Behaviormetrika
https://doi.org/10.1007/s41237-023-00202-5

ORIGINAL PAPER

Approaches to estimating longitudinal diagnostic classification models

Matthew J. Madison[1] · Seungwon Chung[2] · Junok Kim[3] · Laine P. Bradshaw[1]

Received: 1 September 2022 / Accepted: 15 June 2023

Abstract

Recent developments have enabled the modeling of longitudinal assessment data in a diagnostic classification model (DCM) framework. These longitudinal DCMs were developed to provide measures of student growth on a discrete scale in the form of attribute mastery transitions, thereby supporting categorical and criterion-referenced interpretations of growth. Studies employing longitudinal DCMs have used different statistical approaches to model examinee attribute mastery transitions. Yet, there has not been research that systematically compares the potential advantages and shortcomings of these different approaches. Via simulation, this study compares and evaluates the performance of three different approaches to estimating longitudinal DCMs. Results show that performance is similar in terms of classification accuracy and reliability, but practical considerations and the overall goals of the application should guide the choice of modeling approach. Implications of these results are discussed.

Keywords Diagnostic classification models · Cognitive diagnosis models · Longitudinal · Growth · Transition · Estimation

Communicated by Jonathan Templin.

✉ Matthew J. Madison
mjmadison@uga.edu

Seungwon Chung
seungwon.chung@cambiumassessment.com

Junok Kim
junokkim@ucla.edu

Laine P. Bradshaw
laineb@uga.edu

[1] University of Georgia, Athens, GA, USA

[2] Cambium Assessment, Washington D.C. 20007, USA

[3] University of CA – Los Angeles, Los Angeles, CA 90095, USA

Published online: 28 June 2023

1 Introduction

Over the past two decades or so, educational assessment researchers and practitioners have shifted attention to studying how students change or 'grow' over time. These examinations of student growth can illuminate the learning that has occurred over a period of time. Traditionally, classical test theory and item response theory (IRT) approaches have been used to model student growth over time. The CTT and IRT frameworks primarily have been used to provide measures of student growth on a continuous scale in the form of gain scores or latent ability gain scores, respectively. Additionally, student growth percentiles (SGPs; Betebenner 2009) have been used to provide norm-referenced quantifications of student growth. More recently, diagnostic classification models (DCMs; Rupp et al. 2010; see also Bradshaw 2016) have been used to provide measures of student growth on a discrete scale. For DCMs, growth at the individual level is defined as transitions in attribute mastery over time. On a group level, growth for DCMs is defined as changes in overall attribute mastery proportions over time. In these ways, longitudinal DCMs support categorical and criterion-referenced interpretations of growth.

The utilization of longitudinal DCMs in research studies and operational assessment is promising, but due to their very recent development, there is a need for a deeper examination into the application of these models. In particular, the studies employing longitudinal DCMs have used different statistical approaches for estimation. Jurich and Bradshaw (2014) used a calibrate-and-score approach, where pre-test classifications were obtained during the pre-test item calibration, and post-test classifications were obtained by scoring the post-test responses with item parameters fixed according to the pre-test item parameter calibration. Madison and Bradshaw (2018a, b) obtained pre- and post-test classifications and item parameter estimates simultaneously using a latent transition analysis framework (LTA; Collins and Wugalter 1992), which is a longitudinal extension of the latent class model. Additionally, one could specify separate attributes for each time point. This is the analogue of an approach used in the IRT literature, where an additional ability dimension is specified for each individual time point (Paek et al. 2014).

The purpose of this study is to compare these three approaches (calibrate-and-score, latent transition, specifying separate attributes) to estimating longitudinal DCMs. These are not the only approaches to estimating longitudinal DCMs (see Hansen 2013; Wang et al. 2018; Huang 2017), but the three included approaches have been used more widely in published studies and are available in commercial software and therefore, are expected to be most accessible for researchers. First, we describe a general DCM, the log-linear cognitive diagnosis model (LCDM; Henson et al. 2009) and the three approaches for extending the LCDM for longitudinal data. Then, using a simulation study, we compare the classification accuracy and classification reliability of each approach. We close by discussing the advantages and shortcomings of each approach and implications of the results for research and practice.

2 Log-linear cognitive diagnosis model (LCDM)

The LCDM is a general DCM that employs a canonical logit function to link binary item responses to examinee latent traits and item parameters. In the LCDM, the latent traits are the attribute profiles that represent the mastery status patterns across all attributes. In this study, we focus on dichotomous attributes (e.g., mastery/non-mastery). For a test measuring A dichotomous attributes, there are a total of 2^A potential attribute profiles. Each attribute profile is represented by a vector of length A, where each element indicates the mastery status of the corresponding attribute. For example, an attribute profile of [0,1,1] would indicate that the examinee had mastered Attributes 2 and 3, but has not mastered Attribute 1. The LCDM probabilistically classifies each examinee into one of these profiles based on the observed item responses.

The LCDM models the conditional probability of a correct response as a function of the attribute profile of an individual, the attributes measured by the item and the item parameters. The LCDM is a general DCM, where many popular DCMs can be specified by constraining certain item parameters. Using a general DCM like the LCDM allows for a "top–down" approach to finding the most parsimonious DCM. Among other general DCMs (GDINA, de la Torre 2011; GDM, von Davier 2005), we chose the LCDM because of its availability in Mplus (Muthén and Muthén 1998–2017), whose generality makes the longitudinal extensions possible. In the LCDM, the item parameters for a simple structure item include an intercept and a main effect, and for complex items measuring multiple attributes, also include interaction effects of the attributes measured by the item.

Here, we introduce the mathematical form of the LCDM logit response function. Consider an item measuring two attributes, Attribute 1 (α_1) and Attribute 3 (α_3). The probability of a correct response to item i by an examinee with attribute profile $\boldsymbol{\alpha}$ is expressed as:

$$logit\left(X_{ic} = 1|\boldsymbol{\alpha}\right) = \lambda_{i,0} + \lambda_{i,1(1)}\alpha_1 + \lambda_{i,1(3)}\alpha_3 + \lambda_{i,2(1,3)}\alpha_1 \cdot \alpha_3 \tag{1}$$

Here, $\lambda_{i,0}$ is the intercept for item i and represents the log-odds of a correct response for examinees who have mastered neither Attribute 1 nor Attribute 3. The main effect terms, $\lambda_{i,1(1)}$ and $\lambda_{i,1(3)}$, represent the increase in log-odds of a correct response given mastery of Attribute 1 or Attribute 3, respectively. Lastly, the interaction term, $\lambda_{i,2(1,3)}$, represents the change in log-odds of a correct response for examinees who have mastered both Attribute 1 and Attribute 3.

Based on the LCDM described above, this study assessed three approaches for longitudinal settings. The first two approaches draw similarities from IRT, while the third approach is based on the LTA framework. Thus, prior to introducing our approaches, it is worthwhile to briefly consider how growth is modeled in IRT framework (e.g., Paek et al. 2014). In the IRT context, there are primarily two methods of accommodating longitudinal data, separate calibration and concurrent calibration. In separate calibration, the item parameters are estimated separately for each time point (with common item parameters constrained equal) and the latent ability distributions are usually specified as $\theta \sim N(0, 1)$, and then linked through

a linking process (see Kolen and Brennan 2004; von Davier and von Davier 2007 for more details). Then, individual growth is determined by adjusting each ability estimate via linking coefficients with reference to the ability estimate from the initial testing occasion. In concurrent calibration, item parameters and latent abilities for each time point are estimated concurrently. In the next sections, we describe the foundations of each approach to extending the LCDM to longitudinal data. For this study, we focus on the simplest case of two time points (pre-test/post-test) and a common test design. We note, however, that these approaches are not limited to two time points (see Wang et al. 2018), nor a common test design (see, Madison and Bradshaw 2018a). For each of the three approaches described below, growth for individuals is defined as transitions between non-mastery and mastery over time, and growth for the group is defined as changes in overall attribute mastery proportions over time. Finally, within each approach, conditional attribute transition probability matrices can be obtained using the estimated classifications or in the case of the TDCM, directly from transitional parameter estimates.

3 Approach 1—Calibrate-and-score

In Approach 1, calibrate-and-score (CS), the LCDM is fit to the pre-test data (or one of the measurement occasions more generally), items are calibrated, and pre-test classifications and attribute mastery proportions are obtained. Next, post-test classifications and attribute mastery proportions are obtained by scoring post-test item responses with post-test item parameters constrained equal to the calibrated pre-test parameter estimates. With calibrated item parameters, the posterior probability examinee e has an attribute profile $\boldsymbol{\alpha}_k$ at post-test can be computed using their observed post-test item response vector $\boldsymbol{x}_e$:

$$P\left(\boldsymbol{\alpha}_e = \boldsymbol{\alpha}_k | \boldsymbol{x}_e\right) = \frac{\nu_k \prod_{i=1}^{I} \pi_{ki}^{x_{ei}} \left(1 - \pi_{ki}\right)^{1-x_{ei}}}{\sum_{c=1}^{2^A} \nu_c \prod_{i=1}^{I} \pi_{ci}^{x_{ei}} \left(1 - \pi_{ci}\right)^{1-x_{ei}}}. \tag{2}$$

The model estimates include the vector of proportions of examinees in each attribute profile ($\boldsymbol{\nu}_c$). Using the estimated classifications from each time point, marginal attribute conditional transition probability matrices can be obtained in a post hoc fashion. For example, to obtain the overall group's conditional probability of transitioning from non-mastery to mastery ($0 \rightarrow 1$), we can take the number of examinees who were classified as non-masters at the pre-test and classified as masters at the post-test, and divide that number by the number of examinees who were classified as non-masters at the pretest. This calculation is Bayes' Theorem applied to the pre- and post-test classifications:

$$P\left(\alpha_{post} = 1 | \alpha_{pre} = 0\right) = \frac{P\left(\alpha_{pre} = 0 \cap \alpha_{post} = 1\right)}{P\left(\alpha_{pre} = 0\right)}.$$

Assuming simple structure items, the total number of parameters in the CS approach for the pre-test is given by $2n_t + (2^A - 1)$, where n_t is the total number of items at each time point. The first term, $2n_t$, represents the number of item parameters to be estimated (would increase with complex items with multiple main effects and interactions effects). The second term, $2^A - 1$, represents the number of attribute profile proportions to be estimated; the last attribute profile proportion is not estimated as they all must sum to 1. With a common item design, the number of parameters estimated at post-test is $2^A - 1$; only the attribute profile proportions need to be estimated. Admittedly, as noted by others (e.g., Paek et al. 2014; Huang 2017), this approach does not account for the interdependency at pre- and post-test due to the measurement of the same respondents. An advantage of making this sacrifice is that estimation is simplified with fewer parameters being estimated. In the simulation study, we examine the impact of ignoring these dependencies.

4 Approach 2—Concurrent I

Approach 2 is a hybrid of DCM and concurrent-calibration IRT. Here, each time point's items are calibrated concurrently and akin to multidimensional IRT framework, additional attributes are specified to accommodate the additional time points. For example, if there are three measured attributes measured over two occasions, then this approach would consist of six total correlated attributes (three at pre-test, three at post-test). With these attributes specified, pre- and post-items are calibrated jointly, imposing equality constraints on common items. Recall that in the post-test run of Approach 1, the common item parameters were constrained to be equal to their respective pre-test estimates. In Approach 2, on the other hand, all item parameters are estimated concurrently, with common item parameters constrained to be equal.

With only simple structure items assumed invariant over time, the total number of parameters to be estimated in Approach 2 is $2n_t + (2^{2A} - 1)$. Since the number of attributes A is doubled, the number of attribute profiles to be estimated is greater than in Approach 1. While the computational load increases compared to Approach 1, this procedure yields proper estimation since the doubly increased number of attributes directly reflects the same attributes being measured at pre- and post-test by the same respondents. In this approach, transition probabilities can be estimated in a post-hoc fashion using estimated classifications as in Approach 1, or they can be more properly estimated using the estimated attribute profile proportions.

DCMs do not have identification constraints for the latent ability distribution. In concurrent-calibration IRT, only the distribution of θ at post-test is freely estimated, and growth is defined in reference to θ at pre-test, which is typically fixed to be normally distributed with a mean of 0 and standard deviation of 1. With DCMs, the attribute mastery distribution is freely estimated at all time points. This free estimation of the attribute mastery distribution at both time points is permissible because the scale is not arbitrarily defined.

5 Approach 3—Concurrent II

The third approach uses an LTA framework to accommodate the longitudinal item response data. The transition diagnostic classification model (TDCM; Madison and Bradshaw 2018a, b) is specified as a constrained LTA and is statistically equivalent to Approach 2 in terms of examinee classifications, model fit, and the number of parameters estimated; however, due to the LTA structural parametrization, transition probabilities are immediately output, and the secondary application of Bayes' Theorem is not necessary for the estimation of transition probabilities.

In Approach 3, we fit a TDCM to account for the longitudinal data. The TDCM is a constrained LTA with attribute profiles, analogues to the latent classes in LTA, specified at each time point in advance. With LCDM as the measurement model at each time point, the TDCM estimates transition probabilities between different attribute mastery status across testing occasions. Given these transition probabilities, we can evaluate whether learning was successful.

Unlike the first two approaches, TDCM explicitly estimates the attribute profiles at different time points simultaneously. In TDCM, the probability of the item response vector for examinee e is expressed as a function of the item response probabilities, the probability of transition between different attribute profiles across time points, and the probability of membership in a specific profile at the initial occasion. In a pre-test/post-test design with two testing occasions, the probability of item response vector $\boldsymbol{x}_e$, a realization of the random variable $\boldsymbol{X}_e$ is defined as:

$$P(\boldsymbol{X}_e = \boldsymbol{x}_e) = \sum_{c_1=1}^{C} \sum_{c_2=1}^{C} \nu_{c_1} \tau_{c_2|c_1} \prod_{t=1}^{2} \prod_{i=1}^{I} \pi_{ic}^{x_{eit}} \left(1 - \pi_{ic}\right)^{1-x_{eit}}. \tag{3}$$

Here, ν_{c_1} is the probability of belonging to the attribute profile c in the pre-test, $\tau_{c_2|c_1}$ represents the attribute mastery status transition probabilities from pre- to post-test, and x_{eit} is Examinee e's response to Item i in Time t. The item response probabilities, π_{ic}, are estimated with the LCDM. Notice that the item response probabilities are not time dependent; rather, they are assumed equal across time points. In this approach, marginal attribute transition probabilities can be estimated in a post-hoc fashion using the estimated classifications as in Approach 1 and 2, or they can be more properly calculated using the estimated profile level transition probabilities (see Madison and Bradshaw 2018b).

In the TDCM, when considering only simple structure items and assuming item parameter invariance, there are $2n_t$ item parameter estimates, $2 * \left(2^A - 1\right)$ attribute profile proportions, and $2^A(2^A - 1)$ transition probabilities. This yields the same total number of parameters as in Approach 2. The TDCM framework affords some additional flexibility in terms of incorporating covariates to predict examinee transitions; predictors can be included in the TDCM by conditioning transition probabilities and attribute profile proportions on a continuous or categorical predictor (e.g., Madison and Bradshaw 2018a, b; Wang, et al. 2018).

6 Simulation study

To compare the performance of the three approaches under different assessment contexts, we designed a simulation study. The primary manipulated factor was the estimation approach with three levels (Approach 1, 2, and 3). Fixed conditions include the sample size (1000), number of attributes (3), number of time points (2), Q-matrix, attribute mastery and change distributions, and item parameters. We focused on the case of two time points (e.g., pre-test/post-test). There were 15 items with each attribute being measured 7 times total (3 simple structure for each attribute, 6 complex structure items measuring two attributes). Pre-test mastery proportions were fixed at 0.40 for all attributes and the post-test mastery proportions were 0.40, 0.55, and 0.70, for the three attributes, respectively. These pre- and post-test proportions correspond to mastery proportion growth rates of 0, 0.15, and 0.30. The within time-point and between time-point attribute correlations were both fixed at 0.50 to represent a moderate-sized correlation expected in educational assessment contexts (e.g., Bradshaw et al. 2014; Kunina-Habenicht et al. 2009).

For each item, intercepts were fixed at 0.2; main effects were fixed at 2.5 and 1.5 on simple and complex structure items, respectively; and two-way interactions were fixed at 1. The parameter values produce a 0.50 difference in correct response probability between complete masters (examinees mastering none of the required attributes) and complete non-masters (examinees mastering none of the required attributes). These item parameters were chosen to reflect modest and realistic item qualities observed in applied DCM studies (Bradshaw et al. 2014; Madison and Bradshaw 2018a; Kunina-Habenicht et al. 2009; Templin and Hoffman 2013).

Madison and Bradshaw (2018a) demonstrated via simulation that the TDCM is robust to departures from full measurement invariance over time. More specifically, they showed that when item parameter invariance was assumed over time, but item parameter drift (IPD; Goldstein 1983) was present, the TDCM was able to provide accurate and reliable classifications. We wanted to explore this result for the other approaches (1 and 2). Therefore, we included item parameter drift conditions. More specifically, we added three IPD amount conditions (20%, 40%, and 60%). In each IPD amount condition, 1.0 was added or subtracted to the 20%, 40%, and 60% of the model's intercepts and main effects.

Data were generated using R, Version 4.2.2 (R Core Team 2022) and analyzed using Mplus, Version 8 (Muthén and Muthén 1998–2017). Within each modeling approach, the full LCDM with up to two-way interaction effects was estimated. Mplus syntax for each modeling approach is provided at the first author's website. There were 250 replications per condition, creating a total of $12 \times 250 = 3000$ analyses. In the results section next, we summarize results by condition with respect to classification accuracy, classification reliability, and structural parameter recovery.

7 Simulation results

7.1 Classification accuracy

To compute classification accuracy rates, the estimated attribute mastery classifications were compared to the generated attribute mastery classifications. Table 1 shows the classification accuracy rates for each condition. Because the classifications accuracy rates for each attribute were nearly identical, and pre- and post-test classification accuracy rates were nearly identical, accuracy rates were averaged over individual attributes and over pre- and post-test. Overall, classification accuracy rates were strong, consistently greater than 0.912. As expected, classification accuracy rates were identical for Approach 2 and 3. We observed a slight decrease in classification accuracy for Approach 1. Similar to results reported by Madison and Bradshaw (2018a), we observed a small negative effect of ignoring IPD; for all three approaches, comparing 60% IPD to 0% IPD, classification accuracy decreased by approximately 0.02.

7.2 Classification reliability

To compute classification reliability, we employed a longitudinal extension of the DCM reliability metric developed by Templin and Bradshaw (2013; Madison 2019). At a single testing occasion, this reliability metric is interpreted as the correlation between classifications obtained from two independent administrations of the same test. In longitudinal contexts, this reliability metric is interpreted as the correlation between estimated transitions obtained from two independent pre-test/post-test experiments. We applied this reliability metric to the pre- to post-test mastery classifications (e.g., $[0 \rightarrow 1]$) to capture the consistency of these mastery status transitions. Table 2 displays the transition reliabilities for each estimation approach. Because the transitions reliabilities for each attribute were nearly identical, reliabilities were averaged over individual attributes. Overall, classification reliability was high, ranging from 0.892 to 0.931. Similar to classification accuracy, classification reliabilities were identical for Approach 2 and 3. Also similar to results for classification accuracy, we observed a slight decrease in classification reliability for

Table 1 Simulation study classification accuracy rates

IPD amount (%)	Approach 1	Approach 2	Approach 3
0	0.926	0.931	0.931
20	0.923	0.930	0.930
40	0.917	0.924	0.924
60	0.912	0.923	0.923

IPD = item parameter drift, ±1 adjustment to intercept/main effect; Approach 1 = calibrate and score; Approach 2 = specify separate attributes; Approach 3 = transition DCM; accuracy rates averaged over the three attributes and over pre- and post-test

Table 2 Simulation study transition reliability estimates

IPD amount (%)	Approach 1	Approach 2	Approach 3
0	0.906	0.931	0.931
20	0.903	0.929	0.929
40	0.898	0.925	0.925
60	0.892	0.922	0.922

IPD = item parameter drift, ±1 adjustment to intercept/main effect; Approach 1 = calibrate and score; Approach 2 = specify separate attributes; Approach 3 = transition DCM; reliability averaged over the three attributes

Approach 1. Following previous results, this decrease was slight, with an average decrease of 0.03 across conditions.

7.3 Structural parameter recovery

We examined the recovery of two aggregate structural parameters: (1) overall growth in attribute mastery proportions and (2) marginal attribute transition probabilities. Recall that for overall growth in attribute mastery, we generated Attributes 1, 2, and 3 to have growth rates of 0, 0.10, and 0.20. Table 3 shows the average estimated growth in attribute mastery proportion for the three approaches. First, as expected, Approaches 2 and 3 were identical. Next, we observed that as IPD increased, growth in attribute mastery tended to be overestimated. Finally, in comparing the approaches, we observed that when there was no IPD, the calibrate and score approach growth rates were recovered as well as the other approaches. But when IPD was present, the calibrate and score approach tended to overestimate growth slightly more than the other approaches.

Table 4 shows the median absolute deviation for marginal attribute transition probabilities and includes both $\tau_{01} = P\left(\alpha_{post} = 1 | \alpha_{pre} = 0\right)$ and $\tau_{10} = P\left(\alpha_{post} = 0 | \alpha_{pre} = 1\right)$. The other transitions (τ_{00} and τ_{11}) do not need to be

Table 3 Simulation study overall growth in mastery proportion recovery

IPD amount (%)	Approach 1			Approach 2			Approach 3		
	Att1	Att2	Att3	Att1	Att2	Att3	Att1	Att2	Att3
0	0.003	0.103	0.199	0.003	0.099	0.198	0.003	0.099	0.198
20	0.011	0.111	0.215	0.001	0.103	0.207	0.001	0.103	0.207
40	0.023	0.123	0.219	0.017	0.117	0.212	0.017	0.117	0.212
60	0.038	0.138	0.235	0.016	0.120	0.224	0.016	0.120	0.224

IPD = item parameter drift, ±1 adjustment to intercept/main effect; Approach 1 = calibrate and score; Approach 2 = specify separate attributes; Approach 3 = transition DCM; Attributes 1, 2, and 3 had 0, 0.10, and 0.20 growth in attribute mastery, respectively

Table 4 Simulation study transition probability median absolute deviations

Parameter	IPD amount (%)	Approach 1			Approach 2/3		
		Att1	Att2	Att3	Att1	Att2	Att3
τ_{01}	0	0.020	0.015	0.030	0.011	0.012	0.010
	20	0.032	0.019	0.045	0.009	0.011	0.011
	40	0.044	0.042	0.062	0.015	0.016	0.023
	60	0.057	0.068	0.080	0.013	0.020	0.027
τ_{10}	0	0.014	0.021	0.026	0.013	0.014	0.015
	20	0.017	0.026	0.030	0.015	0.014	0.010
	40	0.018	0.033	0.033	0.015	0.016	0.011
	60	0.017	0.036	0.039	0.022	0.021	0.014

IPD = item parameter drift, ±1 adjustment to intercept/main effect; Approach 1 = calibrate and score; Approach 2 = specify separate attributes; Approach 3 = transition DCM; parameter recovery averaged over the three attributes

reported as $\tau_{00} + \tau_{01} = 1$ and $\tau_{10} + \tau_{11} = 1$. In Table 4, we combined Approaches 2 and 3 as we have seen they are identical. Recall that for Approach 1 (calibrate-and-score), marginal attribute transition probabilities are calculated in a post-hoc fashion using the estimated classifications. This is not ideal as it does not account for error in classifications. Approaches 2 and 3 use estimated profile proportions and transition probabilities, respectively, to estimate the marginal attribute transition probabilities. A couple of results are immediately noticeable. First, Approaches 2/3 had better recovery of transition probabilities than Approach 1. Second, for Approach 1, attributes with more growth had worse recovery. For Approaches 2–3, however, recovery was not impacted by attribute growth. Finally, it is apparent that increases in IPD resulted in worse parameter recovery for all approaches, with the decrease in performance more pronounced for Approach 1.

8 Discussion

The utilization of DCMs has recently expanded to longitudinal settings for modeling changes in attribute mastery over time. Using a simulation study, this study compares three different approaches to estimating longitudinal DCMs: calibrate-and-score, specifying separate attributes for each time point, and LTA. In the methods sections, we described how the LTA approach is statistically equivalent to specifying separate attributes for each time point with common item parameter equality constraints. Results from the simulation confirmed this result, with classification accuracy and reliability all being identical for the LTA approach and specifying separate attributes for each time point. From a practical perspective, this equivalency is a key discovery because to date, Mplus has been the only software used in published studies applying the TDCM (Kaya and Leite 2016; Madison and Bradshaw 2018b; Tian et al. 2020). Knowing that the TDCM is statistically equivalent to specifying separate attributes means that the TDCM can be made available in the R packages

(*CDM*, George et al. 2016; *GDINA,* Ma and de la Torre 2020; *mirt,* Chalmers 2012), which are more efficient than Mplus for estimating DCMs. For the calibrate-and-score approach, classification accuracy, classification reliability, and structural parameter recovery were slightly decreased compared to the other two estimation approaches. In the presence of item parameter drift, a misspecification often found in longitudinal studies, we observed that all three methods were robust in terms of classification accuracy and reliability.

While the simulation study showed that all three methods are similar in terms of performance, practical considerations must be considered. First, if the sample size is small relative to the number of attributes and Q-matrix complexity, then the more complex approaches may have estimation issues. As the number of time points and number of attributes increase, the LTA and specifying separate attributes approaches both quickly become extremely complex with exponentially increasing dimensionality. For T time points and A dichotomous attributes, the total number of latent classes in these two approaches is 2^{AT}. In our explorations (limited to Mplus with maximum likelihood estimation), with greater than four dichotomous attributes or greater than two time points, the estimation time and data requirements for these two approaches are not feasible for most applications. In these cases, the calibrate-and-score approach may be the only option. Also, in operational contexts, where items are pre-calibrated, the calibrate-and-score approach is commonly applied. Therefore, it is an appealing result that the calibrate-and-score approach is only slightly, almost negligibly, less accurate and reliable than the other two, more complex approaches. These results are limited to the simulation conditions presented and are not expected to hold in every assessment situation. Specifically, under other types of model misspecification or misfit, the more complex approaches are expected to perform increasingly better than the calibrate-and-score approach. Of course, more research is needed examining these approaches under varying types of tests and study designs, misfit, and model misspecifications.

Another consideration in choosing an estimation approach is the overall goal of the analysis. If the sole purpose is to obtain examinee classifications at multiple time points, then the calibrate-and-score or specifying separate attributes approaches will suffice. But if the application involves the inclusion of covariates to predict examinee transition, the LTA approach provides powerful methodology to evaluate intervention effects in a DCM framework (e.g., Madison and Bradshaw 2018b; Wang et al. 2018).

We note that these are not the only approaches to estimating longitudinal DCMs. Hansen et al. (2016) proposed a hierarchical extension to the LCDM that included continuous dimensions to account for local item dependence over time. Huang (2017) used a multilevel DCM to assess change over time in a DCM framework. Wang et al. (2018) employed a higher-order hidden Markov model. Pan et al. (2020) applied a generalized multivariate growth curve model. Other recent studies have applied measurement and structural model reductions and Bayesian estimation approaches to overcome potential estimation complexities (Chen et al. 2018; Wang et al. 2018; e.g., Zhan et al. 2019). These different approaches have different advantages and limitations depending on characteristics of the sample, design of the assessment, goals of the analysis, and data collection process. We limited our study

to these three approaches because they have been used in published studies and they are publicly available in commonly applied software (see https://www.matthewmadison.com for Mplus syntax), and therefore, are expected to be most accessible more researchers. We hope that this study provides some guidance in the application of these methods. We also hope that this paper brings more attention to DCMs and their utility in longitudinal settings, which may be of interest to researchers desiring psychometric methods that support categorical and criterion-referenced interpretations of growth.

Funding National Science Foundation (Grant no. 2050138); Institute of Education Sciences (Grant no. R305D220020).

Data availability This manuscript has no associated data or the data will not be deposited.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Betebenner DW (2009) Norm- and criterion-referenced student growth. Educ Meas Issues Pract 28(4):42–51

Bradshaw L (2016) Diagnostic classification models. In: Rupp A, Leighton J (eds) Handbook of cognition and assessment. Wiley-Blackwell, Hoboken, pp 297–326

Bradshaw L, Izsák A, Templin J, Jacobson E (2014) Diagnosing teachers' understandings of rational number: building a multidimensional test within the diagnostic classification framework. Educ Meas Issues Pract 33(1):2–14

Chalmers P (2012) *mirt*: A multidimensional item response theory package for the R environment. J Stat Softw 48(6):1–29. https://doi.org/10.18637/jss.v048.i06

Chen Y, Culpepper SA, Wang S, Douglas J (2018) A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. Appl Psychol Meas 42(1):5–23

Collins LM, Wugalter SE (1992) Latent class models for stage-sequential dynamic latent variables. Multivar Behav Res 27:131–157

de la Torre J (2011) The generalized DINA model framework. Psychometrika 76(2):179–199

George AC, Robitzsch A, Kiefer T, Groß J, Ünlü A (2016) The R package CDM for cognitive diagnosis models. J Stat Softw 74(2):1–24. https://doi.org/10.18637/jss.v074.i02

Goldstein H (1983) Measuring changes in educational attainment over time: problems and possibilities. J Educ Meas 33:315–332

Hansen M (2013) Hierarchical item response models for cognitive diagnosis. Unpublished doctoral dissertation. University of California – Los Angeles, Los Angeles

Hansen M, Cai L, Monroe S, Li Z (2016) Limited-information goodness-of-fit testing of diagnostic classification item response models. Br J Math Stat Psychol 69:225–252

Henson RA, Templin JL, Willse JT (2009) Defining a family of cognitive diagnosis models using log-linear models with latent variables. Psychometrika 74(2):191–210

Huang HY (2017) Multilevel cognitive diagnosis models for assessing changes in latent attributes. J Educ Meas 54(4):440–480

Jurich DP, Bradshaw LP (2014) An illustration of diagnostic classification modeling in student learning outcomes assessment. Int J Test 14:49–72

Kaya Y, Leite W (2016) Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: an evaluation of model performance. Educ Psychol Measur 77:369–388

Kolen MJ, Brennan RL (2004) Test equating, scaling, and linking, 2nd edn. Springer-Verlag, New York, NY

Kunina-Habenicht O, Rupp AA, Wilhelm O (2009) A practical illustration of multidimensional diagnostic skills profiling: comparing results from confirmatory factor analysis and diagnostic classification models. Stud Educ Eval 35(2):64–70

Ma W, de la Torre J (2020) *GDINA*: an R package for cognitive diagnosis modeling. J Stat Softw 93(14):1–26. https://doi.org/10.18637/jss.v093.i14

Madison MJ (2019) Reliably assessing growth with longitudinal diagnostic classification models. Educ Meas Issues Pract 38(2):68–78

Madison MJ, Bradshaw LP (2018a) Assessing growth in a diagnostic classification model framework. Psychometrika 83(4):963–990

Madison MJ, Bradshaw LP (2018b) Evaluating intervention effects in a diagnostic classification model framework. J Educ Meas 55(1):32–51

Muthén LK, Muthén BO (1998–2017) Mplus user's guide, 8th edn. Muthén & Muthén, Los Angele

Paek I, Park H-J, Cai L, Chi E (2014) A comparison of three IRT approaches to examinee ability change modeling in a single group anchor test design. Educ Psychol Measur 74(4):659–676

Pan Q, Qin L, Kingston N (2020) Growth modeling in a diagnostic classification model framework—a multivariate longitudinal diagnostic classification model. Front Psychol. https://doi.org/10.3389/fpsyg.2020.01714

R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Rupp AA, Templin J, Henson R (2010) Diagnostic measurement: theory, methods, and applications. Guilford, New York

Templin J, Bradshaw L (2013) Measuring the reliability of diagnostic classification model examinee estimates. J Classif 30(2):251–275

Templin J, Hoffman L (2013) Obtaining diagnostic classification model estimates using Mplus. Educ Meas Issues Pract 32:37–50

Tian W, Zhang J, Peng Q, Yang X (2020) Q-Matrix designs of longitudinal diagnostic classification models with hierarchical attributes for formative assessment. Front Psychol. https://doi.org/10.3389/fpsyg.2020.01694

von Davier M (2005) A general diagnostic model applied to language testing data (Research Report No. RR-05–16). Educational Testing Service, Princeton

von Davier M, von Davier AA (2007) A unified approach to IRT scale linking and scale transformations. Methodol Eu J Res Methods Behav Soc Sci 3(3):115–124

Wang S, Yang Y, Culpepper SA, Douglas J (2018) Tracking Skill Acquisition with cognitive diagnosis models: a higher-order, hidden Markov model with covariates. J Educ Behav Stat 43(1):57–87

Zhan P, Jiao H, Man K, Wang L (2019) Using JAGS for Bayesian cognitive diagnosis modeling: a tutorial. J Educ Behav Stat 44(4):473–503