Extracting Latent Dimensions from Multidimensional Response Timing Data

Guoliang Xu (gx2150@tc.columbia.edu)
Department of Human Development, Teachers College,
Columbia University, New York, NY 10025 USA

James E Corter (jec34@tc.columbia.edu)
Department of Human Development, Teachers College,
Columbia University, New York, NY 10025 USA

Abstract

Computer-based assessments enable the collection of fine-grained response process data through log files. We propose a novel method for extracting latent dimensions from such multidimensional response timing data, based on applying the Weighted MDS (WMDS) model. In our method, dissimilarities among examinees in their response timing vectors are computed, one such matrix for each test item, then WMDS is applied to this collection of matrices. The resulting latent dimensions represent variation among examinees in their patterns of response timing variables, with the dimension weights of the WMDS model reflecting differences across items in the importance of the latent dimensions. Latent dimensions are interpreted via permutation-based importance, correlation analysis and network analysis. Our method is demonstrated using response data from the PISA 2018 Reading and Mathematics assessments. Results show that the extracted latent dimensions are statistically reliable, educationally interpretable, and boost predictive accuracy when used in conjunction with item scores.

Keywords: multidimensional scaling, Individual Differences MDS framework, latent feature extraction and interpretation, response process, educational data mining, PISA

Introduction

In recent years, the increasing prevalence of computer-based problem-solving tasks in large-scale assessments has made it easier to obtain both item response scores and student response process data. In response, proposals have been made to utilize response time and process information in addition to response accuracy, with the goal of improving the diagnosticity of test performance data. Early work in this area investigated the relationship between simple response time and response accuracy, exploring the diagnostic value of response time data for assessing individual performance and cognitive processes (e.g., Thissen, 1983; Wise & DeMars, 2006; Ferrando & Lorenzo-Seva, 2007; Molenaar, 2015; Pokropek, 2016; Wang & Xu, 2015; Ulitzsch et al., 2020; Su & Davison, 2019; Zhan et al., 2018).

To leverage the joint information from both response time and accuracy, various modeling approaches have been proposed. A prominent class of models extends traditional item response theory (IRT) frameworks by explicitly modeling response time, such as the hierarchical modeling approach by van der Linden (2007, 2008), which assumes a joint distribution over accuracy and response time conditioned on latent ability and speed. Other approaches, including mixture models (e.g., Wang & Xu, 2015; Ulitzsch et al., 2020) and diffusion-based models (e.g., Molenaar, 2015; Zhan et al., 2018), aim to classify or interpret response behavior types. Additionally, multidimensional and person-centered approaches have also been explored to better represent individual variation in cognitive strategies or test-taking behavior. For example, Ferrando and Lorenzo-Seva (2007) proposed a measurement factor-analytic model to capture latent traits underlying timing and accuracy, while Pokropek (2016) used a special case of the graded membership model to detect response guessing behaviors across test-takers.

While these prior methods offer valuable insights into the joint modeling of simple response time and accuracy, most still rely on parametric assumptions or impose specific functional forms on the data. Furthermore, the response time methods do not extend naturally to *multidimensional* response time data (an example is given below), nor to other types of response process data, such as keystroke log data. Thus, there is a need for a more general, non-parametric framework capable of robustly extracting interpretable latent dimensions from high-dimensional process data across diverse testing contexts and sample sizes.

To explore latent cognitive structures in a data-driven manner, Tang et al. (2020) introduced a multidimensional scaling (MDS) approach. They computed dissimilarities among low-level action sequences from student response process data for specific items and applied classical MDS to uncover latent dimensions, followed by PCA for interpretation. This geometric method provides a promising direction for modeling individual response processes without strong parametric constraints. Zhang et al. (2022) proposed used a clustering method to extract meaningful patterns of self-directed learning actions from course-level data, diagnosing students in terms of their evolution in behaviors across the semester of a course. These new techniques for deep information extraction on response processes offer scalable data-driven methods for latent variable extraction. But Tang et al.'s method is specifically designed for *sequences* of low-level actions (e.g., keystrokes) and treats all items as exchangeable, neglecting item-level differences in cognitive demands.

To address these limitations, we propose a generalizable, nonparametric method for extracting latent dimensions from multidimensional response time data. Our approach uses weighted multidimensional scaling (MDS), also known as the INDSCAL model (Carroll & Chang, 1970), to account for item-level variation through differential weighting across dimensions. Traditional weighted MDS analyzes multiple proximity matrices simultaneously, each representing a different data "source" (often the sources are individual participants), and jointly estimates a shared latent space together with source-specific weights for the latent dimensions. In our proposed method the data sources are not different individuals (examinees), but different test items. Each proximity matrix consists of the dissimilarities among the examinees. Thus, our method involves the joint estimation of a shared low-dimensional latent space of

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

students based on inter-student distance patterns derived from multidimensional response timing data, and of item-specific dimension weights that capture how each item differentiates students along distinct latent dimensions. In doing so, the model accounts for item heterogeneity and reveals how different items emphasize different cognitive processes or general time allocation patterns. The result is an interpretable representation of student differences that incorporates both overall performance structure and item-specific diagnostic variation. To demonstrate our proposed method, we analyze data from the Mathematics and Reading sections of the PISA 2018 survey.

Methodology

In this section, we describe the proposed method for latent feature extraction from multidimensional response data. In many online assessments, the recorded data includes not only the examinee's responses and corresponding scores but also response process data for each item or question. This data includes metrics such as the time of the first action, the time of the last visit, the number of actions, the total time spent, and the number of visits. For individual items, we denote the time related response process variables (dimensions) as $t_1, t_2, \ldots, t_m$. Assuming n students, M time variables, and Z items, which may be related or independent, the size of the response process timing data expands to $n \times M \times Z$ data points in our final dataset. Thus, efficiently summarizing information from these detailed response characteristics is crucial.

Scaling

To begin, we preprocess the dataset to ensure consistency across response variables with different units. This is achieved by applying the max-min scaling method after any necessary data cleaning. Max-min scaling, also known as normalization by the range, is a technique commonly used in data preprocessing. It is used to transform numerical features into a specific range, typically between 0 and 1. The formula for max-min scaling is:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Here x is the original value of the data point, x′ is the scaled value of the data point, $x_{\min}$ is the minimum value of this variable in the dataset, $x_{\max}$ is the maximum value of this variable in the dataset.

Distance (Similarity) Measure

Euclidean distance, also known as the Minkowski L2 norm, is a widely used metric in multidimensional scaling and clustering analysis. Here, we apply this metric to calculate the distance between the timing vectors for each pair of students, by item. So, for the z-th item, the distance between the timing vectors for students i and j (assuming five timing response variables) is as follows:

$$d_z(i,j) = \sqrt{\sum_{m=1}^{M} \left(t_{im} - t_{jm}\right)^2} \quad (2)$$

In this formula t_{im} represents the value of the m-th timing variable for student i; t_{jm} represents the value of the m-th timing variable for student j; the term $d_z(i,j)$ denotes the Euclidean distance between student i and student j for the z-th item. We apply this metric to compute the inter-student distance matrix for each item. As a result, we obtain Z proximity matrices—one for each test item. Each proximity matrix is of size $n \times n$, capturing the Euclidean distances (based on the response timing variables) between all pairs of the n students. Thus, the results provide an R-dimensional map of the ways in which the response process vectors of examinees differ, on an item-by-item basis. We next discuss how the weighted MDS model can accomplish this.

Weighted Multidimensional Scaling Latent Dimension Extraction Framework

Weighted MDS The weighted MDS method, also known as Individual Differences Multidimensional Scaling (Carroll & Chang, 1970), is an extension of multidimensional scaling (MDS) designed to handle multiple proximity matrices. In its standard formulation, each proximity matrix represents one individual's judgments of similarity among a fixed set of stimuli (e.g., words, images, test items). This method assumes that all individuals perceive these stimuli within a shared underlying multidimensional space, but each individual weights the dimensions differently, capturing individual differences in attention or salience. This results in: a) a common configuration of items in a low-dimensional space; b) a set of individual-specific dimension weights, used to scale the shared space and produce personalized distance representations.

In our study, we apply weighted MDS differently. Instead of having multiple individuals provide judgments among a fixed set of stimuli, we create multiple inter-student proximity matrices, each derived from a different item or task**.** Specifically, each test item yields a symmetric $n \times n$ matrix describing the perceived dissimilarity or distance between pairs of n students based on their multidimensional response timing variable vector for that item. Our goal is to derive a low-dimensional representation of students**,** such that the distances between students in this space reflect their performance or ability differences across items. To accomplish this, we apply the WMDS model in a novel way to allow for item-specific dimension weights, recognizing that different items may emphasize different aspects of student ability. In our approach, each item yields a separate matrix of proximities among students, computed based on the multiple response process indicators**.** The final Z proximity matrices are then input to WMDS to infer a shared student coordinate space, modulated by item-specific weights.

This use of the weighted MDS model allows us to model these multiple proximity matrices by estimating: a) A matrix of shared student coordinates $\mathbf{X} \in \mathbb{R}^{N \times K}$, where each row x_i denotes the position of student i in a $\boldsymbol{K}$-dimensional latent space; b) A matrix of item weights $\mathbf{W} \in \mathbb{R}^{M \times K}$, where each row $w^{(m)}$ reflects how strongly item m loads on each latent dimension. Formally, it allows us to model the predicted distance between student i and j on item m as a weighted Euclidean distance:

$$\hat{d}_{ij}^{(m)} = \sqrt{\sum_{k=1}^{K} w_k^{(m)} (x_{ik} - x_{jk})^2} \quad (3)$$

This formulation allows each item to stretch or compress the common latent space according to its weighting over dimensions. Items that emphasize different cognitive skills induce different "views" of the same student space.

Weighted MDS Optimization Next, to estimate both the student coordinates and the item-specific weights, we use a standard weighted MDS algorithm to minimize the discrepancy between the predicted and observed inter-student distances across all items:

$$\mathcal{L} = \sum_{m=1}^{M} \sum_{i<j} \left(D_{ij}^{(m)} - \hat{d}_{ij}^{(m)} \right)^2 \quad (4)$$

where $D_{ij}^{(m)}$ denotes the observed distance between student i and student j on item m, and $\hat{d}_{ij}^{(m)}$ is the corresponding model-predicted distance computed as a weighted Euclidean distance in the shared latent space. To fit the weighted MDS model to the data, we use the INDSCAL model implementation in the R *smacof* package (de Leeuw & Mair, 2011). The *smacof* package employs the majorization algorithm, an iterative optimization technique designed to minimize the stress function, which quantifies the overall discrepancy between observed and fitted distances. This procedure is well-suited for models incorporating individual or condition-specific variation, such as our weighted MDS setting, since it reliably converges to a locally optimal solution even in high-dimensional spaces.

Latent Dimension Extraction After convergence, the final resulting configuration matrix $\mathbf{X} \in \mathbb{R}^{N \times K}$, returned by the *smacofIndDiff()* function, contains the estimated coordinates of all students in the latent space. Each row x_i serves as a compact representation of student i's relative position, summarizing how this student differs from others across the full set of items.

These final coordinates incorporate the global consistency of student similarity patterns across items, while item differences in the emphasis on specific latent traits are represented through the estimated dimension weights $w^{(m)}$. Thus, the final student coordinates matrix reflects a consensus embedding that integrates all available inter-student distance information while respecting the heterogeneous diagnostic focus of different items. These coordinates thus serve as interpretable indicators of student-level differences in underlying abilities or response patterns.

Validation of the Latent Dimensions

Do the extracted latent dimensions represent important latent abilities or domain-relevant skills or strategies? This question can best be answered by validating the latent dimensions against some external criterion, such as "true" ability. In the application to PISA 2018 described below, true ability can be operationally defined as the student's total test scores (math, reading or math and reading items in PISA 2018), "purified" by excluding those items used to derive the latent features. Use of this purified test score is a conservative approach, because it removes any confounding with the analyzed items. We can compute this criterion score across both domains, Mathematics and Reading, or separately by domain. Predictive performance of the latent dimensions for students are assessed via linear regression using averaged 5-fold cross-validation. Adjusted R^2 was used to evaluate model fit.

Interpretation Framework

To interpret the latent dimensions, we employ a three-part framework: permutation-based variable importance, correlation analysis, and network visualization. Variable importance is assessed using a permutation method (Altmann et al., 2010) implemented with Random Forests (Breiman, 2001), chosen for their ability to capture nonlinear relationships. Model performance is evaluated using mean squared error (MSE). Correlation matrix analysis is used to explore linear associations between latent dimensions and response variables. To further understand the structure of these associations, we conduct network analysis, as follows. Correlations above 0.35 are used to define a binary adjacency matrix, from which we construct an undirected graph where nodes represent features and latent dimensions, and edges represent significant relationships. The resulting network is visualized to highlight central features by color and size.

Application of the Methods to PISA 2018

Initial Validation

The dataset utilized in this study originates from the PISA 2018 Test (https://www.oecd.org/pisa/data/2018database/). Within the PISA assessment, students complete one of three topic combinations: Reading and Mathematics, Reading and Science, or Reading and Global Competence. This study specifically utilizes students completing the first combination, Reading and Mathematics. To simulate a short assessment, such as might arise in a formative classroom assessment, or in the early stages of an adaptive computer-based test, we selected ten PISA items—five from Mathematics and five from Reading. The selection of these ten items was guided by the goal of maximizing the number of students who completed the items, ensuring sufficient sample size for robust statistical analysis. Of the total set of examinees experiencing Reading and Mathematics subtests, 921 completed all ten of the target items. For each item, the dataset records one original response score variable and five response process (timing) variables. The response score variables are labeled with "S" or "C" at the end, indicating whether the answer was correct or incorrect. The five timing variables capture students' interaction behaviors with the item: the time of the first action ("F", timestamp before the first interaction with the item), the time of the last visit ("T", timestamp of the final interaction with the item), the number of actions ("A", interaction count), the total time spent ("TT", sum of time spent on the item), and the number of visits ("V", how many times the item was opened or revisited).

For our latent feature extraction framework, we generate inter-student distance matrices based on the five response time variables, separately for each of the five items under

each domain. These ten matrices form a list used as input to the weighted MDS algorithm.

As an external criterion for validation, we used the total test score for all Pisa math and reading combined (but "purified" by omitting the ten items used for latent feature extraction), purified math total score, and purified reading total score for each student as a proxy for their true achievement. We compared the predictive value of the latent features extracted by the WMDS method to latent features extracted by two alternatives widely used methods, hierarchical clustering and 2-way nonmetric MDS. Figure 1 shows the performance of the methods across several feature extraction scenarios (using 2, 4, 6, 8, and 10 latent features) in predicting the total purified score (combined reading and math). The graph shows that the weighted MDS method (red line) consistently outperforms the other two techniques. Notably, as the number of latent dimensions increases, the Adjusted R-squared (RSQ) for weighted MDS improves significantly. This indicates that weighted MDS captures more meaningful latent information from the response process variables, especially when combined with item scores. F tests indicated that the addition of the latent variables improved the performance of item scores in predicting Total test score. For example, the increase in RSQ by including four latent response process variables over test scores alone was significant, $F(4, 608) = 11.579$, $p<.001$.

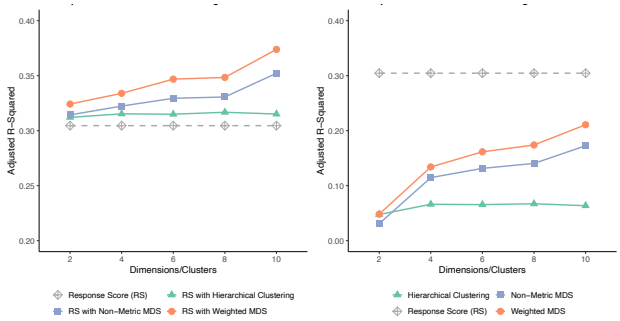

Figure 1: Adjusted RSQ in predicting "purified" Total test score (Reading + Math) from latent dimensions extracted from 10 prediction items(k=2,4,6,8,10). Left panel: using item scores and latent response timing dimensions; Right panel: using latent dimensions only. Dotted line shows adjusted RSQ using item scores alone.

When the method is applied separately to the Mathematics and Reading domains, we observe similar results for predicting pure Math scores (Figure 2). Here, weighted MDS again leads to higher Adjusted RSQ values, particularly when extracting more than four latent dimensions. The increase in RSQ over using item scores alone was again significant, e.g., $F(4,911) = 12.831$, $p<.001$. The comparison highlights the robustness of weighted MDS in isolating and utilizing key latent features relevant to student performance in math. However, for Reading scores (Figure 2), there is little or no advantage for weighted MDS over the alternative methods, and the increase in RSQ for the addition of four latent dimensions over using item scores alone is not significant, $F(4,911) = 1.885$, $p = .111$, although it is nominally the most effective approach as the number of latent dimensions increases. Performance on Reading test items is apparently not as reliant on one or more specific strategies or abilities that leave traces in the response process data.

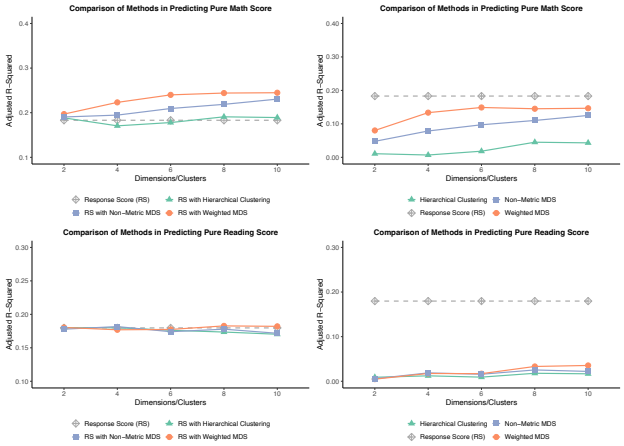

Figure 2: Adjusted RSQ in predicting "purified" domain total test score using latent dimensions extracted from 5 domain test items (k=2,4,6,8,10). Left panel: using item scores and latent dimensions Right panel: using latent response timing dimensions only. Top: Math domain. Bottom: Reading domain.

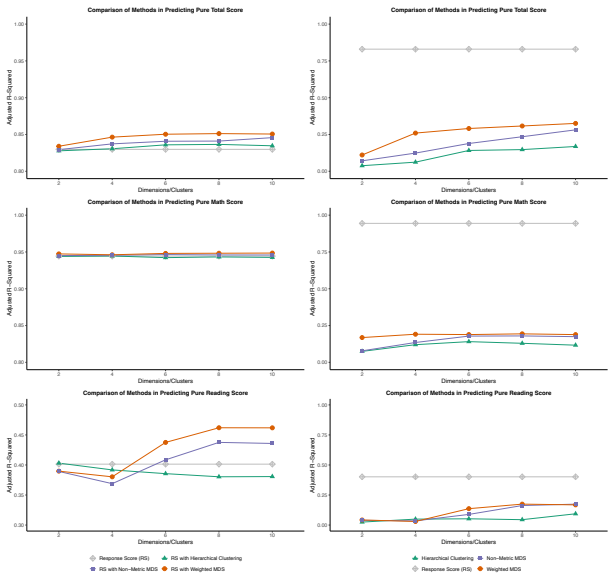

Figure 3: Adjusted RSQ in predicting purified total, math and reading scores (top to bottom) with randomly selected item sets. Left panels: using item scores combined with latent response timing dimensions (k=2,4,6,8,10). Right panels: using latent response timing dimensions only. The grey line shows performance using item scores only.

Validation: Random Selection of Items

In order to ensure generalizability, and make sure that the initial selection of the 10 items did not determine the results, we randomly selected a new set of 10 entirely different items, but again comprised of five items from Math and five from Reading. This alternate dataset included only 280 participants who completed all 10 items.

Using the same methodology as in Case Study 1, we evaluated our approach by using the extracted latent features to predict the purified total score, purified math total score, and purified reading total score. The results, shown in Figure 3, align with our previous findings. Even with a different, randomly chosen item set and reduced sample size, our method demonstrated enhanced performance in predicting the pure total score and pure math total score, and continued to exhibit stable and superior performance for predicting the pure reading total score when six or more dimensions were extracted. These results underscore the robustness of our approach across varying test conditions.

Validation: Varying the Test Length

In this analysis, we aimed to assess the robustness and generalizability of our feature extraction method by increasing the number of test items incrementally from 10 to 20. At each stage of the analysis, two items were added, with the number of prediction items increasing from 10 to 20. We extracted latent features from the response process time data and evaluated their predictive performance in three models: (1) using response scores combined with the extracted latent time features, (2) using only response scores, and (3) using only the extracted latent time features. In one approach, we extracted four latent dimensions across all item counts (Figure 4, left panel). As an alternative approach, we adapted the number of extracted features to the number of items analyzed (e.g., 10 dimensions for 10 items, 12 dimensions for 12 items).

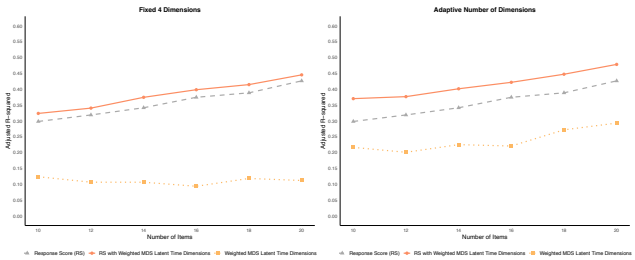

Figure 4: Adjusted RSQ in predicting Total test score with varying numbers of items, ranging from 10 to 20. With each increment, one additional reading and one additional math item are added.

In summary, across all configurations—whether extracting four, ten, or the adaptively optimal number of dimensions—the proposed method consistently demonstrated the added value of latent time features in predicting student performance. The latent dimensions derived from response process times consistently contributed to better predictions, underscoring the robustness and effectiveness of our feature extraction approach.

Latent Feature Interpretation

We examined the extracted latent dimensions using data from the initial validation, in which four dimensions (B1–B4) were derived. These dimensions appear to reflect distinct student strategies, skill mastery levels, or general response patterns related to response process (timing). First, network analysis was used to examine relationships among the response timing variables, score variables, and the latent dimensions. As shown in Figure 5, latent dimensions B1 and B2 were more strongly associated with reading-related timing variables, while B3 and B4 were linked to math-related ones, suggesting that the dimensions capture unique domain-specific response patterns.

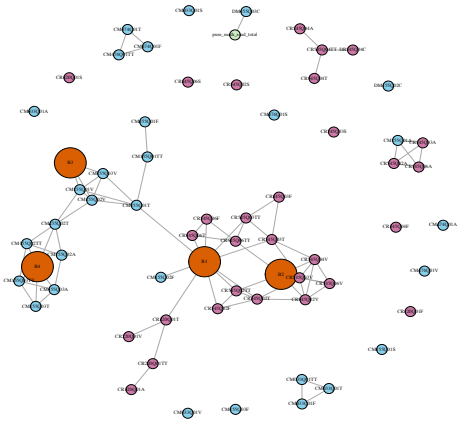

Figure 5: Network analysis of latent dimensions B1–B4 and associated variables. The green node represents the purified total score. Purple and blue nodes denote reading- and math-related timing and score variables, respectively.

In Figure 6, the variable importance results confirmed the network analysis patterns: reading-related variables were the strongest predictors of B1 and B2, while math-related variables were most predictive of B3 and B4.

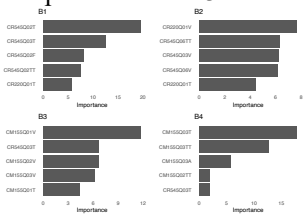

Figure 6: Permutation-based variable importance analysis.

We employed correlation matrix analysis to further examine the relationships between the latent dimensions, the response process variables, and score variables (both individual item scores and purified total scores). To simplify the analysis, we averaged the response timing variables of the same type within the context of reading and math domains. This process resulted in ten general variables: RT, RF, RA,

RTT, RV for reading, and MT, MF, MA, MTT, MV for mathematics. Correlations of these variables with the latent factors are shown in Table 1.

Table 1: Simple correlations between four latent dimensions (B1-B4) and observed response variables

Variable	B1	B2	B3	B4
Timing Summary Variable: **RT (Reading: time to last visit)**	**-0.774**	-0.001	0.083	-0.135
Timing Summary Variable: **RF (Reading: time to 1st action)**	**-0.617**	0.290	-0.108	-0.209
Timing Summary Variable: **RA (Reading: # of actions)**	-0.177	0.143	-0.092	0.055
Timing Summary Variable: **RTT (Reading: total time)**	**-0.555**	**0.543**	-0.064	-0.055
Timing Summary Variable: **RV (Reading: # of visits)**	0.295	**0.542**	-0.241	0.062
Timing Summary Variable: **MT (Math: time to last visit)**	-0.412	0.075	-0.154	**0.601**
Timing Summary Variable: **MF (Math: time to 1st action)**	-0.476	0.140	-0.046	0.151
Timing Summary Variable: **MA (Math: # of actions)**	-0.042	0.047	0.216	**0.516**
Timing Summary Variable: **MTT (Math: total time)**	-0.349	0.297	0.189	**0.556**
Timing Summary Variable: **MV (Math: # of visits)**	0.148	0.349	**0.593**	-0.039
Accuracy Score Variable: **Summed item scores – Read (5)**	0.000	-0.076	0.145	0.268
Accuracy Score Variable: **Summed item scores – Math (5)**	0.133	-0.023	0.135	0.375
Accuracy Score Variable: **Summed item scores – Total (10)**	0.127	-0.030	0.143	0.384
Accuracy Score Variable: **Purified Reading Total**	0.096	-0.033	0.141	0.221
Accuracy Score Variable: **Purified Math Total**	0.210	0.019	0.109	0.237
Accuracy Score Variable: **Purified Total Score**	0.204	0.001	0.144	0.276

Note that both B1 and B2 are linked to reading domains, but they are correlated with different aspects of the response process. B1 had strong negative correlations with the time of the last visit (RT, r=−0.77) and the time of the first action (RF, r=−0.62), suggesting that it captures quicker response behavior in reading tasks. In contrast, B2 was more strongly related to the total time spent (RTT, r=0.54) and the total number of visits (RV, r=0.54). Math-related timing variables were correlated more strongly with dimensions B3 and B4: B3 showed a positive correlation with the total number of visits (MV, r= 0.59), while B4 was associated with the time of the last visit (MT, r=0.601), the number of actions (MA, r=0.52), and the total time spent on math items (MTT, r=0.56). Thus, B1 and B2 are both linked to reading domains, but they capture different behavioral strategies, just as B3 and B4 do for math domains. Table 2 summarizes these interpretations.

Table 2: Latent dimension interpretation summary

Latent Dimensions	Characteristic Response Behavior	Key Features
B1 Rapid work w/ checking (Reading)	Strong Negative with: RT (Reading: last visit), RF (Reading: 1st action), RTT (Reading: total time) Negative with: MT (Math: last visit), MF (Math: 1st action) Positive with: RV (Reading: # of visits)	Short reading times, early first action Relation to performance: Weak positive
B2 Uncertainty / Persistence (Reading)	Strong Positive with: RTT (Reading: total time), RV (Reading: # of visits) Positive with: MTT (Math: total time), MV (Math: # of visits)	Long Reading times, many visits Relation to performance: Near-zero
B3 Math vs. Reading Focus	Strong Positive with: MV (Math: # of visits) Positive with: MTT (Math: total time), MA (Math: # of actions) Negative with: RV (Reading: # of visits)	Many math visits and actions Relation to performance: Weak positive
B4 Effort / Persistence (Math)	Strong Positive with: MT (Math: last visit), MTT (Math: total time), MA (Math: # of actions)	Long math time, many actions Relation to performance: Positive

Discussion

A framework is described for latent feature extraction using the transposed formulation of the Weighted MDS method, aimed at uncovering patterns in response process timing data. The framework uses weighted MDS to extract and interpret meaningful latent dimensions from the data. The Weighted MDS method accounts for variability in response strategy across items by assigning different weights to each dimension.

In evaluations of the method across various item sets, dimensionalities, and numbers of items, the extracted latent dimensions consistently demonstrated high information content in predicting total score and math performance. Furthermore, this "nonparametric" method performed consistently well across both large and moderate sample sizes. However, gains were minimal or non-existent for reading scores, possibly due to domain-specific variation in cognitive processes (Maddox, 2023) or to exogeneous factors.

Our interpretation framework for an application to mixed sets of reading and math items that extracted four latent factors showed that reading-related variables were critical for B1 and B2, while math-related variables were more important for B3 and B4. The specific pattern of correlations suggests straightforward interpretations for each dimension. These results suggest different metacognitive strategies or behaviors and differential allocation of effort across domains.

We believe that our framework has potential for use in practical applications in cognitive and educational assessment. For example, in computerized adaptive testing (CAT), our approach can help address the "cold start" problem in CAT, where early items need to be selected on minimal information (the latent variables provide a crucial "boost"). Specifically, integrating response time feature variables can provide a more precise assessment of a student's initial ability, enabling more accurate selection of subsequent items.

More generally, our method, which handles multivariate timing data with Individual Differences a weighted MDS approach, significantly enhances the diagnostic accuracy of short assessments with small sample sizes. With the increasing use of formative assessments in classrooms, there is growing demand for non-parametric assessment methods that can operate effectively under small-sample, high-dimensional conditions (e.g., Chiu & Köhn, 2019). From a cognitive standpoint, our framework can be employed to identify and code different time strategies used by students during assessments, offering insights into their test-taking behaviors. Thus, the method might support the development of personalized learning strategies and targeted interventions by identifying specific patterns in student responses that might be missed by traditional assessment methods.

Regarding limitations, we note that the current findings rely on specific datasets, and the variability of response process variables across different assessments and contexts warrants caution. Thus, further investigation is needed to establish broad generalizability for the proposed method.

References

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, *35*(3), 283–319. https://doi.org/10.1007/BF02310791

Chiu, C.-Y., & Köhn, H.-F. (2019). Nonparametric Methods in Cognitively Diagnostic Assessment. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (pp. 107–132). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_5

de Leeuw, J., & Mair, P. (2011). *Multidimensional Scaling Using Majorization: SMACOF in R*. https://escholarship.org/uc/item/9z64v481

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An Item Response Theory Model for Incorporating Response Time Data in Binary Personality Items. *Applied Psychological Measurement*, *31*(6), 525–543. https://doi.org/10.1177/0146621606295197

Maddox, B. (2023). The uses of process data in large-scale educational assessments. *OECD Education Working Papers*, (286), 0_1-23.

Molenaar, D. (2015). The Value of Response Times in Item Response Modeling. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3–4), 177–181. https://doi.org/10.1080/15366367.2015.1105073

Pokropek, A. (2016). Grade of Membership Response Time Model for Detecting Guessing Behaviors. *Journal of Educational and Behavioral Statistics*, *41*(3), 300–325. https://doi.org/10.3102/1076998616636618

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using Response Time to Detect Item Preknowledge in Computer-Based Licensure Examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38–47. https://doi.org/10.1111/emip.12102

Su, S., & Davison, M. L. (2019). Improving the Predictive Validity of Reading Comprehension Using Response Times of Correct Item Responses. *Applied Measurement in Education*, *32*(2), 166–182. https://doi.org/10.1080/08957347.2019.1577247

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent Feature Extraction for Process Data via Multidimensional Scaling. *Psychometrika*, *85*(2), 378–397. https://doi.org/10.1007/s11336-020-09708-3

Thissen, D. (1983). Timed Testing: An Approach Using Item Response Theory. In D. J. Weiss (Ed.), *New Horizons in Testing* (pp. 179–203). Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50019-6

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). Using Response Times for Joint Modeling of Response and Omission Behavior. *Multivariate Behavioral Research*, *55*(3), 425–453. https://doi.org/10.1080/00273171.2019.1643699

van der Linden, W. J. (2007). A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2008). Using Response Times for Item Selection in Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5–20. https://doi.org/10.3102/1076998607302626

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

Wang, S., Zhang, S., Douglas, J., & Culpepper, S. (2018). Using Response Times to Assess Learning Progress: A Joint Model for Responses and Response Times. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 45–58. https://doi.org/10.1080/15366367.2018.1435105

Wise, S. L., & DeMars, C. E. (2006). An Application of Item Response Time: The Effort-Moderated IRT Model. *Journal of Educational Measurement*, *43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 262–286. https://doi.org/10.1111/bmsp.12114

Zhang, M., Du, X., Hung, J.-L., Li, H., Liu, M., & Tang, H. (2022). Analyzing and Interpreting Students' Self-regulated Learning Patterns Combining Time-series Feature Extraction, Segmentation, and Clustering. *Journal of Educational Computing Research*, *60*(5), 1130–1165. https://doi.org/10.1177/07356331211065097