

### Online supplementary materials

#### Meta-analysis

**Dataset.** This section provides additional information regarding the search process used to identify, screen, classify, and include studies in the meta-analysis. I followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines, which are freely available online: <http://www.prisma-statement.org>. Prior to beginning the search process, I established a list of keywords to be used in library-housed online databases. The keywords included terms used to describe compromise categories ('compromise categories', 'merged categories', 'mixed categories', 'intermediate categories'), terms used to describe the bilingual population of interest ('early learners', 'early bilinguals', 'simultaneous bilinguals'), and well as terms commonly used to describe voice-timing studies ('VOT', 'voice-onset time'). Every possible combination of the three categories of search terms was used ( $n = 40$  individual queries) in each of the following six databases: ERIC, Science Direct, Linguistics and Language Behavior Abstracts, PsycINFO, ProQuest Dissertations and Theses, and FirstSearch. This resulted in 240 independent search queries. The following table provides a breakdown of the number of records identified in each database for each search query.

Keyword	Database					
	ERIC	FirstSearch	LLBA	ProQuest	PsychINFO	Science Direct
mixed categories	934	1959	10540	11280	0	4411
merged categories	48	119	2792	4257	0	2550
compromise categories	46	212	1890	3611	0	1848
intermediate categories	852	1274	8271	10504	0	3876
mixed categories + early learners	493	6	4684	6240	48	1016
merged categories + early learners	170	0	945	1879	30	403
mixed categories + early bilinguals	215	1	3690	5733	217	652
merged categories + early bilinguals	93	0	762	1737	58	221
compromise categories + early learners	172	0	827	1938	58	320
mixed categories + early learners + VOT	504	0	310	581	42	41
merged categories + early learners + VOT	180	0	124	257	12	25
intermediate categories + early learners	419	10	4287	5953	107	902
compromise categories + early bilinguals	87	0	684	1776	57	217
mixed categories + early bilinguals + VOT	219	0	313	569	44	74
merged categories + simultaneous bilinguals	41	0	1272	2368	339	335
merged categories + early bilinguals + VOT	98	0	127	266	12	39
intermediate categories + early bilinguals	202	0	2684	4940	104	441
merged categories + simultaneous bilinguals	16	0	281	801	311	118
compromise categories + early learners + VOT	182	0	69	177	58	7
intermediate categories + early learners + VOT	430	0	281	574	19	70
compromise categories + early bilinguals + VOT	92	0	69	174	11	11
compromise categories + simultaneous bilinguals	13	0	236	808	328	125
mixed categories + simultaneous bilinguals + VOT	47	0	170	287	14	42
intermediate categories + early bilinguals + VOT	208	0	258	546	20	58
merged categories + simultaneous bilinguals + VOT	21	0	79	151	1	21
intermediate categories + simultaneous bilinguals	27	0	877	1967	243	216
mixed categories + early learners + voice-onset time	117	0	309	496	2	381
merged categories + early learners + voice-onset time	39	0	118	214	2	105
compromise categories + simultaneous bilinguals + VOT	18	0	46	101	1	8
mixed categories + early bilinguals + voice-onset time	50	0	317	495	3	279
merged categories + early bilinguals + voice-onset time	22	0	121	216	3	77
intermediate categories + simultaneous bilinguals + VOT	33	0	137	270	3	33
compromise categories + early learners + voice-onset time	38	0	73	151	2	146
intermediate categories + early learners + voice-onset time	73	0	286	510	3	251
compromise categories + early bilinguals + voice-onset time	19	0	68	138	3	94
mixed categories + simultaneous bilinguals + voice-onset time	17	0	172	256	3	173
intermediate categories + early bilinguals + voice-onset time	38	0	268	480	3	166
merged categories + simultaneous bilinguals + voice-onset time	12	0	72	122	2	46
compromise categories + simultaneous bilinguals + voice-onset time	10	0	45	81	2	56
intermediate categories + simultaneous bilinguals + voice-onset time	18	0	142	248	2	97

*Figure 7.* Number of records of identified studies across six library-housed databases as a function of search query.

There were 153,860 records identified through database searching and 27 additional ancestry studies identified through Google and Google scholar. After removing duplicates and irrelevant hits the study pool contained 148 records. The 148 full-text articles and dissertations were assessed for eligibility using the criteria explained in section 2.1 of the manuscript. A total of 68 appeared to meet the established inclusion criteria, however, 48 were removed. Thus the final dataset included 20 studies. The specific reasons for exclusion are provided in the table below.

Reason for exclusion	n	% of discarded
3-way contrast	5	10
Duplicate data	2	4
L3 or child participants	15	31
Incomplete or missing data	10	21
No control group	16	33
<b>Total</b>	48	100

As observed in the table, of the discarded studies, 5 (10%) examined languages that included 3-way voicing distinctions, 2 (4%) presented duplicate data previously published, 15 (31%) investigated something different from what was originally believed, such as L3 speech or child bilinguals, 10 (21%) had incomplete or missing data (typically SD was not reported), and 16 (33%) of the studies did not include a control group.

**Analysis.** This section provides additional tables that complement the meta-analysis reported in the *Results* section of the manuscript. This study employs Bayesian Data Analysis (BDA) for quantitative inferential statistics. Specifically, a cross-classified Bayesian meta-analysis was conducted by fitting the study data with the multilevel regression model formulated below:

$$\begin{aligned}
 \text{SMD}_i &\sim \text{Normal}(\theta_i, \sigma_i = \text{se}_i) \\
 \theta_i &\sim \text{Normal}(\mu, \tau) \\
 \mu &\sim \text{Normal}(0, 1) \\
 \tau &\sim \text{HalfCauchy}(0, 1)
 \end{aligned}$$

BDA implies the use of Bayesian *credible intervals*—and other metrics—for statistical inferences. A Bayesian model calculates a posterior distribution, i.e., a distribution of plausible parameter values, given the data, a data-generating model, and any prior information we have about those parameter values. Posterior distributions are computationally costly. For this reason, the Hamiltonian Markov Chain Monte Carlo algorithm is used to obtain a sample that includes thousands of values from the posterior distribution. In practical terms, what this means is that the

model does not calculate a single point estimate for an effect  $\beta$  like in a traditional frequentist framework, but rather it draws a sample of 4,000 plausible values for  $\beta$ . This allows the researcher to quantify her uncertainty regarding  $\beta$  by summarizing the distribution of those values. Generally, the present study uses four statistics to describe the posterior distribution: (1) the mean, (2) the highest density credible interval (HDI), (3) the proportion of the HDI that falls within a Region of Practical Equivalence (ROPE), and (4) the Maximum Probability of Effect (MPE). The mean provides a point estimate for the distribution. The 95% highest density credible interval provides bounds for the effect. The ROPE designates a region of practical equivalence around a point null value and calculates the proportion of the HDI that falls within this interval. The MPE calculates the proportion of the posterior distribution that is of the median's sign (or the probability that the effect is positive or negative).

If, for instance, a hypothesis states that  $\beta > 0$ , we judge there is *compelling evidence* for this hypothesis if the mean point estimate is a positive number, if the 95% HDI of  $\beta$  does not contain 0 and is outside the ROPE, and the posterior  $P(\beta > 0)$  is close to one. Together these four statistics allow us to quantify our uncertainty and provide an intuitive interpretation of any given effect. Consider a case in which the posterior mean of  $\beta$  is 100 and the 95% HDI is [40, 160]. The interval tells us that we can be 95% certain the *true* value of  $\beta$  is between 40 and 160, given the data, our model, and our prior information. Furthermore, the interval allows us to specify areas of uncertainty. In this example, we can conclude that the effect is almost certain to be positive. The lower interval value of 40 tells us that 95% of the plausible values are greater than 40. We also note that the interval covers a wide range of values, thus we also conclude that we are not very certain about the size of the effect. This type of interpretation is not possible under a frequentist paradigm.

For more information regarding how BDA differs from more traditional frequentist analyses, see Kimball, Shantz, Eager, and Roy (2016). Additionally, Kruschke and Liddell (2018) and Nicenboim and Vasishth (2016) provide tutorials designed for linguists, and Schoot and Depaoli (2014) provides a general presentation of reporting BDA results for the Social Sciences.

In the present study, the model utilizes calculated effect sizes and standard error to sample from a posterior distributions of plausible values for a pooled effect of “compromise” categories with regard to VOT. The structure of the grouping variables is described as cross-classified because each study belongs to multiple clusters in the model. The posterior distribution of the grouping variables provide useful information about different aspects of the individual studies. The following table provides a summary of the posterior distribution of the model presented in section 2.1:

	Parameter	Estimate	95% HDI
<b>Population-level effects</b>			
	Intercept	−0.132	[−0.708, 0.468]
	Lexical stress	−0.14	[−0.766, 0.483]
	Analystic strategy	0.182	[−0.657, 1.035]
<b>Group-level effects</b>			
Pooling method	/k/	−0.163	[−0.874, 0.202]
	/p/	−0.021	[−0.456, 0.369]
	/pt/	0.108	[−0.373, 0.924]
	/ptk/	0.061	[−0.344, 0.586]
	/t/	−0.001	[−0.429, 0.405]
Study	Amengual (2011)	−0.039	[−0.661, 0.575]
	Antoniou et al. (2010)	0.267	[−0.169, 0.73]
	Brown & Copple (2018)	0.093	[−0.618, 0.858]
	Flege (1991)	0.258	[−0.345, 0.96]
	Flege & Eefting (1987a)	−0.18	[−0.864, 0.457]
	Flege & Eefting (1987b)	−0.537	[−1.128, −0.011]
	Flege & Eefting (1988)	0.135	[−0.484, 0.782]
	Fowler et al. (2008)	−0.443	[−1.033, 0.069]
	Hazen & Boulakia (1993)	−0.436	[−1.132, 0.15]
	Jones (2020)	0.034	[−0.678, 0.754]
	Kim (2011)	0.284	[−0.316, 0.987]
	Knightly et al. (2003)	0.107	[−0.481, 0.723]
	Kupisch & Lleo (2017)	−0.346	[−1.154, 0.32]

Lein et al. (2016)	−0.035	[−0.815, 0.695]
Liman (2013)	0.262	[−0.442, 1.105]
MaCleod (2005)	0.255	[−0.449, 1.069]
Magloire & Green (1999)	0.118	[−0.547, 0.808]
Schmidt & Flege (1996)	0.327	[−0.195, 0.916]
Simonet & Casillas (2014)	−0.216	[−0.809, 0.343]
Sundara et al. (2006)	0.035	[−0.591, 0.661]
<b>Pooled effect</b>	−0.132	[−0.708, 0.468]

## Production of coronal stops

This section provides additional tables and figures that complement the analysis of coronal stop production reported in the *Results* section of the manuscript. The first table provides a list of the stimuli used in the production task.

Spanish		English	
/d/	/t/	/d/	/t/
daba	taberna	dagger	tapioca
dado	tabla	dakota	tabloid
daga	tabú	daltonian	taboo
daltónico	tactil	damage	tacit
dama	tamaño	damnation	tackle
dañar	también	damper	tactics
danés	tampoco	dancette	tambourine
danesa	tanque	dancing	tanker
danino	tanto	danielle	tantrum
daño	taza	danseur	tattoo
danza	tabaco	dapper	tamper
danzar	taco	dazzle	tablet

The next table summarizes the posterior distribution of the Bayesian regression models.

Outcome	Language	Parameter	$\beta$	95% HDI	MPE	ROPE %	ROPE
VOT	Spanish	Intercept	-23.188	[-31.839, -14.888]	1	0	[-5.115, 5.115]
		Group	0.78	[-6.977, 8.604]	0.579	0.844	[-5.115, 5.115]
		Phoneme	-38.865	[-46.555, -31.766]	1	0	[-5.115, 5.115]
		Item rep.	-1.02	[-2.726, 0.675]	0.89	1	[-5.115, 5.115]
		Group $\times$ Phoneme	1.213	[-4.842, 8.058]	0.646	0.898	[-5.115, 5.115]
	English	Intercept	45.199	[38.657, 51.812]	1	0	[-4.595, 4.595]
		Group	-7.005	[-13.015, -0.951]	0.988	0.19	[-4.595, 4.595]
		Phoneme	-31.743	[-37.77, -25.363]	1	0	[-4.595, 4.595]
		Item rep.	-0.21	[-1.574, 1.133]	0.622	1	[-4.595, 4.595]
		Group $\times$ Phoneme	-5.161	[-11.02, 0.18]	0.966	0.41	[-4.595, 4.595]
Relative VOT	Spanish	Intercept	0.211	[0.189, 0.233]	1	0	[-0.014, 0.014]
		Group	-0.001	[-0.018, 0.015]	0.576	0.945	[-0.014, 0.014]
		Phoneme	0.1	[0.071, 0.127]	1	0	[-0.014, 0.014]
		Item rep.	0.002	[-0.003, 0.007]	0.825	1	[-0.014, 0.014]
		Group $\times$ Phoneme	-0.004	[-0.028, 0.017]	0.658	0.792	[-0.014, 0.014]
	English	Intercept	0.292	[0.272, 0.312]	1	0	[-0.013, 0.013]
		Group	0.004	[-0.013, 0.02]	0.697	0.898	[-0.013, 0.013]
		Phoneme	-0.101	[-0.119, -0.083]	1	0	[-0.013, 0.013]
		Item rep.	-0.002	[-0.005, 0.002]	0.829	1	[-0.013, 0.013]
		Group $\times$ Phoneme	0.01	[-0.004, 0.025]	0.911	0.699	[-0.013, 0.013]

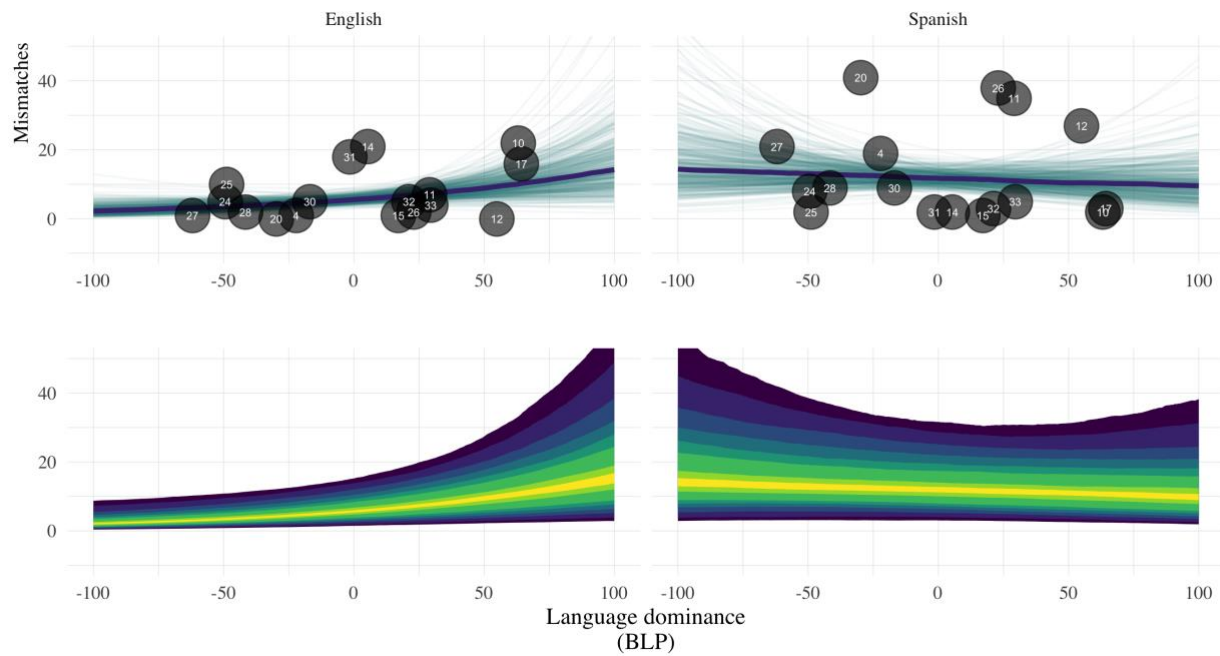
And the following table provides a summary of the ANOVA models.

Outcome	Language	Effect	<i>F</i>	<i>df</i> <sub>1</sub>	<i>df</i> <sub>2</sub>	<i>MSE</i>	<i>p</i>	$\hat{\eta}^2_G$
VOT	Spanish	Group	0.28	1	23	772.37	.59	0.007
		Phoneme	116.22	1	23	540.44	< .001	0.675
		Group $\times$ Phoneme	0.45	1	23	540.44	.51	0.008
	English	Group	5.98	1	23	372.96	.02	0.124
		Phoneme	139.85	1	23	314.65	< .001	0.736
		Group $\times$ Phoneme	3.93	1	23	314.65	.05	0.073
Relative VOT	Spanish	Group	0.19	1	23	0	.67	0.003
		Phoneme	75.88	1	23	0.01	< .001	0.684
		Group $\times$ Phoneme	0.44	1	23	0.01	.51	0.012
	English	Group	0.39	1	23	0	.54	0.009
		Phoneme	213.28	1	23	0	< .001	0.808
		Group $\times$ Phoneme	2.02	1	23	0	.17	0.038



### Performance mismatches

An exploratory model analyzing the number of performance mismatches as a function of language and language dominance was fit using a Bayesian regression model. The analysis did not provide compelling evidence of a relationship between language dominance and the amount of performance mismatches produced by the early bilinguals.



*Figure 7.* Performance mismatches as a function of language and language dominance scores. The top panels plots 400 draws from the posterior distribution of the model and the bottom panels provide posterior predictive intervals.

## Reproducibility information

### About this document

This document was written in RMarkdown using papaja (Aust & Barth, 2020).

### Session info

```

setting  value
version  R version 4.0.3 (2020-10-10)
os       macOS Catalina 10.15.6
system   x86_64, darwin17.0
ui       X11
language (EN)
collate  en_US.UTF-8
ctype    en_US.UTF-8
tz       America/New_York
date     2020-10-28

               loadedversion      date
abind          1.4-5 2016-07-21
afex           0.28-0 2020-09-20
arrayhelpers   1.1-0 2020-02-04
assertthat     0.2.1 2019-03-21
backports      1.1.10 2020-09-15
base64enc      0.1-3 2015-07-28
bayesplot      1.7.2 2020-05-28
bayestestR     0.7.2 2020-07-20
beeswarm       0.2.3 2016-04-25
bookdown       0.21 2020-10-13
boot           1.3-25 2020-04-26
bridgesampling 1.0-0 2020-02-26
brms           2.14.0 2020-10-08
Brodingnag     1.2-6 2018-08-13
callr          3.5.1 2020-10-13
car            3.0-10 2020-09-29
carData        3.0-4 2020-05-22
cellranger     1.1.0 2016-07-27
class          7.3-17 2020-04-26
cli            2.1.0 2020-10-12
cluster        2.1.0 2019-06-19
coda           0.19-4 2020-09-30
codetools      0.2-17 2020-10-17
colorspace     1.4-1 2019-03-18
colourpicker   1.1.0 2020-09-14
CompQuadForm   1.4.3 2017-04-12
crayon         1.3.4 2017-09-16
crosstalk      1.1.0.1 2020-03-13
curl           4.3 2019-12-02
data.table     1.13.2 2020-10-19
DEoptimR       1.0-8 2016-11-19

```

desc	1.2.0	2018-05-01
devtools	2.3.2	2020-09-18
digest	0.6.26	2020-10-17
diptest	0.75-7	2016-12-05
distributional	0.2.1	2020-10-06
dmetar	0.0.9000	2020-09-22
dplyr	1.0.2	2020-08-18
DT	0.16	2020-10-13
dygraphs	1.1.1.6	2018-07-11
ellipsis	0.3.1	2020-05-15
emmeans	1.5.1	2020-09-18
esc	0.5.1	2019-12-04
estimability	1.3	2018-02-11
evaluate	0.14	2019-05-28
fansi	0.4.1	2020-01-08
farver	2.0.3	2020-01-16
fastmap	1.0.1	2019-10-08
flexmix	2.3-17	2020-10-12
forcats	0.5.0	2020-03-01
foreign	0.8-80	2020-05-24
fpc	2.2-8	2020-09-19
fs	1.5.0	2020-07-31
generics	0.0.2	2018-11-29
ggbeeswarm	0.6.0	2017-08-07
ggdist	2.2.0	2020-07-12
ggplot2	3.3.2	2020-06-19
ggrepel	0.8.2	2020-03-08
ggridges	0.5.2	2020-01-12
glue	1.4.2	2020-08-27
gridExtra	2.3	2017-09-09
gtable	0.3.0	2019-03-25
gtools	3.8.2	2020-03-31
haven	2.3.1	2020-06-01
here	0.1	2017-05-28
highr	0.8	2019-03-20
hms	0.5.3	2020-01-08
htmltools	0.5.0	2020-06-16
htmlwidgets	1.5.2	2020-10-03
httpuv	1.5.4	2020-06-06
igraph	1.2.6	2020-10-06
inline	0.3.16	2020-09-06
insight	0.9.6	2020-09-20
jsonlite	1.7.1	2020-09-07
kernlab	0.9-29	2019-11-12
knitr	1.30	2020-09-22
later	1.1.0.1	2020-06-05
lattice	0.20-41	2020-04-02
lifecycle	0.2.0	2020-03-06
lme4	1.1-23	2020-04-07
lmerTest	3.1-2	2020-04-08
loo	2.3.1	2020-07-14

magic	1.5-9	2018-09-17
magrittr	1.5	2014-11-22
markdown	1.1	2019-08-07
MASS	7.3-53	2020-09-09
Matrix	1.2-18	2019-11-27
matrixStats	0.57.0	2020-09-25
mclust	5.4.6	2020-04-11
memoise	1.1.0	2017-04-21
meta	4.15-1	2020-10-02
metafor	2.4-0	2020-03-19
metaviz	0.3.1	2020-04-09
mime	0.9	2020-02-04
miniUI	0.1.1.1	2018-05-18
minqa	1.2.4	2014-10-09
modeltools	0.2-23	2020-03-05
MuMIn	1.43.17	2020-04-15
munsell	0.5.0	2018-06-12
mvtnorm	1.1-1	2020-06-09
netmeta	1.2-1	2020-04-16
nlme	3.1-149	2020-08-23
nloptr	1.2.2.2	2020-07-02
nnet	7.3-14	2020-04-26
numDeriv	2016.8-1.1	2019-06-06
openxlsx	4.2.2	2020-09-17
papaja	0.1.0.9997	2020-10-21
patchwork	1.0.1.9000	2020-10-10
pillar	1.4.6	2020-07-10
pkgbuild	1.1.0	2020-07-13
pkgconfig	2.0.3	2019-09-22
pkgload	1.1.0	2020-05-29
plyr	1.8.6	2020-03-03
png	0.1-7	2013-12-03
poibin	1.5	2020-01-08
prabclus	2.3-2	2020-01-08
prettyunits	1.1.1	2020-01-24
processx	3.4.4	2020-09-03
promises	1.1.1	2020-06-09
ps	1.4.0	2020-10-07
puniform	0.2.2	2020-06-20
purrr	0.3.4	2020-04-17
R6	2.4.1	2019-11-12
Rcpp	1.0.5	2020-07-06
RcppParallel	5.0.2	2020-06-24
readr	1.4.0	2020-10-05
readxl	1.3.1	2019-03-13
remotes	2.2.0	2020-07-21
reshape2	1.4.4	2020-04-09
rio	0.5.16	2018-11-26
rlang	0.4.8	2020-10-08
rmarkdown	2.5	2020-10-21
robustbase	0.93-6	2020-03-23

robvis	0.3.0	2019-11-22
rprojroot	1.3-2	2018-01-03
rsconnect	0.8.16	2019-12-13
rstan	2.21.2	2020-07-27
rstantools	2.1.1	2020-07-06
rstudioapi	0.11	2020-02-07
scales	1.1.1	2020-05-11
sessioninfo	1.1.1	2018-11-05
shiny	1.5.0	2020-06-23
shinyjs	2.0.0	2020-09-09
shinytan	2.5.0	2018-05-01
shinythemes	1.1.2	2018-11-06
StanHeaders	2.21.0-6	2020-08-16
statmod	1.4.35	2020-10-19
stringi	1.5.3	2020-09-09
stringr	1.4.0	2019-02-10
svUnit	1.0.3	2020-04-20
testthat	2.3.2	2020-03-02
threejs	0.3.3	2020-01-21
tibble	3.0.4	2020-10-12
tidybayes	2.1.1	2020-06-19
tidyr	1.1.2	2020-08-27
tidyselect	1.1.0	2020-05-11
usethis	1.6.3	2020-09-17
V8	3.2.0	2020-06-19
vctrs	0.3.4	2020-08-29
vipor	0.4.5	2017-03-22
withr	2.3.0	2020-09-22
xfun	0.18	2020-09-29
xml2	1.3.2	2020-04-23
xtable	1.8-4	2019-04-21
xts	0.12.1	2020-09-09
yaml	2.2.1	2020-02-01
zip	2.1.1	2020-08-27
zoo	1.8-8	2020-05-02