Computerized adaptive testing without IRT for flexible measurement and prediction

van der Ark, L.A.; Smits, N.

DOI
10.1007/978-3-031-10370-4_19

Publication date
2023

Document Version
Final published version

Published in
Essays on Contemporary Psychometrics

Chapter 19
Computerized Adaptive Testing Without IRT for Flexible Measurement and Prediction

L. Andries van der Ark and Niels Smits

Abstract In education, testing procedures can be lengthy. The long duration takes up precious time and affects the quality of responses, possibly resulting in a biased diagnosis or wrong treatment. The problem can be reduced using computer adaptive testing (CAT). However, three issues prevent the use of traditional CAT: (1) the type of tests and questionnaires we focus on do not allow for the construction of large item banks, (2) the test data are usually not (approximately) unidimensional, and (3) the aim of the researchers may not only be measurement but also prediction. We propose a flexible generalization of CAT to accommodate these three issues, coined FlexCAT. First, FlexCAT estimates the (discrete) density of item-score vectors (denoted $\mathbf{p}$) using any convenient model that provides a good description of $\mathbf{p}$; this need not be an IRT model. Second, FlexCAT estimates test scores from $\hat{\mathbf{p}}$. In contrast to traditional CAT, the test score need not be a latent trait but can also be the total score, ordinal scores such as percentiles, or external criteria that the test aims to predict. We introduce FlexCAT for the case that a latent class model is used to estimate $\mathbf{p}$, and the total score is used as a test score. Using a real-data example, we compare the accuracy of FlexCAT and traditional CAT. Finally, we discuss the challenges FlexCAT still faces.

19.1 Introduction

In education, testing procedures can be lengthy. Especially for respondents who are unable to focus for long time periods, such as very young students or students in special needs education, standard educational tests pose a problem. When students get tired or distracted, they may resort to careless responding, or they may decide to stop the test procedure, possibly resulting in a biased test result or incorrect follow-

L. A. van der Ark (✉) · N. Smits
Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands
e-mail: L.A.vanderark@uva.nl; n.smits@uva.nl

L. A. van der Ark et al. (eds.), *Essays on Contemporary Psychometrics*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-031-10370-4_19

up treatment. The test time can be reduced using computer adaptive testing (CAT; e.g., Magis et al., 2017; Wainer, 2000). However, CAT requires a large item bank, approximately unidimensional test data, and a latent trait with a known (typically a normal) distribution. Many tests, especially *typical performance tests*, do not allow for the construction of large item banks, as there are only a limited number of things one can ask a respondent. Also, many tests produce test data that are not approximately unidimensional. For example, there may be a dominant dimension and one or more nuisance dimensions. Finally, tests measuring certain phenomena typically produce a latent trait that has a skewed distribution (for some examples, see Molenaar et al., 2012). For such tests, traditional CAT may be suboptimal.

Consider the School Attitude Questionnaire Internet (SAQI, Vorst, 2006; also see, Psi Testuitgevers, n.d.), a test for students aged 9–16 years. The 160 trichotomous items measure motivation, well-being, and self-confidence with respect to going to school. The SAQI consists of ten scales. The SAQI provides scores at the scale level, aggregated scale level (i.e., motivation, well-being, and self-confidence), and at the overall level (a total score). The administration of 160 items may take more than 2 h, which can be strenuous for young students. Using a CAT could be helpful to reduce the response burden. However, the requirements of a CAT pose a problem. For constructs such as motivation, well-being, and self-confidence, it is infeasible to write enough items to fill a large item bank, as there is only a limited number of questions one can ask on these topics. Also, the SAQI aggregated-level scores "motivation," "well-being," and "self-confidence," and the SAQI total score are the sum of multiple scale scores. As a result these scores are multidimensional. Also, even several SAQI scale-level scores are multidimensional. As traditional CAT assumes that the data are unidimensional, traditional CAT may produce biased estimates of the SAQI scores, and this bias may also be present in other typical-performance tests and possibly also in some maximum-performance tests.

In this chapter, we propose an alternative view on CAT, coined *FlexCAT*, that allows for the use of more flexible models than item response theory (IRT) models, which are traditionally used in CAT. First, we briefly describe the five building blocks of a traditional CAT. Second, we introduce FlexCAT using the same five building blocks. Third, using SAQI item scores, we compare the accuracy of FlexCAT and traditional CAT. Finally, we discuss the challenges of FlexCAT that must be resolved.

19.2 Traditional CAT

CAT procedures are iterative procedures. The algorithms for CAT have often been described as containing five building blocks (e.g., Wainer, 2000; Weiss & Kingsbury, 1984). Figure 19.1 shows a flow diagram of the five building blocks in an iterative CAT procedure that also fits FlexCAT. Building blocks "calibration" and "starting level" are grouped together in the *preliminary phase*, as the calibration and determining the starting level take place before the item administration. The

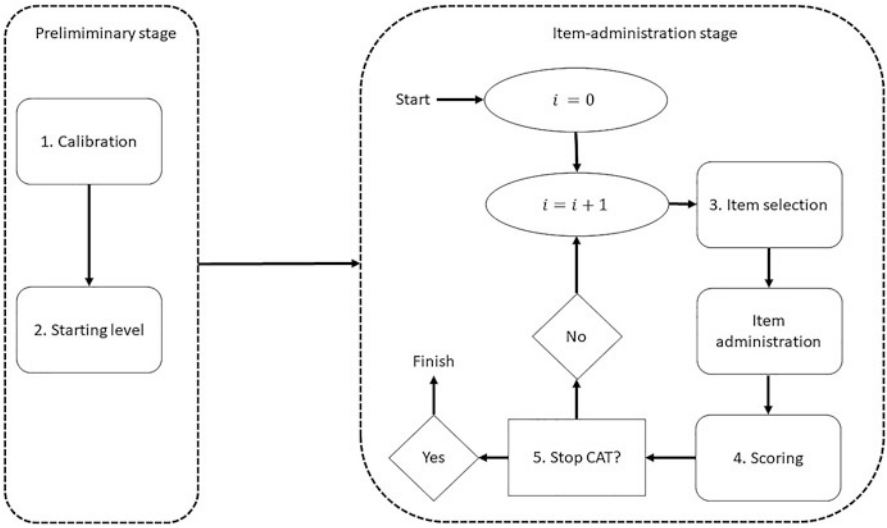

Fig. 19.1 Flow diagram of the five numbered building blocks of CAT in an iterative process (indicated by *i*): (1) calibration, (2) starting level, (3) item selection, (4) scoring, and (5) the decision whether to stop the CAT. Administering an item to a respondent (indicated by "Item administration") is not part to the CAT algorithm and therefore not considered a building block

remaining building blocks are grouped together in the *item administration* phase, these building blocks are part of the measurement procedure of a single respondent.

Calibration First, the items of the complete test (the "item bank") should be calibrated under an IRT model to obtain the item parameters that feed the CAT-algorithm. The selected IRT model should match the item format (e.g., Edelen & Reeve, 2007). Suppose the two-parameter logistic model is used to model dichotomously scored items. The probability that a randomly chosen respondent with latent trait score θ has a response X_j of 1 on item j is given by

$$P\left(X_j = 1|\theta\right) = \frac{e^{\alpha_j(\theta-\delta_j)}}{1 + e^{\alpha_j(\theta-\delta_j)}}, \tag{19.1}$$

where α_j is the item's slope parameter and δ_j is its location parameter.

Starting Level Usually, there is no information available about the respondent before the administration of the first item, and therefore some provisional estimate of the latent trait is required at the start of the CAT (Wainer, 2000). Most often, the average of the latent trait in the population is taken as a starting point, and the item that is most informative for that value is thus selected. Once the starting level has been determined, the item administration stage starts.

Item Selection Once an estimate of the respondent's latent trait has been obtained, a new item is selected that is most informative about this estimate. Let a prime denote the first derivative. Then, for item j, Fisher information

$$I_j(\theta) = \frac{[P'(X_j=1|\theta)]^2}{P(X_j=1|\theta)P(X_j=0|\theta)}, \tag{19.2}$$

may be used to quantify measurement quality as a function of the latent trait. From the items that have not yet been administered, the item with the highest information at the current estimate $\hat{\theta}$ is selected. The selected item is administered to the respondent, and the resulting item score is obtained.

Scoring After obtaining the item score, the CAT updates the estimate of the respondent's latent trait value. There are two popular latent trait estimation methods. Maximum likelihood (ML) estimates θ as the value with the highest likelihood of producing the observed responses (Thissen, 1991). By contrast, Bayesian estimation adds to this likelihood a prior distribution of the latent trait, such as the standard normal distribution (e.g., Embretson & Reise, 2000). Bayesian estimation can and ML estimation cannot provide an estimate for perfect response patterns. Let $f(\theta)$ denote the prior distribution of θ, and let $L(\theta)$ denote the likelihood function. One Bayesian method, expected a posteriori (EAP), takes the average of the posterior distribution of the latent trait, that is,

$$\hat{\theta}_{\text{EAP}} = \frac{\int \theta f(\theta) L(\theta) d\theta}{\int f(\theta) L(\theta) d\theta}. \tag{19.3}$$

Stopping Rule The CAT algorithm alternately administers items and updates the estimate of the respondent's latent trait score until the item pool is exhausted unless a termination criterion is specified, such as a pre-specified level of measurement precision. This criterion is met when the respondent's standard error of θ is small enough. The standard error when using EAP estimation is given by

$$SE\left(\hat{\theta}_{\text{EAP}}\right) = \sqrt{\frac{\int \left(\theta - \hat{\theta}_{\text{EAP}}\right)^2 f(\theta) L(\theta) d\theta}{\int f(\theta) L(\theta) d\theta}}. \tag{19.4}$$

19.3 General Concept of FlexCAT

The main differences between FlexCAT and traditional CAT are in the building blocks calibration and starting level. The other building blocks also differ between FlexCAT and traditional CAT, but these differences are merely adaptations that are required because the building blocks calibration and starting level are rather different. Therefore, we discuss these two building blocks first.

19.3.1 Calibration

In FlexCAT, the calibration step entails the estimation of the *density of the item-score vectors* using a large sample. An item-score vector is a vector containing scores on all items. Suppose a test consists of J items, indexed by j ($j = 1, \ldots, J$), and suppose that item j has $C_j + 1$ response categories, $0, \ldots, c, \ldots C_j$. Then the number of possible item-score vectors equals $V = \prod_j (C_j + 1)$. For simplicity, but without loss of generalizability, we assume that all items have the same number of categories, that is, $C_j = C$ for all j. As a result, the number of possible item-score vectors equals

$$V = \prod\nolimits_j (C + 1) = (C + 1)^J. \tag{19.5}$$

Let X_j denote the integer score on item j, with realization x_j ($x_j \in \{0, \ldots, c, \ldots, C\}$. Let $\mathbf{r}_v = (x_{v1}, \ldots, x_{vJ})^T$ ($v = 1, \ldots, V$) denote the vth item-score vector. The item-score vectors can be collected in a $V \times J$ matrix $\mathbf{R} = \left(\mathbf{r}_1^T, \ldots, \mathbf{r}_V^T\right)$. The density of the item-score vectors, collected in the $V \times 1$ vector $\mathbf{p} = (P(\mathbf{r}_1), \ldots, P(\mathbf{r}_V))$, plays a central role in the calibration step.

In traditional CAT, it is assumed that an IRT model generates $\mathbf{p}$. Using Eq. 19.1 and the property of local independence, it follows that

$$\begin{aligned} P(\mathbf{r}_v) &= P(X_1 = x_{v1}, \ldots, X_J = x_{vJ}) = \int \prod_j P\left(X_j = x_{vj}|\theta\right) f(\theta)\, d\theta \\ &= \int \prod_j \left[\frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}}\right]^{x_{vj}} \left[1 - \frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}}\right]^{1-x_{vj}} f(\theta)\, d\theta. \end{aligned} \tag{19.6}$$

Estimation of $\mathbf{p}$ in traditional CAT (Eq. 19.6) thus requires estimating the item parameters α_j and δ_j and the distribution of the latent trait, $f(\theta)$.

The first notion of FlexCAT is that it has no assumptions on the process that may have generated $\mathbf{p}$, and the procedure is completely data driven. Vector $\mathbf{p}$ can be estimated using any *convenient* model that provides a *good description* of the item-score vector density. In this chapter, "convenient" means that $\mathbf{p}$ can be estimated directly from the test data, without the test constructor providing additional information (e.g., the number of dimensions or distributional assumptions). A "good description" is used pragmatically and means that the estimated item-score density, $\hat{\mathbf{p}}$, describes the associations in the test data so well that it provides a useful tool for measurement and prediction.

Hence, in FlexCAT the calibration stage consists of finding an estimate of $\mathbf{p}$ with a model of choice. Besides IRT models, candidate models for estimating $\mathbf{p}$ include the latent class model (LCM; e.g., Vermunt et al., 2008; Linzer, 2011; Van Buuren & Eggen, 2017), the divisive LCM (Van der Palm et al., 2016), kernel estimation methods (e.g., Li & Racine, 2003), and decision trees (e.g., Ho, 1995; Yan et al., 2004).

19.3.2 Starting Level

In traditional IRT, the estimated latent trait (e.g., $\hat{\theta}_{\text{EAP}}$, Eq. 19.4) is used as a score to communicate the measurement of a respondent. The starting level—when there is no information about the respondent yet—is the average latent trait level. The second notion of FlexCAT is that any score that can be derived from $\mathbf{p}$ can be used to communicate measurement results. Hence, for FlexCAT, determining which score will be used in the CAT procedure is part of the building block "starting level." Besides $\hat{\theta}_{\text{EAP}}$, a possible candidate is the *total score* (or equivalently, the mean item score) as most tests use the total score for measurement. Both the estimated latent trait and the total score can also be transformed to standard scores, percentile scores, or stanines to facilitate communication and interpretation. These adapted test scores can also be used as scores in FlexCAT. If the goal of the test is selection or prediction, a response variable could be a useful score. Examples of response variables include treatment (yes, no), placement (several nominal categories), or selection (selected, not selected). Note that when FlexCAT is used for prediction, the response variable (Y) must be included in the calibration model. For example, if a ten-item test should predict whether or treatment is effective ($Y = 1$) or not ($Y = 0$), then the vth item-score vector used for estimating $\mathbf{p}$ should be $\mathbf{r}_v = (X_{v1} = x_{v1}, \ldots, X_{v10} = x_{v10,}, Y_v = y_v)$. LCMs and decision trees can easily incorporate response variables while calibrating items, but this is more difficult for standard IRT models.

19.3.3 Item Selection, Scoring, and Stopping Rule

In FlexCAT, the item-administration stage—item selection, scoring (or more accurately updating the score), and stopping rules—are essentially the same as for traditional CAT. However, based on the choices made during building blocks "calibration" and "starting level," the building blocks in the item administration stage may have to be adapted. For example, when using the LCM for calibration and the total score for measurement, Fisher information (Eq. 19.2) is unavailable, and alternatives should be developed. Also, for the—discrete—total score, a stopping rule based on the modal value may be preferred over a stopping rule based on standard errors of the score (Eq. 19.4). As the building blocks in the item administration stage should be adapted depending on the choices made for calibration model and score, FlexCAT is more like an umbrella term for different types of CAT.

19.4 FlexCAT Using the Latent Class Model and the Total Score

19.4.1 Calibration

As a showcase, we show the estimation of item-score vector density $\mathbf{p}$ using the LCM with W latent classes denoted LCM(W). Let Ξ denote the categorical latent variable having Wcategories (classes). The parameters of LCM(W) are the *class weights* $\pi_w \equiv P(\Xi = w)$ $(w = 1, \ldots, W)$ and the *conditional item score probabilities* $\pi_{j(c)|w} \equiv P(X_j = c|\Xi = w)$ $(j = 1, \ldots, J; c = 0, \ldots, C; w = 1, \ldots, W)$. Under the LCM($W$)

$$P\left(X_j = x_j\right) = \sum\nolimits_w \pi_w \pi_{j(c_j)|w}. \tag{19.7}$$

LCMs assume that item scores are locally independent given the score on Ξ, that is,

$$P\left(\mathbf{r}_v\right) = P\left(X_1 = x_1, \ldots, X_J = c_J\right) = \sum\nolimits_w \prod\nolimits_j \pi_w \pi_{j(c_j)|w} \tag{19.8}$$

(cf. Eq. 19.6).

Table 19.1 shows a constructed small example with three dichotomous items. It is assumed that the estimated parameters of the LCM(2) provide a good description of the data. Hence, $\hat{\mathbf{p}}$ is derived from the parameters of the LCM(2) (see note in Table 19.1). For density estimation using the LCM, two issues are important.

Table 19.1 Example of LCM(2) parameter estimates for three dichotomous items, the matrix containing the $V = 8$ possible item-score vectors ($\mathbf{R}$), and the estimated density of the item-score vectors ($\hat{\mathbf{p}}$), which is derived from the latent class parameters (see note)

Latent class parameters			$\mathbf{R}$	$\hat{\mathbf{p}}$
$\hat{\boldsymbol{\pi}}_w$	$\hat{\boldsymbol{\pi}}_{j\|1}$	$\hat{\boldsymbol{\pi}}_{j\|2}$		
$\begin{pmatrix} .2 \\ .8 \end{pmatrix}$	$\begin{pmatrix} .3 & .7 \\ .2 & .8 \\ .1 & .9 \end{pmatrix}$	$\begin{pmatrix} .6 & .4 \\ .9 & .1 \\ .6 & .4 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} .2604 \\ .1836 \\ .0336 \\ .0624 \\ .1756 \\ .1404 \\ .0304 \\ .1136 \end{pmatrix}$

Note. $\hat{p}_1 = \hat{\pi}_1 \times \left(\hat{\pi}_{1(0)|1} \times \hat{\pi}_{2(0)|1} \times \hat{\pi}_{3(0)|1}\right) + \hat{\pi}_2 \times \left(\hat{\pi}_{1(0)|2} \times \hat{\pi}_{2(0)|2} \times \hat{\pi}_{3(0)|2}\right) = .2 \times (.3 \times .2 \times .1) + .8 \times (.6 \times .9 \times .6) = .2604$, $\hat{p}_2 = \hat{\pi}_1 \times \left(\hat{\pi}_{1(0)|1} \times \hat{\pi}_{2(0)|1} \times \hat{\pi}_{3(1)|1}\right) + \hat{\pi}_2 \times \left(\hat{\pi}_{1(0)|2} \times \hat{\pi}_{2(0)|2} \times \hat{\pi}_{3(1)|2}\right) = .2 \times (.3 \times .2 \times .9) + .8 \times (.6 \times .9 \times .4) = .1836$, etc.

Goodness of Fit If the LCM is used as a density estimation tool, the interpretation of the latent classes is not particularly important (Vermunt et al., 2008; also, see Linzer, 2011). Therefore, issues that are important in traditional latent class analysis, such as local optima (e.g., McCutcheon, 2002), obtaining a modest number of latent classes to facilitate interpretation, and identifiability (e.g., Goodman, 1974), are not so important for the LCM as a density estimation tool, as long as the estimated density captures the higher-order interactions well. If the number of latent classes, W, is too small, the density is underfitted, which means that important associations or interactions are possibly ignored in the estimated density. If W is too large, the density may be overfitted; that is, the density estimate contains certain random fluctuations that are sample specific. Determining the correct W is typically done using information criteria, such as AIC (e.g., Bozdogan, 1987) or BIC (Schwarz, 1978). For increasing numbers of W (starting with $W = 1$), the information criterion is computed for LCM(W), and the LCM(W) that produces the lowest value of the information criterion is selected as a density estimator. It is well known that AIC tends to overestimate W, and BIC tends to underestimate W (e.g., Lukociene & Vermunt, 2010). Vermunt et al. (2008, p. 378) noted that overfitting is less problematic than underfitting, and for now, we advocate using AIC to determine W. However, this is an issue that should be investigated further, as there are many alternative information criteria and also indices for local fit (e.g., Nagelkerke et al., 2016).

Computational Feasibility The size of the $V \times 1$ vector $\mathbf{p}$ can increase dramatically. For example, for the SAQI ($J = 160$ items, $C + 1 = 3$ categories), $V = 3^{160} \approx 2.18 \times 10^{76}$ (cf. Equation 19.5), which is computationally infeasible. As the number of free parameters in the LCM equals $W - 1 + W \times J \times C$, for the SAQI, $\hat{\mathbf{p}}$ is estimated using $W - 1 + W \times 480$ parameters. For $W = 200$, which is a large number of latent classes (e.g., see example in Vermunt et al., 2008), the number of parameters is less than 100,000, which is computationally feasible, although the density estimation procedure may be slow. Standard software (e.g., **poLCA**; Linzer & Lewis, 2011; or **Latent GOLD**, Vermunt & Magidson, 2013) can be used to estimate $\mathbf{p}$.

19.4.2 Starting Level

At the starting level, the density of the selected score is estimated. Here we use total score $X_+ = \sum_j X_j$. For J items, each having item scores $0, 1 \ldots, C$, there are $H = JC + 1$ possible total scores, indexed by h ($h \in \{0, 1, \ldots, H - 1\}$). Let $\mathbf{x}_+ = (0, \ldots, H - 1)^T$ be an $H \times 1$ vector containing all possible total scores. The density of the total scores can be collected in an $H \times 1$ vector $\mathbf{p}_{X_+} = (P[X_+ = 0], \ldots, P[X_+ = H - 1])^T$. Let $\mathbf{Q}$ be a $V \times H$ design matrix that relates $\mathbf{p}$ to $\mathbf{p}_{X_+}$, and let $\mathbf{r}_+ = (r_{+1}, \ldots, r_{+v}, \ldots, r_{+V})^T$ be a $V \times 1$ vector containing the total scores of the item-score vectors in $\mathbf{R}$; that is, $\mathbf{r}_+ = \mathbf{R} \cdot \mathbf{1}$. For the elements of $\mathbf{Q}$, simple matrix algebra shows that $q_{v,\,h+1} = 1$ if $r_{+v} = h$ and $q_{v,\,h+1} = 0$ otherwise,

Table 19.2 Continuation of the example in Table 19.1 showing the relation between the estimated item-score vector density $\hat{\mathbf{p}}$ and total-score density $\hat{\mathbf{p}}_{X_+}$. Item-score vectors ($\mathbf{R}$) and their estimated density ($\hat{\mathbf{p}}$) are taken from Table 19.1. The total scores produced by the item-score vectors are in $\mathbf{r}_+ = \mathbf{R} \cdot \mathbf{1}$. Design matrix $\mathbf{Q}$ is derived from $\mathbf{r}_+$ (see text). Vector $\mathbf{x}_+$ contains all possible total scores. Total-score-density equals $\hat{\mathbf{p}}_{X_+} = \mathbf{Q}^{\mathbf{T}}\hat{\mathbf{p}}$

$\mathbf{R}$	$\hat{\mathbf{p}}$	$\mathbf{r}_+$	$\mathbf{Q}$	$\mathbf{x}_+$	$\hat{\mathbf{p}}_{X_+}$
$\begin{pmatrix} 0&0&0 \\ 0&0&1 \\ 0&1&0 \\ 0&1&1 \\ 1&0&0 \\ 1&0&1 \\ 1&1&0 \\ 1&1&1 \end{pmatrix}$	$\begin{pmatrix} .2604 \\ .1836 \\ .0336 \\ .0624 \\ .1756 \\ .1404 \\ .0304 \\ .1136 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 1 \\ 2 \\ 2 \\ 3 \end{pmatrix}$	$\begin{pmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \end{pmatrix}$	$\begin{pmatrix} .2604 \\ .3928 \\ .2332 \\ .1136 \end{pmatrix}$

for $h = 0, \ldots, H-1$. It follows that $\mathbf{p}_{X_+} = \mathbf{Q}^{\mathbf{T}}\mathbf{p}$ (and $\hat{\mathbf{p}}_{X_+} = \mathbf{Q}^{\mathbf{T}}\hat{\mathbf{p}}$). Table 19.2 continues the example from Table 19.1 and illustrates $\hat{\mathbf{p}}_{X_+} = \mathbf{Q}^{\mathbf{T}}\hat{\mathbf{p}}$.

19.4.3 Selecting the Next Item

Just as CAT, FlexCAT is an iterative process (Fig. 19.1). The starting level can be seen as iteration $i = 0$, where no item has yet been administered. Iteration i ($i = 1, 2, \ldots$) starts with the selection of the ith item. As noted earlier, Fisher information (Equation 19.2) is unavailable here. A possible strategy for selecting the ith item for respondent n is searching for the item that provides as much information as possible on respondent n's expected total score.

At the start of iteration i, there are $i-1$ items that have already been administered to respondent n, whereas the remaining $G = J - i + 1$ items, indexed by g ($g = 1, \ldots, G$), have not yet been administrated to respondent n. Let $\mathbf{r}^{n,\,i-1}$denote the item-score vector of respondent n at iteration $i-1$; that is, $\mathbf{r}^{n,\,i-1}$ contains $i-1$ observed item scores obtained in the previous iterations and G missing item scores. Similarly, let $\mathbf{r}^{n,i-1}_{X_g=c}$ denote the item-score vector of respondent n at iteration $i-1$, assuming that respondent n will obtain score c on item g in iteration i. Let $P(X_g = c|\,\mathbf{r}^{n,\,i-1})$ denote the probability that respondent n will obtain score c on item g in iteration i, let $E(X_+|\,\mathbf{r}^{n,\,i-1})$ denote the expected total score at iteration $i-1$ for respondent n, and let $E\left(X_+|\mathbf{r}^{n,i-1}_{X_g=c}\right)$ denote the expected total score at iteration $i-1$ for respondent n assuming that respondent n will obtain score c on item g in iteration i. A possible way to express the additional value of item g in iteration i on respondent n's expected total score is

$$\Delta_g^{n,i} = \sum_c P\left(X_g = c|\mathbf{r}^{n,i-1}\right)\left|E\left(X_+|\mathbf{r}^{n,i-1}_{X_g=c}\right) - E\left(X_+|\mathbf{r}^{n,i-1}\right)\right|. \qquad (19.9)$$

The absolute difference between $E\left(X_+|\mathbf{r}^{n,i-1}_{X_g=c}\right)$ and $E(X_+|\,\mathbf{r}^{n,\,i\,-\,1})$ in Equation 19.9 is the effect of having $X_g = c$ on the expected total score; this effect is weighed by the probability that $X_g = c$ actually occurs. $\Delta_g^{n,i}$ is then the sum of these weighed effects over all response categories of item g. The item that produces the highest value $\Delta_g^{n,i}$ is selected as the next item to be administrated. $\Delta_g^{n,i}$ can be computed relatively easily. Let $\mathbf{a}^{i\,-\,1,\,n}$ be an indicator vector of length V, with $a_v^{i-1,n} = 1$if the vth item-score vector in $\mathbf{R}$ is still admissible given respondent n's responses in the previous $i - 1$ iterations, and $a_v^{i-1,n} = 0$, otherwise. Similarly, let $\mathbf{a}^{i-1,n}_{X_g=c}$be an indicator vector of length V, with $a^{i-1,n}_{v|X_g=c} = 1$if the vth item-score vector in $\mathbf{R}$ is still admissible given respondent n's responses in the previous $i - 1$ iterations and given that respondent n would obtain item score $X_g = c$ if item g were to be administered in iteration i; and $a^{i-1,n}_{v|X_g=c} = 0$, otherwise. Table 19.3 shows an example to illustrate $\mathbf{a}^{i\,-\,1,\,n}$ and $\mathbf{a}^{i-1,n}_{X_g=c}$.

Let $\mathbf{x} \circ \mathbf{y}$ denote the Hadamard or elementwise product of vectors $\mathbf{x}$ and $\mathbf{y}$, and let $\left[\frac{\mathbf{x}}{\mathbf{y}}\right]$ be a vector that consist of the elementwise division of $\mathbf{x}$ by $\mathbf{y}$. For example, for $\mathbf{x} = [3, 2]$ and for $\mathbf{y} = [1, 2]$, then $\mathbf{x} \circ \mathbf{y} = [3, 4]$, and $\left[\frac{\mathbf{x}}{\mathbf{y}}\right] = [3, 1]$. The terms in Eq. 19.9 can be expressed as

$$P\left(X_g = c|\mathbf{r}^{i-1,n}\right) = \frac{\mathbf{1}^T\left[\mathbf{a}^{i-1,n}_{X_g=c} \circ \mathbf{p}\right]}{\mathbf{1}^T\left[\mathbf{a}^{i-1,n} \circ \mathbf{p}\right]}, \tag{19.10}$$

$$E\left(X_+|\mathbf{r}^{i-1,n}\right) = \mathbf{x}_+^T\,\mathbf{Q}^T\left[\frac{\mathbf{a}^{i-1,n} \circ \mathbf{p}}{\mathbf{1}\mathbf{1}^T\left(\mathbf{a}^{i-1,n} \circ \mathbf{p}\right)}\right], \tag{19.11}$$

Table 19.3 Example showing design vectors $\mathbf{a}^{i\,-\,1,\,n}$ and $\mathbf{a}^{i-1,n}_{X_g=c}$ for respondent n in iteration $i = 2$, who has endorsed item 3 ($X_3 = 1$) in iteration 1

$\mathbf{R}$	$\mathbf{a}^{1,\,n}$	$\mathbf{a}^{1,n}_{X_1=0}$	$\mathbf{a}^{1,n}_{X_1=1}$	$\mathbf{a}^{1,n}_{X_2=0}$	$\mathbf{a}^{1,n}_{X_2=1}$
$\begin{pmatrix} 0\,0\,0 \\ 0\,0\,1 \\ 0\,1\,0 \\ 0\,1\,1 \\ 1\,0\,0 \\ 1\,0\,1 \\ 1\,1\,0 \\ 1\,1\,1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$

Note: As respondent n has endorsed item 3 in iteration 1, all item-score vectors in $\mathbf{R}$ containing $X_3 = 0$ are inadmissible in iteration 2; hence, the corresponding elements in $\mathbf{a}^{1,\,n}$ are zeroes. In $\mathbf{a}^{1,n}_{X_1=0}$, the additional constraint is that $X_1 = 0$, leaving only two admissible item-score vectors. A similar logic applies to $\mathbf{a}^{1,n}_{X_1=1}$, $\mathbf{a}^{1,n}_{X_2=0}$, and $\mathbf{a}^{1,n}_{X_2=1}$

and

$$E\left(X_+|\mathbf{r}^{i-1,n}_{X_g=c}\right) = \mathbf{x}_+^T\,\mathbf{Q}^T\left[\frac{\mathbf{a}^{i-1,n}_{X_g=c}\circ\mathbf{p}}{\mathbf{1}\mathbf{1}^T\left(\mathbf{a}^{i-1,n}_{X_g=c}\circ\mathbf{p}\right)}\right]. \tag{19.12}$$

We have provided Eqs. 19.10, 19.11, and 19.12 in matrix notation, so they are consistent with our computer code in the vector-based programming language R (R Core Team, 2021). In the Appendix, we elaborate on these equations. As $\mathbf{x}_+^T$ and $\mathbf{Q}$ are fixed design matrices, and $\mathbf{p}$ has been estimated in the preliminary stage (Fig. 19.1) and remains fixed in the item administration stage, Eqs. 19.10, 19.11, and 19.12 show that only design vectors $\mathbf{a}^{i-1,n}$ and $\mathbf{a}^{i-1,n}_{X_g=c}$ require modification for computing $\Delta_g^{n,i}$ (Eq. 19.9). In the running example, at iteration 1 (no items have been administered), Eq. 19.9 results in $\Delta_1^{n,1} = 0.612$, $\Delta_2^{n,1} = 0.544$, and $\Delta_3^{n,1} = 0.660$. Hence, item 3 would be selected as the first item to be administered to all respondents.

19.4.4 Scoring

After a new item has been selected, the item is administered to the respondent (Fig. 19.1). Once the respondent has provided the score to the selected item, the estimated score density has to be updated from $\hat{\mathbf{p}}^{i-1,n}_{X_+}$ to $\hat{\mathbf{p}}^{i,n}_{X_+}$. Suppose that respondent n has obtained score c on item g in iteration i, then $\mathbf{a}^{i,n}$ is an $V \times 1$ indicator vector, with $a_v^{i,n} = 1$if the vth item-score vector in $\mathbf{R}$ is still admissible given respondent n's responses to the previously administered i items, and $a_v^{i,n} = 0$, otherwise.

Vector $\mathbf{a}^{i,n}$ can be updated from $\mathbf{a}^{i-1,n}$ by setting the elements in $\mathbf{a}^{i-1,n}$ that correspond to response patterns in which $X_g = c$ to 0. The item-score vector density and total-score density are updated using

$$\hat{\mathbf{p}}^{i,n} = \left[\frac{\mathbf{a}^{i-1,n}\circ\mathbf{p}}{\mathbf{1}\mathbf{1}^T\left(\mathbf{a}^{i-1,n}\circ\mathbf{p}\right)}\right] \tag{19.13}$$

and

$$\hat{\mathbf{p}}^{(i,n)}_{X_+} = \mathbf{Q}^{\mathrm{T}}\hat{\mathbf{p}}^{(i,n)} \tag{19.14}$$

19.4.5 Stopping Rule

As a possible stopping rule, FlexCAT may be terminated if the modal value of $\hat{\mathbf{p}}_{X_+}^{i,n} > c$; that is, $\max\left(\hat{\mathbf{p}}_{X_+}^{i,n}\right) > c$, where $0 \leq c \leq 1$. If $c < \max\left(\hat{\mathbf{p}}_{X_+}^{0,n}\right)$, FlexCAT stops before any item has been administered. If $c = 1$, all items will be administered. For all remaining values of c, it holds that if c becomes larger, the precision of the score estimate increases, but the expected number of administered items increases as well. We stress that alternative stopping rules may be used as well. For example, one may compute the expected sum score $E(X_+|\,\mathbf{r}^{i,n})$ (cf. Equation 19.9) and use its standard deviation as a measure of precision.

19.4.6 Small Example

Table 19.4 shows the iterative procedure for the running example based on the LCM(2) in Table 19.1, using the stopping rule $\max\left(\hat{\mathbf{p}}_{X_+}^{i,n}\right) > .9$. At iteration $i = 0$, Table 19.4 shows the matrix of item-score vectors ($\mathbf{R}$; taken from Table 19.1), design matrix $\mathbf{a}^{0,n}$, estimated item-score vector density $\hat{\mathbf{p}}$ (taken from Table 19.1), the best estimate of the item-score vector density for respondent n at iteration 0 ($\hat{\mathbf{p}}^{0,n}$), transformation matrix $\mathbf{Q}$ (taken from Table 19.2), and the estimated score density ($\hat{\mathbf{p}}_{X_+}^{0,n}$). Note that $\mathbf{a}^{0,n} = \mathbf{1}$ shows that all item-score vectors are still admissible. Also note that $\hat{\mathbf{p}}^{0,n} = \hat{\mathbf{p}}$ as there is no information yet on respondent n in iteration 0. As $\max\left(\hat{\mathbf{p}}_{X_+}^{0,n}\right) = .3929 < .9$, FlexCAT continues.

At iteration $i = 1$, $\boldsymbol{\Delta}^{1,n} = \left(\Delta_1^{1,n}, \Delta_2^{1,n}, \Delta_3^{1,n}\right)^T$ (Equation 19.9) has the highest value at $\Delta_3^{1,n}$; hence, item 3 is selected as the new item and presented to respondent n. Respondent n obtains item score $X_3 = 1$. Hence design matrix $\mathbf{a}^{1,n}$ has all elements that pertain to item-score vectors for which $X_3 = 0$ set to zero, resulting in updates of the item-score vector density ($\hat{\mathbf{p}}^{1,n}$) and total-score density ($\hat{\mathbf{p}}_{X_+}^{1,n}$). As $\max\left(\hat{\mathbf{p}}_{X_+}^{1,n}\right) = .4056 < .9$, the CAT continues. At iteration 2, item 1 is selected, and respondent n obtains item score $X_1 = 1$. As $\max\left(\hat{\mathbf{p}}_{X_+}^{2,n}\right) = .5527 < .9$, the CAT continues. At iteration 3, all items have been administered, necessarily leading to $\max\left(\hat{\mathbf{p}}_{X_+}^{3,n}\right) = 1 > .9$, so FlexCAT terminates, and the expected (and real) score equals 2.

Table 19.4 Iterative procedure for the running example based on the LCM(2) in Table 19.1. For details see text

i	$\boldsymbol{\Delta}^{i,n}$	g	X_g	$\mathbf{R}$	$\mathbf{a}^{i,n}$	$\mathbf{p}$	$\mathbf{p}^{i,n}$	$\mathbf{lQ}$	$\mathbf{p}_{X_+}^{i,n}$	S
0				$\begin{pmatrix} 0&0&0\\ 0&0&1\\ 0&1&0\\ 0&1&1\\ 1&0&0\\ 1&0&1\\ 1&1&0\\ 1&1&1 \end{pmatrix}$	$\begin{pmatrix} 1\\1\\1\\1\\1\\1\\1\\1 \end{pmatrix}$	$\begin{pmatrix} .2604\\ .1836\\ .0336\\ .0624\\ .1756\\ .1404\\ .0304\\ .1136 \end{pmatrix}$	$\begin{pmatrix} .2604\\ .1836\\ .0336\\ .0624\\ .1756\\ .1404\\ .0304\\ .1136 \end{pmatrix}$	$\begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&1&0\\ 0&0&0&1 \end{pmatrix}$	$\begin{pmatrix} .2604\\ .3928\\ .2332\\ .1136 \end{pmatrix}$	N
1	$\begin{pmatrix} .6120\\ .5440\\ \mathbf{.6600} \end{pmatrix}$	3	1		$\begin{pmatrix} 0\\1\\0\\1\\0\\1\\0\\1 \end{pmatrix}$		$\begin{pmatrix} 0\\ .3672\\ 0\\ .1248\\ 0\\ .2808\\ 0\\ .2272 \end{pmatrix}$		$\begin{pmatrix} 0\\ .3672\\ .4056\\ .2272 \end{pmatrix}$	N
2	$\begin{pmatrix} \mathbf{.5966}\\ .5530\\ - \end{pmatrix}$	1	1		$\begin{pmatrix} 0\\0\\0\\0\\0\\1\\0\\1 \end{pmatrix}$		$\begin{pmatrix} 0\\0\\0\\0\\0\\ .5528\\ 0\\ .4472 \end{pmatrix}$		$\begin{pmatrix} 0\\0\\ .5527\\ .4472 \end{pmatrix}$	N
3	$\begin{pmatrix} -\\ \mathbf{.4944}\\ - \end{pmatrix}$	2	0		$\begin{pmatrix} 0\\0\\0\\0\\0\\1\\0\\0 \end{pmatrix}$		$\begin{pmatrix} 0\\0\\0\\0\\0\\1\\0\\0 \end{pmatrix}$		$\begin{pmatrix} 0\\0\\1\\0 \end{pmatrix}$	Y

Note: i = iteration; $\boldsymbol{\Delta}^{i,n}$ = vector of deltas (see text); g = selected item; X_g = respondent's n score on the selected item; for $\mathbf{R}$, $\mathbf{a}^{i,n}$, $\hat{\mathbf{p}}$, $\hat{\mathbf{p}}^{i,n}$, $\mathbf{Q}$, and $\hat{\mathbf{p}}_{X_+}^{i,n}$, see text; S = Stop?; N = No (i.e., $\max\left(\hat{\mathbf{p}}_{X_+}^{i,n}\right) \leq .9$); Y = Yes (i.e., $\max\left(\hat{\mathbf{p}}_{X_+}^{i,n}\right) > .9$)

19.5 Comparing FlexCAT and Traditional CAT

We compared the outcomes of traditional CAT to the outcomes of FlexCAT using LCM and the total score as described above. The study serves as an illustration of how different types of CAT can be compared.

19.5.1 Method

Data We used the scores from 4211 Belgian students aged between 9 and 19 (53% women) to the 16 items of the SAQI scale Leertaakgerichtheid (Orientation Towards Learning Task). The data had been deidentified, and respondents with missing item scores had been removed before we obtained the data. As dichotomous item scores were easiest to handle in a traditional CAT, we coded the response "that is true" to item score 1, and responses "that is sometimes true" and "that is not true" to item score 0. We investigated the dichotomous item scores using the Mokken scale analysis (e.g., Sijtsma & Van der Ark, 2017) and found no violations of unidimensionality, local independence, or monotonicity. The scalability coefficient for the entire scale was $H = .427$ ($SE = .007$), suggesting a "medium scale" using Mokken's (1971) benchmarks.

Simulation Design The original data were split randomly in a training set (80% of the item-score vectors, $N = 3369$) used for calibration and a validation set (20% of the item score vectors, $N = 842$). The two types of CAT were applied to each of the 842 item-score vectors in the validation set. The response to an item in the CAT equaled a respondent's actual response in the data. As a result, the data obtained from each CAT procedure was a 842×16 matrix. Items administered in the CAT had scores equal to the item scores in the data, and items not administered in the CAT had missing values.

For FlexCAT, we used the settings as described in this chapter. As a stopping rule, we used $\max\left(\mathbf{p}_{X_+}^{i,n}\right) > c$, using the following values of c : .90, .85, .80, and .75. The LCM was estimated using the R-package **poLCA** (Linzer & Lewis, 2011). For the remainder, we used our own computer code. For the calibration of traditional CAT, we used the two-parameter logistic model. In the traditional CAT, the average percentage of administered items was set approximately equal to the average percentage of administered items of FlexCAT by finetuning the required standard error in traditional CAT's stopping rule. This allowed us to compare the quality of the measurement under an equal level of response burden. Both the calibration and the iterative item administration of the CAT were conducted using the R-package **mirt** (Chalmers, 2012) for calibration, and the R-package **mirtCAT** (Chalmers, 2016) for running the traditional CAT with default settings.

Dependent Variables As the scores used in traditional CAT (estimated latent trait value) and FlexCAT (estimated total score) were incomparable, we used the stanines of the respective scores to compare the two types of CAT: More specifically, we reported the percentage of respondents for which the stanine estimated using CAT was equal to the stanine computed from the complete data, the percentage of respondents for which the difference between the two stanines was 1, and the percentage of the respondents for which the difference between the two stanines was greater than 1. In addition we compared computed the correlation between a respondent's estimated and real score.

19.5.2 Results and Discussion

Depending on the stopping rule, for FlexCAT, the median percentage of administered items ranged between 75% (12 items) and 87.5% (14 items) and for traditional CAT between 75% (12 items) and 100% (16 items). For FlexCAT, the distribution of the number of administered was approximately symmetric (Table 19.5, upper panel) and skewed to the left for traditional CAT (Table 19.5, lower panel). These skewed distributions indicate that, compared to FlexCAT, a large proportion of the respondents in the traditional CAT required relatively few items, and a large proportion of the respondents require all items. FlexCAT showed smaller differences between the actual stanine and the expected stanine than traditional CAT (Table 19.6, middle columns), whereas the correlation between the actual scores and estimated scores where very high for both types of CAT (Table 19.6, last column).

For this example, results showed that FlexCAT and traditional CAT are both doing well, and although FlexCAT performed a bit better, the differences were not overwhelming. This can be expected as we found no violations of unidimensionality, local independence, and monotonicity for this scale, which suggests that a two-parameter logistic model can estimate the item-score vector density rather well. The percentage of items that were administered was less than typically expected in CAT,

Table 19.5 The percentage items of administered in FlexCAT and the corresponding percentage of items administered in traditional CAT, for the SAQI scale Leertaakgerichtheid (Orientation Towards Learning Task)

CAT	*c*	Min (%)	First quartile (%)	Second quartile (%)	Third quartile (%)	Max (%)
FlexCAT	.90	75.0	81.2	87.5	93.8	100.0
	.85	62.5	75.0	81.2	87.5	100.0
	.80	56.2	75.0	81.2	87.5	100.0
	.75	50.0	68.8	75.0	81.2	100.0
Trad. CAT	.90	56.3	68.8	100.0	100.0	100.0
	.85	50.0	62.5	93.8	100.0	100.0
	.80	50.0	56.3	81.2	100.0	100.0
	.75	43.8	50.0	75.0	100.0	100.0

Table 19.6 Difference between actual stanine and estimated stanine for FlexCAT and traditional CAT for the SAQI scale Leertaakgerichtheid (Orientation Towards Learning Task) and the correlation between the estimated and actual score

CAT	c	Difference			Correlation
		0	1	>1	
FlexCAT	.90	98%	2%	0%	.998
	.85	96%	4%	0%	.997
	.80	93%	7%	0%	.995
	.75	91%	9%	0%	.993
Trad. CAT	.90	92%	8%	0%	.996
	.85	93%	7%	0%	.994
	.80	89%	11%	0%	.992
	.75	87%	13%	0%	.991

which may be due to the rather strict stopping rules. Finally, it may be noted that FlexCAT was rather slow: In the simulations, processing a single respondent took approximately 40 s, compared to less than 1 s for a traditional CAT. As the number of items increase, computation time increases too, so for larger data sets, FlexCAT may be too slow.

19.6 Discussion

We proposed a generalization of CAT, coined FlexCAT, and we conjecture that FlexCAT will be useful for tests and questionnaires that do not meet the requirements of IRT models, tests, and questionnaires that are used for both measurement and prediction, and tests and questionnaires that have different measurement levels and items with different numbers of response categories. In a first example concerning the SAQI scale "Orientation Towards Learning Task," we used the LCM to estimate the density of the item-score vectors (**p**), and we used the total score for measurement, finding slightly better results for FlexCAT. The similarity could explain the quality of the scale, which showed no violations of the IRT-model assumptions. However, when multiple scales of SAQI should be administered, then FlexCAT has the additional advantage over traditional CAT: Item scores from scales that already have been administered may help predict the total score of a scale that still has to be administered, and thus reducing the response burden. The percentage of administered items was higher than expected, which suggests that stopping rules and other settings of FlexCAT should be thoroughly investigated. This chapter is merely the start of FlexCAT, and many things need to be investigated before FlexCAT can be used.

The LCM is an attractive candidate to estimate **p**. We are not the first ones to apply the LCM tot CAT. Cheng (2009) and Wang et al. (2012) used the LCM for a CAT for cognitive diagnostic models, which can be conceived as an LCM with 2^Q latent classes, where Q is the number of attributes required to make a test. From a FlexCAT perspective, these authors estimated **p** using the LCM(2^Q) and used the same 2^Q classes weights as measurement scores. Similarly, Van Buuren and Eggen

(2017) estimated **p** using the LCM with a small number of latent classes and used expected class membership as the measurement score. Our use of LCMs in the SAQI example was different, in the sense that we used the LCM as a convenient device to obtain an accurate estimate of **p**, and we were not interested in the number of latent classes, class weights, or parameter identifiability. We were not the first to use the LCM as a density estimation method either. Van der Palm et al. (2016) used the divisive LCM to estimate discrete densities.

Before the LCM can be used as an off-the-shelf density estimator for FlexCAT, the following problems need to be resolved. First is the *curse of dimensionality* problem. As the number of items increases, the order of vector **p**, which is C^J, increases exponentially. For, example, for $J = 130$ items having $C = 5$ ordered response categories, $C^J \approx 7.3 \times 10^{90}$. As 7.3×10^{90} is more than 1 billion times the commonly accepted number of particles in the observed universe, these numbers are beyond the computational limits that are physically possible (cf. Lloyd, 2000). Estimating **p** for this test using the LCM(200) requires $(W - 1) + W \times J \times (C - 1) = 199 + 200 \times 130 \times 4 = 104,199$ free parameters. This is not a computational problem, even for a regular laptop, but the huge model makes FlexCAT very slow, possibly too slow for a sound administration. The administration of a ten-item CAT required 40 seconds, and the computation time increases as the number of items increases. This is one of the main issues that must be investigated. In addition, local optima (e.g., Shireman et al., 2016) may have a large effect on the estimates. Second, choices for goodness of fit criteria, item selection rules, and stopping rules need to be investigated.

Other models can also be used to estimate **p**. From a FlexCAT perspective, Yan et al. (2004) used decision trees to estimate **p** and the total score for measurement. Recently, Gonzalez (2021) provided machine-learning techniques for individual diagnostic assessment. Implementation of other models requires an adaptation of "building blocks" 2, 3, 4, and 5, which will lead to new challenges. Various choices of density-estimation models and scores for FlexCAT must be compared using both simulated and real-life data, to learn which choices tend to work well.

Probably the biggest challenge is the application of FlexCAT in real-life CAT administrations. In addition to a well-working FlexCAT, in which optimal choices have been made with respect to the density estimation, the score, item-selection rules, and stopping rules, it requires fast and user-friendly software and training programs for test administrators.

A.1 Appendix

Consider Eq. 19.10:

$$P\left(X_g = c|\mathbf{r}^{i-1,n}\right) = \frac{\mathbf{1}^T\left[\mathbf{a}^{i-1,n}_{X_g=c} \circ \mathbf{p}\right]}{\mathbf{1}^T\left[\mathbf{a}^{i-1,n} \circ \mathbf{p}\right]}. \tag{19.10}$$

In the numerator, $\mathbf{a}_{X_g=c}^{i-1,n} \circ \mathbf{p}$ produces a $V \times 1$ vector $\mathbf{p}^{**}$ where $p_v^{**} = p_v$ if $a_{X_g=c}^{i-1,n} = 1$ and $p_v^{**} = 0$ otherwise; that is, those probabilities from $\mathbf{p}^T$ are selected that pertain to item-score vectors that are still admissible for respondent n, given the $i-1$ previous item scores and given that score c on item g has been obtained in iteration i. Pre-multiplying $\mathbf{p}^{**}$ with a unit vector sums up the admissible probabilities producing $P(\mathbf{r}^{n,i-1}, X_g = c)$. Analogously, in the denominator, $\mathbf{a}^{i-1,n} \circ \mathbf{p}$ produces a $V \times 1$ vector $\mathbf{p}^*$ where $p_v^* = p_v$ if $a^{i-1,n} = 1$ and $p_v^* = 0$ otherwise; that is, those probabilities from $\mathbf{p}^T$ are selected that pertain to item-score vectors that are still admissible for respondent n, given the $i-1$ previous item scores. Pre-multiplying $\mathbf{p}^*$ with a unit vector sums up the admissible probabilities producing $P(\mathbf{r}^{n,i-1})$. The ratio of $P(\mathbf{r}^{n,i-1}, X_g = c)$ and $P(\mathbf{r}^{n,i-1})$ equals $P(X_g = c|\,\mathbf{r}^{i-1,n})$.

In Eq. 19.11,

$$E\left(X_+|\mathbf{r}^{i-1,n}\right) = \mathbf{x}_+^T\,\mathbf{Q}^T\left[\frac{\mathbf{a}^{i-1,n} \circ \mathbf{p}}{\mathbf{1}\mathbf{1}^T\left(\mathbf{a}^{i-1,n} \circ \mathbf{p}\right)}\right], \tag{19.11}$$

the numerator of the last term results in vector $\mathbf{p}^*$ (cf. denominator of Eq. 19.10), whereas the denominator equals $\mathbf{1}\mathbf{1}^T\mathbf{p}^*$, which is a $V \times 1$ vector with each element equal to $\sum_v p_v^*$. Hence the last term of Eq. 19.11 is the $V \times 1$ vector of rescaled probabilities of admissible item-score vectors $\left[\frac{p_1^*}{\sum_v p_v^*}, \frac{p_2^*}{\sum_v p_v^*}, \ldots, \frac{p_V^*}{\sum_v p_v^*}\right]^T = \mathbf{p}^{n,i-1}$ (e.g., Table 19.4), Hence, Eq. 19.11 reduces to

$$E\left(X_+|\mathbf{r}^{n,i-1}\right) = \mathbf{x}_+^T.\mathbf{Q}^T.\mathbf{p}^{n,i-1} = \mathbf{x}_+^T.\mathbf{p}_{X_+}^{n,i-1}, \tag{A.1}$$

where $\mathbf{p}_{X_+}^{n,i-1}$ is the density of the total scores given the admissible item-score vectors. Because $\mathbf{x}_+^T.\mathbf{p}_{X_+}^{n,i-1} = \sum_{h=0}^{H-1} h\ P\left(X_+ = h|\mathbf{r}^{n,i-1}\right) = E\left(X_+|\mathbf{r}^{n,i-1}\right)$, Eq. 19.11 is true. Equation 19.12 follows a very similar logic.

References

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370. https://doi.org/10.1007/BF02294361

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(1), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*, 1–38. https://doi.org/10.18637/jss.v071.i05

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619–632. https://doi.org/10.1007/s11336-009-9123-2

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(5), 5–18. https://doi.org/10.1007/s11136-007-9198-0

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for Psychologists*. Psychology Press. https://doi.org/10.4324/9781410605269

Gonzalez, O. (2021). Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychological Methods, 26*(2), 236–254. https://doi.org/10.1037/met0000317

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*(2), 215–231. https://doi.org/10.2307/2334349

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282). https://doi.org/10.1109/ICDAR.1995.598994

Li, Q., & Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis, 86*(2), 266–292. https://doi.org/10.1016/S0047-259X(02)00025-8

Linzer, D. A. (2011). Reliable inference in highly stratified contingency tables: Using latent class models as density estimators. *Political Analysis, 19*(2), 173–187. https://doi.org/10.1093/pan/mpr006

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of statistical software, 42*(1), 1–29. https://doi.org/10.18637/jss.v042.i10

Lloyd, S. (2000). Ultimate physical limits to computation. *Nature, 406*(6799), 1047–1054. https://doi.org/10.1038/35023282

Lukociene, O., & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 241–250). Springer. https://doi.org/10.1007/978-3-642-01044-6_22

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages* `catr` *and* `mstr`. Springer. https://doi.org/10.1007/978-3-319-69218-0

McCutcheon, A. L. (2002). Basic concepts and procedures in single-and multiple-group latent class analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 56–88). Cambridge University Press. https://doi.org/10.1017/CBO9780511499531

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. De Gruyter.

Molenaar, D., Dolan, C. V., & De Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika, 77*(3), 455–478. https://doi.org/10.1007/S11336-012-9273-5

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2016). Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology, 46*(1), 252–282. https://doi.org/10.1177/0081175015581379

Psi testuitgevers. (n.d.). *SAQI vertaald* [SAQI translated]. Retrieved August 19, 2021, from https://www.psitestuitgevers.nl/producten/saqi_svl/vertaald/

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://www.jstor.org/stable/2958889

Shireman, E. M., Steinley, D., & Brusco, M. J. (2016). Local optima in mixture modeling. *Multivariate Behavioral Research, 51*(4), 466–481. https://doi.org/10.1080/00273171.2016.1160359

Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*(1), 137–158. https://doi.org/10.1111/bmsp.12078

Thissen, D. (1991). *MULTILOG user's guide* [Computer software]. Scientific Software.

Van Buuren, N., & Eggen, T. H. (2017). Latent-class-based item selection for computerized adaptive progress tests. *Journal of Computerized Adaptive Testing, 5*(2). https://doi.org/10.7333/jcat.v5i2.62

Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2016). Divisive latent class modeling as a density estimation method for categorical data. *Journal of Classification, 33*(1), 52–72. https://doi.org/10.1007/s00357-016-9195-5

Vermunt, J. K., & Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax*. Statistical Innovations Inc. https://www.statisticalinnovations.com/wp-content/uploads/LGtechnical.pdf

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of categorical data using latent class analysis. *Sociological Methodology, 38*(1), 369–397. https://doi.org/10.1111/j.1467-9531.2008.00202.x

Vorst, H. C. M. (2006). *School attitude questionnaire – Internet (SAQI)*. Libbe Mulder. https://hdl.handle.net/11245/1.272122

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Erlbaum.

Wang, C., Chang, H.-H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods, 44*(1), 95–109. https://doi.org/10.3758/s13428-011-0143-3

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Yan, D., Lewis, C., & Stocking, M. (2004). Adaptive testing with regression trees in the presence of multidimensionality. *Journal of Educational and Behavioral Statistics, 29*(3), 293–316. https://doi.org/10.3102/10769986029003293