frontiers in Psychology

GENERAL COMMENTARY
published: 30 October 2018
doi: 10.3389/fpsyg.2018.01988

Commentary: On the Importance of the Speed-Ability Trade-Off When Dealing With Not Reached Items

Steffi Pohl[1]* *and Matthias von Davier*[2]

[1] Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany, [2] National Board of Medical Examiners, Philadelphia, PA, United States

Keywords: missing values, response time, not reached items, speed-ability trade-off, time limit, speed-accuracy

A Commentary on

On the Importance of the Speed-Ability Trade-Off When Dealing With Not Reached Items
by Tijmstra, J., and Bolsinova M. (2018). Front. Psychol. 9:964. doi: 10.3389/fpsyg.2018.00964

In their 2018 article, (T&B) discuss how to deal with not reached items due to low working speed in ability tests (Tijmstra and Bolsinova, 2018). An important contribution of the paper is focusing on the question of how to define the targeted ability measure. In this note, we aim to add further aspects to this discussion and to propose alternative approaches.

Edited by:
Ioannis Tsaousis,
University of Crete, Greece

Reviewed by:
Yong Luo,
National Center for Assessment in Higher Education, Saudi Arabia

***Correspondence:**
Steffi Pohl
steffi.pohl@fu-berlin.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *31 July 2018*
Accepted: *27 September 2018*
Published: *30 October 2018*

Citation:
Pohl S and von Davier M (2018) Commentary: On the Importance of the Speed-Ability Trade-Off When Dealing With Not Reached Items. Front. Psychol. 9:1988. doi: 10.3389/fpsyg.2018.01988

CHALLENGES IN ESTIMATING OPTIMAL ABILITY

Ignoring the Dimensional Structure

To show effects of too low working speed, T&B (p. 6) consider a model combining effective working speed and optimal ability

$$P_i\left(X=1|\theta^*,\tau\right)=\frac{1}{1+\exp\left(-\alpha_i\left[\theta_p^*+\gamma_p\left(\tau^*-\tau_p\right)-\beta_i\right]\right)}. \quad (1)$$

T&B assume two respondent groups: Compliers with $\tau_p=\tau^*$ and non-compliers with lower than optimal working speed, i.e., $\tau^*<\tau_p$ which implies $\gamma_p\left(\tau^*-\tau_p\right)<0$ if $\gamma_p>0$. We refer to this group as slow non-compliers (slowNCs).

For compliers (with $\tau^*=\tau_p$), the model in (1) reduces to a one-dimensional IRT model since $\gamma_p\left(\tau^*-\tau_p\right)=0$. For non-compliers, defining $\alpha_{1i}=\alpha_i$, $\alpha_{2ip}=-\gamma_p\alpha_i$ and $\beta_{ip}^*=\alpha_i\left(\gamma_p\tau^*-\beta_i\right)$, a person-specific two-dimensional IRT model depending on the speed-ability trade-off (SAbT) parameter γ_p results, i.e.,

$$P_i\left(X=1|\theta^*,\tau\right)=\frac{1}{1+\exp\left(-\left[\alpha_{1i}\theta^*+\alpha_{2ip}\tau+\beta_{ip}^*\right]\right)}. \quad (2)$$

Apart from specific experimental settings, which are rarely feasible to implement in large-scale assessments, in practice this model cannot be estimated, so T&B resort to fixing γ_p to a constant for their simulations. This specifies a regular two-dimensional IRT for simulation, and using a unidimensional model for analysis will of course result in biased ability estimates, which can be quantified as follows

$$E\left(\hat{\theta}_p\right) = \theta_p^* + \gamma_p\left(\tau^* - \tau_p\right). \quad (3)$$

Only compliers with $\tau^* = \tau_p$ or respondents with $\gamma_p = 0$ would obtain unbiased person parameter estimates from a unidimensional model. Thus, bias is not a result of how missing responses are treated, but due to ignoring the dimensional structure.

Respondents Faster Than Optimal

T&B only consider non-compliance as lower speed than optimal. However, most of the non-complying respondents show higher speed than optimal. Even respondents who manage responding to all items within the time limit will not have speed $\tau_p = \tau^*$, but $\tau_p > \tau^*$. This was noted by Kuhn and Ranger (2015) and shown in our own empirical data analyses (up to 70% of respondents without missing values finish the test some time before the time limit; Pohl, 2018; Pohl et al., under review; Ulitzsch et al., under review). Thus, a third group is needed in this discussion, which we will call faster non-compliers (fastNCs). Note that fastNCs—who will likely reach all items—will also receive biased estimates according to Equation (3). Hence, the issue of estimating optimal ability cannot solely be solved by focusing on the treatment of missing values.

EVALUATION OF MISSING DATA APPROACHES

Assumption on the Missing Data Process

When evaluating the performance of approaches for estimating optimal ability, one must consider a more realistic missing data mechanism including that (a) there is fastNC and (b) not reached items also occur due to quitting. In fact, in low stakes assessments quitting seems to be the main reason for not reached items (up to 90% of not reached items are due to quitting, see Pohl, 2018; Pohl et al., under review; Ulitzsch et al., under review). This will alter the results.

Performance of the Missing Data Treatments

T&B conclude that incorrect scoring shows the best results compared to other approaches. First, T&B's result seems somewhat surprising since the finding on the performance of incorrect scoring stands in stark contrast to other published research on this approach (Lord, 1974; De Ayala et al., 2001; Rose et al., 2010; Pohl et al., 2014) which show that incorrect scoring results in highly biased parameter estimates whenever missing values do not only occur on otherwise incorrect responses. Second, note that scoring missing values as incorrect results in a different definition of the target ability for different subgroups. For slowNCs with missing values, scoring these as incorrect results in an overcorrection for speed while aiming at estimating optimal ability. For compliers and fastNCs no corrections for speed are made, as there are no missing data, but instead effective ability is estimated.

DISCUSSION OF PROPOSED SOLUTIONS

We appreciate the solutions proposed by T&B and want to add further aspects for consideration:

Non-speeded power tests rely on respondents (a) being aware of their own SAbT function and (b) being highly motivated to optimize performance. The first assumption is unlikely to hold in many applications. The second assumption may hold in high stakes assessments, while in low stakes assessments, for which the missing data approaches have been suggested, empirical data (e.g., Cosgrove and Cartwright, 2014; Pohl et al., under review); suggest otherwise. Also note that this solution requires moving from measuring optimal ability for a given time limit and instead opt for measuring effective ability given the chosen speed.

Item-level time limits help respondents to manage time and reduce variability in chosen speed. However, note that this solution (a) cannot resolve the issue of differences in speed across respondents as there will still be fastNCs and (b) induces other problems, as for example increased item omit rates or rapid guessing.

AN ALTERNATIVE SOLUTION

One may conjecture that effective speed and effective ability more closely mirror real life behavior, which is typically the goal in large scale assessments (OECD, 2017). These may even be better predictors for later outcomes than optimal ability: In everyday situations there is no information on optimal speed but persons typically chose their speed given external time limits.

Pohl et al., under review and Ulitzsch et al., under review suggest describing performance of respondents by the *profile of all dimensions of performance*: effective ability, effective speed, and test endurance (as a measure of quitting behavior) and to use these dimensions for evaluating and comparing performance. This allows developing a richer description of differences in performance and to disentangle the different aspects involved. This also allows explaining differences in performance (e.g., Sachse et al., in preparation). If stakeholders are interested in only one score per domain, as for example for country rankings, we suggest using a constructive approach and decide either empirically (through prediction of key outcomes) or by means of a validity argument how to combine ability, speed, and test endurance by developing a composite score that reflects the combination one wants to focus on. One advantage of such an approach would be that this composite is the same for all respondents (not just for those with missing values). Note that this solution also works for omitted responses; these just need a slightly different modeling approach (Ulitzsch et al., 2018; Ulitzsch et al., under review).

AUTHOR CONTRIBUTIONS

SP wrote the first draft of the manuscript including the general outline of argumentation. MvD discussed these with SP and added further ideas. SP and MvD both revised the manuscript.

FUNDING

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the project Using response times to account for missing data in competence tests (Grant No. PO1655/3-1) as well as part of the project Analyzing relations between latent competencies and context information in the National Educational Panel Study within the Priority Programme 1646: Education as a Lifelong Process (Grant No. PO1655/2-1).

REFERENCES

Cosgrove, J., and Cartwright, F. (2014). Changes in achievement on PISA: the case of Ireland and implications for international assessment practice. *Large-Scale Assess. Educ.* 2:2. doi: 10.1186/2196-0739-2-2

De Ayala, R. J., Plake, B. S., and Impara, J. C. (2001) . The impact of omitted responses on the accuracy of ability estimation in item response theory. *J. Educ. Meas.* 38, 213–234. doi: 10.1111/j.1745-3984.2001.tb01124.x

Kuhn, J.-T., and Ranger, J. (2015). Measuring speed, ability, or motivation: a commentary on Goldhammer (2015). *Measurement* 13, 173–176. doi: 10.1080/15366367.2015.1105065

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika* 39, 247–264. doi: 10.1007/BF02291471

OECD (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, Revised Edition*. Paris: PISA, OECD Publishing.

Pohl, S. (2018). *Using Response Times to Model Missing Values in Competence Tests*. Invited talk at the Department of Methodology and Statistics. Tilburg University.

Pohl, S., Gräfe, L., and Rose, N. (2014). Dealing with omitted and not reached items in competence tests - Evaluating approaches accounting for missing responses in IRT models. *Educ. Psychol. Measur.* 74, 423–452. doi: 10.1177/0013164413504926

Rose, N., von Davier, M., and Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Res. Rep. Ser.* 2010:i–53. doi: 10.1002/j.2333-8504.2010.tb02218.x

Tijmstra, J., and Bolsinova, M. (2018). On the importance of the speed-ability trade-off when dealing with not reached items. *Front. Psychol.* 9:964. doi: 10.3389/fpsyg.2018.00964

Ulitzsch, E., Pohl, S., and von Davier, M. (2018). "Using nonresponse times to account for omitted items in competence tests," in *Presentation at the 19. Annual Meeting of the National Council on Measurement and Education (NCME)* (Washington, DC).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.