An Item Response Theory based open-source online assessment of vocabulary skills in 3 to 8-year-old children

Manuel Bohn[1], Julia Prein[1], Daniel Haun[1], & Natalia Gagarina[2]

[1] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2] Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany

Author Note

We thank Susanne Mauritz for help with the data collection.

The authors made the following contributions. Manuel Bohn: Conceptualization, Formal Analysis, Writing - Original Draft Preparation, Writing - Review & Editing; Julia Prein: Conceptualization, Software, Writing - Original Draft Preparation, Writing - Review & Editing; Daniel Haun: Conceptualization, Writing - Review & Editing; Natalia Gagarina: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Manuel Bohn, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. E-mail: manuel_bohn@eva.mpg.de

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

An Item Response Theory based open-source online assessment of vocabulary skills in 3 to 8-year-old children

Introduction

Dunn and Dunn (1965)

paucity of vocabulary measures for older children in general and in particular for high-quality ones.

Item-pool generation

The initial item pool consisted of 32 items taken from the German CLT (Haman et al., 2017; Haman, Łuniewska, & Pomiechowska, 2015) and 20 new items. The addition of new items was necessary due to ceiling effects for monolingual 5-year-olds in the previous version. New items were generated in line with the construction of the original CLT in a stepwise process. First, we compiled a list of age-of-acquisition ratings for 3,928 German words from various sources (Birchenough, Davies, & Connelly, 2017; Łuniewska et al., 2019; Schröder, Gemballa, Ruppin, & Wartenburger, 2012). From this list, we selected 20 words based on the following criteria: words should refer to concepts that could easily and unambiguously be depicted in a drawing, age-of-acquisition ratings should be spread equally between six and ten years of age, and words should have comparable complexity indices (see Haman et al., 2017). The so-selected 20 words served as additional target words in the item pool (total of 52 items). For each target word, we selected three distractors. The first distractor was unrelated to the target word but was chosen to have a comparable rated age-of-acquisition. The second distractor was semantically related to the target word (e.g. ruin – fortress; elk – mammoth). The third distractor was phonetically similar to the target, that is ... (e.g. Gazelle [eng.: gazelle] – Libelle [eng.: dragonfly]). The full list of targets and distractors can be found in the associate online repository. Finally, an artist (same as for the original

CLT items) drew pictures representing each target and distractor words. This procedure ensured that the original CLT and the newly generated items formed a homogeneous pool.

Task design and implementation

The task was programmed in `JavaScript` and `HTML` and presented as a website which could be opened in any modern web browser. In addition to participants' responses, we recorded webcam videos[1]. Both files were sent to a server after the study was finished. The task started with several instruction pages that explained to parents the task and how they should assists their child if needed. The task can be accessed via the following link: https://ccp-odc.eva.mpg.de/clt-extended/.

On each trial (see Figure 1), participants saw four pictures and heard a verbal prompt (pre-recorded by a native German speaker) asking them to select one of the pictures (prompt: "Zeige mir [target word]"; eng.: "Show me [target word]"). The verbal prompt was automatically played in the beginning of the trial but could also be replayed by clicking on a loudspeaker button. Selected pictures were marked via a blue frame. Participants moved on to the next trial by clicking on a button at the bottom of the screen. If children could not select the pictures themselves (via mouse click or tapping on the touch screen), they should point to the screen and parents should select the pointed-to picture.

The positioning of the pictures (target and distractors) was counterbalanced so that the target picture appeared equally often in each corner and no more than twice in the same corner. We generated two versions of the task with different item orders. Each order was created so that trial number and age-of-acquisition ratings were correlated with $r = .85$. We assumed that this would make later trials more difficult, but not perfectly so.

[1] Due to access rights issues, webcam recording was not possible when participants used iOS devices.

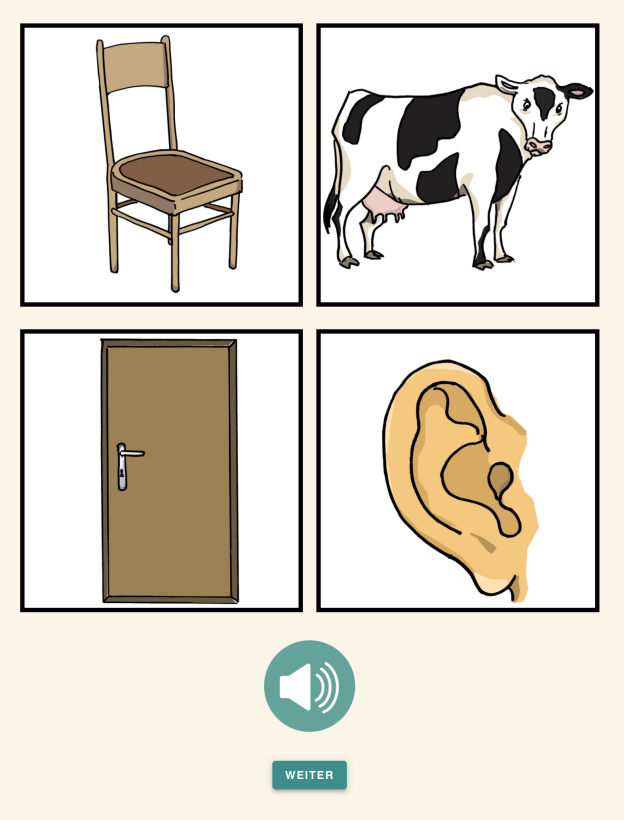

Figure 1. Screenshot from the task. On each trial, participants heard a word and were asked to pick out the corresponding picture. Verbal prompts could be replayed by pressing the loudspeaker button.

Item selection

The goal of the item selection process was to find the minimal subset of items necessary to measure vocabulary skills on an individual level. As a first step, we collected data for the full 52-item task from a large sample of children in the target age range. Next, we determined which IRT model best fit the data and used this model to estimate the item parameters (difficulty and discrimination). We removed items that showed differential item functioning (DIF) when the data was split either by sex or by trial order. Finally, we used a simulated annealing process (Kirkpatrick, Gelatt Jr, & Vecchi, 1983) to determine the size of the reduced tasl and to select the items. Data collection was pre-registered at:

https://osf.io/qzstk. The pre-registered sample size was based on recommendations found in the literature (Morizot, Ainsworth, & Reise, 2007). The datasets generated during the current study as well as the analysis code are available in the following repository: https://github.com/ccp-eva/vocab.

Participants

Participants were recruited via database of children whose parents volunteered to participate online studies on child development. Parents received and email with a short study description and a personalized link. After one week, parents received a reminder if they had not already taken part in the study. Response rate to invitations was ~50%. The final sample included a total of 581 children (n = 307 girls) with a mean age of 5.63 (range: 3.01 – 7.99). Participants were randomly assigned to one of the two task versions. Data was collected between February and May 2022.

Descriptive results

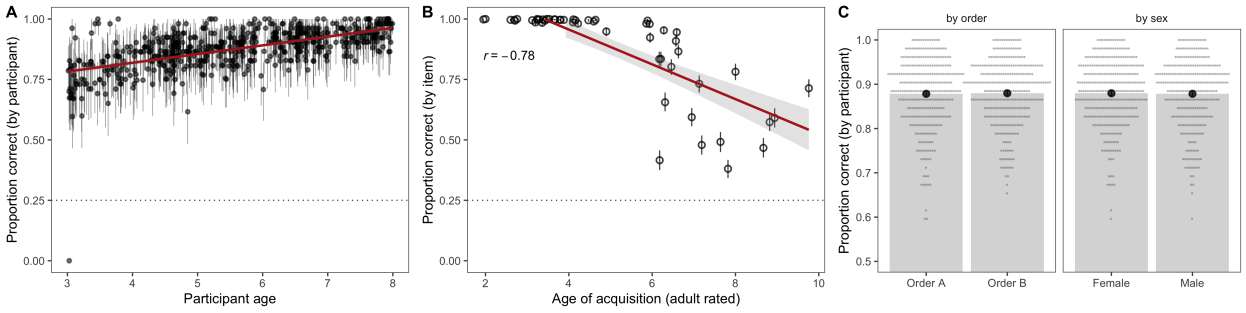

Figure 2. Descriptive results of the task. A: Proportion of correct responses (with 95% CI) for each participant by age. B: Proportion of correct responses (with 95% CI) for each item by rated age-of-acquisition of the target word. C: Proportion of correct responses (with 95% CI) by trial order (left) and sex (right).

On a participant level, performance in the full task (52 items) steadily increased with age (Figure 2A). On an item level, performance was above chance (25%) for all items.

Table 1

Model comparison

Model	ELPD	SE(ELPD)	ΔELPD	SE(ΔELPD)
3PL	-6,089.51	80.89	0.00	0.00
2PL	-6,124.12	81.01	-34.61	8.60
1PL (Rasch)	-6,233.70	82.13	-144.19	18.20

Note. ELPD = expected log posterior density, SE = standard error, ELPD differences are in comparison to the 3PL model. Higher ELPD values indicate better model fit.

Furthermore, the average proportion of correct responses was negatively correlated with age-of-acquisition ratings (Figure 2B). These descriptive results replicate well-known results in the literature and thereby validate the overall approach. Figure 2C shows that there were – on average – no differences between participants who received order A and order B as well as between female and male participants. This result suggests that these grouping variables are suitable to investigate differential item functioning (see below).

Item response modelling

IRT models were implemented in a Bayesian framework in `R` using the `brms` package (Bürkner, 2017, 2019). Given the binary outcome of the data, we used logistic models to predict the probability of a correct answer based on participant's latent ability and item characteristics (difficulty and discrimination). All models had converging chains and provided a good fit to the data. For details about prior and MCMC settings, please see the analysis script in the associated online repository. We compared models using Bayesian approximate leave-one-out cross-validation (Vehtari, Gelman, & Gabry, 2017) based on differences in expected log posterior density (ELPD) and the associated standard error (SE).

Table 2

Model comparison to assess DIF

Model	ELPD	SE(ELPD)	ΔELPD	SE(ΔELPD)
3PL split by sex	-6,065.58	80.94	0.00	0.00
3PL	-6,089.51	80.89	-23.93	8.38
3PL split by order	-6,090.34	80.75	-24.75	9.27

Note. ELPD = expected log posterior density, SE = standard error, ELPD differences are in comparison to the 3PL model. Higher ELPD values indicate better model fit.

As a first step, we compared three models with increasing complexity: a 1PL (Rasch) model which assumed that items differ in difficulty but have the same discrimination parameter (1), a 2PL model which addiotionally allowed items to have different discrimination parameters, and a 3Pl model which further added a guessing parameter of 0.25. Table 1 shows that the 3PL model provided – by far – the best fit. For the following item selection procedure, we therefore used the item parameters (difficulty and discrimination) estimated in the 3PL model.

Differential item functioning

As a first step in the item selection process, we removed items that showed differential item functioning (DIF). DIF refers to situations in which items show differential characteristics for subgroups that have otherwise the same overall score (Holland & Wainer, 2012). To assess DIF for the present task, we followed the procedure suggested by Bürkner (2019) and fit two extended 3PL models (one for trial order and one for sex) which estimated separate item characteristics for each subgroup. As an overall assessment of DIF we compared these extended models to the basic 3PL. We found no indication for DIF based on

trial order but did so for sex (see Table 2). To decide which items to remove, we computed the difference between mean estimates for male and female participants for each item and excluded those items for which the absolute difference was larger than two standard deviations of all differences. Four items had to be excluded based on this procedure (see Figure 3).

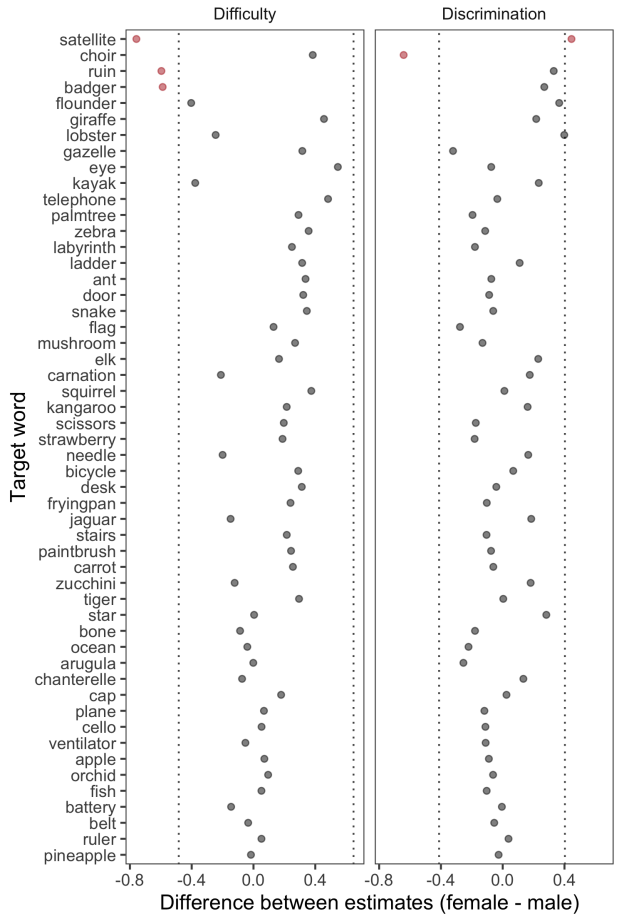

Figure 3. Differential item functioning. Difference between estimates for female and male participants for the two item parameters. Dashed lines show cut-off points. Red points indicate items that were excluded.

Simulated annealing

The goal of this last step of the item selection process was to select a smaller subset of items that nevertheless allow for precise measurement. The basis for this selection process was a score which we defined to capture three important characteristics that the items of any subset should have. First, items should be equally spaced across the latent ability space. This characteristic ensures that the task is suited for different ability levels and thus for a broader range of ages. We quantified the spread of any given subset as the standard deviation of the distance (in difficulty estimates) between adjacent items. Lower values indicate smaller distances and thus an overall more equal spacing. Second, items should have maximum discrimination. That is, we preferred items that distinguished well between narrowly defined regions of the latent ability. Discrimination parameters were divided by 2 to put them on a scale comparable to the standard deviations of the distances. Third, difficulty estimates should have narrow credible intervals. The idea behind this characteristic was that many of the easier items had very wide credible intervals because most of the participants answered correctly. Of those items we sought to select the ones with more precise difficulty estimates. For scaling purposes, difficulty estimates were divided by 6.

We used simulated annealing (Kirkpatrick et al., 1983) to find the optimal items for any given size of the subset. This process ...

We applied simulated annealing to subsets ranging from 5 to 40 items. For each (optimal) subset we then computed the correlation between performance based on the subset and based on the full task This allowed us to assess how well the subset was able to capture variation between individuals in comparison to the full task. Figure 4A shows how the correlation between subset and full task increase with an increasing number of items in the subset. The resulting curve leveled-off at around 20 items in that adding additional items to the subset did not incerease the correlation any further. We therefore concluded that 20 items would be the ideal size of the subset.

When running the simulated annealing procedure for 20 items 100 times, it always returned the same item selection. We therefore chose this subset of items for the reduced task. Figure 4B shows the item parameters for the selected items and Figure 4C shows their item characteristic curves.

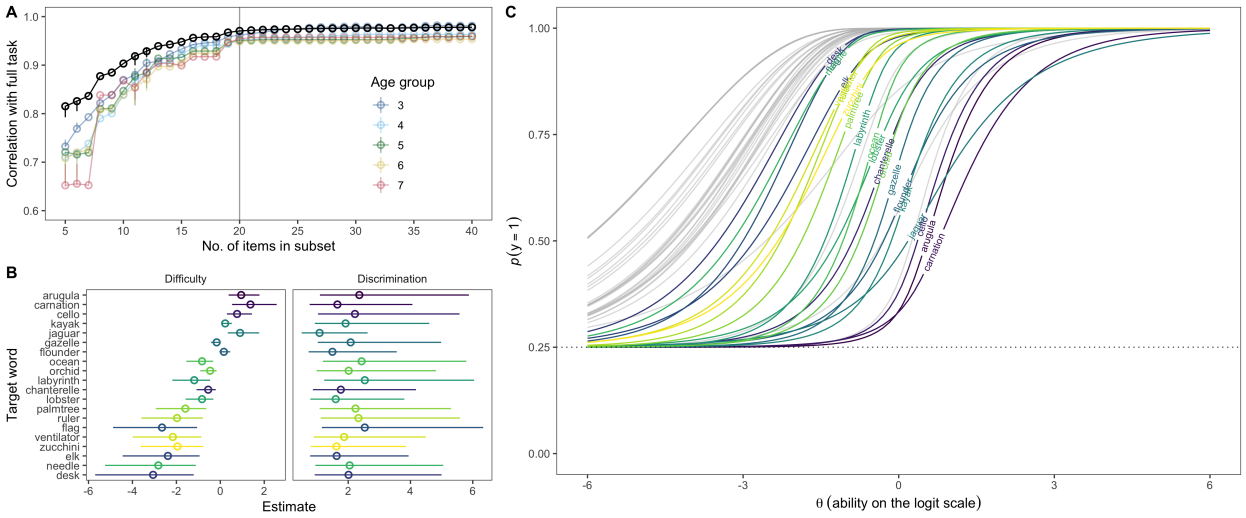

Figure 4. Item selection process. A) Correlation between reduced and full task (52 items). Points show mean correlation based on 100 iterations. Vertical lines show the range of correlations in cases when they differed between iterations. Black lines and points show correlations for the full sample and colored points and lines show correlations by age group. B) Item parameters for the selected 20 items estimated based on the 3PL model. C) Item characteristic curves for all 52 items, with excluded items in grey and selected items in color.

Discussion

References

Birchenough, J. M., Davies, R., & Connelly, V. (2017). Rated age-of-acquisition norms for over 3,200 german words. *Behavior Research Methods*, *49*(2), 484–501.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bürkner, P.-C. (2019). Bayesian item response modeling in r with brms and stan. *arXiv Preprint arXiv:1905.09501*.

Dunn, L. M., & Dunn, L. M. (1965). *Peabody picture vocabulary test*.

Haman, E., Łuniewska, M., Hansen, P., Simonsen, H. G., Chiat, S., Bjekić, J., … others. (2017). Noun and verb knowledge in monolingual preschool children across 17 languages: Data from cross-linguistic lexical tasks (LITMUS-CLT). *Clinical Linguistics & Phonetics*, *31*(11-12), 818–843.

Haman, E., Łuniewska, M., & Pomiechowska, B. (2015). Designing cross-linguistic lexical tasks (CLTs) for bilingual preschool children. *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*, 196–240.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

Kirkpatrick, S., Gelatt Jr, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Łuniewska, M., Wodniecka, Z., Miller, C. A., Smolik, F., Butcher, M., Chondrogianni, V., … others. (2019). Age of acquisition of 299 words in seven languages: American english, czech, gaelic, lebanese arabic, malay, persian and western armenian. *PloS One*, *14*(8), e0220611.

Morizot, J., Ainsworth, A., & Reise, S. (2007). *Toward modern psychometrics: Application of item response theory models in personality research in robins RW, fraley RC, & krueger RF (eds.), Handbook of research methods in personality psychology (pp. 407–421)*. New York, NY: Guildford Press.[Google Scholar].

Schröder, A., Gemballa, T., Ruppin, S., & Wartenburger, I. (2012). German norms for

197 semantic typicality, age of acquisition, and concept familiarity. *Behavior Research*
198 *Methods*, *44*(2), 380–394.
199 Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using
200 leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.