Mining Process Data for Engagement

A Five-Level Behavioral Taxonomy from TIMSS 2023

Hamdollah Ravand, Susanne Frick, Farshad Effatpanah Hesari, Olga Kunina-Habenicht, Purya Baghaei & **Philipp Doebler**

Beyond Results: Process Data and AI in Educational Assessment
September 24, 2026

Funding Acknowledgement

We are grateful that the project „Beyond the Score: A multi-level and cross-grade analysis of process data in TIMSS 2023 for equity, instruction and item design” is funded by a Research and Development Grant from the International Association for the Evaluation of Educational Achievement (Stichting IEA Secretariaat Nederland) to P. Doebler.

Co-Authors

Hamdollah Ravand
Engagement

Susanne Frick
Process Data, IRT for Engagement, Sequences, Test Taking Styles

Farshad Effatpanah-Hesari
Cross-Grade, Equity, Item Design

Olga Kunina-Habenicht
IRT, Assessment, Item Design

Purya Baghaei
IEA liason, TIMMS 2023 Standards, Policy value

Philipp Doebler
Principal Investigator

Research Questions

What does process data reveal about test-taking behavior?

1. Rule-based inference: Disengaged or not?
2. Is there more than one engaged state/strategy?
3. Good ways of measurement model integration (Bezirhan et al., 2021; Ulitzsch et al., 2020)?
4. What predicts (dis-)engagement?
5. Policy value of process data and item design recommendations?

Beyond binary (dis-)engagment

What is rapid responding?

- An establised definiton of random guessing (RG): A response time (RT) < 10 % of mean RT (**Normative Threshold** [NT]; Wise & Ma, 2012)
- NT adopted in TIMSS literature (e.g., Bulut & Gorgun, 2019; Papanastasiou et al., 2026)
- Constructed response: Essentially no guessing
- Better global label for very fast responses: **Rapid Responding (RR)**
- Related concepts: motivation, engagement
- Related behaviors: item revisits, idling, skipping, skimming (strategic reviewing), quitting

Rapid responding and engagement

- Engagement: Extent to which examinees invest the time, effort, and cognitive resources needed to respond meaningfully to assessment tasks
- RR could be an indicator of engagement, but is **not** identical
- Correct RR is maybe just recall of a well-known fact, a very simple calculation, or a response produced using a familiar procedure
- Especially: A correct RR on a CR item could indicate **high** engagment!

From “engaged vs. disengaged” to heterogeneous engagement

CONVENTIONAL APPROACH

Rapid responding → disengaged
No rapid responding → engaged

⚠ The “engaged” group is treated as homogeneous.

THE PROPOSED APPROACH

Outcome-stratified rapid responding (RR_incorrect) identifies disengagement.

Revisit behavior then differentiates three engaged response processes:
1. Error-Revising
2. Correct-Response Reviewing
3. Direct Completion

HOW THE FIVE CATEGORIES ARE DEFINED

STAGE 1: RR_incorrect

- > 30% → Severely Disengaged
- 15%-30% → Moderately Disengaged

STAGE 2: RR_incorrect ≤ 15%

- Frequently revisits incorrect items: **Error-Revising Engaged**
- Frequently revisits earlier correct responses: **Correct-Response Reviewing**
- No revisits: **Direct Completion Engaged**

Details of the Taxonomy

- $\text{RR\_incorrect} = \frac{\#\text{RR among incorrect responses}}{\#\text{incorrect responses}}$ (**outcome stratification**; Lundgren & Eklöf, 2020)
- Compare with Wise & Kong's (2005) RTE = (#items - #RR)/#items
- RR_incorrect >30% → **Severely Disengaged**
- RR_incorrect ≤30% & RR_incorrect >15% → **Moderately Disengaged**
- Cut-off choices for severe and moderate disengagement (e.g., 30% and 15%) are arbitrary to a degree (sensitivity analysis)
- For remaining engaged categories calculate the mean number of revisits on correct (corR) and incorrect items (incR)
- incR ≥ 1 → **Error-Revising Engaged**
- corR ≥ 1 → **Correct-Response Reviewing**
- Remaining cases: **Direct Completion Engaged**

} Cut-off value of **1** has implications

Empirical Findings

Four main findings

TIMSS 2023 mathematics ; Grade 4: N = 330,853 (60 countries)
Grade 8: N = 289,897 (43 countries)

1 DISENGAGEMENT RISES WITH GRADE

Grade 4: 1.6% disengaged
Grade 8: 5.2% disengaged

Direct Completion: 80.4% → 76.1%
Error-Revising: 17.7% → 18.5%
Correct-Response Reviewing: 0.3% → 0.2%

2 ITEM TYPE MATTERS

Strongest and most consistent predictor of RR: Multiple-choice much higher rapid-responding odds than constructed-response

Grade 4 OR = 11.15 (MC vs CR)
Grade 8 OR = 9.09 (MC vs CR)

3 "ENGAGED" IS HETEROGENEOUS

Among conventionally engaged students: Error-Revising > Direct Completion (Grade 4: +31.69 points, d = .31; Grade 8: +24.96 points, d = .23)

Engaged subtypes also differed in SES; Grade 8 differed in gender composition.

4 PREDICTION DEPENDS ON GRADE

- Grade 4: 5-level taxonomy adds information beyond RR_incorrect ($\Delta\chi^2 = 825.96$, $p < .001$).
- Grade 8: taxonomy does not outperform simpler measures; RTE performs best ($R^2 = .023$ vs $.014$).

VALIDITY

- Discriminant validity: near-zero correlations with achievement and SES
- Predictive validity: engagement associated with achievement
- Known-groups validity: demographic differences supported

ROBUSTNESS

Agreement for different thresholds was strong (weighted $\kappa = .863–.999$ across alternative RR_incorrect cutoffs; 10%–30%)

What can the new taxonomy contribute?

IF THE GOAL IS PREDICTION

Use RTE / RR_incorrect.

The five-level taxonomy improved prediction in Grade 4, but not Grade 8. For older students, simpler indicators remain more appropriate.

IF THE GOAL IS CHARACTERIZATION

Use the taxonomy.

It reveals qualitatively different response-process patterns among students who would otherwise all be labeled “engaged.”

THREE IMPLICATIONS

MEASUREMENT

Response-process evidence can enrich validity arguments beyond binary effort indicators.

ASSESSMENT DESIGN

Item format matters for engagement behavior; process data can inform design and monitoring.

INTERPRETATION

The categories are descriptive behavioral profiles—not established psychological states.

IRT and the Five Level Taxonomy

Measurement models and engagement

- Measurement models for process data go beyond response times:
  - Bezirhan et al. (2021): Revisting as a strategy indicator
  - Ulitzsch et al. (2020): Quitting indicating fatigue/demotivation
- Typically: Continuous uni- or multi-dimensional latent trait models
- Modelling Perspective: Revisits and number of items completed before quitting are both **counts.** What to do with counts?

- Substantive question: Relationship of these complementary perspectives to the Five Level Taxonomy of (dis-)engagement?

Example: TIMSS 2023, United Arab Emirates, Booklet 10

- *J*=2,654 grade 8 students (50.8% female; SES: M=0.07 (SD=0.24)), *I*=25 dichotomized items
- Revisits: M=0.64, SD = 0.87; 960 (36.2%) students do not revisit
- Revisits on items-level: range of means 0.39-1.15, with median = 0 for all items, range of item specific maxima 8-169
- Skimming: Some students skip a lot of items before revisting (26 (1.0%) skip more than 75% of items(!) before revisiting

Five-Level Taxonomy

Severely Disengaged	Moderately Disengaged	Error-Revising Engaged	Correct-Response Reviewing	Direct Completion Engaged
78 (2.9%)	154 (5.8%)	611 (23.0%)	3 (0.1%)	1,808 (68.1%)

Hierarchical Speed-Accuracy-Revisits (SAR) Model (Bezirhan et al., 2021)

Person $j = 1, \ldots, J$, item/block $i = 1, \ldots, I$ with maximal score M_i

$$
\begin{aligned}
X_{ij} &\sim \text{Binomial}\big(M_i,\ \text{logit}^{-1}(u_j - b_i)\big)\,, \\
\log T_{ij} &\sim \mathcal{N}\big(\xi_i - t_j,\ a_i^{-2}\big)\,, \\
R_{ij} &\sim \text{Poisson}\{\exp(r_j - \gamma_i)\}\,.
\end{aligned}
$$

Person-level latent traits:

$$
\boldsymbol{\eta}_j = (u_j, t_j, r_j)' \sim \mathcal{N}_3(\mathbf{0}, \Sigma_P), \qquad \Sigma_P = D_\sigma\, \Omega_P\, D_\sigma.
$$

Priors:

$$
\sigma_{P,k} \sim \text{Half-Cauchy}(0, 4), \quad \Omega_P \sim \text{LKJ}(1), \quad b_i, \xi_i, \gamma_i \sim \mathcal{N}(0, 100^2), \quad a_i^{-1} \sim \text{Half-Cauchy}(0, 4).
$$

Example: TIMSS 2023, United Arab Emirates, Booklet 10

- Reimplementation of Bezirhan et al.'s (2021) model in stan
- Fit with stan (4 chains, each with 1250 draws warmup and 750 samples)

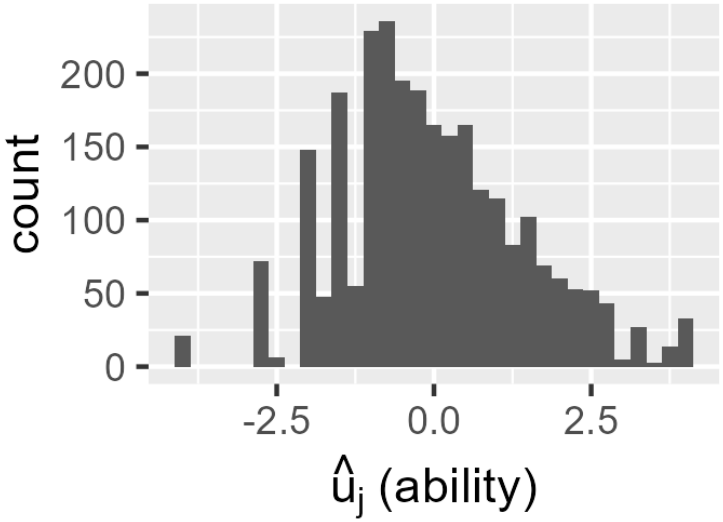
Correlations of latent variables (posterior means)

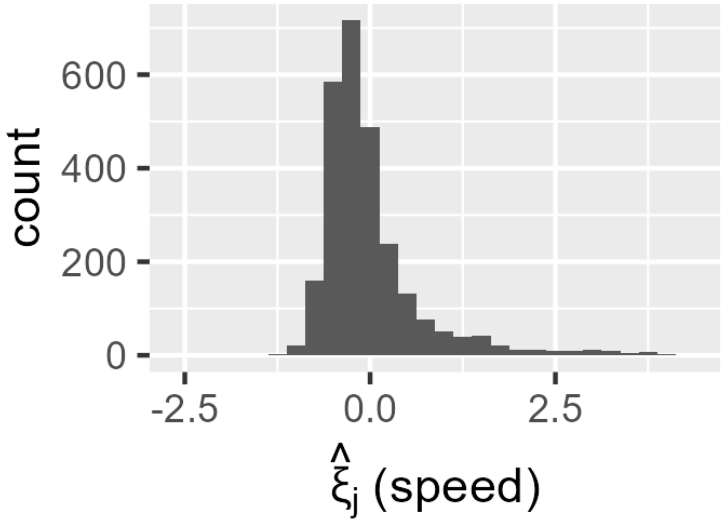
	speed	revisits
ability	-.201	.113
speed	1.000	.448

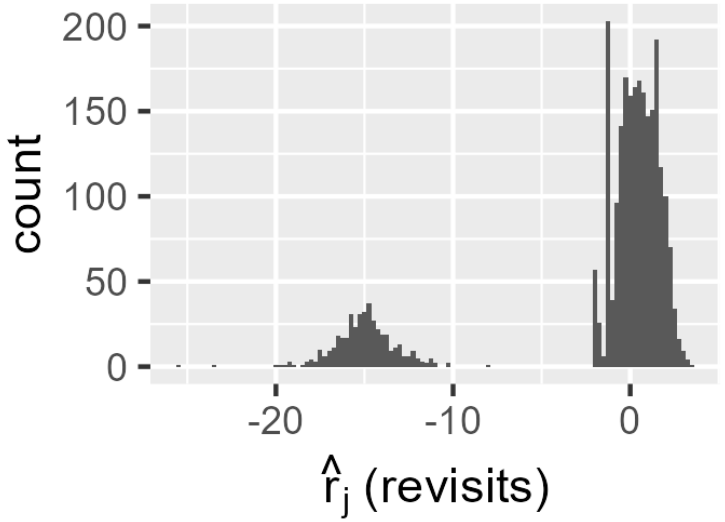
- Strategies vary between students: Skimming reflected by positive correlation of revisits and speed
- Revisits tend to improve ability scores.
- High ability scores tend to coincide with taking time on first visit.

Bezirhan et al.: Latent Traits

- MAP person par. estimation based on posterior mean item parameters

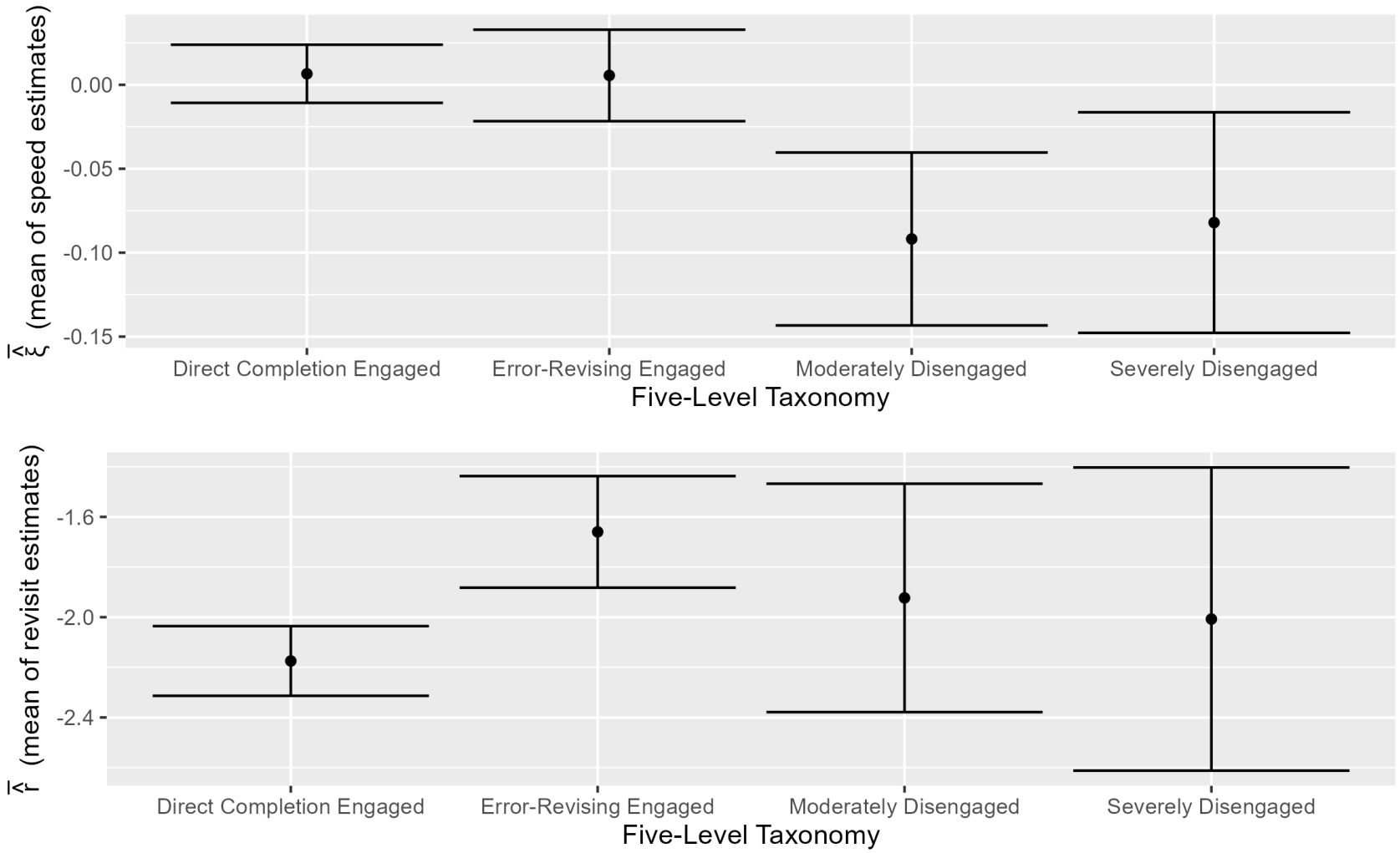






- Bimodal distribution of revisits indicating a mixture of „never revisiters“ and others

Bezirhan et al.: Latent Traits and Five Level Taxonomy

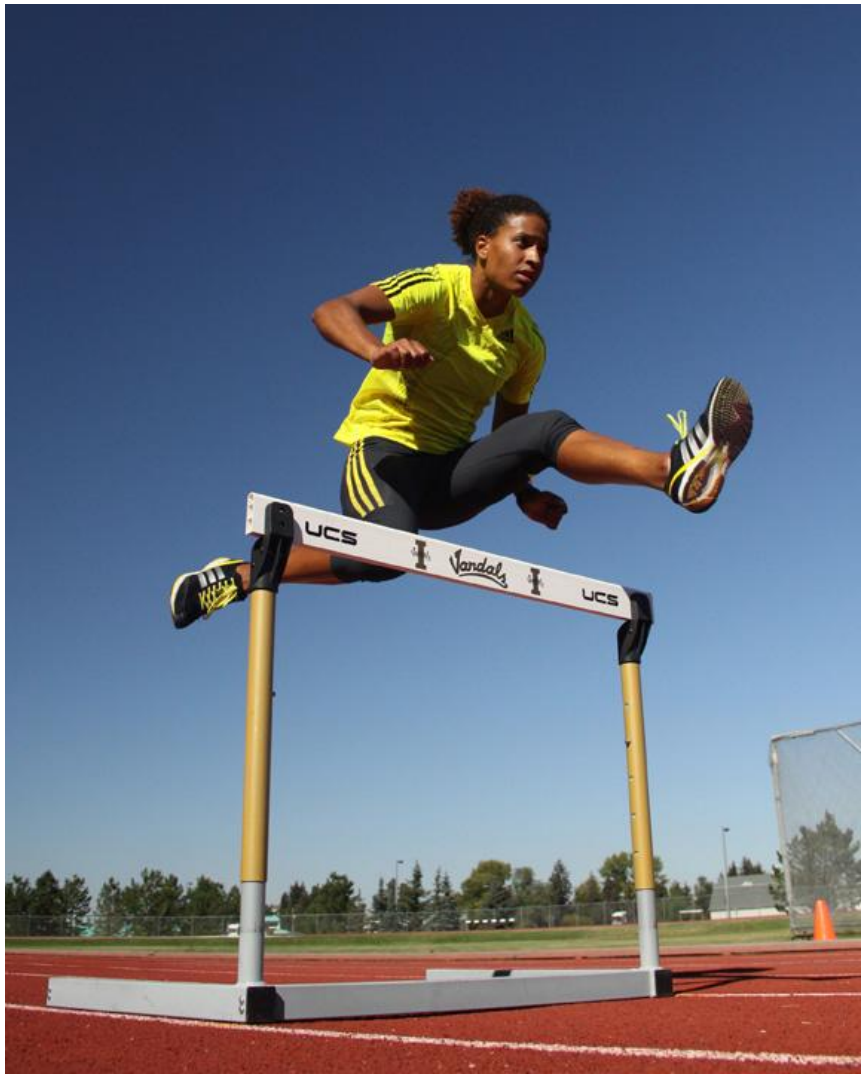


Patterns consistent with our expectations

Zero counts are informative!

- If outside factors imply that behavior is not possible, the zero is uninformative (e.g., no revisit possible due to different implementations).
- But: For many (test-taking) behaviors this is not the case!

- What's the association of the **number of items with revisits** (# non-zeros) and their **intensity** (# of revisits)?
- Is a **latent hurdle process** plausible?



Bayesian negative binomial hurdle IRT model (Doebler, Frick, Demko, Petersen, in prep.)

- Key idea: Model the transition from zero to a count > 0 like in IRT for dichotomous responses.

- Let $\pi_i := P(Y_i > 0|\theta_1)$ and $\text{logit}(\pi_i) = \delta_i + \theta_1$ (1PL case) or $\text{logit}(\pi_i) = \delta_i + \alpha_i\theta_1$ (2PL case)

- For a count Y_i with conditional expectation μ_i assume $\ln(\mu_i) = d_i + \theta_2$ or $\ln(\mu_i) = d_i + a_i\theta_2$, and a conditional Negative Binomial distribution (need to estimate ϕ_i)

- There are two correlated person parameters, θ_1 (breadth; jump the hurdle) and θ_2 (intensity; counts beyond hurdle)!
  - $\rho > 0$: high breadth and high intensity tend to coincide
  - $\rho < 0$: low breadth and high intensity tend to coincide

Bayesian negative binomial hurdle IRT model (Doebler, Frick, Demko, Petersen, in prep.)

For person p and item i, $Y_{pi} \in \{0, 1, 2, \ldots\}$:

$$\pi_{pi} := P(Y_{pi} > 0 \mid \theta_{1p}), \qquad \operatorname{logit}(\pi_{pi}) = \delta_i + \theta_{1p}$$

$$Y_{pi} \mid Y_{pi} > 0 \sim \operatorname{NegBin}(\mu_{pi}, \phi_i), \qquad \log(\mu_{pi}) = d_i + \theta_{2p}.$$

For $y_{pi} > 0$, the NB log-likelihood is

$$\ell_{\mathrm{NB}} = \log \Gamma(y_{pi} + \phi_i) - \log \Gamma(\phi_i) - \log(y_{pi}!) + \phi_i \log \frac{\phi_i}{\phi_i + \mu_{pi}} + y_{pi} \log \frac{\mu_{pi}}{\phi_i + \mu_{pi}}.$$

The two latent traits are jointly distributed:

$$\begin{pmatrix} \theta_{1p} \\ \theta_{2p} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right].$$

Example: TIMSS 2023, United Arab Emirates, Booklet 10

- Implementation of hurdle model in brms (4 chains, 1000 draws warmup,1000 samples)
- Baseline: Bayesian NegBin IRT model (cf. Hung, 2012) without hurdle
- Hurdle model fits considerably better (Δelpd = 3464.9 (SE=98.7))

Parameter	Estimate	SE	95% CrI
Fixed effects			
Intercept_Y	−0.83	0.06	[−0.94, −0.72]
Intercept_ϕ	0.88	0.18	[0.53, 1.23]
Intercept_{hu}	1.14	0.10	[0.95, 1.33]
Person-level variation			
$\text{SD}(\theta_1)$	2.39	0.05	[2.30, 2.48]
$\text{SD}(\theta_2)$	1.03	0.03	[0.98, 1.09]
ρ_{θ_1,θ_2}	0.53	0.02	[0.57, 0.49]
Item-level variation			
$\text{SD}(d_i)$	0.21	0.04	[0.15, 0.29]
$\text{SD}(\log \phi_i)$	0.79	0.15	[0.54, 1.14]
$\text{SD}(\delta_i)$	0.42	0.06	[0.32, 0.56]
$\text{Cor}(d_i, \log \phi_i)$	−0.13	0.23	[−0.56, 0.33]
$\text{Cor}(d_i, \delta_i)$	−0.86	0.09	[−0.97, −0.64]
$\text{Cor}(\log \phi_i, \delta_i)$	0.52	0.16	[0.14, 0.79]

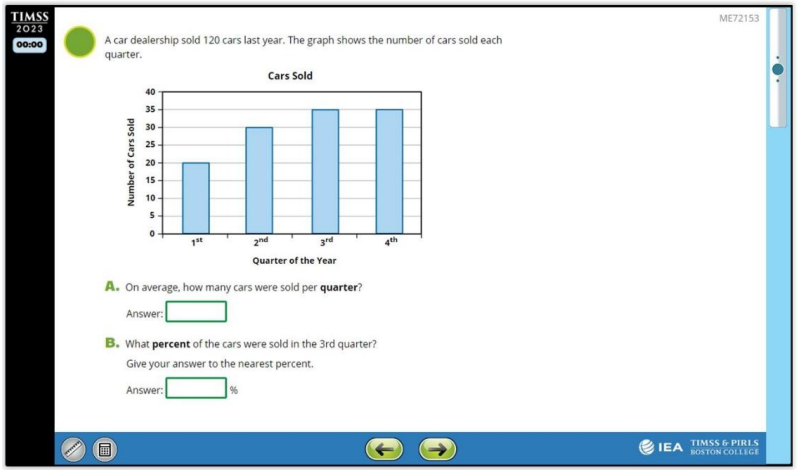
$N = 62{,}093$ observations, 2,654 persons, and 25 items.


Correct-Response Reviewing
Direct Completion Engaged
Error-Revising Engaged
Moderately Disengaged
Severely Disengaged
$\hat{\theta}_1$ breadth (hurdle)
$\hat{\theta}_2$ intensity (counts > 0)
level
Correct-Response Reviewing
Direct Completion Engaged
Error-Revising Engaged
Moderately Disengaged
Severely Disengaged
probs
99%
95%
80%
50%

Limitations of the Five Level Taxonomy and the IRT-based Validation

- Interpretation of levels: Are there three useful engaged categories?
- Construction of levels: Threshold for „correct response reviewing“ debatable; potentially further features

- Conditional independence in count data models is questionable given sequential presentations (with navigation arrows)



Source: von Davier & Mullis (2022, October). TIMSS 2023: Progress Towards a Fully Digital Assessment: Project Update. IEA General Assembly, Split. https://www.iea.nl/sites/default/files/2022-10/GA63_TIMSS%202023.pdf

Summary and Outlook

- Binary measures of Rapid Responding (or Rapid Guessing) like RTE tell us whether effort is insufficient.
- The Five Level Taxomony tells us how engagement is expressed.
- The Taxonomy is reflected in latent variable models.

- A measurement model including item-level RR and other process is within reach.

References

Bezirhan, U., von Davier, M., & Grabovsky, I. (2021). Modeling item revisit behavior: The hierarchical speed–accuracy–revisits model. Educational and Psychological Measurement, 81(2), 363-387.

Bulut, O., & Gorgun, G. (2023, July 4). Utilizing Response Time for Scoring the TIMSS 2019 Problem Solving and Inquiry Tasks. https://doi.org/10.31234/osf.io/zc98s

Gorgun, G., & Bulut, O. (2023). Incorporating test-taking engagement into the item selection algorithm in low-stakes computerized adaptive tests. Large-scale Assessments in Education, 11(1), 27.

Bulut, O., Liu, J. X., & Bulut, H. C. (2026). Disengagement matters: A response-time informed approach to scoring low-stakes assessments. International Journal of Assessment Tools in Education, 13(1), 123-144.

Hung, L. F. (2012). A negative binomial regression model for accuracy tests. Applied Psychological Measurement, 36(2), 88-103.

Lundgren, E., & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort in large-scale assessments. International Journal of Testing, 20(4), 271–293. https://doi.org/10.1080/15305058.2020.1738107

Papanastasiou, E. C., Michaelides, M. P., & Gkolia, A. (2026). Beyond Rapid Guessing: Evaluating Test Engagement With Successful and Unsuccessful Time Management Through Process Data in TIMSS 2019. Journal of Psychoeducational Assessment, 07342829261427119.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A multiprocess item response model for not-reached items due to time limits and quitting. Educational and Psychological Measurement, 80(3), 522-547.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. Applied Measurement in Education, 18(2), 163-183.

Wise, S. L., & Ma, L. (2012, April). Setting response time thresholds for a CAT item pool: The normative threshold method. In annual meeting of the National Council on Measurement in Education, Vancouver, Canada (pp. 163-183).

Wise, S. L., & Smith, L. F. (2011, April). A model of examinee test-taking effort [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, New Orleans, LA, United States.

Appendix

TIMSS 2023 Public Use File and Restriced Use File

- PUF sufficient for basic Five Level Taxonomy
- From RUF, additional variables:
  - Total number of actions
  - Actions at first and last attempt
  - Response changes
  - Quitting indicators