Article

Empirical Priors in Polytomous Computerized Adaptive Tests: Risks and Rewards in Clinical Settings

Applied Psychological Measurement
2023, Vol. 47(1) 48–63

DOI: 10.1177/01466216221124091
journals.sagepub.com/home/apm

Niek Frans[1,2], Johan Braeken[3,1], Bernard P. Veldkamp[4], and Muirne C. S. Paap[1,2]

Abstract

The use of empirical prior information about participants has been shown to substantially improve the efficiency of computerized adaptive tests (CATs) in educational settings. However, it is unclear how these results translate to clinical settings, where small item banks with highly informative polytomous items often lead to very short CATs. We explored the risks and rewards of using prior information in CAT in two simulation studies, rooted in applied clinical examples. In the first simulation, prior precision and bias in the prior location were manipulated independently. Our results show that a precise personalized prior can meaningfully increase CAT efficiency. However, this reward comes with the potential risk of overconfidence in wrong empirical information (i.e., using a precise severely biased prior), which can lead to unnecessarily long tests, or severely biased estimates. The latter risk can be mitigated by setting a minimum number of items that are to be administered during the CAT, or by setting a less precise prior; be it at the expense of canceling out any efficiency gains. The second simulation, with more realistic bias and precision combinations in the empirical prior, places the prevalence of the potential risks in context. With similar estimation bias, an empirical prior reduced CAT test length, compared to a standard normal prior, in 68% of cases, by a median of 20%; while test length increased in only 3% of cases. The use of prior information in CAT seems to be a feasible and simple method to reduce test burden for patients and clinical practitioners alike.

[1]Department of Research and Innovation, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway
[2]The Nieuwenhuis Institute for Educational Research, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands
[3]Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo, Oslo, Norway
[4]Department of Research Methodology, Measurement and Data Analysis, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands

Corresponding Author:
Niek Frans, The Nieuwenhuis Institute for Educational Research, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Rozenstraat 38, Groningen 9712TJ, The Netherlands.
Email: n.frans@rug.nl

Keywords
computerized adaptive test, prior information, polytomous items, patient reported outcome measurement information system, measurement efficiency, item response theory

Computerized adaptive testing (van der Linden & Glas, 2000) is a powerful tool to administer tests that are tailor-made for the participant. By using item response theory (IRT; see Lord, 1980) to administer items that are matched on the participant's estimated trait level, Computerized Adaptive Tests (CATs) provide reliable trait estimates with considerably fewer items compared to static linear versions of the same test (Chang & van der Linden, 2003). Given that the trait level of a participant is unknown before any items have been administered, it is common practice to assume an average trait level as the starting estimate of each participant. However, other sources of information about the participant are frequently available (e.g., demographical characteristics of the participant, or scores from previous test administrations), and may be used to obtain a more accurate starting estimate. Several studies have explored how these sources of information can be used to improve CAT estimation and efficiency (e.g., Matteucci & Veldkamp, 2013; van der Linden, 1999).

These studies show that the inclusion of prior information about a participant can reduce both test length and estimation bias. Van der Linden (1999) concludes that the general use of prior information in educational assessment appears to be inhibited solely by the assumption that including information on prior test scores in performance assessment may be unfair to students. However, the use of prior information may be more acceptable—and thus more likely to be implemented—in clinical measurement, where tests are typically used as a diagnostic instrument rather than as a measure of aptitude (Matteucci & Veldkamp, 2013). In such settings, it is common practice to include information provided by the patient (regarding past experiences) or by multiple sources, in the assessment procedure. For example, Achenbach (2006) posits that data from multiple informants is essential in the assessment of psychopathology and personality.

Thus far, studies investigating the use of prior information in CAT have focused on the benefits of empirical prior information in settings commonly found in educational contexts, where CATs are often supported by large item banks of dichotomous items (Matteucci & Veldkamp, 2013; van der Linden, 1999). It remains unclear how the uniformly positive message of using empirical prior information in CAT will generalize to applied clinical settings. Item banks used in clinical practice are typically much smaller, based on highly informative polytomous items, and aimed to measure conceptually narrow pathological constructs (Reise & Waller, 2009). These characteristics often lead to very short CATs and item banks that only provide adequate information regarding the pathological range of the latent trait scale (Reise & Waller, 2009). In addition, while previous studies have focused mainly on the benefits of using prior information in ideal settings, little is known about the potential risks when prior information is not perfectly accurate. Overconfidence in inaccurate prior information may in fact increase test length and/or lead to severely biased final trait estimates, by selecting an incorrect starting point or introducing bias in the trait estimation process, and administering items that do not match the participant's trait level. In this paper, we explore both the potential reward and risk of using empirical prior information in circumstances resembling realistic clinical CAT settings.

Using Empirical Prior Information in CAT

Various authors have discussed how prior information may be included in latent trait estimation in IRT and CAT (e.g., van der Linden, 1999; Zwinderman, 1991). In CAT, it is quite common to rely

on an empirical Bayes paradigm to estimate a person's latent trait. Such a Bayesian estimation paradigm avoids the drawbacks of simple maximum likelihood estimation (e.g., infinite estimates for all incorrect/correct response patterns) for CAT, by including a prior distribution on the latent trait to supplement the likelihood of the observed response data in order to approximate the posterior distribution of a person's latent trait (Bock & Mislevy, 1982). Two common empirical Bayes estimators are the Expected A Posteriori (EAP) and the Maximum A Posteriori (MAP), which use either the mean or the mode of the posterior distribution of the latent trait as a point estimate, respectively. In absence of empirical prior information, the common default prior distribution is the standard normal distribution $\widehat{\theta}_p^{(0)} \sim N(0,1)$. However, with available empirical prior information, an empirical prior distribution can be defined in various ways (see, for example, He et al., 2019; van der Linden, 1999); for instance, by making the mean and variance of the normal distribution person specific $\widehat{\theta}_p^{(0)} \sim N(\mu_p, \sigma_p^2)$. In such a personalized empirical prior, the mean μ_p reflects the location of person p on the latent trait, and the variance σ_p^2 reflects the level of uncertainty of the information on that location. The inverse of the prior variance is also known as the precision (i.e., $1/\sigma_p^2$) of the prior.

Bayesian estimators such as EAP and MAP are so-called shrinkage estimators that pull the person's latent trait estimate $\widehat{\theta}_p$ away from the maximum likelihood estimate towards the location μ_p of the prior. The degree of shrinkage depends on the relative precision of the prior compared to the amount of information present in the data. Hence, EAP and MAP estimates are biased by definition, unless the test is perfectly reliable or the location μ_p of the prior is identical to the participant's true location θ_p on the latent trait (Kolen & Tong, 2010). This slight increase in bias is typically counterbalanced by overall lower error variance.

In an educational context, Matteucci and Veldkamp (2013) illustrated how, compared to a default standard normal prior, the adoption of a personalized empirical prior distribution with a location close to the person's true trait level leads to shorter test lengths in fixed-precision CATs. Furthermore, even the estimation bias was reduced; with largest effects for extreme θ participants, where the mismatch with the default standard normal prior was greatest, and whose estimates would otherwise be shrunk towards zero.

These results suggest that using empirical prior information can be highly rewarding. However, it is important to note that these advantages are dependent on the prior information giving a precise and unbiased initial estimate of the latent trait (van der Linden, 1999). There is a risk that the location μ_p of the prior might substantially diverge from the participant's true latent trait θ_p, making our initial estimate in fact severely biased. Moreover, since item selection in CAT is conditional on the trait estimate, an incorrect initial estimate could lead to the selection of an item that is not very informative for the estimation of the person's latent trait, and in the worst case could lead the CAT astray. An initial misstart by the participant, that is not responding in line with their true latent trait value, could lead to a string of mismatched items making a fixed-precision CAT unnecessarily long, and potentially inducing substantial estimation bias in a short fixed-length CAT (e.g., Chang & Ying, 2008; Rulison & Loken, 2009). However, if a test is sufficiently long and reliable, the influence of the prior on the trait estimate will eventually dissipate (van der Linden & Pashley, 2010). The purpose of a CAT is to provide a relatively short, accurate test, whereby one should bear in mind that, especially in clinical settings, due to the small number of items, the risks associated with using a mismatched prior might be substantial when using a CAT.

Current Study

In sum, whereas previous studies have focused on the rewards for CAT—in terms of accuracy and efficiency—that result from using an unbiased personalized empirical prior in educational contexts, we focus on both the risks and rewards of using empirical priors in a clinical CAT setting. Since CATs are increasingly used in the domain of health measurement, and the inclusion of prior information may be highly applicable in this context, exploring the impact of utilizing prior information in clinical CATs is particularly relevant. Since item banks in clinical contexts generally consist of a relatively small number of highly informative polytomous items, it is unclear whether the uniformly positive message of using empirical prior information in CAT will generalize to applied clinical settings. To explore the risks and rewards of a personalized empirical prior in a clinical context, we conducted a simulation study using item banks that were simulated based on characteristics found in the Patient Reported Outcome Measurement Information System (PROMIS; Reeve et al., 2007), one of the most ambitious and widely known CAT applications in health care. A second study was conducted using item banks and empirical priors that have been simulated based on a structured clinical interview to assess personality functioning (Hummelen et al., 2021). By using realistic examples and personalized empirical priors of varying quality, we explored both the risks and rewards of using personalized empirical priors as compared to a generic empirical prior or a commonly used standard normal prior.

Simulation Study 1

In this study, we compared the performance of a fixed-precision CAT using a personalized empirical prior and two generic priors that represent a general and clinical population with fixed prior location μ. We varied the quality of the personalized empirical prior by systematically changing the degree of bias in the location of the prior distribution and the precision of the prior.

Item Bank

To ensure that simulated item banks had realistic properties, item bank sizes and item parameters were based on characteristics found in empirical item banks in PROMIS, similar to a recent study by Paap et al. (2019).[1] We simulated 100 item banks for two different sizes (i.e., length $N = \{30, 60\}$) that represent moderately sized and larger item banks, currently found in the PROMIS adult measures database[2]. Item bank size was varied, as smaller item banks generally have fewer informative items at any given trait location, which may make them less effective at mitigating the influence of a biased prior. Each item bank consisted of polytomous items with five response categories calibrated under the Graded Response Model (GRM; Samejima, 1996). The fourth category threshold parameter for all items was sampled from a normal distribution, with mean 2.2 and variance 0.16. The stepwise distances towards the other category thresholds in an item were sampled from a log-normal distribution with mean 0.75 and variance 1.44. Item discrimination parameters were sampled from a truncated normal distribution with location and scale of 3.5 and 1.0, respectively, on the interval [1.5, 5]. The resulting test information functions for simulated item banks of length $N = 30$ and of length $N = 60$ can be found in the online supplement. For average item banks of both lengths, information is maximized for trait values of $\theta_p = 1.1$ (i.e., higher degrees of pathology), which is consistent with the fact that PROMIS item banks are calibrated in such a way that a trait value of zero represents trait characteristics in the general population (Reeve et al., 2007).

Simulees

To evaluate the performance of the different CATs across the latent trait space, the true θ values of simulees were generated according to a grid ranging from −1.0 to 3.0 in increments of 0.5. These θ values represent the target population of the PROMIS instruments used as a basis for the simulation, and were matched to the item bank information curve to ensure that the vast majority of generated item banks would supply sufficient information for the CAT to reach the fixed-precision stopping threshold. At each of the nine θ grid points, 100 simulees were located, resulting in a total of $p = 900$ simulees for each item bank.

CAT Administration

All CAT simulations were run in R version 4.1.1 (R Core Team, 2021) with the mirtCAT package version 1.11 (Chalmers, 2016), and customized scripts for the setup of the priors, the data simulation, and the statistical analyses of the CAT results. The scripts can be found online at https://github.com/Niek-F/CAT_Empirical_Prior.

To initialize each CAT, the most informative item in the item bank given the prior location μ was selected as the starting item. Likewise, subsequent items were selected based on the maximum Fisher information criterion. Maximum A Posteriori estimation was used to estimate the location of the simulee on the latent trait scale.

Stopping rule and constraints. A fixed-precision stopping criterion was used, with the following threshold for the standard error of the trait estimate: $SE(\widehat{\theta}_p) \leq 0.316$ (i.e., roughly corresponding to a required local reliability of $1–.316^2 = .90$); in combination with a minimum number of items constraint. Babcock and Weiss (2012) suggest that setting a minimum number of 15–20 dichotomous items might prevent fixed-precision CATs from terminating before converging on the true trait estimate. Considering that a 5-category polytomous item can be viewed as a collection of 4 dichotomous pseudo-items, and that items in a clinical context tend to be more informative than in an educational context, we set the minimum number of items to be administered in the CAT at 2, 3, or 4 items; alongside a CAT administration without such a constraint (i.e., minimum number of items = 1). Thus, we have a total of four different stopping rule variants in our study.

Prior conditions. Three different priors were used in each CAT: (i) a *personalized* prior $\widehat{\theta}_p^{(0)} \sim N(\mu_p, \sigma_p^2)$ with a location that is based on the participant's true trait; (ii) a *generic clinical prior* $\widehat{\theta}_p^{(0)} \sim N(1, 1)$, which represents the clinical target population of the item banks and has a prior location that is closely aligned with the trait value for which the average item bank provides maximum information; and (iii) a *generic default standard normal* prior $\widehat{\theta}_p^{(0)} \sim N(0, 1)$, with a location that represents average trait values in a non-clinical population. Including these three prior conditions facilitates direct comparison of CAT performance under a person-specific empirical prior with both a generic empirical prior that takes a pathological value as an initial assumption of the person's trait and a commonly used standard normal prior.

The accuracy of the personalized prior was manipulated by varying the degree of bias in the location of the prior distribution (i.e., $\mu_p = \theta_p + \text{bias}_0$; with $\text{bias}_0 = \{-2, -1, 0, 1, 2\}$). The selected values represent a bias of one or two standard deviations from the mean, under the generic prior distributions that are used for comparison purposes. The precision of the personalized prior was manipulated by varying the variance of the distribution (i.e., $\sigma_p^2 = \{1.00, 0.50, 0.25\}$). The highest variance of 1.00 equaled the variance of the generic priors and was included to assess the

isolated effect of bias in the personalized prior. The lowest variance was set higher than the fixed-precision CAT's stopping criterion (otherwise, the CAT would immediately stop, before any item was administered).

Each simulated item bank was used to generate a new dataset $\boldsymbol{Y}$ consisting of item responses of the $p = 900$ simulees on all N items[3]. For each simulee we ran 5 (bias in personalized prior location) × 3 (precision of personalized prior) × 4 (min. item constraint) = 60 CAT administrations with a personalized prior; and 2 (generic priors) × 4 (min. item constraint) = 8 CAT administrations with a generic prior. Our experimental design with respect to the prior was within-subject (i.e., same item response data $\boldsymbol{Y}$ for the different CAT administrations per replication of an item bank), facilitating comparisons at an individual level across prior conditions. Item bank size was the only between-subject experimental factor (100 replicated item banks per item bank size, with each replication having its own new item response datasets $\boldsymbol{Y}$).

Evaluation criteria. The outcome measures were first computed at the individual level, and then summarized at the θ-grid level by calculating the mean for the 100 simulees on each of the 9 latent grid points. These summary measures represent the expected values at the grid level per replicated item bank. Variation in these summary measures reflects variation due to differences in the underlying item bank replications. Outcomes were reported in terms of the grand mean and standard deviation (*SD*) over item bank replications for central tendency, and spread of these grid-level outcome measures, respectively, supplemented by figures in which the error bars depict the range [min, max] of values.

Convergence and test length. The proportion of CATs that did not reach the fixed precision stopping threshold, and hence did not converge, was computed. The variance of the estimates in most fixed-precision CATs will result in a near-constant, due to the CATs terminating at the stopping precision $\mathrm{SE}(\widehat{\theta}_p) \leq 0.316$. Because the standard errors of the CAT estimates contain little to no variance, we did not include a variance-based outcome measure, but instead looked at differences in test length. Both the absolute test length n, and the relative test length were computed to evaluate CAT efficiency, as a more efficient fixed-precision CAT will reach the stopping precision using fewer items (i.e., shorter test lengths). Before summarizing at the θ-grid level, the relative test length was calculated as a ratio $n_p^{(Empirical)}/n_p^{(Generic)}$ for the personalized prior conditions, using the number of items administered under each generic prior in the denominator given the same CAT stopping rule.

Estimation bias. The accuracy of a simulee's latent trait estimate $\widehat{\theta}_p$ was assessed in terms of its estimation bias $BIAS(\widehat{\theta}_p) = \widehat{\theta}_p - \theta_p$. For the personalized prior conditions, absolute estimation bias was also expressed relative to the two generic priors as $\Delta|BIAS(\widehat{\theta}_p)| = |BIAS(\widehat{\theta}_p^{(Empirical)})| - |BIAS(\widehat{\theta}_p^{(Generic)})|$.

Results

Convergence

When supported by the larger item bank ($N = 60$), nearly all CATs (>99.9%), regardless of what prior was used, reached the required fixed-precision stopping criterion. For the smaller item banks ($N = 30$), 2.3% of CATs ran to bank depletion. Nearly all of these cases (89.8%) occurred under the personalized prior CATs, specifically for simulees with an extreme θ value (i.e., $\theta_p = 3.0$), that, combined with prior bias, led to an extreme prior location (e.g., $bias = 2$, $\theta_p = 3$, $\mu_p = 5$).

Test Length

When using the generic standard normal prior, average test length between simulations varied between 2.1 and 4.1 items for the larger ($N = 60$) item bank, with a mean of 2.6 items. Tests were slightly longer when the smaller ($N = 30$) item bank was used, and varied between 2.1 and 5.9 items with an average of 2.9 items. Using the generic clinical prior generally resulted in tests that were nearly equally long between both item bank sizes. The average test length for the large item bank was 2.8 items, and 3.1 items for the smaller item bank, when utilizing the generic clinical prior. In line with expectations, test length varied as a function of the true θ value, especially in the smaller item bank. Compared to simulees with trait values at the center of the distribution ($\theta_p = 1$), average test length was nearly three times longer at the lowest end of the latent trait scale ($\theta_p = -1$), and roughly two times longer at the highest end of the latent trait scale ($\theta_p = 3$). For the larger item bank, tests were around two times longer at either end of the scale. This is in accordance with the test information function of the item bank having less information in those areas of the scale.

Figure 1 shows the average reduction in test length for the personalized prior condition, relative to a generic standard normal prior. Since both generic priors performed highly comparable in terms of test length, only the comparison with the standard normal prior is shown for both item banks. The top central panel of Figure 1 shows that, given the same prior variance ($\sigma_p^2 = 1$), an unbiased personalized prior did not noticeably reduce test length, compared to a standard normal prior. When utilizing an unbiased personalized prior, average test lengths for the smaller item bank

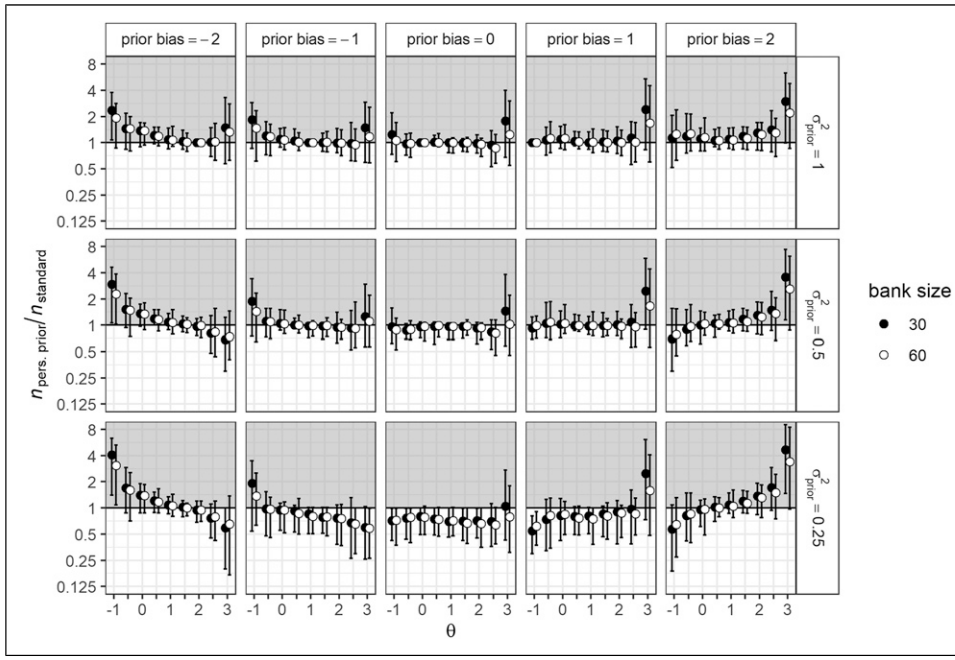

Figure 1. Average relative differences in test length between the generic default and personalized prior conditions for different θ values. Split by item bank size, prior variance, and prior bias. *Note.* Unit-distances on the *y*-axis are log-transformed to reflect the inverse equivalence of a test that is twice as long and a test that is half as long. *Y*-values lower than 1 (white area) indicate that tests are shorter, when using a personalized prior relative to a standard normal prior. The error bars indicate the range of replication averages over 100 simulees.

were up to 1.8 times longer for simulees with $\theta_p = -1$ or $\theta_p = 3$, as the item bank contained few informative items for these trait values.

When the location of the personalized prior was biased, relative test length increased slightly for most θ values (Figure 1, top row). As expected, tests were generally shorter when the variance of the personalized prior σ_p^2 was lower. For example, an unbiased personalized prior reduced test length by a factor of 1.3 when prior variance $\sigma_p^2 = 0.25$, compared to prior variance $\sigma_p^2 = 1.00$. However, a more precise prior led to a drastic increase in test length, when bias in the prior location moved the initial estimate towards an extreme location on the latent trait (e.g., *bias* $= 2$, $\theta_p = 3$, $\mu_p = 5$, bottom right panel of Figure 1). Equivalently, bias in the prior location that resulted in a starting estimate closer towards the center of the latent trait distribution, compared to the location of the generic default prior (e.g., *bias* $= -1$, $\theta_p = 3$, $\mu_p = 2$), resulted in empirical prior CATs that were generally shorter in test length compared to a generic default prior CAT. In essence, bias in the prior location led to relatively shorter/longer tests, when this bias brought the starting estimate closer to/further away from the trait value areas for which the item bank provided maximum information.

Estimation Bias

Average estimation bias varied from −0.34 to −0.13, with a mean of −0.09, when using the generic default prior. That this default prior resulted in negatively biased estimates was to be expected, since trait estimates were shrunk towards the location of the prior (i.e., 0). A generic clinical prior did result in more unbiased estimates (mean estimation bias = 0.00) that varied

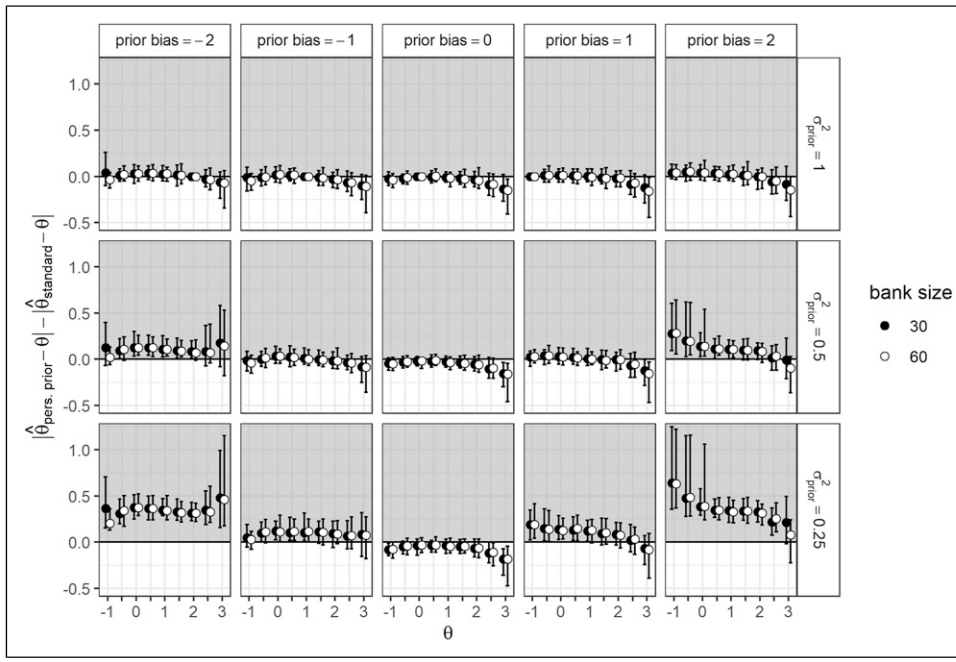

Figure 2. Average differences in absolute estimation bias between the generic default and personalized prior conditions for different θ values. Split by item bank size, prior variance, and prior bias. *Note.* A positive value on the *y*-axis indicates that the estimate using a standard normal prior is less biased than the estimate using a personalized prior.

between −0.23 and 0.22. As expected, both population priors resulted in an estimation bias trend that pulled the estimates towards their respective location (i.e., 0 for the generic default prior and 1 for the clinical prior); consequently, the most extreme estimation bias was found for θ values that were furthest from the prior location. In contrast, mean estimation bias under an *unbiased* personalized prior (i.e., with location set at the true θ value) was near-zero and ranged from −0.03 to 0.02 across the θ-grid. In line with expectations, when increasing bias in the prior location, estimates were increasingly biased towards the location of the prior. The overall effect was symmetrical for positive and negative bias in the prior location, with a mean estimation bias of −0.33 when the prior bias was −2 and −0.16 when prior bias was −1. Mean estimation bias did not differ across bank sizes for the three prior conditions.

There was a mean reduction in absolute estimation bias of 0.05 ($SD = 0.05$) compared to the generic default prior, and a mean reduction of 0.04 ($SD = 0.03$) compared to the generic clinical prior. Figure 2 shows the difference in absolute bias between the personalized prior and a standard normal prior for both item bank sizes[4]. As shown in Figure 2, the differences in estimation bias between CATs supported by an unbiased personalized prior and the generic standard normal prior CAT were somewhat larger for higher θ values ($\theta_p = 3.0; \Delta|BIAS(\widehat{\theta}_p)|$: mean = −0.16, $SD = 0.02$). The figure also shows how absolute average estimation bias under the personalized prior CATs varied as a function of bias in the prior's location. This negative effect of prior bias on estimation bias was severely amplified when the variance of the prior was lower. In line with expectations, when comparing the bottom left and top right panels in Figures 1 and 2, we can see that trait estimates were particularly biased for θ_p values that were associated with relatively short tests.

Influence of Test Length Constraints

Changing the constraint on the minimum number of items to be administered during the CAT influenced the results in two ways. Average test length increased as the required minimum number of items in a CAT increased. Consequently, average absolute estimation bias decreased as the average test length increased. These two effects can be seen in Table 1 for the two generic priors and the unbiased personalized prior used in this study. Although the unbiased personalized prior initially slightly outperformed the other two conditions in terms of test length and estimation bias,

Table 1. Effect of Setting a Minimum Number of Items on Average Absolute Estimation Bias and Average Test Length for Three Prior Conditions and Effect of Bias in Prior Location and Prior Precision on Test Length and Estimation Bias When Administering a Minimum of Four Items Before Terminating the CAT.

	Between conditions comparison											
	Default prior $N(0, 1)$				Clinical prior $N(1, 1)$				Personalized prior $N(\theta_p, \sigma_p^2)$			
	Test length		Abs. est. bias		Test length		Abs. est. bias		Test length		Abs. est. bias	
Min. items	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
—	2.61	0.70	0.132	0.101	2.84	0.98	0.112	0.076	2.26	0.73	0.012	0.012
2	2.61	0.70	0.132	0.101	2.84	0.97	0.112	0.076	2.38	0.64	0.009	0.012
3	3.26	0.43	0.104	0.092	3.45	0.69	0.090	0.070	3.21	0.45	0.004	0.006
4	4.13	0.28	0.080	0.075	4.26	0.45	0.071	0.059	4.15	0.34	0.003	0.003

Note. Min. items = minimum number of items administered. Abs. est. bias = absolute estimation bias. σ_p^2 = variance of the personalized prior, outcomes in this table are aggregated over the different variance values.

Table 2. Average Test Length and Absolute Estimation Bias for CATs Utilizing a Personalized Prior $\widehat{\theta}_p^{(0)} \sim N(\mu_p + bias, \sigma_p^2)$ with a Minimum of 4 Items, Split by Prior Variance and Bias in Prior Location.

	$\sigma_p^2 = 1.00$				$\sigma_p^2 = 0.50$				$\sigma_p^2 = 0.25$			
	Test length		Abs. est. bias		Test length		Abs. est. bias		Test length		Abs. est. bias	
Prior bias	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
−2	4.53	1.17	0.111	0.024	4.57	1.61	0.216	0.045	4.91	2.57	0.403	0.075
−1	4.34	0.74	0.051	0.012	4.30	0.73	0.100	0.018	4.23	0.69	0.188	0.035
0	4.26	0.51	0.004	0.002	4.15	0.29	0.003	0.002	4.05	0.10	0.004	0.004
1	4.33	0.68	0.054	0.014	4.29	0.66	0.102	0.019	4.21	0.62	0.190	0.037
2	4.54	1.09	0.113	0.027	4.53	1.48	0.223	0.055	4.81	2.28	0.411	0.084

Note. Min. items = minimum number of items administered. Abs. est. bias = absolute estimation bias.

the differences between the prior conditions diminished as the minimum number of items increased.

When setting a minimum of four items, CAT efficiency under an unbiased personalized prior was comparable to that under a clinical prior. However, Table 2 shows that there were substantial differences in estimation bias depending on bias in the prior location and precision of the personalized prior. In the least favorable scenarios of a precise and highly biased prior, average absolute estimation bias remained substantial, even with a minimum number of four administered items. Under the same constraint, reducing prior precision substantially reduced estimation bias resulting from a biased personalized prior, but simultaneously eliminated any advantage that a personalized prior provided in terms of the CAT test length.

Simulation Study 2

Design

The basis for the second study was a clinical structured interview measuring impairment of personality functioning in which a so-called *global score* between zero and four—scored by the clinician prior to the main interview—can be used as empirical information to support the CAT. In clinical practice, this global score is obtained by asking several screener questions, and is used as a level of reference to guide further interview questions (Hummelen et al., 2021). This particular example was selected because the global score in this interview already functions as an informal personalized prior in the item selection process. Due to general data protection rules, real data were not used. Instead, this simulation study was based on previously acquired clinical data (Hummelen et al., 2021).

Each simulee was administered two fixed-precision CATs that differed only in the prior distribution used during the initialization and estimation process of the CAT: either the default standard normal prior, or an empirical prior based on the simulee's global score. The CAT algorithmic settings and statistical software were the same as for Study 1, with one exception: we did not include a constraint on the minimum number of items in this simulation, as the first simulation showed that such a constraint mainly reduced the benefits of utilizing an empirical prior.

Table 3. Expected Frequencies of Global Score and θ_p for the Simulated Sample Size $p = 5000$, Based on the Clinical Sample of Hummelen et al. (2021) and Prior Parameters for Each Global Score.

Global score	θ rounded to the nearest integer							Expected frequency	Prior parameters	
	−3	−2	−1	0	1	2	3		μ_p	σ_p^2
0	63	21	21	0	0	0	0	105	−2.4	0.640
1	21	105	732	188	0	0	0	1046	−1.0	0.359
2	0	42	272	1673	356	0	0	2343	0.0	0.339
3	0	0	0	439	711	272	42	1464	0.9	0.597
4	0	0	0	0	0	21	21	42	2.5	0.250

Item Bank and Simulees

We used an existing polytomous item bank consisting of 12 items scored on a 5-point scale. The items were calibrated in a clinical sample with the GRM (see Hummelen et al. (2021) for more details on the item bank), where item discrimination parameters ranged from 2.03 to 3.03, with a mean of 2.47. The true latent trait values of $p = 5000$ simulees were sampled from a standard normal distribution $\theta_p \sim N(0, 1)$. Note that a θ-value of 0 corresponds to the average in a clinical population in this case. As such, this simulation did not include a second generic prior condition, since a generic clinical prior coincides with a default standard normal prior. The full item bank was sufficiently informative for all simulees on the interval θ = [–1.96, 2.64] to be measured in line with the required fixed precision $SE(\widehat{\theta}_p) \leq 0.316$.

Global Score and Empirical Prior

The empirical prior was simulated to mimic the observed relation between the global score and θ estimate in the clinical sample of Hummelen et al. (2021). Table 3 shows the expected frequencies for each integer θ value and global score in a sample of $p = 5000$.

First, the global score of each simulee was obtained by sampling from the expected frequencies in Table 3, conditional on the simulee's true θ value rounded to the nearest integer. After a global score was assigned, the empirical prior $\widehat{\theta}_p^{(0)}$ was assumed to be normally distributed, with mean μ_p and variance σ_p^2 equal to the mean and variance of the θ distribution for the assigned global score (see Table 3).

Evaluation Criteria

The contrast between the empirical and default prior in terms of resulting test length and estimation bias was evaluated on the same criteria used in the previous simulation study.

Results

Convergence

Ninety-eight percent of the CATs using the default standard normal prior reached the required fixed precision. Out of the 104 CATs (2%) that ran to bank depletion, nearly all cases had estimated θ values for which the item bank did not contain sufficient information. In contrast, only 56 CATs

(1%) using the empirical prior ran to bank depletion without reaching the required fixed precision. Fifty of these cases (89%) overlapped with the non-convergent cases under the default prior. The remaining 6 cases (11%) all had a global score of 0 and an average trait value of −2.7 (range $\theta_p = [-3.6, -1.8]$).

Test Length

Overall, the range of the number of items administered varied from 3 to 12 items, with a median of 5 items and a median absolute deviation (*MAD*) of 1 item. Out of the 5000 simulees, 68% had a shorter test length under the empirical prior than under the default prior, 29% simulees had an equal test length, and 3% had a longer test length. The within-subject difference in test length Δn_p showed that CATs utilizing an empirical prior ranged from a test length that is 6 items longer to a test length that is 8 items shorter than the same CATs using a default prior, with a median reduction in test length of 20% or 1 item (*MAD* = 0). Figure 3 shows how this reduction in test length is related to the global score that informed the location and variance of the empirical prior distribution. The reductions could be observed for simulees with global score equal to 1, 2, or 4. For the latter, reduction in test length ranged from −3 to 5 items with median of 2 (*MAD* = 0), whereas for the former two, reduction in test length ranged from −1 to 7 or 8 items with a median of 1 (*MAD* = 0). In contrast, for simulees with a global score equal to 0 or 3, coinciding with less precise empirical priors (cf. Table 3), test length was highly similar between the empirical prior and default prior CATs. Similar to the previous simulation, one can see a small cluster of simulees in the scatterplot of Figure 3 for which the CAT under the default prior had a much shorter test length than the same CAT under the empirical prior. These five cases all had a global score of zero, which assigned them an empirical prior location of −2.4, a value outside of the θ-range with sufficient information in the item bank.

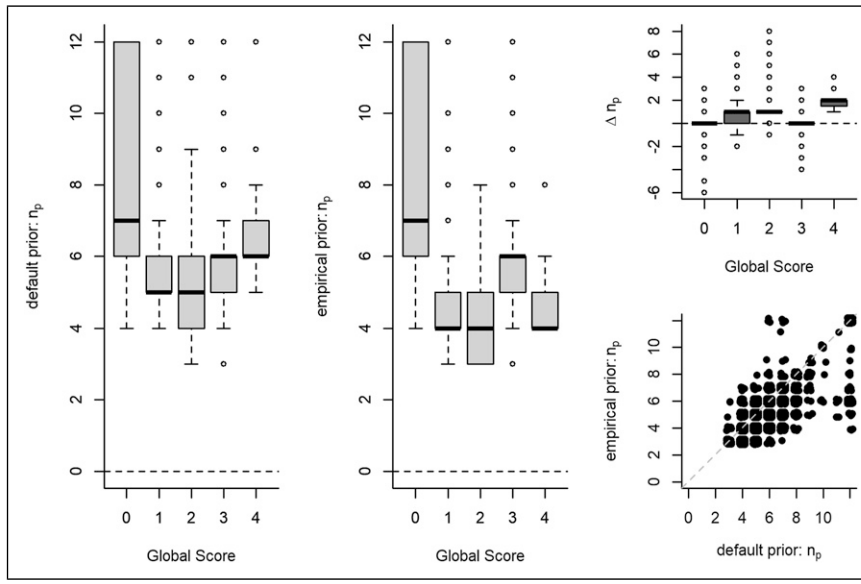

Figure 3. Test length n_p as a function of global score under a default and under an empirical prior. *Note.* The number of simulees is not equally distributed across global scores (see Table 3). Δn_p is the within-subject difference in test length between the default and the empirical prior condition.

Estimation Bias

Overall, the estimation bias in the simulees' latent trait value varied from $BIAS(\widehat{\theta}_p) = -1.10$ to 1.22, with a median of 0.00 ($MAD = 0.21$) for the default prior CATs, and $BIAS(\widehat{\theta}_p) = -1.16$ to 1.32, with a median of -0.02 ($MAD = 0.30$) for the empirical prior CATs. Out of the 5000 simulees, 51% had smaller absolute estimation bias under the empirical prior than under the default prior, while the reverse was true for the remaining 49%. The within-subject difference in absolute estimation bias $\Delta|BIAS(\widehat{\theta}_p)|$ between the default prior CAT and the empirical prior CAT ranged from -0.68 to 0.88, with a median of 0.00 ($MAD = 0.14$). Hence, the use of an empirical prior did not appreciably change the estimation bias in $\widehat{\theta}_p$ when compared to the default prior.

Discussion

Our simulation studies show that using a precise empirical prior in a clinical fixed-precision CAT may lead to substantial gains in test efficiency. In many of the simulated scenarios, CATs supported by a personalized prior performed equivalently or better than CATs supported by a default standard normal prior. The potential gains to CAT efficiency were similar in size to the gains associated with using more complex models like multidimensional CAT in clinical contexts (e.g., Bass et al., 2015; Paap et al., 2019), and to the gains found in previous simulations using an empirical prior in educational contexts (Matteucci & Veldkamp, 2013). In the ideal scenario of an unbiased and highly precise prior for every participant, our first simulation showed a 30% reduction in test length. Under more realistic conditions, where there is a certain risk of bias in the prior location, our second simulation showed that a personalized prior still reduced average test length by 20%, while estimation bias was similar to the standard normal prior condition.

Although these rewards look promising, it seems that the substantial benefits in terms of CAT efficiency associated with the use of a precise empirical prior can simultaneously be linked to the highest risks, when the location of a precise empirical prior is severely biased. Our results indicate that, contrary to initial tentative expectations about fixed-precision CATs (e.g., Chang & Ying, 2008), the trait estimate was not likely to recover sufficiently from a biased starting location before the CAT is terminated. Moreover, when the item bank contained few informative items at the prior location, a large number of items were needed before the CAT stopping precision was reached, resulting in far longer tests compared to a generic prior. This was particularly noticeable in smaller item banks where few informative items were available. In contrast, when the biased prior location was set in an information-rich area of the item bank, CAT length was much shorter compared to a generic prior, but estimates were severely biased, as only a few items were administered to counter the biased starting location.

Our results show that the risk of estimation bias from a biased prior can be ameliorated in two ways: (i) by reducing the precision of the prior or (ii) by imposing a constraint on the minimum number of items administered in the CAT. However, both of these measures dissolved any advantage associated with the use of a personalized prior in terms of test length reduction. For example, utilizing an unbiased personalized prior with low precision (i.e., $\sigma_p^2 = 1$ in the first simulation, or $\sigma_p^2 = 0.6$ in the second simulation) did not increase CAT efficiency over CATs utilizing a standard normal prior. Equivalently, despite providing a better starting location, a generic empirical prior with low precision did not improve CAT efficiency or reduce estimation bias, compared to CATs utilizing a standard normal prior. Increasing the minimum number of items administered in the CAT likewise diminished the differences between the prior conditions in terms of test length and estimation bias. In short, using a low precision prior, or setting a constraint on the minimum number of items are unsuitable, if one wants to preserve the benefits of a personalized prior, while reducing the risks of biased estimates.

The PROMIS item banks that form the basis for the item bank characteristics in the first study are part of the most widely known CAT applications in health care. Combined with the shorter clinical interview used in the second simulation, these findings are presumably relevant to a wide range of clinical applications. However, given that the effect of a personalized prior on CAT efficiency depends on the availability of informative items at the prior location, it is important to carefully examine the properties of the item banks used, before attempting to generalize these conclusions to other clinical item banks. Due to the small size of the item banks typical for applied clinical settings, a percentage of CATs in our studies failed to reach the stopping precision before depleting the item bank. This problem may be common in clinical contexts, and since the percentage was small enough not to greatly influence the conclusions, we retained these cases in our analyses.

Conclusion

Our results show that utilizing prior information in CAT is a relatively simple method to increase CAT efficiency that aligns with current clinical assessment practice. Although an average absolute reduction in test length of 1–4 items might be considered negligible in the context of educational measurement, this may still be considered a relevant reduction in the context of clinical assessment. During the diagnostic phase, the clinician typically administers a range of instruments, so time is very precious. If less time is used for the assessment procedure, this will directly impact the length of time patients spend on waiting lists. The instrument referred to in the second simulation typically takes 1–2 hours to complete; a 1-item reduction therefore equates to a reduction of about 5–10 minutes in administration time, which is considerable. The results of these simulations provide a more complete picture of the risks and rewards of empirical priors in applied CAT scenarios, and show that in general: (i) using a precise empirical prior can be rewarding in terms of test length reduction; (ii) there is a risk that the latent trait estimate in a fixed-precision CAT will not recover from a biased prior, particularly if this prior is highly precise.

Although there are risks involved, our second simulation showed that a personalized prior could provide substantial benefits with minimal risk under less-than-ideal circumstances. Similarly, the first simulation showed that a personalized prior performs comparably to a default standard normal prior in the less extremely disadvantaged scenarios. If the quality of prior information is adequately incorporated in the precision of a personalized prior, the risks of a biased starting location will be largely mitigated, while benefits to CAT efficiency will be retained. Personalized priors provide an easily implementable way to increase efficiency in clinical CATs.

Acknowledgments

We would like to thank the Center for Information Technology of the University of Groningen for their support, and for providing access to the Peregrine high performance computing cluster.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: FRIPRO Young Research Talent grant for the last author (Grant no. NFR 286893), awarded by the Research Council of Norway.

Data Availability Statement

The scripts that support the findings of this study are openly available at https://github.com/Niek-F/CAT_Empirical_Prior

ORCID iDs

Niek Frans https://orcid.org/0000-0001-6684-0684
Johan Braeken https://orcid.org/0000-0002-2119-3222
Muirne C. S. Paap https://orcid.org/0000-0002-1173-7070

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Note that in the study by Paap et al. (2019), higher θ values reflected better functioning. In the current study, higher theta values reflect higher levels of pathology, which is more consistent with the use of assessment instruments in clinical practice.
2. Bank sizes for this database can be found at http://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis/list-of-adult-measures
3. Responses to each item were sampled based on the θ value for each simulee and the sampled item parameters.
4. Relative estimation bias of the personalized prior compared to a generic clinical prior was similar to that compared to a standard normal prior and was therefore omitted from Figure 2. While there was no overall difference in estimation bias between the bank sizes, the different bank sizes are shown to facilitate comparisons between Figures 1 and 2.

References

Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, *15*(2), 94–98. https://doi.org/10.1111/j.0963-7214.2006.00414.x

Babcock, B., & Weiss, D. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, *1*(1), 1–18. https://doi.org/10.7333/1212-0101001

Bass, M., Morris, S., & Neapolitan, R. (2015). Utilizing multidimensional computer adaptive testing to mitigate burden with patient reported outcomes. *AMIA Annual Symposium Proceedings. AMIA Symposium*, *2015*, 320–328.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1–39. https://doi.org/10.18637/jss.v071.i05

Chang, H.-H., & van der Linden, W. J. (2003). Optimal stratification of item pools in a-stratified computerized adaptive testing. *Applied Psychological Measurement*, *27*(4), 262–274. https://doi.org/10.1177/0146621603027004002

Chang, H.-H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, *73*(3), 441–450. https://doi.org/10.1007/S11336-007-9047-7

He, Q., Veldkamp, B. P., Glas, C. A. W., & van den Berg, S. M. (2019). Combining text mining of long constructed responses and item-based measures: A hybrid test design to screen for posttraumatic stress disorder (PTSD). *Frontiers in Psychology*, *10*, 2358. https://doi.org/10.3389/fpsyg.2019.02358

Hummelen, B., Braeken, J., Buer Christensen, T., Nysæter, T. E., Selvik, S. G., Walther, K., Pedersen, G., Eikenaes, I., & Paap, M. C. S. (2021). A psychometric analysis of the structured clinical interview for the DSM-5 alternative model for personality disorders module I (SCID-5-AMPD-I): Level of personality functioning scale (SCID-5-AMPD-I). *Assessment*, *28*(5), 1320–1333. https://doi.org/10.1177/1073191120967972

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, *29*(3), 8–14. https://doi.org/10.1111/j.1745-3992.2010.00179.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Matteucci, M., & Veldkamp, B. P. (2013). On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency. *Statistical Methods and Applications*, *22*(2), 243–267. https://doi.org/10.1007/s10260-012-0216-1

Paap, M. C. S., Born, S., & Braeken, J. (2019). Measurement efficiency for fixed-precision multidimensional computerized adaptive tests: Comparing health measurement and educational testing using example banks. *Applied Psychological Measurement*, *43*(1), 68–83. https://doi.org/10.1177/0146621618765719

R Core Team. (2021). R: A language and environment for statistical computing *(4.1.1)*. R Foundation for Statistical Computing. https://www.r-project.org/

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J.-s., & Cella, D., on behalf of the PROMIS cooperative group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care*, *45*(5 SUPPL 1), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*(1), 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, *33*(2), 83–101. https://doi.org/10.1177/0146621608324023

Samejima, F. (1996). The graded response model. In W. J. Van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Springer.

van der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, *23*(1), 21–29. https://doi.org/10.1177/01466219922031149

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Kluwer Academic Publishers. https://doi.org/10.1007/0-306-47531-6

van der Linden, W. J., & Pashley, P. J. (2010). Elements of adaptive testing. In W. J. Van der Linden, & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). Springer. https://doi.org/10.1007/978-0-387-85461-8

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*(4), 589–600. https://doi.org/10.1007/BF02294492