

The title

Karolina Muszyńska¹ & XXX²

¹ University of Warsaw, Faculty of Psychology

² Stanford University

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. Karolina Muszyńska:
Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing;
XXX: Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to Karolina Muszyńska,
University of Warsaw, Faculty of Psychology, ul. Stefana Banacha 2D, 02-097 Warsaw,
Poland. E-mail: karolina.muszynska@psych.uw.edu.pl

Abstract

15

16 One or two sentences providing a **basic introduction** to the field, comprehensible to a
17 scientist in any discipline. Two to three sentences of **more detailed background**,
18 comprehensible to scientists in related disciplines. One sentence clearly stating the **general**
19 **problem** being addressed by this particular study. One sentence summarizing the main
20 result (with the words “**here we show**” or their equivalent). Two or three sentences
21 explaining what the **main result** reveals in direct comparison to what was thought to be
22 the case previously, or how the main result adds to previous knowledge. One or two
23 sentences to put the results into a more **general context**. Two or three sentences to
24 provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

25

Keywords: keywords

26

Word count: X

The title

Methods

(to paste from google doc)

Participants

(to paste from google doc) Mention the bilinguals and multilinguals

Material

(to paste from google doc)

Procedure

(to paste from google doc)

Data analysis

Results

Psychometric properties of the two CAT-CDIs

Our first aim was to examine whether CAT-CDIs in American English and Polish demonstrate comparable psychometric properties. To that end, we revisit the psychometric properties reported for the American English CAT-CDI (word production) in Kachergis et al. (2022) and compare those to the data from Polish CAT-CDI (Words and Sentences).

We found similarly strong correlations in the two languages between the abilities estimated from CDI-CAT and full CDI scores (American English and Polish: $r = .86$), the abilities estimated from the CDI-CAT and abilities estimated from full CDI (American

English and Polish: $r = .92$), and the abilities estimated from the full CDI and the full
 CDI scores (American English: $r = .95$, Polish: $r = 0.94$). The abilities estimated from the
 CDI-CAT and the full CDI scores were also strongly correlated within individual age
 groups (see Table 2).

Table 1

*American English: Correlations between ability estimated by CAT-CDI and ability
 estimated from full CDI by children's age*

	[15,18)	[18,21)	[21,24)	[24,27)	[27,30)	[30,33)	[33,36]
r ability CAT vs full CDI	0.95	0.85	0.82	0.83	0.59	0.84	0.86
N	26	22	26	30	28	24	48

Table 2

*Polish: Correlations between ability estimated by CAT-CDI and ability estimated
 from full CDI by children's age*

	[18,21)	[21,24)	[24,27)	[27,30)	[30,33)	[33,36]
r ability CAT vs full CDI	0.8	0.94	0.91	0.89	0.95	NA
N	29	22	16	23	22	1

The Polish validation study included 28 data from bi- and multilingual families.
 Though it is a small group, we decided to explore their correlation coefficients
 (non-parametric Spearman's rho) and found these were similar to those found for Polish
 monolingual children (see Table 3 in Supplementary Materials).

Table 3

Supplementary Material: Table S1 - Spearman’s correlations for monolingual and multilingual children in the Polish dataset

lang_group	r	n	correlation
monolingual	0.92	85	Ability from CDI-CAT ~ full CDI score
multilingual	0.90	28	Ability from CDI-CAT ~ full CDI score
monolingual	0.92	85	Ability from CDI-CAT ~ ability from full CDI
multilingual	0.90	28	Ability from CDI-CAT ~ ability from full CDI
monolingual	1.00	85	Ability from full CDI ~ full CDI score
multilingual	1.00	28	Ability from full CDI ~ full CDI score

We also looked at the mean squared error between the abilities as estimated by CAT-CDI and from the full CDI. The mean squared error in English was 0.55 ($Mdn = 0.17$, $SD = 1$), and in Polish it was 0.19 ($Mdn = 0.08$, $SD = 0.45$). We also looked at the children for whom the estimates from the CAT-CDI and full CDI diverged extremely, i.e. their difference between the errors was 1.5 SD from the mean. There were 15 such cases (7.35%) in the English dataset and 4 cases (1.96%) in the Polish dataset. All participants in both datasets showed higher ability estimates on the CDI-CAT compared to the full CDI. If the full CDI is considered the baseline, this suggests that parents may have overestimated their child’s vocabulary on the CDI-CAT, potentially responding “yes – produces” to more items than expected based on full CDI estimates (as suggested by Kachergis, et al. 2022). An alternative explanation is that, for these participants, the full CDI may have underestimated the child’s true ability. Notably, all Polish participants with large discrepancies completed the full CDI in unusually short times (their completion times were among the shortest 5% in the sample) suggesting their responses may have been rushed or less attentive. This could have led to lower ability estimates from the full CDI.

69 Supporting this interpretation, their CDI-CAT scores had acceptable measurement errors
70 (below or equal to 0.1 for Polish), indicating reliable ability estimation by the CDI-CAT, in
71 contrast to the full CDI. However, this pattern did not appear in the English dataset,
72 where only 2 participants who showed extreme discrepancy also showed very short
73 administrations of the full CDI.

Table 4

production	sex_full	age_full	order	fullTheta	fullTheta_SE	catTheta	catTheta_SE	sq_err	full_cat_diff	extreme_discre
97.00	Female	27.00	full_first	-0.14	0.04	1.20	0.17	1.81	-1.34	yes
8.00	Male	17.00	cat_first	-1.58	0.16	-0.23	0.16	1.82	-1.35	yes
158.00	Male	35.00	full_first	0.14	0.04	1.62	0.16	2.17	-1.47	yes
0.00	Male	34.00	full_first	-2.90	0.43	-1.48	0.38	2.01	-1.42	yes
132.00	Female	21.00	cat_first	0.02	0.04	1.75	0.17	2.99	-1.73	yes
165.00	Male	20.00	full_first	0.18	0.04	1.48	0.17	1.71	-1.31	yes
47.00	Female	28.00	full_first	-0.57	0.06	1.86	0.17	5.90	-2.43	yes
14.00	Male	20.00	full_first	-1.27	0.12	0.01	0.16	1.64	-1.28	yes
124.00	Female	28.00	cat_first	0.00	0.04	1.30	0.17	1.68	-1.30	yes
210.00	Female	26.00	cat_first	0.33	0.03	1.85	0.17	2.31	-1.52	yes
5.00	Female	26.00	cat_first	-1.79	0.19	-0.42	0.19	1.87	-1.37	yes
177.00	Male	28.00	full_first	0.22	0.03	1.62	0.18	1.98	-1.41	yes
470.00	Male	36.00	full_first	1.14	0.03	2.82	0.35	2.83	-1.68	yes
253.00	Male	35.00	cat_first	0.48	0.03	1.83	0.16	1.83	-1.35	yes
287.00	Male	23.00	full_first	0.58	0.03	1.91	0.16	1.78	-1.33	yes

We also re-calculated the mean squared error without the cases of extreme discrepancy, which yielded a MSE of 0.44 ($Mdn = 0.29$, $SD = 0.44$) in English and MSE of 0.12 ($Mdn = 0.07$, $SD = 0.14$) in Polish.

Item properties in the two CAT-CDIs

Our second aim was to analyze similarities and differences in IRT item properties and item selection in CAT in English and Polish.

There are 679 items in the English CAT-CDI and 666 items in the Polish CAT-CDI. For both sets of items, the items' difficulty and discrimination parameters were calculated using IRT 2 parameter model (these included separate samples, see Kachergis et al. 2022 and Krajewski et al. (in preparation)). An item's difficulty indicates the ability level at which there is a 50% probability that a participant will respond correctly. It includes negative and positive values, with 0 indicating medium difficulty, thus reflecting the range of values of ability, i.e. theta. An item's discrimination indicates how well it distinguishes between individuals with slightly different ability levels—especially those near that difficulty point. Of these two parameters, item difficulty is of greater interest to the present paper as it is directly linked to ability and as discrimination power is more about how good the item is at measuring, rather than what it is measuring.

English items are more difficult than the Polish items, $\Delta M = -1.87$, 95% CI $[-2.06, -1.69]$, $t(1227.61) = -20.09$, $p < .001$ (English: min = -7.16, max = 4.45, $M = -2.19$, $Mdn = -2.21$, $SD = 1.98$; Polish: min = -4.34, max = 4.41, $M = -0.32$, $Mdn = -0.43$, $SD = 1.41$). Notably, this was true even for a subset of 390 items common to both languages - these items still proved to be more difficult in English than in Polish: $M_D = -1.68$, 95% CI $[-1.82, -1.54]$, $t(389) = -23.11$, $p < .001$. This difference in mean difficulty may be influenced by the characteristics of the samples used to estimate the IRT models. In English, item difficulty was calculated based on a broader sample of children

aged 12–36 months (spanning the CDI:WG, CDI:WS, and CDI-III), whereas the Polish data came from a sample of narrower age range of 18–36 months, corresponding to the CDI:WS. As a result, item difficulty in English was estimated using a relatively younger sample, for whom certain items may have been more challenging—thus appearing more difficult—compared to the older Polish sample. Still the item difficulty for common items in the two languages was positively and moderately correlated: $r = .65$, 95% CI $[.58, .70]$, $t(388) = 16.68$, $p < .001$.

Table 5

category_pl	rho	n
quantifiers	0.11	10
clothing	0.22	19
connecting_words	0.23	8
locations	0.31	6
pronouns_demonstrative	0.32	4
places	0.32	7
vehicles	0.45	9
furniture_rooms	0.47	17
action_words	0.50	75
body_parts	0.52	17
prepositions	0.55	10
games_routines	0.57	12
time_words	0.57	8
household	0.57	30
outside	0.58	20
sounds	0.60	6
food_drink	0.61	39
descriptive_words (adjectives)	0.62	19
people	0.67	16
animals	0.68	29
toys	0.74	15
descriptive_words (adverbs)	0.80	4
question_words	0.91	10



Figure 1. The relative positioning of the items by semantic category (colored) plotted in the context of the full item pool (grey): (A) English, (B) Polish.

We also wanted to do check whether items in particular CDI semantic categories show related parameter values across the two languages. We performed a series of Spearman’s rank correlations (on the items common to CDI-CATs in Polish and English) for each CDI semantic category (see Table @ref(tab:d_by_cdi_category_tab)). The rank correlations coefficients vary by category, but for half of the categories the correlations are moderate to strong (0.52 to 0.91). Figure 1 shows the relative positioning of the items by semantic category (colored) plotted in the context of the full item pool (grey) in the two languages. It can be seen from the figure that many categories (e.g., sounds) show a similar distribution of items relative to the whole item pool across Polish and English.

Item selection in the two CAT-CDIs

It is to be expected that a CDI-CAT will not need to administer all the items in the item bank. In fact, in the validation study the English CAT-CDI used 251 items (36.91%) and similarly, the Polish CAT-CDI used 258 items (38.74%). By design, a CDI-CAT selects

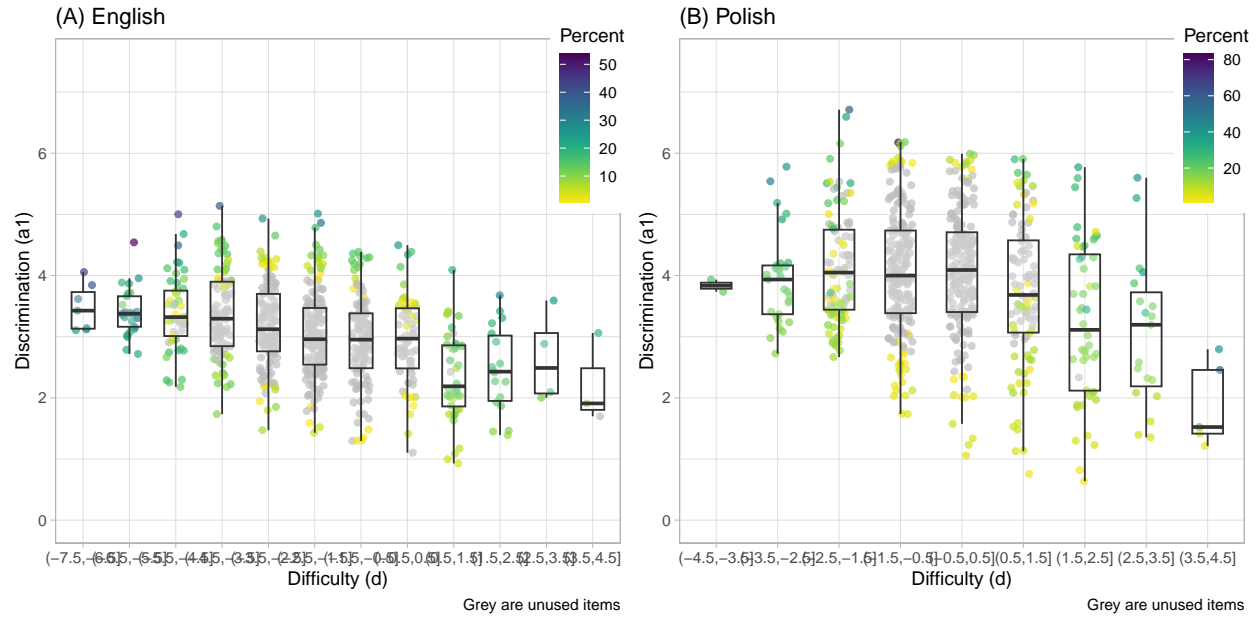


Figure 2. How often a given item was used in CAT-CDI administrations in the validation study in (A) English and (B) Polish. The items (points) are colored by the percentage of their appearance in the CAT-CDI administrations. Items colored in grey are items never used in any of the CAT-CDI administrations.

items that are most informative for each participant. This means it draws from a subset of available items—typically those matched to the participant’s current ability estimate in terms of difficulty, and with high discrimination, meaning they effectively distinguish between individuals with abilities close to that estimate. Figure 2 plots the items by their difficulty and discrimination parameters and colors the items (points) by how often they were used in CAT-CDI administrations in English and Polish. A few findings are of note here. First, both CAT-CDIs used items of very low discrimination (i.e, items that do not discriminate well between ability levels). Often that was because for a given difficulty level there were not enough items and CAT-CDI had to use all the items available (this is particularly true of the items on the two ends of difficulty). However, more surprisingly, it is also true of items of medium difficulty, where items of higher discrimination were available (but were not chosen by the CAT algorithm). [NOTE TO US: I emailed Phil Chalmers,

the author of *mirtcat* and *mirt* packages in R to ask if he has any idea why this is so].

Finally, some items keep being shown in many CAT administrations. In English, 3 items appear in more than half of administrations - “*long*” in 64% of administrations, “*make*” in 54% of administrations, and “*last*” in 53% of administrations. They are also high discrimination items in their difficulty bins. “Long” is additionally one of the starting items, i.e. a first item shown to the parent, chosen so that there is high probability the parent can respond positively. In Polish, 3 items appear in more than half of administrations - “*szukać*” (to look for) appearing in 83% of administrations, “*znaleźć*” (to find) appearing in 61% of administrations, and “*babcia*” (grandmother) appearing in 60% of administrations. All three items show very high discrimination. “Szukać” (to look for) and “znaleźć” (to find) are also in top 5 discrimination items in general. “Babcia” (grandmother) is one of the starting items.

To check whether the CDI items common in both languages may show similar frequencies in CAT-CDI administrations, we ran Spearman’s rank correlation and found that the items frequencies in administrations were weakly correlated at $\rho = r_s = .37$, $S = 6,420,909.29$, $p < .001$.

CAT-CDIs’ usefulness in research and practice

Discussion

1. Correlations strong in both languages (overall and in age-bins; in PL: multi and mono correlations similar and strong).
2. MSE after removing cases with extreme discrepancies.
3. Extreme discrepancies - mixed results?

References

- Auguie, B. (2017). *gridExtra: Miscellaneous functions for "grid" graphics*. Retrieved from <https://CRAN.R-project.org/package=gridExtra>
- Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R Markdown*. <https://doi.org/10.32614/CRAN.package.papaja>
- Barth, M. (2023). *tinylabels: Lightweight variable labels*. Retrieved from <https://cran.r-project.org/package=tinylabels>
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, 71(5), 1–39. <https://doi.org/10.18637/jss.v071.i05>
- Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., . . . Borges, B. (2024). *Shiny: Web application framework for r*. Retrieved from <https://CRAN.R-project.org/package=shiny>
- Garnier, Simon, Ross, Noam, Rudis, Robert, . . . Cédric. (2023). *viridis(Lite) - colorblind-friendly color maps for r*. <https://doi.org/10.5281/zenodo.4678327>
- Garnier, Simon, Ross, Noam, Rudis, Robert, . . . Cédric. (2024). *viridis(Lite) - colorblind-friendly color maps for r*. <https://doi.org/10.5281/zenodo.4679423>
- Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3), 1–25. Retrieved from <https://www.jstatsoft.org/v40/i03/>
- Kassambara, A. (2023). *Ggpubr: 'ggplot2' based publication ready plots*. Retrieved from <https://CRAN.R-project.org/package=ggpubr>
- Krajewski, G. (2025). *Multilada: MultiLADA's little helpers*. Retrieved from <https://github.com/gkrajewski/Multilada>

- Müller, K. (2020). *Here: A simpler way to find your files*. Retrieved from <https://CRAN.R-project.org/package=here>
- Müller, K., & Wickham, H. (2025). *Tibble: Simple data frames*. Retrieved from <https://CRAN.R-project.org/package=tibble>
- R Core Team. (2025). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. New York: Springer. Retrieved from <http://lmdvr.r-forge.r-project.org>
- Slowikowski, K. (2024). *Ggrepel: Automatically position non-overlapping text labels with 'ggplot2'*. Retrieved from <https://CRAN.R-project.org/package=ggrepel>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <https://ggplot2.tidyverse.org>
- Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from <https://CRAN.R-project.org/package=forcats>
- Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from <https://CRAN.R-project.org/package=stringr>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. Retrieved from <https://CRAN.R-project.org/package=dplyr>
- Wickham, H., & Henry, L. (2025). *Purrr: Functional programming tools*. Retrieved from <https://CRAN.R-project.org/package=purrr>
- Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data*. Retrieved from <https://CRAN.R-project.org/package=readr>

Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data*. Retrieved from
<https://CRAN.R-project.org/package=tidyr>

R (Version 4.4.3; R Core Team, 2025) and the R-packages *dplyr* (Version 1.1.4; Wickham, François, Henry, Müller, & Vaughan, 2023), *forcats* (Version 1.0.0; Wickham, 2023a), *ggplot2* (Version 3.5.2; Wickham, 2016), *ggpubr* (Version 0.6.0; Kassambara, 2023), *ggrepel* (Version 0.9.6; Slowikowski, 2024), *gridExtra* (Version 2.3; Auguie, 2017), *here* (Version 1.0.1; Müller, 2020), *lattice* (Version 0.22.7; Sarkar, 2008), *lubridate* (Version 1.9.4; Grolemund & Wickham, 2011), *mirt* (Version 1.44.0; Chalmers, 2012, 2016), *mirtCAT* (Version 1.14; Chalmers, 2016), *Multilada* (Version 0.7.0; Krajewski, 2025), *papaja* (Version 0.1.3; Aust & Barth, 2024), *purrr* (Version 1.0.4; Wickham & Henry, 2025), *readr* (Version 2.1.5; Wickham, Hester, & Bryan, 2024), *shiny* (Version 1.10.0; Chang et al., 2024), *stringr* (Version 1.5.1; Wickham, 2023b), *tibble* (Version 3.3.0; Müller & Wickham, 2025), *tidyr* (Version 1.3.1; Wickham, Vaughan, & Girlich, 2024), *tidyverse* (Version 2.0.0; Wickham et al., 2019), *tinylab* (Version 0.2.5; Barth, 2023), *viridis* (Garnier et al., 2023; Version 0.6.5; Garnier et al., 2024) and *viridisLite* (Version 0.4.2; Garnier et al., 2023)