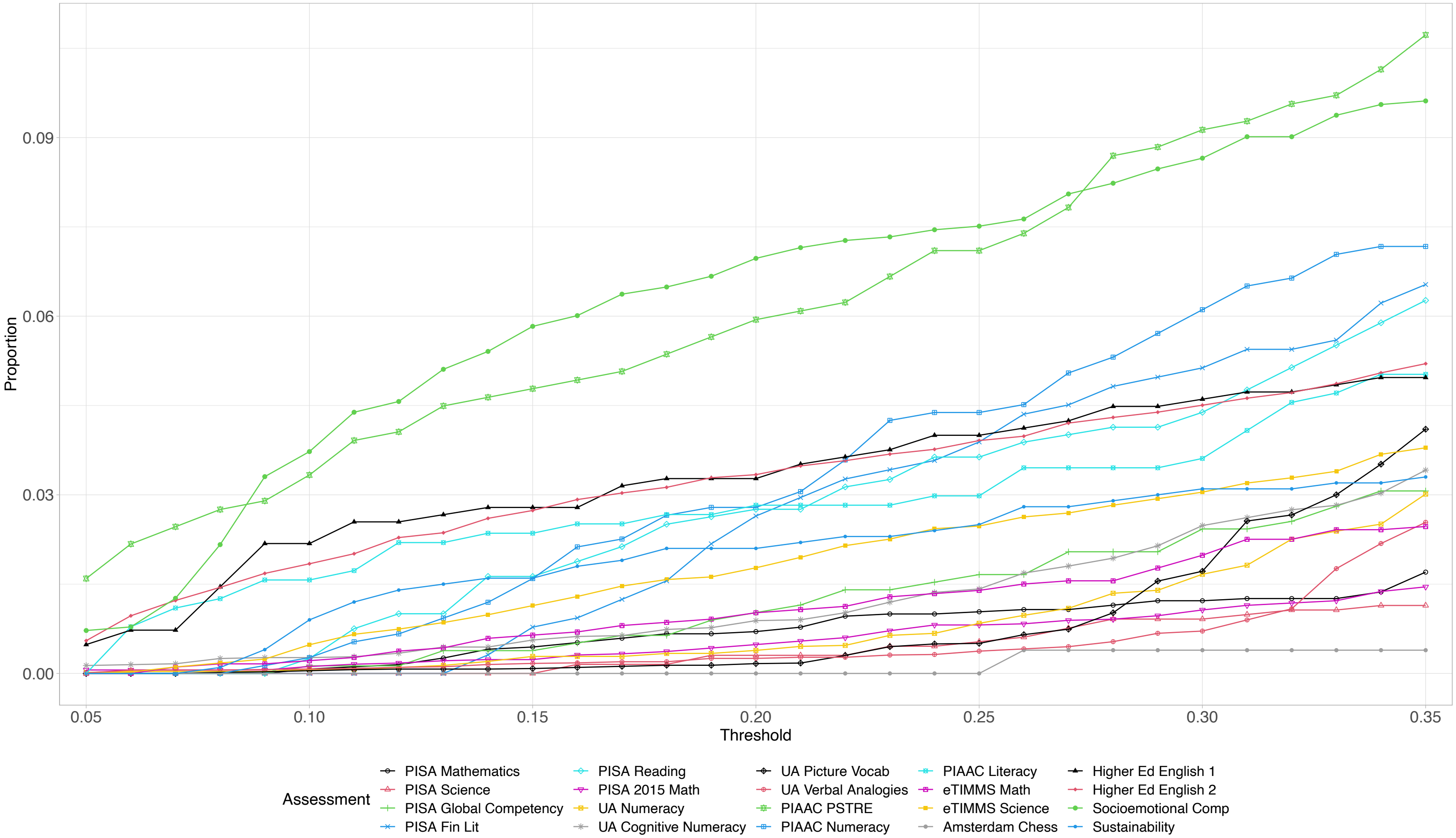


Figure 5. Proportion of examinees who engaged in rapid guessing behaviors on all items.

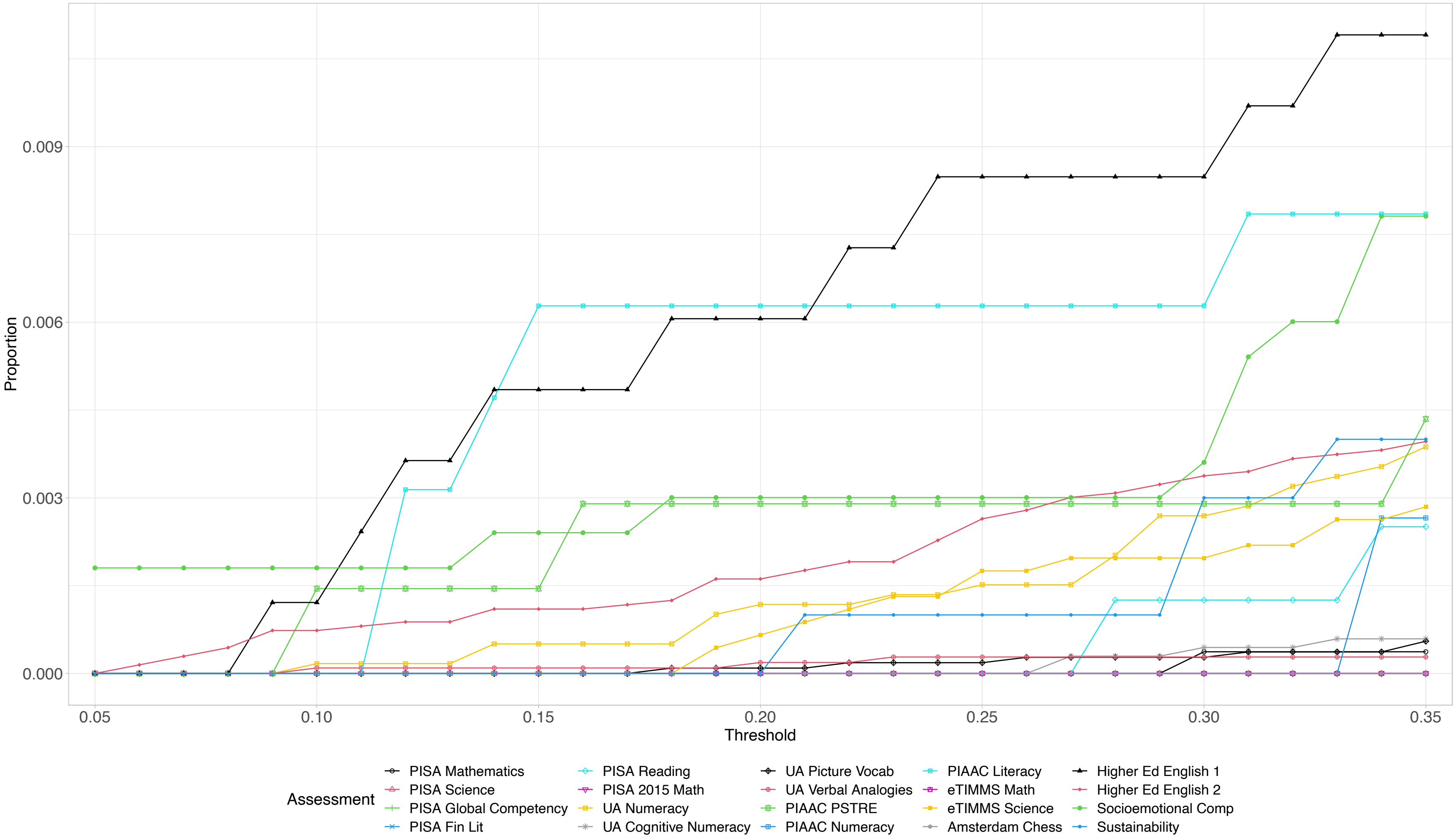


Figure 6. Comparative fit index as a function of scoring method and threshold value.

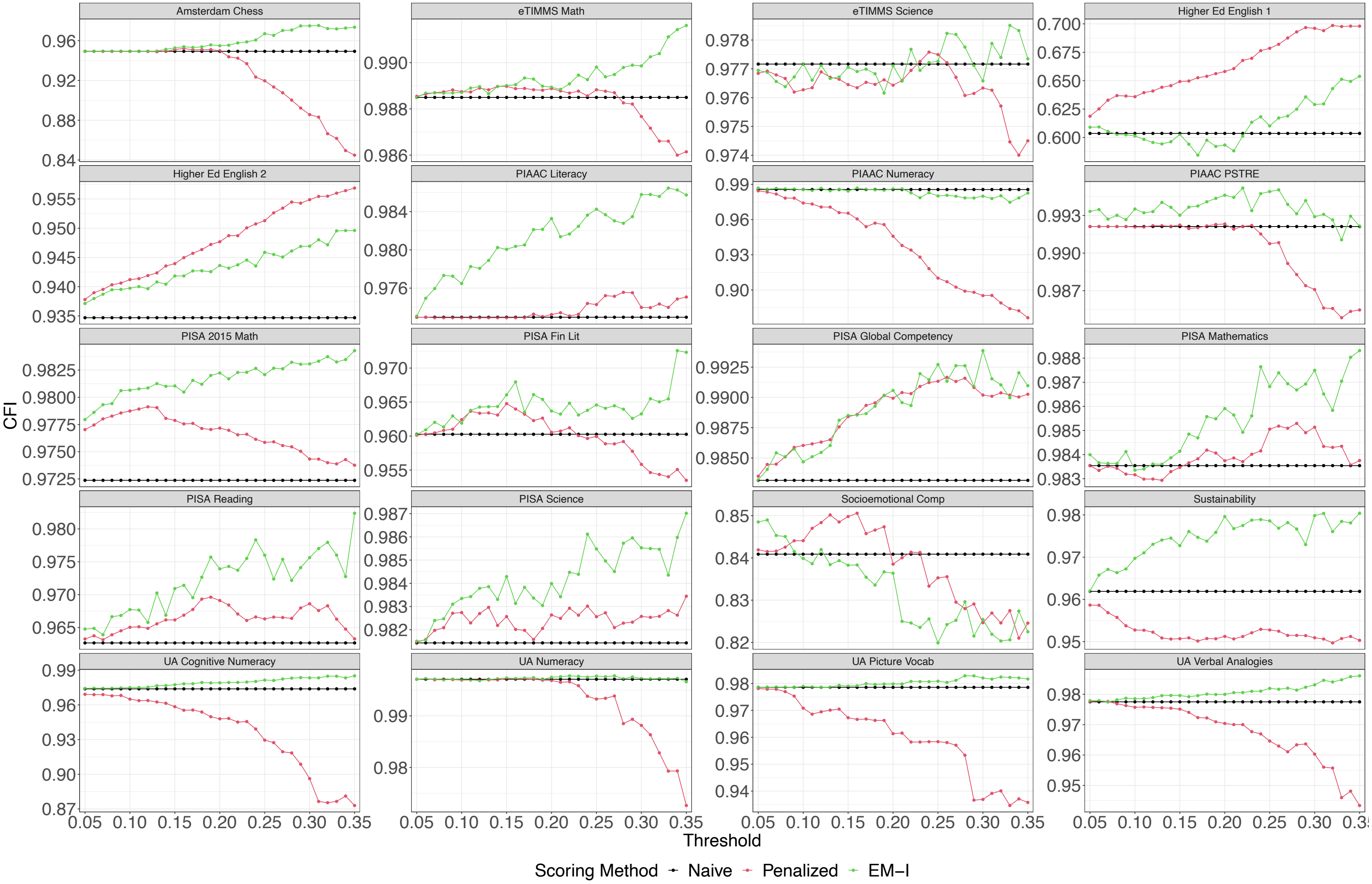


Figure 7. Tucker-Lewis fit index as a function of scoring method and threshold value.

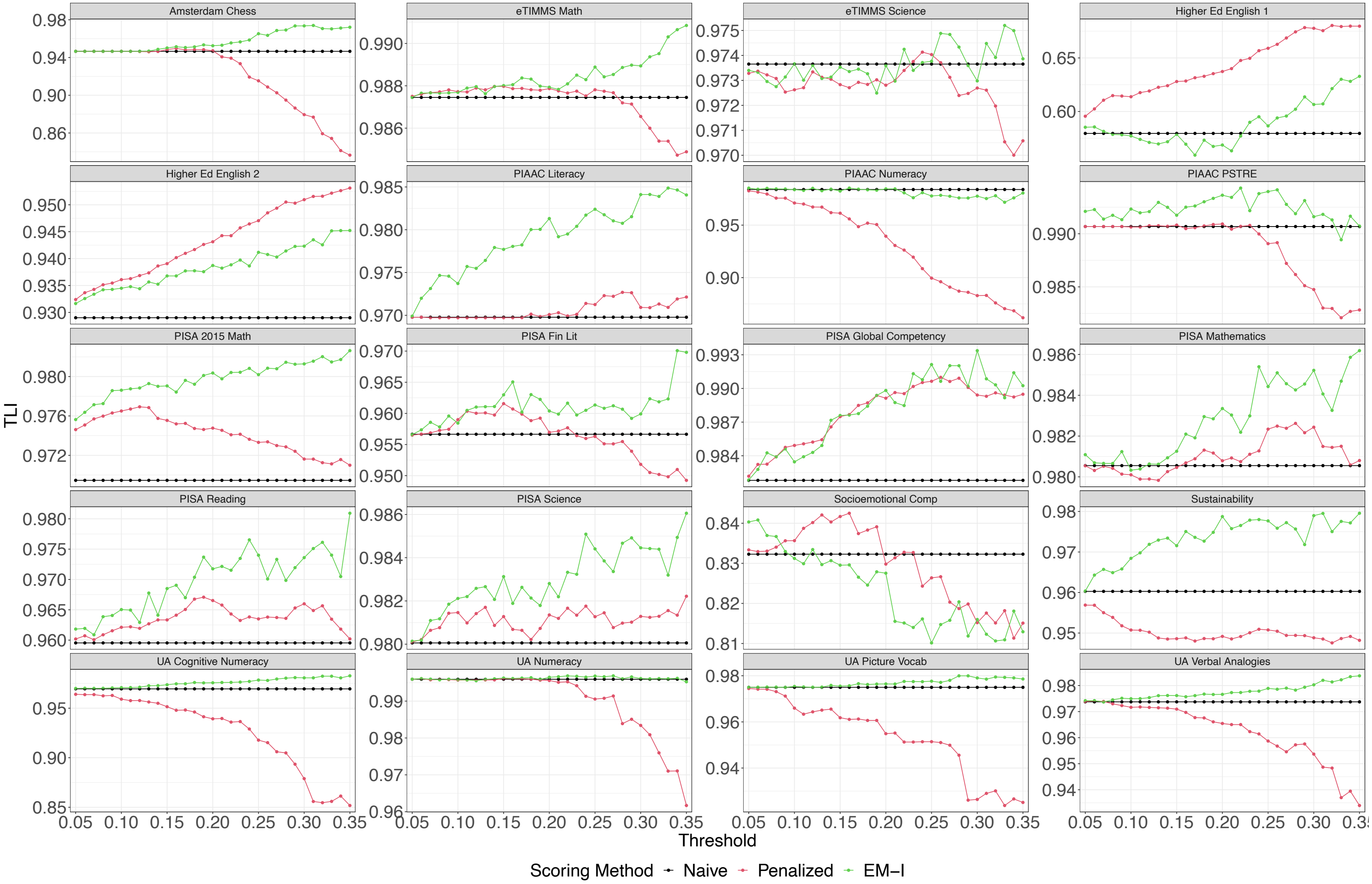


Figure 8. Root mean square error of approximation as a function of scoring method and threshold value.

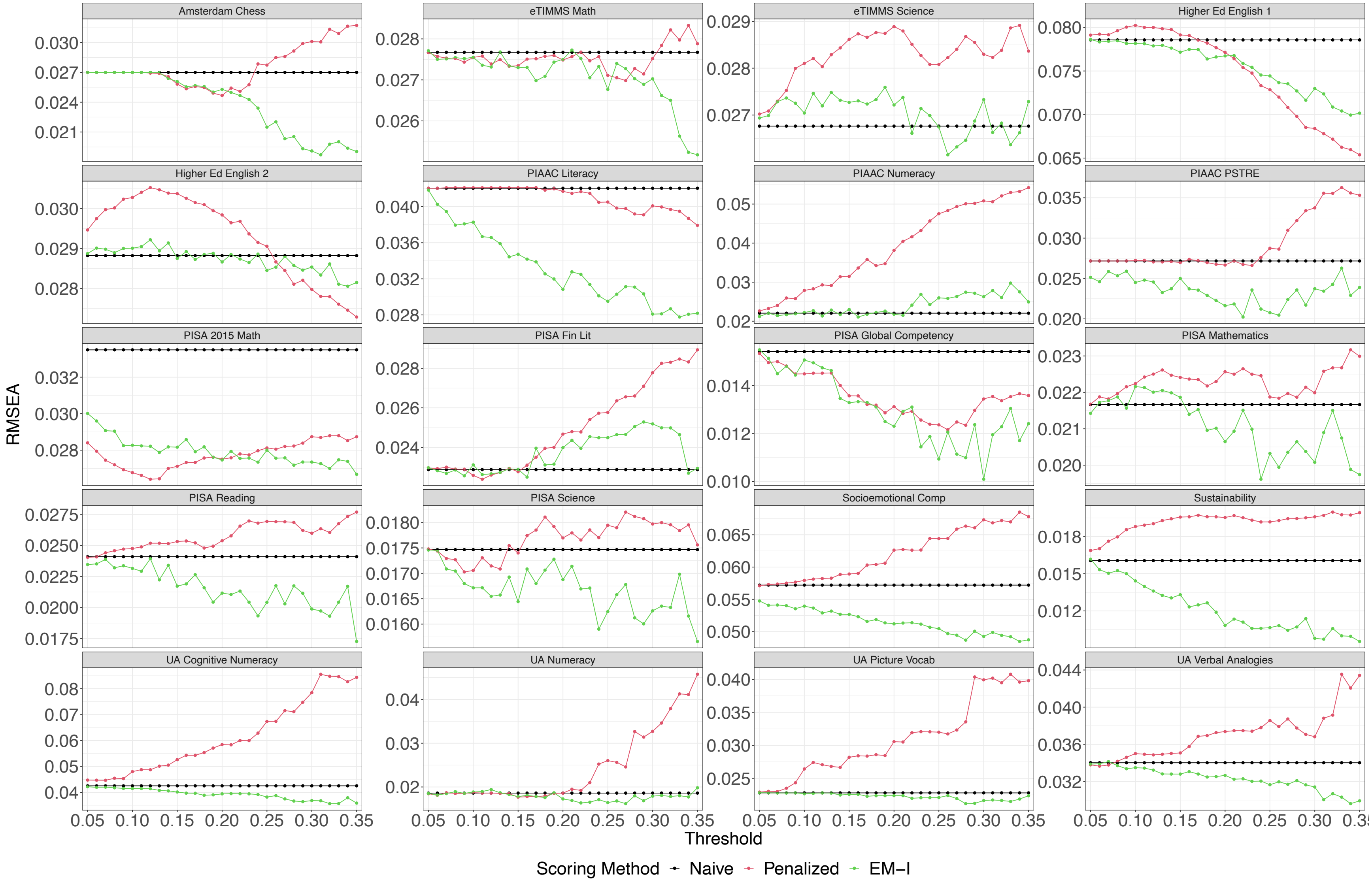
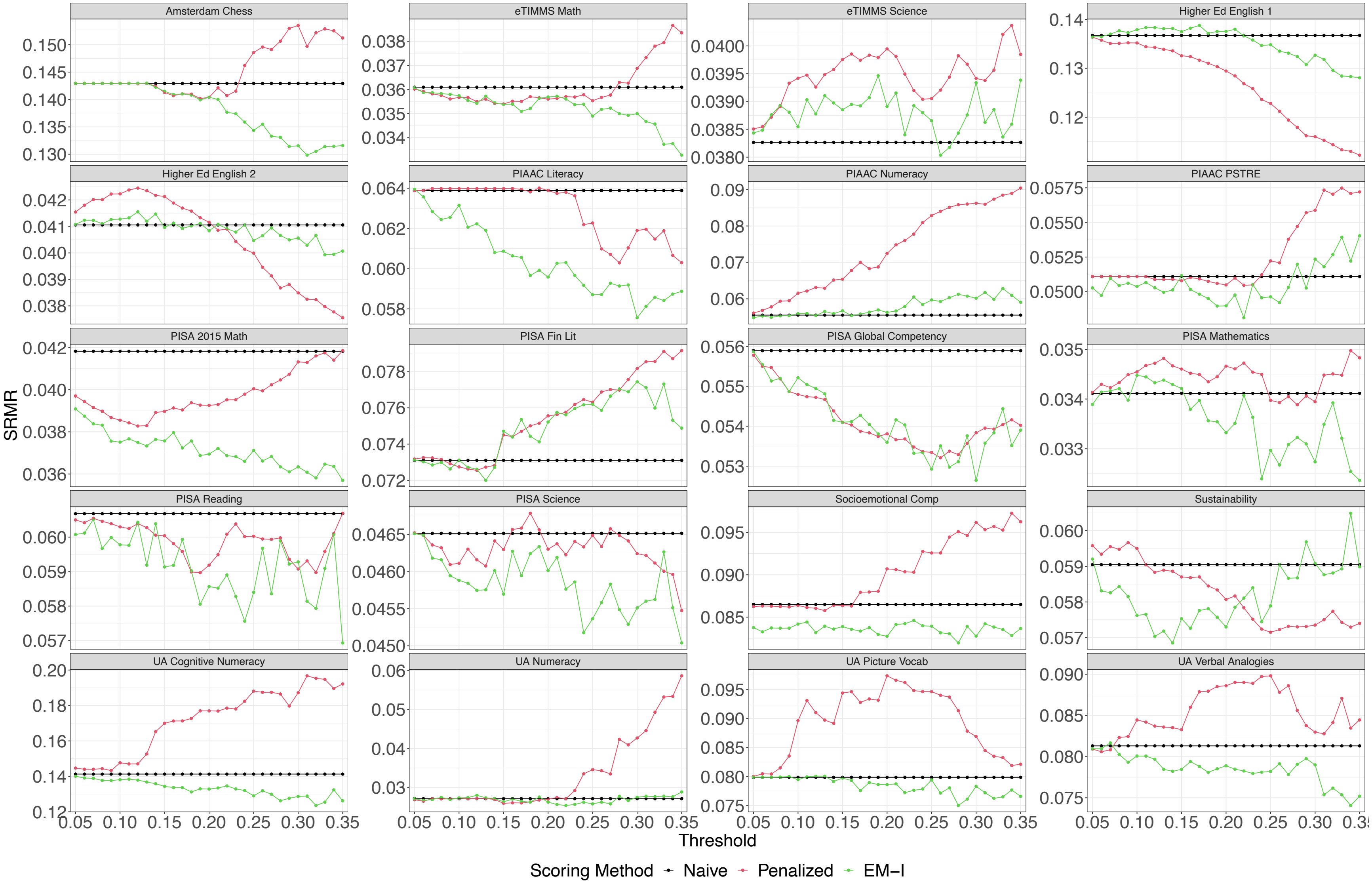
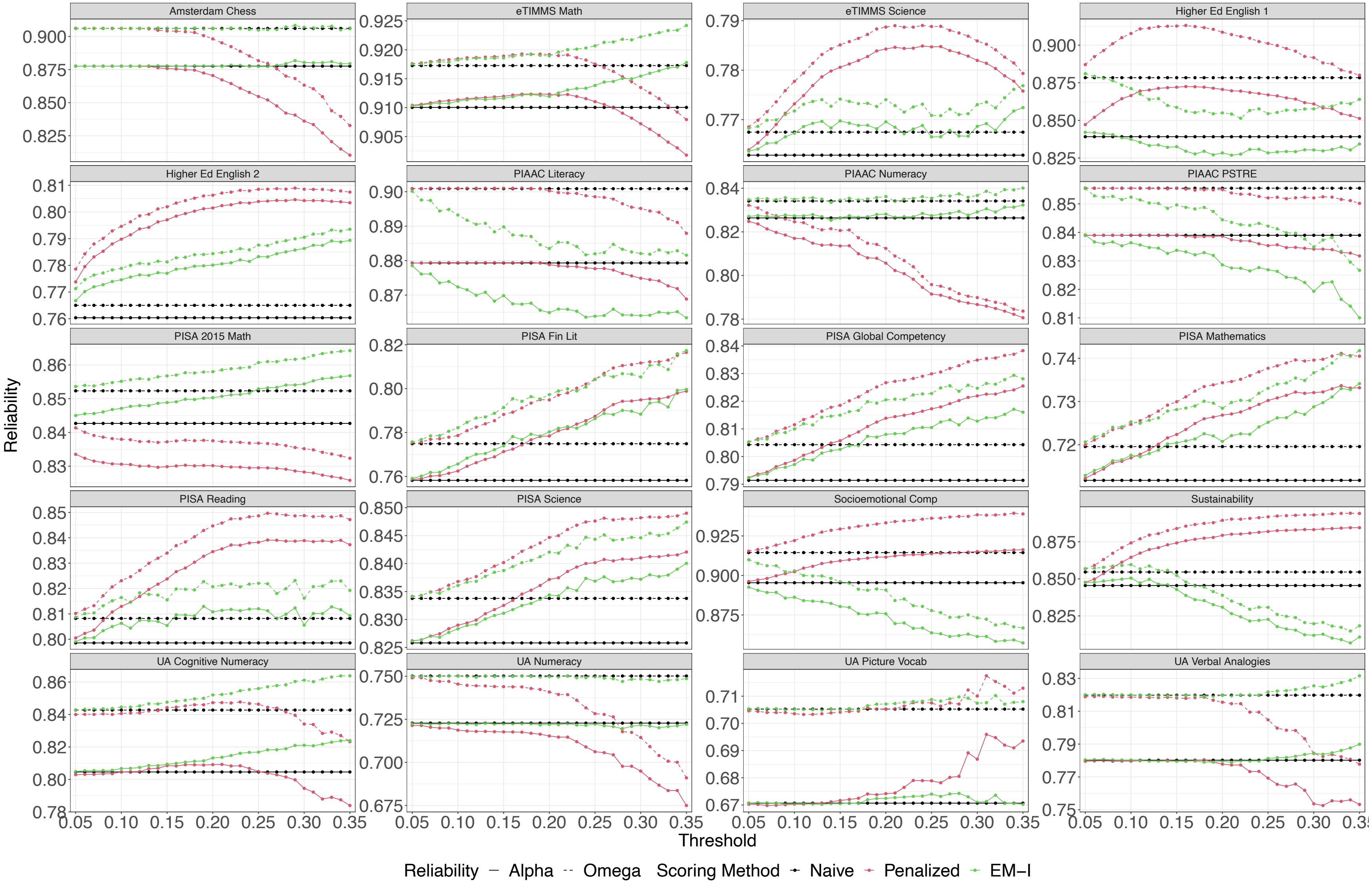


Figure 9. Standardized root mean residual as a function of scoring method and threshold value.



]

Figure 10. Reliability (coefficient alpha and coefficient omega) as a function of scoring method and threshold value.


Amsterdam Chess
eTIMMS Math
eTIMMS Science
Higher Ed English 1
Higher Ed English 2
PIAAC Literacy
PIAAC Numeracy
PIAAC PSTRE
PISA 2015 Math
PISA Fin Lit
PISA Global Competency
PISA Mathematics
PISA Reading
PISA Science
Socioemotional Comp
Sustainability
UA Cognitive Numeracy
UA Numeracy
UA Picture Vocab
UA Verbal Analogies
Reliability
Threshold
Reliability — Alpha -- Omega Scoring Method • Naive • Penalized • EM–I