Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Sequence mining of cognitive strategies in a computerized neuropsychological task

[†] M. de M. Rodrigues ,[*†] A. L. O. Serpa
† Computational Intelligence in Psychometrics and Epidemiology Research, Mackenzie Presbyterian University, São Paulo, Brazil
*Corresponding author. Email: alexandre.serpa@mackenzie.br

Abstract
The study examines cognitive strategies during problem-solving using the Tower of London (ToL-BR) task. A sample of 592 participants aged between 12 and 29 years completed a computerized ToL-BR test, which required arranging coloured spheres on pegs to match a target configuration with the fewest moves possible. The Longest Common Subsequence (LCS) method, a sequence mining technique, was applied to compare each participant's sequence of moves with an optimal solution. Two main metrics were derived: similarity (the degree of match with the optimal sequence) and efficiency (a normalized ratio reflecting the economy of movements). Results revealed distinct age-related differences. Young adults (18–24 and 25–29 years) consistently showed higher similarity and efficiency scores compared to teenagers (12–17 years), particularly on items requiring complex look-ahead planning. Statistical analyses using Linear Mixed-Effects Models (LMM) confirmed these developmental trajectories, indicating a clear improvement in sequence planning with age. In contrast, differences between male and female participants were non-significant across the task, indicating that these developmental trajectories operate independently of sex. Overall, the findings highlight the effectiveness of sequence mining in uncovering subtle cognitive processes and emphasize that executive function assessments, particularly those measuring planning and problem-solving, should account for age-related variability.

Keywords: Psychometrics, Executive functions, Data science

1. Introduction

Executive functions such as planning and problem solving are critical for daily functioning, yet traditional analyses of tasks. This study investigates planning strategies and efficiency in the Tower of London (ToL) task—a widely utilized measure of executive functions—through sequence mining techniques.

The distinction between automatic and deliberate cognitive processes, a dual-processing framework widely applied in psychology (Kahneman, 2012), provides the theoretical foundation for understanding executive functions. Automatic processes operate rapidly as pre-established responses to routine stimuli, whereas deliberative processes—closely linked to executive functions—enable individuals to adapt behaviors, generate new alternatives, and engage in planning. According to Diamond's (2013) influential model, core executive functions include working memory, inhibitory control, and cognitive flexibility, which serve as the basis for higher-order abilities such as planning (Malloy-Diniz et al., 2008).

Tim Shallice (1982) proposed the Supervisory Attentional System (SAS) model to explain the interaction, suggesting the SAS is recruited for novel or demanding tasks like planning, decision-making, and error monitoring, while routine behaviour uses automatic action schemes. Shallice's original Tower of London task tested this model; patients with left anterior cerebral cortex lesions performed worse, supporting the SAS dysfunction linked to planning impairment. Building on this foundation, the Tower of London (ToL-BR) emerged as a neuropsychological test specifically designed to evaluate deliberative cognitive processes.

Digital-based large-scale assessments now capture not only item responses but also rich, timestamped interaction traces — "process data" —that open new possibilities for measuring how examinees approach tasks and solve problems. This shift transforms testing from a static measure of correctness into a dynamic resource for understanding learning and performance: process data can reveal

strategy use, temporal patterns, and intermediate behaviours that are invisible to response-only models, offering complementary information for assessment development, fairness checks, and instructional feedback (He & von Davier, 2015).

Despite their promise, process data present several methodological challenges. They are voluminous, heterogeneous, and structurally sequential, which complicates direct incorporation into traditional psychometric models (He et al., 2019; Ulitzsch et al, 2022). To address these, researchers adapted and combined approaches from sequence mining, natural language processing, and machine learning. Work on ngram and mini-sequence representations enables the decomposition of long action streams into interpretable features and the identification of discriminative action patterns via feature-selection techniques (He & von Davier, 2015). Complementary pairwise sequence-similarity measures, including longest common subsequence (LCS), provide distance-based summaries of navigation and timing patterns that can be used for clustering or as predictors in supervised models (He et al., 2019; He et al, 2021; Ulitzsch et. al., 2023).

Sequence-mining and pairwise sequence-similarity measures have become central tools for summarizing and comparing examinee action streams. One widely used similarity measure is the Longest Common Subsequence (LCS), which quantifies the length of the longest sequence of actions that appears, in the same order (but not necessarily consecutively), in both sequences being compared. In process-data applications, LCS provides an interpretable distance-like measure of commonality between action sequences and has been used alongside measures such as Dynamic Time Warping (DTW) and n-gram features to cluster behaviours, detect common strategies, and predict task success (He et al., 2019; He et al., 2021).

This research investigates the sequential process of problem-solving to identify and characterize underlying cognitive strategies. Beyond simply capturing these mental approaches, the study examines how they vary across diverse demographic groups. The primary objective is to quantitatively evaluate these strategies by analysing the response patterns of young adults on the computerized ToL-BR test. The ToL is a well-established instrument for assessing executive functions—specifically planning and working memory. By focusing on young adults, we aim to establish a baseline understanding of mature cognitive mechanisms.

2. Material and methods

2.1 Data

The study sample initially comprised 647 participants, of whom N=592 provided complete data for the sequence mining analysis and were included in the final study (M=20.0, SD=8.02). Participants were recruited through convenience sampling by several researchers from different cities, like Belo Horizonte, São Paulo, Florianópolis and Recife, in Brazil. Inclusion criteria required participants to be within the 12–29 age range and to have no self-reported history of neurological disorders.

To investigate the developmental trajectory of planning strategies, the sample was stratified into three developmental groups: Adolescents, Young Adults I, and Young Adults II. These age cut-offs were established based on neurobiological frameworks of executive function maturation, specifically the structural and functional stabilization of the prefrontal cortex, the primary neural substrate for planning, which typically occurs in the mid-20s (Diamond, 2013). Table 1 provides the demographic breakdown of the analyzed sample.

Table 1. Participant demographics information

Developmental Group	Age Range	n	% of Total	Sex (Female / Male)
Adolescents	12–17	219	37.0%	151 / 68
Young Adults I	18–24	272	45.9%	219 / 53
Young Adults II	25–29	101	17.1%	67 / 34
Total	**12–29**	**592**	**100%**	**437 / 155**

2.2 Instruments

2.2.1 Computerized Tower of London (ToL-BR)

The computerized Tower of London (ToL-BR) by Serpa et al. (2024) is a visuospatial problem-solving task requiring participants to move three coloured spheres across three pegs to match a target configuration using the minimum number of moves. Trials displayed the start and goal arrangements, with participants executing one-sphere moves until the goal was achieved optimally. The full task included 19 items, varying in optimal move count and structural complexity to assess planning and executive control. Table 2 details item difficulty (estimated by IRT) and minimum required moves.

Table 2. Item difficulty parameters and minimum required movements for the ToL-BR.

Item	Difficult (category)	Theta	Minimal required movements
1	Medium easy	-0.565	3
2	Easy	-1.369	3
3	Medium easy	-0.547	4
4	Medium hard	0.562	4
5	Medium	-0.478	5
6	Medium hard	0.568	5
7	Medium hard	0.777	5
8	Easy	-1.082	5
9	Hard	0.963	5
10	Medium hard	0.910	5
11	Medium	0.019	6
12	Medium	0.387	6
13	Medium hard	0.516	6
14	Medium	0.438	6
15	Medium	-0.161	7
16	Medium	-0.415	7
17	Medium	0.204	7
18	Medium	-0.116	8
19	Medium easy	-0.609	9

2.3. Data analysis

Descriptive statistics were calculated for demographics, performance measures and raw log data from the ToLBR were processed using the R statistical environment (R Core Team, 2025). Test takers' actions were detected by using the LCS method. This method detects differences between a predefined reference sequence (RS) and the sequence of actions recorded for a test taker on a given item, hereafter the observed sequence (OS). Following He et al. (2021), we computed, for each item, a distance measure between the observed and reference sequences. To summarize general behavioural patterns, we focused on two primary dimensions, similarity and efficiency, defined and implemented according He et al (2021). To characterize the distribution of performance, we reported the standard deviation (SD) for both metrics per item to reflect the inter-individual variability for each specific task configuration.

- **Similarity:** quantifies the extent to which an individual's action sequence overlaps with the reference sequence (or with the closest reference sequence when multiple optimal references exist).
- **Efficiency:** captures the behavioral economy of movement and is operationalized as a normalized ratio. It equals the length of the reference sequence divided by the length of the observed sequence ($Efficiency = Length(RS) \div Length(OS)$), with values closer to 1.0 reflecting fewer redundant actions.

Each TOL move was encoded as a two-character token combining colour and peg position (e.g., B3 R2 G2 B1 G1, where B = Blue, R = Red, G = Green, and 1–3 are peg positions) and stored as an ordered

vector or space-separated string in the study database. Sequence analysis compared each observed vector to the task's reference (optimal) sequence using the LCS algorithm, implemented via the 'qualV' package (Jachner et al., 2007) in R. The LCS is the longest list of tokens that appear in the same relative order in both sequences (for example, between the observed B3 R2 G2 B1 G1 and the reference B3 R2 G1 B1 G1 the LCS is B3, R2, B1, G1, length = 4). We used the raw LCS length and a normalized similarity score (LCS length divided by the shorter sequence length) as primary measures of sequencing accuracy.

For items with multiple optimal solutions (ties), the reference sequence that yielded the highest LCS length for the participant's observed sequence was selected. This ensures that the similarity score reflects the closest alignment to any valid optimal strategy.

To account for the hierarchical structure of the data (responses nested within participants) and to address the limitations of multiple independent tests, data were analyzed using Linear Mixed-Effects Models (LMM) with the lme4 (Bates et al., 2026) and lmerTest (Kuznetsova et al., 2026) packages. Age (as a continuous variable) and sex were entered as fixed effects. Random intercepts were specified for both Participants (to control for inter-individual variability) and Items (to control for varying structural complexity across the task). Model fit was evaluated using the Restricted Maximum Likelihood (REML) criterion, and degrees of freedom were estimated via Satterthwaite's method. To unpack the anticipated item-level variance and specifically test the hypothesis regarding cognitive load, a secondary, targeted LMM was subsequently conducted. This restricted model isolated a subset of high-complexity items, defined *a priori* as those requiring a minimum of 6 moves. This subset approach allows for the evaluation of developmental effects under high 'look-ahead' planning demands while entirely avoiding the inflation of Type I error associated with multiple item-wise comparisons.

3. Results

The study analysed data from a total sample of 592 participants to evaluate cognitive planning strategies. As shown in Table 3, items with higher structural complexity, such as Items 12 and 16, exhibited lower mean similarity scores and higher performance variability compared to simpler items. A preliminary analysis indicated a strong positive correlation between Similarity and Efficiency scores across items ($r > .90$, $p < .001$). Despite this collinearity, both metrics were retained for subsequent analyses to capture distinct facets of executive planning: the cognitive adherence to the optimal solution path (Similarity) and the behavioural economy of movement (Efficiency).

Table 3. Mean and standard deviation for similarity and efficiency for all items

Item	Mean Similarity	SD Similarity	Mean Efficiency	SD Efficiency
1	0.907	0.194	0.872	0.243
2	0.875	0.260	0.853	0.292
3	0.824	0.282	0.786	0.310
4	0.753	0.255	0.693	0.296
5	0.853	0.248	0.819	0.275
6	0.728	0.253	0.671	0.284
7	0.696	0.264	0.637	0.289
8	0.944	0.161	0.925	0.186
9	0.686	0.251	0.625	0.278
10	0.695	0.282	0.641	0.305
11	0.803	0.280	0.758	0.294
12	0.812	0.245	0.766	0.268

13	0.718	0.270	0.667	0.288
14	0.744	0.297	0.698	0.311
15	0.844	0.236	0.810	0.256
16	0.863	0.237	0.828	0.254
17	0.765	0.269	0.722	0.284
18	0.907	0.175	0.875	0.196
19	0.782	0.208	0.663	0.253

3.1. Developmental Differences in Planning Strategies

Developmental trajectories in planning strategies were first evaluated globally using a Linear Mixed-Effects Model (LMM). When averaging performance across all items, Age exhibited a marginal positive trend on sequence Similarity (β = 9.66 X 10^{-4}, t(515.6) = 1.74, p = .081) and did not reach significance for Efficiency (β = 8.79 X 10^{-4}, t(477.7) = 1.40, p = .162). Importantly, the global LMM revealed substantial variance associated with the Item random effect, indicating that planning performance fluctuates significantly depending on the structural complexity of the problem. Because the global model dilutes the developmental effect by averaging highly straightforward items with demanding ones, we conducted a targeted LMM restricted to items requiring high planning depth (i.e., items requiring $\geq$ 6 minimum moves, corresponding to Items 11 and 19 in our processed subset).

Within this high-complexity subset, the LMM confirmed a significant main effect of Age on both cognitive and behavioral metrics. Age significantly predicted sequence Similarity (β = 0.002, t(522.3) = 2.33, p = .020) as well as movement Efficiency (β = 0.002, t(463.5) = 2.29, p = .022). These findings demonstrate that when the cognitive demand for look-ahead planning is maximized, young adults consistently exhibit higher alignment with optimal sequences and execute their plans with significantly fewer redundant moves compared to adolescents. Furthermore, consistent with the global model, sex and clinical group remained non-significant in the high-complexity subset (p > .05), confirming a purely developmental trajectory.

A noteworthy finding pertains to the role of items like Item 11. Although psychometrically classified as having medium difficulty (θ = 0.019), it emerged as highly sensitive to developmental changes. This suggests that developmental sensitivity in sequence planning may not be strictly synonymous with IRT-based item difficulty, but rather with the specific 'look-ahead' demands of the task.

3.2. Sex Differences

Potential sex-based disparities in planning performance were evaluated simultaneously within the LMM framework, controlling for age and item-level variance. The model revealed no significant main effect of sex on either cognitive Similarity (t(667.7) = -0.41, p = .678) or behavioral Efficiency (t(680.7) = -0.22, p = .825). The inclusion of Clinical Group as an exploratory fixed effect also yielded non-significant results for both metrics (p > .30). Given these findings, sex does not appear to act as a primary moderator of executive planning performance on this computerized task, indicating that the observed developmental trajectories operate independently of participant sex.

4. Conclusion

The findings demonstrate age-related improvements in cognitive strategies, which emerge most clearly when task structural demands are maximized. By isolating a subset of high-complexity items requiring extended look-ahead sequences, the targeted sequence mining analysis confirmed that young adults exhibit significantly higher strategic alignment (Similarity) and execution economy (Efficiency) compared to adolescents. These developmental trajectories were supported by Linear Mixed-Effects Models (LMM), accounting for item-level complexity and individual variability.

A primary limitation is the absence of latency metrics, such as the distinction between initial planning time and execution time. While sequence mining captures the strategic routes taken by participants,

integrating timestamps alongside LCS-based metrics in future research is essential to dissociate cognitive 'ability to plan' from the behavioral 'ability to execute'.
Furthermore, the high collinearity between Similarity and Efficiency indices suggests a potential lack of discriminant validity between the two metrics within this paradigm. This may stem from task-design features that could influence the way the examinee is interacting with the task. Future research addressing these features and detangling the effect of these characteristics on examinees' behavior could contribute to understanding this relationship.
The study highlights the utility of sequence mining (LCS) in capturing nuanced cognitive processes and reinforces the importance of considering age-related variability in planning efficiency when assessing executive functions.

Acknowledgement

Competing Interests A. L. O. S. receive royalties from Vetor Editora (Giunti Psychometrics Group) for the computerized version of the Tower of London (TOL-BR) published in Brazil.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N., Tanaka, E., Jagan, M., Boylan, R. D., & Ly, A. (2026). *lme4: Linear mixed-effects models using 'Eigen' and S4* (Version 2.0-1) [Computer software]. CRAN. https://doi.org/10.32614/CRAN.package.lme4

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurevpsych-113011-143750

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. Computers & Education, 166, 104170. https://doi.org/10.1016/j.compedu.2021.104170

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with N-grams. In *Quantitative Psychology Research* (pp. 173–190). Springer International Publishing. https://doi.org/10.1007/978-3-319-19977-1_13

Jachner, S., van den Boogaart, K. G. & Petzoldt, T. (2007). Statistical Methods for the Qualitative Assessment of Dynamic Models with Time Delay (R Package qualV). Journal of Statistical Software, 22(8), 1-30. doi:10.18637/jss.v022.i08

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2026). *lmerTest: Tests in linear mixed effects models* (Version 3.2-1) [Computer software]. CRAN. https://doi.org/10.32614/CRAN.package.lmerTest

Malloy-Diniz, L. F., Martins, C. C., Nassif, E. P., Levy, A. M., Leite, W. B., & Fuentes, D. (2008). Planning abilities of children aged 4 years and 9 months to 8 ½ years: Effects of age, fluid intelligence and school type on performance in the Tower of London test. Dementia and Neuropsychologia, 2(1), 26-30. https://doi.org/10.1590/S1980- 57642009DN20100006

R Core Team (2025). _R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Shallice, T. (1982). Specific impairments of planning. *Philos. Trans. R. Soc. B Biol. Sci. 298*, 199–209. doi: 10.1098/rstb.1982.0082

Ulitzsch, E., He, Q., & Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *47*(1), 3–35. https://doi.org/10.3102/10769986211010467

Ulitzsch, E., Ulitzsch, V., He, Q., & Ludtke, O. (2023). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods*, *55*(3), 1392–1412. https://doi.org/10.3758/s13428-022-01844-1