1   The title

2   Karolina Muszyńska[1] & To fill in later on[2]

3   [1] University of Warsaw, Faculty of Psychology

4   [2] To fill in later on

5   Author Note

12   Abstract

13   Abstract here

14   *Keywords:* keywords

15   Word count: X

16   The title

17   **Methods**

18 **Participants**

19   In the American English validation study, the participants were 204 parents of

20 children aged 15 to 36 months ($M = 26.21$, $Mdn = 26$, SD = 6.37). There were 103 girls in

21 the sample (50.49%). The sample was ethnically diverse, including White, Black,

22 Latino/Hispanic, Asian, Native American, and other races/ethnicities. None of the children

23 were reported to hear other languages apart from English. The average number of years of

24 education attained by the primary caregiver was 15.82 (SD = 2.27).

25   In the Polish validation study, the participants were 113 parents of children aged 18

26 to 34 months ($M = 24.73$, $Mdn = 24$, SD = 4.47). There were 49 girls in the sample

27 (43.36%). The sample was White (reflecting the limited ethnic diversity within the Polish

28 population). However, there were 28 children multilingual with Polish (as their home

29 language) and some other (majority) language. [todo: add info on parental edu]

30 **Materials**

31   The data was gathered with CAT-CDIs in American English and Polish. The

32 American English CAT-CDI was created from the items that were common to three CDIs:

33 Words and Gestures, Words and Sentences and CDI-III. With these items, two CAT-CDIs

34 were created, one for word production and another for word comprehension (see Kachergis

35 et al. 2022 for more details). The Polish CAT-CDIs were created following a largely similar

36 procedure as the ones in American English, but there was a separate CDI-CAT developed

37 for each CDI version (CDI: Words and Gestures and CDI: Words and Sentences) and for

38 each CDI:WG subscale (word production, word comprehension, gesture use) (Krajewski et

39 al. in preparation). Here we show data from the validation study of American English

CAT-CDI word production (i.e. 679 items) and Polish CAT-CDI version of WS: Words and Sentences (i.e. 666 items). There are 395 words that appear in both CDI-CAT language versions.

Procedure

Though in both validation studies the parents were asked to fill in both the full CDI and the CAT-CDI, the procedure differed between the two validation studies. In the validation of American English CDI-CAT, parents were given both CDIs in one sitting, but the order of the CDI versions was counterbalanced. In the Polish validation study, the order of the CDIs was fixed, with the full CDI always first, and then, after 1–30 days ($M =$ 8.31, $Mdn = 8$, SD = 4.90) parents were asked to fill in the CAT-CDI. Both studies were done fully online and obtained the approval of adequate ethics committees.

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

64 The Polish validation study included 28 data from bi- and multilingual families.

65 Though it is a small group, we decided to explore their correlation coefficients

66 (non-parametric Spearman's rho) and found these were similar to those found for Polish

67 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 then zoomed in on 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

and larger discrepancies between estimates from full and CAT CDI. Supporting this

interpretation, their CDI-CAT scores had acceptable measurement errors (below or equal

to 0.1 for Polish), indicating reliable ability estimation by the CDI-CAT, in contrast to the

full CDI. This was partially true for the American sample - though only 2 participants had

similarly short full CDIs, all but two showed relatively low measurement errors (more on

measurement error in Section XX).

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 is on the

same scale as ability with negative values indicating difficulty items, values around 0

indicating medium difficulty, and positive values indicating easy items. 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 4

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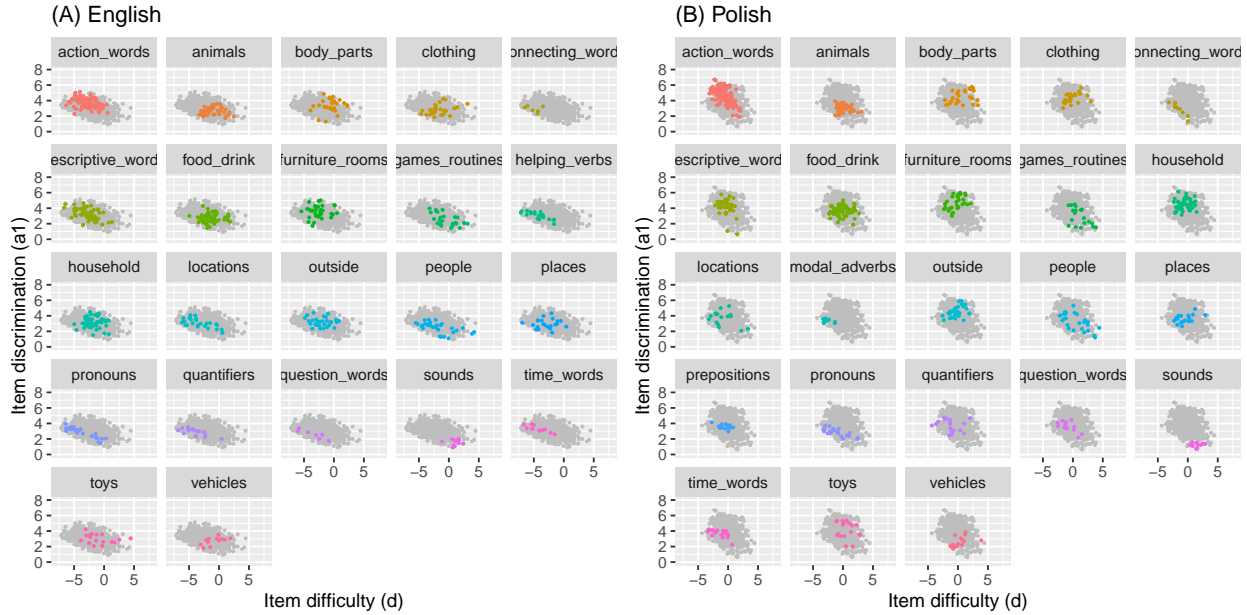
Figure 1. The relative positioning of the items by semantic category (colored) plotted in the context of the full item pool (grey): (A) English, (B) Polish.

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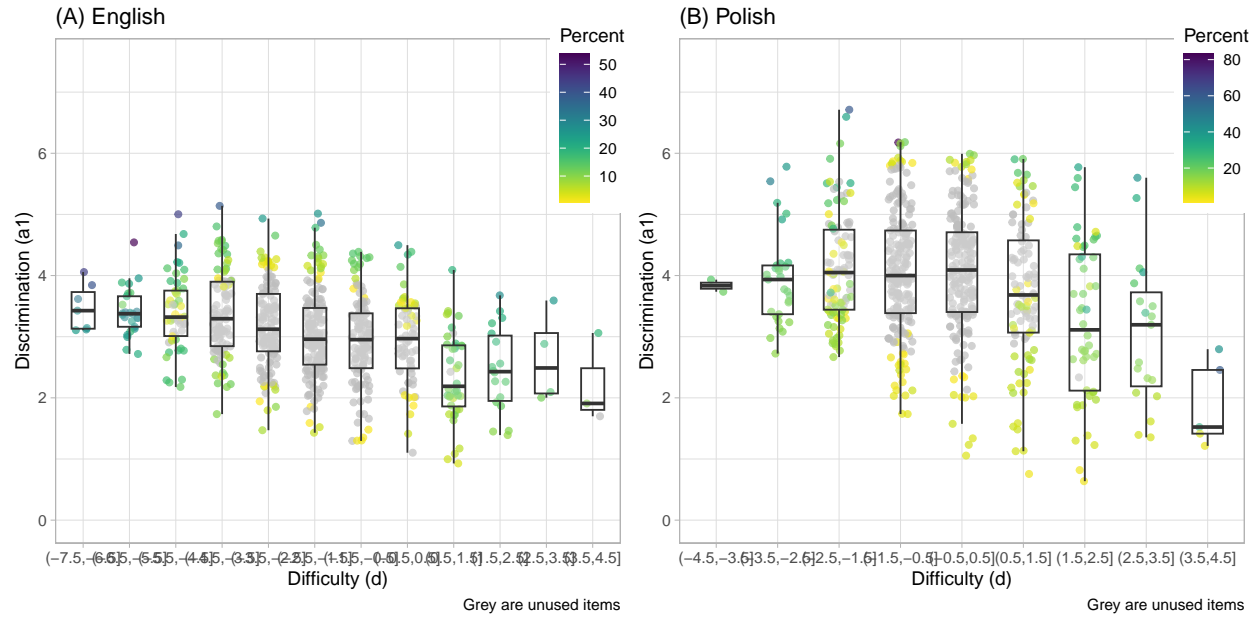
Figure 2. How often a given item was used in CAT-CDI administrations in the validation study in (A) English and (B) Polish. The items (points) are colored by the percentage of their appearance in the CAT-CDI administrations. Items colored in grey are items never used in any of the CAT-CDI administrations.

135 items that are most informative for each participant. This means it draws from a subset of

136 available items–typically those matched to the participant's current ability estimate in

137 terms of difficulty, and with high discrimination, meaning they effectively distinguish

138 between individuals with abilities close to that estimate. Figure 2 plots the items by their

139 difficulty and discrimination parameters and colors the items (points) by how often they

140 were used in CAT-CDI administrations in English and Polish. A few findings are of note

141 here. First, both CAT-CDIs used items of very low discrimination (i.e, items that do no

142 discriminate well between ability levels). Often that was because for a given difficulty level

143 there were not enough items and CAT-CDI had to use all the items available (this is

144 particularly true of the items on the two ends of difficulty). However, more surprisingly, it

145 is also true of items of medium difficulty, where items of higher discrimination were

146 available (but were not chosen by the CAT algorithm).

147   Finally, some items keep being shown in many CAT administrations. In English, 3

148 items appear in more than half of administrations - *"long"* in 64% of administrations,

149 *"make"* in 54% of administrations, and *"last"* in 53% of administrations. They are also

150 high discrimination items in their difficulty bins. "Long" is additionally one of the starting

151 items, i.e. a first item shown to the parent, chosen so that there is high probability the

152 parent can respond positively. In Polish, 3 items appear in more than half of

153 administrations - *"szukać"* (to look for) appearing in 83% of administrations, *"znaleźć"* (to

154 find) appearing in 61% of administrations, and *"babcia"* (grandmother) appearing in 60%

155 of administrations. All three items show very high discrimination. "Szukać" (to look for)

156 and "znaleźć" (to find) are also in top 5 discrimination items in general. "Babcia"

157 (grandmother) is one of the starting items.

158   To check whether the CDI items common in both languages may show similar

159 frequencies in CAT-CDI administrations, we ran Spearman's rank correlation and found

160 that the items frequencies in administrations were weakly correlated, $r_s = .37$,

161 $S = 6,420,909.29$, $p < .001$. However, the two CAT-CDIs differed in their potential length,

162 as the English CAT-CDI was set to administer 25–50 items and the Polish CAT-CDI could

163 administer up to 75 items (no minimum was set, but in practice the minimum items

164 administered were 13). As an item is more likely to appear in longer tests (simply due to

165 test length), then frequency of the Polish items could be inflated.

CAT-CDIs' usefullness in research and practice

167   One of the key metrics for CAT-CDI's usefulness is whether it can shorten the CDI

168 administration compared to the full CDI, but still obtain a reliable estimate of child's

169 lexical ability. We have found that CAT-CDIs were significantly shorter to administer: the

170 English CAT-CDI median time was 342s (~5.7 minutes) (median full CDI time was 1466s

171 (~24.43 minutes)) and the Polish CAT-CDI median time was 117s (~1.95 minutes) (median

172 full CDI time was 1240s (~20.67 minutes)). Even though the CAT-CDI in Polish seemed

173 shorter (compared to English CAT-CDI), the number of items administered in the two

174 language versions of CAT-CDI was comparable: English $M = 31.42$, $Mdn = 25$ (25–50

175 items), Polish $M = 34.71$, $Mdn = 23$ (13–75 items).

176   The two CAT-CDIs differed in their settings as to when to finish the administration.

177 Specifically, the Polish CAT-CDI was set so that the SEM as low as 0.1 be reached (with as

178 few items as possible), and if the SEM could not be lowered to or below 0.1, the CAT-CDI

179 stopped after administering a maximum of 75 items. The English CAT-CDI on the other

180 hand was set to administer a maximum of 50 items, but the administration could be

181 stopped earlier if the SEM of 0.15 or lower was reached. In other words, the Polish

182 CAT-CDI prioritized as low SEM as possible, while the English CAT-CDI prioritized

183 shorter test length. This difference resulted in slightly lower mean SEM in Polish

184 CAT-CDI, compared to English CAT-CDI $\Delta M = 0.06$, 95% CI $[0.05, 0.07]$,

185 $t(257.25) = 11.36$, $p < .001$ but comparable mean number of items administered,

186 $\Delta M = -3.29$, 95% CI $[-7.74, 1.16]$, $t(137.00) = -1.46$, $p = .147$.

187   In both CAT-CDIs, we have been able to reliably estimate lexical ability for majority

188 of children. For 93% of the American participants, the SEM was 0.2 or lower and for 93%

189 of Polish participants, the SEM was below 0.15 and for 94 participants (83.19%) the Polish

190 CAT-CDI SEM was 0.1 or lower. The remaining particiapants, i.e. those with higher

191 standard errors of measurement (in other words, lower reliability of ability estimates) were

192 typically those at the extremes of the ability scale–children with either relatively low or

193 relatively high lexical skills (see Figure 3).

194   Next: - Parental (in)consistency.

Discussion

196   In points for now:

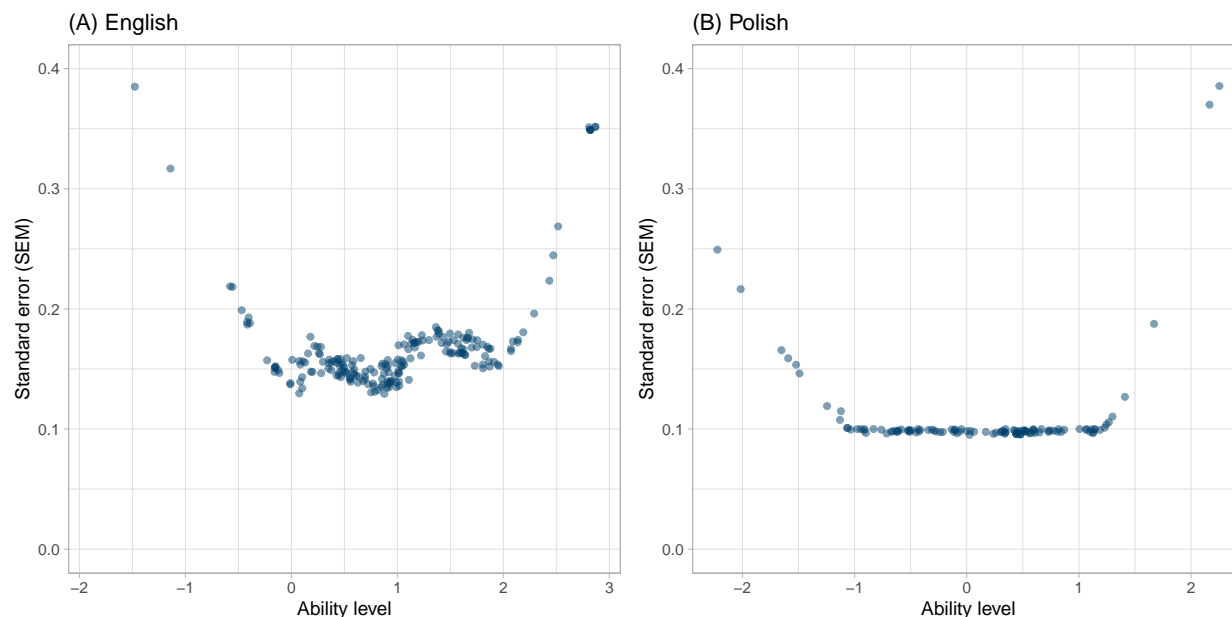197   1. *Strong correlations with full CDI.*

Figure 3. Standard Error of Measurement (SEM) by ability estimate in (A) English and (B) Polish.

The CAT-CDIs were strongly correlated with the full CDI versions in both languages, which supports the overall validity of the CAT approach. These correlations held for both monolingual and multilingual participants. However, the Polish sample was small, so further research is needed—and planned—to confirm these findings.

2. *Cases of extreme divergence.*

CAT-CDI scores diverged notably from full CDI estimates for only a small number of children in both language groups. In the Polish sample all divergent cases had low SEM, and their full CDIs were unusually short–maybe the full form underestimated their ability? In the English sample, their full CDIs were not very short, but SEM was relatively low, which is good.

3. *Item paramteres across languages.*

Item parameters (i.e., difficulty and discrimination) showed moderate correlations across languages. These likely stem from variations in the calibration samples. Still,

211   some semantic categories showed consistent patterns across languages, suggesting

212   underlying structural similarities in lexical development.

213 4. *Item selection.*

214   Each CAT version used c.a. half of the full CDI item pool. Few items were overused

215   (mostly those with highest discrimination in the item pool or starting items). But we

216   also saw items being chosen for CAT that wouldn't be considered very informative

217   (i.e., of low discrimination). That's something to explore further, I wrote to Phil

218   Chalmers (author of mirt and mirtCAT), maybe he'll be able to help?

219 5. *Efficiency and Reliability.*

220   CAT-CDIs were significantly shorter than the full CDIs while maintaining high

221   reliability for most children. The few less reliable estimates were for children at the

222   ends of the ability range (children with either very low or very high lexical skills).

223   Comparing Polish and English samples, the ability ranges with low SEM differed (see

224   Figure 3) - this again might be a result of the calibration samples and item

225   parameters? This is a good place to stress the importance of well-balanced, diverse

226   calibration datasets. Also, while the CAT-CDI is efficient and useful for screening or

227   initial identification ("pomiar przesiewowy" in Polish), it may be less suited for

228   diagnostic purposes where high precision is needed across the full ability range.

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

256 Müller, K. (2020). *Here: A simpler way to find your files*. Retrieved from

257   https://CRAN.R-project.org/package=here

258 Müller, K., & Wickham, H. (2025). *Tibble: Simple data frames*. Retrieved from

259   https://CRAN.R-project.org/package=tibble

260 R Core Team. (2025). *R: A language and environment for statistical computing*. Vienna,

261   Austria: R Foundation for Statistical Computing. Retrieved from

262   https://www.R-project.org/

263 Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. New York: Springer.

264   Retrieved from http://lmdvr.r-forge.r-project.org

265 Slowikowski, K. (2024). *Ggrepel: Automatically position non-overlapping text labels with*

266   *'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggrepel

267 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New

268   York. Retrieved from https://ggplot2.tidyverse.org

269 Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors)*.

270   Retrieved from https://CRAN.R-project.org/package=forcats

271 Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations*.

272   Retrieved from https://CRAN.R-project.org/package=stringr

273 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., . . .

274   Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43),

275   1686. https://doi.org/10.21105/joss.01686

276 Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A*

277   *grammar of data manipulation*. Retrieved from

278   https://CRAN.R-project.org/package=dplyr

279 Wickham, H., & Henry, L. (2025). *Purrr: Functional programming tools*. Retrieved from

280   https://CRAN.R-project.org/package=purrr

281 Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data*. Retrieved

282   from https://CRAN.R-project.org/package=readr

283 Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data.* Retrieved from

284   https://CRAN.R-project.org/package=tidyr

285   R (Version 4.4.3; R Core Team, 2025) and the R-packages *dplyr* (Version 1.1.4;

286 Wickham, François, Henry, Müller, & Vaughan, 2023), *forcats* (Version 1.0.0; Wickham,

287 2023a), *ggplot2* (Version 3.5.2; Wickham, 2016), *ggpubr* (Version 0.6.0; Kassambara, 2023),

288 *ggrepel* (Version 0.9.6; Slowikowski, 2024), *gridExtra* (Version 2.3; Auguie, 2017), *here*

289 (Version 1.0.1; Müller, 2020), *lattice* (Version 0.22.7; Sarkar, 2008), *lubridate* (Version 1.9.4;

290 Grolemund & Wickham, 2011), *mirt* (Version 1.44.0; Chalmers, 2012, 2016), *mirtCAT*

291 (Version 1.14; Chalmers, 2016), *Multilada* (Version 0.7.0; Krajewski, 2025), *papaja* (Version

292 0.1.3; Aust & Barth, 2024), *purrr* (Version 1.0.4; Wickham & Henry, 2025), *readr* (Version

293 2.1.5; Wickham, Hester, & Bryan, 2024), *shiny* (Version 1.10.0; Chang et al., 2024), *stringr*

294 (Version 1.5.1; Wickham, 2023b), *tibble* (Version 3.3.0; Müller & Wickham, 2025), *tidyr*

295 (Version 1.3.1; Wickham, Vaughan, & Girlich, 2024), *tidyverse* (Version 2.0.0; Wickham et

296 al., 2019), *tinylabels* (Version 0.2.5; Barth, 2023), *viridis* (Garnier et al., 2023; Version

297 0.6.5; Garnier et al., 2024) and *viridisLite* (Version 0.4.2; Garnier et al., 2023)