**A Longitudinal Nonparametric Classification Method for Assessing Change in
Latent Attributes Over Time**

Yaohui Liu, Qipeng Chen, Huiying Xu, Peida Zhan*

Department of Psychology, College of Teacher Education, Zhejiang Normal University

#This is a preprint, probably not the final version; the manuscript is still under review.

---

* Corresponding author: Peida Zhan, Email: pdzhan@gmail.com

A Longitudinal Nonparametric Classification Method for Assessing Change in Latent Attributes Over Time

Abstract

This study proposed a longitudinal Hamming distance discrimination (Long-HDD) method to improve the application of longitudinal cognitive diagnosis in practical teaching by introducing a simple computation and less time-consuming nonparametric classification method—HDD—into longitudinal diagnostic data processing. Compared with the HDD, the proposed method represents correlation or dependence between adjacent time points of the same student using Hamming distance in anticipation of using information from the previous time point to improve the classification accuracy at the current time point. A simulation study was conducted to explore the performance of the proposed method in longitudinal diagnostic data analysis and to compare the performance of the proposed method with the HDD and a parametric longitudinal diagnostic classification model. The findings suggest that (1) the Long-HDD can provide high classification accuracy in longitudinal diagnostic data analysis; (2) compared with the parametric model, the Long-HDD is almost unaffected by sample size and performs better than the parametric model in small sample sizes; and (3) the Long-HDD consumes much less computing time than the parametric model. Overall, the Long-HDD is well suited to analyzing longitudinal diagnostic data and can provide speedy diagnostic feedback due to its convenient computation, which is especially significant in small-scale assessments at the classroom and school levels.

Keywords cognitive diagnosis; nonparametric classification; longitudinal assessment, Hamming distance

Education is a long-term endeavor, and students are constantly learning new knowledge and skills as they go through the education process. By tracking students' knowledge and skills over time, it is not only useful to assess the effectiveness of teaching methods but also to map out students' learning trajectories, which facilitates the optimization of the teaching process and the implementation of remedial teaching.

Measuring students' development relies on longitudinal data collected over multiple measures of latent constructs along the developmental trajectory (Wang et al., 2018). In recent years, longitudinal cognitive/learning diagnosis, which evaluates students' latent attributes (e.g., knowledge and skills) and identifies their strengths and weaknesses over a period, has gradually attracted the attention of researchers and practitioners (e.g., Tang & Zhan, 2021; Ren et al., 2021; Wu, 2019; Zhan, 2020a). Several longitudinal diagnostic classification models (DCMs) have been proposed to provide methodological support for data analysis in longitudinal cognitive diagnosis (for a review, see Zhan, 2020b), such as the higher-order latent structural model-based models (e.g., Huang, 2017; Lee, 2017; Zhan, Jiao, Liao, & Li, 2019; Zhan & He, 2021) and the latent transition analysis-based models (e.g., Chen et al., 2018; Li et al., 2016; Madison & Bradshaw, 2018; Wang et al., 2018). The results of these studies suggest that longitudinal DCMs as parametric models can diagnose the individual and overall developmental trajectory of students.

Nevertheless, the practical application of parametric models may encounter some obstacles, especially in small-scale educational projects (e.g., class- or school-level assessments) where diagnostic tracking is most appropriate. For example, there are always concerns that the cognitive assumptions (e.g., conjunctive or disjunctive) that applied in the selected DCMs are incorrect. In addition, large sample size is typically required to obtain accurate parameter estimation and classification. Furthermore, a certain level of programming ability is required to estimate the parameters of complex models. Moreover, parametric models usually require excessive computing time,

especially for complex longitudinal DCMs using the Markov Chain Monte Carlo (MCMC) algorithm (e.g., Chen et al., 2018; Huang, 2017; Li et al., 2016; Wang et al., 2018; Zhan & He, 2021). Those obstacles increase the difficulty of using parametric models and are not conducive to their application and promotion in the practical teaching environment.

To avoid those obstacles, some nonparametric classification methods (NCMs) have been proposed (e.g., Chiu, Douglas, et al., 2009; Chiu & Douglas, 2013; Chiu, et al., 2018; Wang & Douglas, 2015). Compared to the parametric DCMs, NCMs have no requirement for large sample sizes because no parameters are estimated and they make no assumptions about population distribution. Specifically, the application of NCMs gets rid of the dependence on sample size and can be used even when the sample size is 1. In addition, NCMs did not need to consider the parametric models' non-convergence issue and performed well with a variety of cognitive assumptions (e.g., conjunctive and disjunctive) (Chiu & Douglas, 2013; Wang & Douglas, 2015; Chiu et al., 2018; Akbay, 2016). Furthermore, NCMs have a low requirement for programming ability and can give speedy classification, providing for timely diagnostic feedback, due to their computational simplicity.

However, to our knowledge, no study has attempted to use NCMs to analyze longitudinal cognitive diagnostic data; namely, the psychometric performance of NCMs in tracking student development is unknown. Given the demand for longitudinal cognitive diagnosis in practical teaching and the simplicity of NCMs, it is worth experimenting with integrating the two to see how effectively it tracks student development. To this end, the primary goal of this article is to propose a longitudinal NCM that can track changes in the mastery statuses of latent attributes.

The rest of this article starts with a brief review of two methods relevant to this study. One is the longitudinal higher-order deterministic-inputs, noisy "and" gate (Long-DINA) model (Zhan, Jiao, Liao, & Lei, 2019), which is a representative parametric

longitudinal DCM; another one is the Hamming distance discrimination (HDD) for cross-sectional diagnostic data (Chiu & Douglas, 2013), which is a representative NCM with good diagnostic classification function (e.g., Chiu & Douglas, 2013; Chiu et al., 2018). The proposed longitudinal NCM—longitudinal HDD (Long-HDD)—is then presented, followed by a simulation study evaluating the psychometric properties of the proposed method. Finally, the authors summarize their findings and discuss possible directions for future research.

Background

Long-DINA Model

To make this study more focused, this section briefly introduces the Long-DINA model that will be used later as a parametric model for comparison in the simulation study. Previous studies have shown that the Long-DINA model has good psychometric performance in analyzing longitudinal diagnostic data (Pan & Zhan, 2020; Zhan, 2020a; Zhan, Jiao, Liao, & Li, 2019).

A graphical representation of the Long-DINA model for two-time points is shown in Figure 1. Let y_{nit} be the item response of person n ($n = 1, \dots, N$) to item i ($i = 1, \dots, I$) at time point t ($t = 1, \dots, T$). The Long-DINA model can be expressed as follows:

First-order:

$$P(y_{nit} = 1 | \boldsymbol{\alpha}_{nt}, \boldsymbol{g}_{it}, s_{it}) = g_{it} + (1 - g_{it} - s_{it}) \prod_{k=1}^K \alpha_{nkt}^{q_{ikt}}, \quad (1)$$

Second-order:

$$P(\alpha_{nkt} = 1 | \theta_{nt}, \beta_k, \delta_k) = \frac{\exp(\beta_k \theta_{nt} - \delta_k)}{1 + \exp(\beta_k \theta_{nt} - \delta_k)}, \quad (2)$$

Third-order:

$$\boldsymbol{\theta}_n = (\theta_{n1}, \dots, \theta_{nT})' \sim MVN_T(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (3)$$

where g_{it} and s_{it} are the guessing and slipping parameter for item i at time point t , respectively; $\alpha_{nt} = (\alpha_{n1t}, \dots, \alpha_{nKt})'$ denotes person n 's attribute profile at time point t , $\alpha_{nkt} \in \{0, 1\}$; q_{ikt} is the element in an I -by- K polytomous Q-matrix at time point t ; θ_{nt} is person n 's general ability at time point t ; β_k and δ_k are the slope and difficulty parameters of attribute k on all-time points, respectively, since the same latent structure is assumed to be measured at different time points; $\mu = (\mu_1, \dots, \mu_T)'$ is the mean vector, and Σ is the variance-covariance matrix.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ \vdots & \ddots & \\ \sigma_{1T} & \dots & \sigma_T^2 \end{bmatrix}, \quad (4)$$

where σ_{1T} is the covariance of the first and T th general abilities. As a starting and reference point for subsequent time points, θ_{n1} is constrained to follow a standard normal distribution. The mean values and variances of θ_{nt} ($t \geq 2$) are free to estimate. When $T = 1$, the Long-DINA model is reduced to the higher-order DINA model (de la Torre & Douglas, 2004).

Figure 1

A Graphical Representation of the Long-DINA Model.

Note, θ = higher-order ability; α = latent attribute; Y = observed item response; $t = t$ -th test time point; $n = n$ -th student.

HDD Method

If not explicitly stated, the single term HDD refers to the cross-sectional HDD throughout the study. The HDD is a simple and common NCM that compares the Hamming distance between ideal response patterns (IRPs) and observed response patterns (ORPs). The HDD uses the pattern of attributes corresponding to the IRP closest to the ORP as the attribute mastery pattern (AMP) of the student (Chiu & Douglas, 2013). Hamming distance is the number of elements that differ between two vectors of the same length at the corresponding positions. It can be denoted as:

$$d_{hm}(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^I |A_i - B_i|, \quad (5)$$

where $|A_i - B_i| = 1$ indicates the i -th element in $\mathbf{A}$ is not the same as the i th element in $\mathbf{B}$, while $|A_i - B_i| = 0$ for otherwise. For instance, the Hamming distance between (1, 1, 0, 1) and (1, 0, 0, 1) is 2.

Both ORP and IRP can be considered as vectors containing only 0 or 1 with a length equal to the number of items. The difference between ORP and IRP is that ORP is the result under the influence of some stochastic elements (e.g., guessing and slipping), while IRP is not. In the HDD, if K attributes are measured by I items in a test, there are a total of 2^K AMP, and the corresponding 2^K IRPs can be obtained through the Q-matrix. By comparing the Hamming distance between a student's ORP and all IRPs, the attribute pattern corresponding to the IRP with the minimum Hamming distance from the ORP is considered as the student's AMP. The Hamming distance between student n 's ORP and c -th IRP ($c = 1, 2, \dots, 2^K$) can be expressed as:

$$d_{hm}(ORP_n, IRP_c) = \sum_{i=1}^I |y_{ni} - y_{ci}^*|, \quad (6)$$

where y_{ni} is the observed response of student n to item i , and y_{ci}^* is the ideal response of a student with c -th IRP to item i .

For example, see Figure 2, $K = 3$ attributes are measured in $I = 5$ items. There are $2^3 = 8$ possible attribute patterns, each of which will generate a corresponding IRP based on

the Q-matrix. Suppose there are two students 1 and 2 with $ORP_1 (1, 1, 0, 0, 0)$ and $ORP_2 (1, 1, 1, 0, 1)$. By comparing the Hamming distances between the student's ORP and all 8 IRPs, an IRP set can be obtained, in which the IRP with the minimum Hamming distance to the ORP is retained. When only one IRP is included in the IRP set, the attribute pattern corresponding to this IRP is used as the AMP of the student (e.g., AMP_2 in Figure 2). Multiple IRPs may, however, be included in an IRP set if more than one IRP has the minimum Hamming distance from the OPR (e.g., student 1's IRP set in Figure 2), which is common in short tests and low-quality tests. This situation is called a tie and can be broken by randomly selecting one of the IRPs or by assigning different weights to different attributes or attribute patterns (Wang & Douglas, 2015). Here, the attribute pattern (1, 0, 0) that corresponding to the IRP (1, 0, 0, 0, 0) is randomly selected as the AMP of student 1.

It should be noted that the HDD can also be used to analyze longitudinal diagnostic data, which is similar to using cross-sectional parametric models to analyze longitudinal diagnostic data (Zhan, 2020a). Specifically, there are two different computing strategies with the same result (see the online Appendix for proof): one is to analyze the data at different time points separately and repeatedly using the HDD (i.e., HDD with separate computing strategy); the other is to analyze the data at all time points as a holistic cross-sectional data using the HDD (i.e., HDD with simultaneity computing strategy).

Figure 2*An Example of HDD.*

Note, HDD = Hamming distance discrimination; K = number of attribute; ORP = observed response pattern; IRP = ideal response pattern.

Long-HDD Method

In longitudinal studies, multiple latent variables representing the same latent constructs measured at different time points are frequently highly correlated. In the Long-DINA model, such correlation is handled by the multivariate normal distribution (see Equation 3). Similarly, in NCMs, how to handle the correlation between AMPs at adjacent time points without involving model parameters is a problem that needs to be resolved.

Currently, in the analysis of process data in virtual assessments, distance methods were used to handle the correlation between different action sequences, such as the

Hamming distance (Bergner et al., 2014) and the Edit distance (Hao et al., 2015). When comparing the distance between different students' problem-solving action sequences, the closer the distance, the more consistent their action sequences are and the closer their abilities are. In other words, the closer the distance between the action sequences, the higher the correlation between them. Inspired by that, in the proposed method, the Hamming distance is used again to represent the correlation between the AMPs of the same student at different time points. We assume that the AMPs of the same student at adjacent time points are highly correlated, therefore the Hamming distance between his or her AMPs on two consecutive time points should also be the smallest.

The proposed Long-HDD attempts to represent correlation or dependence between adjacent time points of the same student using Hamming distance in anticipation of using information from the previous time point to improve the classification accuracy at the current time point. The correlation or dependence between adjacent time points of the same student can be expressed as:

$$d_{hm}(AMP_{nt}, \alpha_{nc(t+1)}^*) = \sum_{k=1}^K |\alpha_{nkt} - \alpha_{nck(t+1)}^*|, \quad (7)$$

where $\alpha_{nc(t+1)}^*$ is the attribute pattern corresponding to c -th IRP in the IRP set with minimum Hamming distance with ORP of student n at time point $t + 1$, α_{nkt} is the student n 's mastery state of attribute k at time point t (i.e., k -th attribute in student n 's AMP at time point t), and $\alpha_{nck(t+1)}^*$ is the k -th attribute in the attribute pattern corresponding to c -th IRP in the IRP set of student n at time point $t + 1$.

As demonstrated in Figure 3, at a given time point, the computation of Long-HDD is consistent with that of the HDD (see Equation 6). The primary distinction is that to determine the student's AMP at time point 2, the Hamming distance between each appropriate attribute pattern in the IRP set and his or her AMP at time point 1 must be calculated (see Equation 7); then use the attribute pattern with the minimum Hamming distance as his or her AMP at time point 2. Overall, the proposed method uses several

Hamming distances to reduce the range of viable AMPs for the student, which is often helpful when ties occur. When $T = 1$ or at the first time point, the Long-HDD is equivalent to the HDD.

Figure 3
A Graphical Representation of the Long-HDD Method.

Note, Long-HDD = longitudinal Hamming distance discrimination; K = number of attribute; ORP = observed response pattern; IRP = ideal response pattern.

Simulation study

Design and Data Generation

The purpose of the simulation study is twofold: (a) to explore the performance of the Long-HDD in longitudinal diagnostic data analysis, that is, whether it can provide

accurate classification in tracking the development and (b) to compare the classification accuracy of the Long-HDD, HDD with a separated computing strategy, and Long-DINA model.

Five factors were manipulated. First, sample size $N = 25, 50, 100$, and 300 . Considering the applicability of NCMs, this study focused on small-scale assessments. Second, test length at each time point $I = 25$ and 50 . Third, the number of attributes at each time point $K = 3$ and 5 . Fourth, the number of time points $T = 2$ and 3 . Fifth, analysis method $M = \text{Long-HDD, HDD, and Long-DINA}$.

In the Q-matrix at each time point, the first K items that make up the reachability/identity matrix were set as anchor items and were fixed across all time points (Zhan, Jiao, Liao, & Li, 2019). Except for the anchor items, the attribute patterns examined by the other items were extracted with equal probability from the available patterns comprising more than or equal to 2 attributes.

Considering that in practice the guessing probability and slipping probability of items may be negatively correlated (Zhan, Jiao, Liao, & Bian, 2019), the abnormal response probabilities were generated from a bivariate normal distribution at each time point, as follows:

$$\begin{pmatrix} \text{logit}(g_{it}) \\ \text{logit}(s_{it}) \end{pmatrix} \sim N \left(\begin{pmatrix} -2.197 \\ -2.197 \end{pmatrix}, \begin{pmatrix} 1 & -0.6 \\ -0.6 & 1 \end{pmatrix} \right), \quad (8)$$

where g_{it} and s_{it} are the guessing and slipping probabilities of item i at time point t , respectively. This setting leads the guessing and slipping probabilities for all items to follow a positively skewed distribution (mean ≈ 0.1 , minimum ≈ 0.01 , and maximum ≈ 0.6). Noted that the guessing and slipping probabilities of anchor items were fixed at 0.1 .

At the first time point, the true attribute pattern of each student was randomly chosen from all possible 2^K patterns with equal probability. The generation of true attribute pattern of each student at subsequent time points referred to the approach

proposed by Li et al. (2016). Specifically, for each attribute, its transition probability between adjacent time points was set to be the same and as follow:

$$P(\alpha_{nkt} \rightarrow \alpha_{nk(t+1)}) = \begin{cases} P(0 \rightarrow 0) = 0.8 \\ P(0 \rightarrow 1) = 0.2 \\ P(1 \rightarrow 0) = 0.05 \\ P(1 \rightarrow 1) = 0.95 \end{cases} \quad (9)$$

Finally, based on the generated latent attribute patterns, abnormal response probabilities, and Q-matrix, the DINA model was utilized to generate the observed data at each time point. Fifty datasets were generated for each simulated condition.

Analysis

Fifty replications were implemented in each simulated condition. The HDD method is repeatedly used to analyze the data at each time point. The Long-HDD is used to analyze the data at each time point one by one because it needs to use the classification results from the previous time point to estimate the AMP of the current time point. The Long-DINA model is used to analyze the longitudinal data, this involves the reintegration of response data from multiple time points into one large response matrix, which is then analyzed as a whole. For the Long-DINA model, the MCMC algorithm was used to estimate the parameters (Zhan, Jiao, Ma, & Wang, 2019). Two chains were sampled simultaneously and 8,000 iterations were run for each chain, with the first 5,000 iterations used for burn-in. Convergence was achieved according to the potential scale reduction factor < 1.1 criteria (Brooks & Gelman, 1998).

The attribute pattern correct classification rate (PCCR) and the longitudinal PCCR (LPCCR) (Zhan, Jiao, Liao, & Li, 2019) were computed to evaluate the classification accuracy: $PCCR_t = \sum_{r=1}^R \sum_{n=1}^N I[\hat{\alpha}_{ntr} = \alpha_{nt}] / NR$ and $LPCCR = \sum_{r=1}^R \sum_{n=1}^N I[\forall t, \hat{\alpha}_{ntr} = \alpha_{nt}] / NR$, where $I[\cdot]$ is an indicator function, N is the sample size, R is the total number of replication, t is the t -th time point, $\hat{\alpha}_{nt}$ and α_{nt} is the estimated AMP and true AMP for person n at time point t , respectively.

Result

Figures 1 and 2 display the PCCR at each time point and the LPCCR of three methods across all simulated conditions. First, some general findings. As the number of examined attributes increases and the number of time points increases, the classification accuracy of all three methods decreases. By contrast, as the test length increases, the classification accuracy of all three methods increases.

In addition, sample size had little effect on the HDD and Long-HDD. And, under all conditions, the performance of HDD was inferior to that of the Long-HDD, indicating that it is vital to take into account the correlation or dependence between latent attributes at adjacent time points while analyzing longitudinal data.

Furthermore, when the Long-HDD and Long-DINA were compared, the comparative advantages of the former were more clear when the sample size was small, and the latter only showed the classification accuracy matching the former when the sample size climbed to 100 or more.

Table 1 presents the mean computing time of three methods across 50 replications (take the condition of $T = 3$ and $I = 25$ as an example). They presented in the following order in computing time: Long-DINA > Long-HDD > HDD. Not surprisingly, the time consumed by the two nonparametric methods for diagnostic classification was much less than that of the parametric model. In addition, compared to the HDD, the Long-HDD requires additional Hamming distance calculation, so the total computing time slightly increased.

Overall, the results of the simulation study show that (1) the Long-HDD can provide high classification accuracy in longitudinal diagnostic data analysis; (2) compared with the Long-DINA model, the Long-HDD is almost unaffected by sample size and performs better than the Long-DINA model in small sample size; and (3) the Long-HDD consumes much less computing time than the Long-DINA model.

Figure 4

Classification Accuracy of Long-HDD, HDD, and Long-DINA Model ($T = 2$ time points).

Note, T = number of time points; N = sample size; I = test length; K = number of attributes; HDD = Hamming distance discrimination; Long-HDD = longitudinal HDD; Long-DINA = longitudinal deterministic-inputs, noisy “and” gate model; PCCR = attribute pattern correct classification rate; LPCCR = longitudinal PCCR.

Figure 5

Classification Accuracy of Long-HDD, HDD, and Long-DINA Model (T = 3 time points).

Note, T = number of time points; N = sample size; I = test length; K = number of attributes; HDD = Hamming distance discrimination; Long-HDD = longitudinal HDD; Long-DINA = longitudinal deterministic-inputs, noisy “and” gate model; PCCR = attribute pattern correct classification rate; LPCCR = longitudinal PCCR.

Table 1*Computing Time of Long-HDD, HDD, and Long-DINA (T=3 and I=25).*

K	N	HDD	Long-HDD	Long-DINA
3	25	1.532E-04	2.378E-04	4.478E+03
	50	1.421E-04	2.083E-04	1.497E+03
	100	8.715E-05	1.193E-04	6.897E+02
	300	7.104E-05	9.535E-05	3.208E+02
5	25	1.472E-04	2.164E-04	3.769E+03
	50	9.145E-05	1.275E-04	1.200E+03
	100	7.685E-05	1.058E-04	5.798E+02
	300	7.570E-05	1.013E-04	2.656E+02

Note. T = number of time points; N = sample size; I = test length; K = number of attributes; HDD = Hamming distance discrimination; Long-HDD = longitudinal HDD; Long-DINA = longitudinal deterministic-inputs, noisy “and” gate model. The expression aEb is expressed as $a \times 10^b$. For example, 1.532E-04 is expressed as 1.532×10^{-4} , 4.478E+03 is expressed as 4.478×10^3 . The units are seconds.

Summary and Discussion

This study proposed the Long-HDD to improve the application of longitudinal cognitive diagnosis in practical teaching by introducing a simple computation and less time-consuming nonparametric method—HDD—into longitudinal diagnosis data processing. Compared with the cross-sectional HDD, the proposed method represents correlation or dependence between adjacent time points of the same student using Hamming distance in anticipation of using information from the previous time point to improve the classification accuracy at the current time point. The proposed method not only reduces the difficulty of longitudinal diagnostic data analysis but also broadens the application range of NCMs. A simulation study was conducted to explore the performance of the proposed method in longitudinal diagnostic data analysis, and to compare the performance of the proposed method with the cross-sectional HDD and Long-DINA model. The findings suggest that the Long-HDD is well suited to analyzing longitudinal diagnostic data, and it can provide speedy diagnostic feedback due to its

convenient computation, which is especially significant in small-scale assessments at the classroom and school levels.

Despite these promising results, further studies are still needed. First, the Long-HDD is an extension of the HDD, thus the limitations of the HDD also exist in the Long-HDD. For example, only the simplest Hamming distance was considered in the HDD. Specifically, at a given time point, when there are multiple IRPs with the minimum Hamming distance (i.e., tie existed), the Long-HDD randomly selects an attribute pattern corresponding to the IRP as the student's AMP. In practice, different attributes may have different weights, therefore, the weighted HDD (Chiu & Douglas, 2013) can be incorporated into the Long-HDD in further studies. Second, this study assumed that the attributes follow the conjunctive condensation rule, namely, students would have a high correct item response probability only if they have mastered all examined attributes by that item. This assumption may constrain the practical application of the Long-HDD to some extent. In the future, the general NCM (Chiu et al., 2018) can be incorporated into the Long-HDD to improve its applicability under different cognitive assumptions. Third, the transition probability between adjacent time points was set to be the same for each attribute. Although we believe that this setting has no bearing on the conclusions of this study, the specific performance of the Long-HDD under shifty transition probabilities warrants additional exploration.

Last but not least, the proposed Long-HDD only works in theory when the IPR set contains multiple IRPs (i.e., tie existed). When there is just one IRP in the IPR set, the Long-HDD is equivalent to the HDD with a separated computing strategy. Wang and Douglas (2015) found that increasing the number of items would reduce the occurrence of ties. However, our results found that increasing the test length from 25 to 50 did not seem to reduce the advantage of the Long-HDD over the HDD with a separated computing strategy. We believe that in addition to the test length, the occurrence of the tie is also affected by the complexity of the Q-matrix and the quality of items.

References

- Akbay, L. (2016). Relative efficiency of the nonparametric approach on attribute classification for small sample cases. *Journal Of European Education*, 6(1).
- Bergner, Y., Shu, Z., & von Davier, A. A. (2014). Visualization and confirmatory clustering of sequence data from a simulation-based assessment task. *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 177–184).
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455
- Chen, Y., Culpepper, S. A., Wang, S., & Douglas, J. (2018). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied psychological measurement*, 42(1), 5-23.
- Chiu, C. Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, 30(2), 225-250.
- Chiu, C. Y., Douglas, J. A., & Li, X. D. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74(4), 633–665.
- Chiu, CY., Sun, Y. & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika* 83, 355–375.
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333-353.
- Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining*, 7(1), 33–50.
- Huang, H. Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *Journal of Educational Measurement*, 54(4), 440-480.
- Lee, S. Y. (2017). *Growth curve cognitive diagnosis models for longitudinal assessment*. University of California, Berkeley.
- Li, F. M., Cohen, A., Bottge, B., & Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement*, 76(2), 181-204.
- Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika*, 83(4), 963-990.
- Pan, Y., & Zhan, P. (2020). The impact of sample attrition on longitudinal learning diagnosis: A prolog. *Frontiers in psychology*, 11, 1051.
- Ren, H., Xu, N., Lin, Y., Zhang, S., & Yang, T. (2021). Remedial teaching and learning from a cognitive diagnostic model perspective: Taking the data distribution characteristics as an example. *Frontiers in Psychology*, 12.
- Tang, F., & Zhan, P. (2021). Does diagnostic feedback promote learning? Evidence from a longitudinal cognitive diagnostic assessment. *AERA Open*, 7, 23328584211060804.

- Wang, S. Y., & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika*, 80(1), 85–100.
- Wang, S. Y., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*, 43(1).
- Wu, H. M. (2019). Online individualised tutor for improving mathematics learning: a cognitive diagnostic model approach. *Educational Psychology*, 39(10), 1218-1232.
- Zhan, P. (2020a). A Markov estimation strategy for longitudinal learning diagnosis: Providing timely diagnostic feedback. *Educational and Psychological Measurement*, 80(6), 1145-1167.
- Zhan, P. (2020b). Longitudinal learning diagnosis: minireview and future research directions. *Frontiers in Psychology*, 11, 1185.
- Zhan, P., & He, K. (2021). A longitudinal diagnostic model with hierarchical learning trajectories. *Educational Measurement: Issues and Practice*.
- Zhan, P., Jiao, H., Liao, D., & Li, F. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics*, 44(3), 251-281.
- Zhan, P., Jiao, H., Liao, M., & Bian, Y. (2019). Bayesian DINA modeling incorporating within-item characteristic dependency. *Applied Psychological Measurement*, 43(2), 143-158.
- Zhan, P., Jiao, H., Man, K., & Wang, L. (2019). Using JAGS for bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics*, 44(4), 473-503.

Online Appendix

Consistency of HDD Results Under Two Estimation Strategies

Let's say we have two-time points, each of which measures K attributes with I items. First, we need to explain some characters that we'll use later. ORP_1 , ORP_2 , AMP_1 , AMP_2 , IRP set, IRP set 1, and IRP set 2 represent observed response pattern at time point 1, observed response pattern at time point 2, attribute mastery pattern of a certain student at time point 1, attribute mastery pattern of the same student at time point 2, the set of IRP(s) screened with the shortest Hamming distance from ORP , the set of IRP(s) screened with the shortest Hamming distance from ORP_1 at time point 1, and the set of IRP(s) screened with the shortest Hamming distance from ORP_2 at time point 2, respectively.

HDD with Separated Computing Strategy

In this case, we can get 2^K IRPs at each time point and the length of each IRP is equal to I . By comparing the Hamming distance between ORP and IRPs at each time point. IRP set 1 and IRP set 2 can be obtained. Assuming that the Hamming distance between $IRPr_1$ which is randomly selected from the IRP set 1 and ORP_1 is s_1 , and the Hamming distance between $IRPr_2$ which is randomly selected from the IRP set 2 and ORP_2 is s_2 , then the total Hamming distance between (ORP_1, ORP_2) and (IRP_1, IRP_2) is s_1+s_2 . For example, the ORP_1 is (0,1,1), ORP_2 is (0,1,0), IRP_1 is (1,1,1) and IRP_2 is (1,1,1), then the Hamming distance between (ORP_1, ORP_2) (i.e., (0,1,1,0,1,0)) and (IRP_1, IRP_2) (i.e., (1,1,1,1,1,1)) which is 3 equals the distance between ORP_1 and IRP_1 which is 1 add the distance between ORP_2 and IRP_2 which is 2. As shown in Figure S1, if IRP set 1 has n elements and IRP set 2 has m elements, then the probability of (AMP_1, AMP_2) is $1/nm$.

HDD with Simultaneity Computing Strategy

In this case, the longitudinal data were analyzed as a whole data, and the Q-matrix is a bit different. Assuming there are 5 items at each time point, three attributes, and two-time points, its Q matrix can be expressed as Figure S2. $I_1, I_2, \dots, I_5$ are items measured at the first time point, and $I_6, I_7, \dots, I_{10}$ are items measured at the second time points. K_{11}, K_{21}, K_{31} represent the three attributes measured at time point 1, and K_{12}, K_{22}, K_{32} represent the same three attributes measured at time point 2.

Figure S3 shows the flow of using the HDD with simultaneity computing strategy in longitudinal data. With two time points, we can get 2^{2K} attribute patterns which result in 2^{2K} IRPs. Similarly, an IRP set can be obtained in which the element keeps the smallest Hamming distance with (ORP_1, ORP_2) . Because hamming distance calculates the number of different corresponding bits between two strings of equal length. The Hamming distance between ORP and IRP at different times is independent of each other. It is easy to get that the distance between IRPr and (ORP_1, ORP_2) which is s_1+s_2 equals the distance between $(i_1, i_2, \dots, i_l)$ and (ORP_1) add the distance between $(i_{l+1}, \dots, i_{2l})$ and (ORP_2) . IRPr is randomly selected from the IRP set which equals all combinations of IRP set 1 and IRP set 2 (in Figure S1) (i.e., $w = nm$). Finally, the probability of (AMP_1, AMP_2) is also $1/nm$.

Figure S1*HDD with A Separated Computing Strategy.*

Figure S2*Longitudinal Q-Matrix for Two Time Points.*

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}
K_{11}	1				1					
K_{21}		1				1				
K_{31}			1	1	1					
K_{12}						1				1
K_{22}							1		1	1
K_{32}								1	1	

Note, blank means "0".

Figure S3*HDD with A Simultaneity Computing Strategy.*