Online Supplement to: Evidence on the validity and reliability of automatically generated propositional reasoning items: A multilingual study of the challenges of automatic generation of verbal items

Daniela Gühne and Philipp Doebler
TU Dortmund University

David M. Condon
University of Oregon

Fang Luo
Beijing Normal University

Luning Sun
University of Cambridge

Author Note

Daniela Gühne and Philipp Doebler, Department of Statistics, TU Dortmund University, Dortmund, Germany; David M. Condon, Department of Psychology, University of Oregon, Eugene, United States; Fang Luo, Faculty of Psychology, Beijing Normal University, Beijing, China; Luning Sun, The Psychometrics Centre, Cambridge Judge Business School, University of Cambridge, Cambridge, United Kingdom.

The work of Fang Luo was supported by the National Key R&D Program of China (Grant No. 2018YFC0810602). The work by Daniela Gühne and Philipp Doebler was supported by grant DO 1789/1-1 from the German Research Foundation (DFG). The work of Luning Sun was supported by the Economic and Social Research Council, UK (ESRC; Grant No. ES/L016591/1).

Correspondence concerning this article should be addressed to Philipp Doebler, Department of Statistics, TU Dortmund University, Dortmund, Germany.

Email: doebler@statistik.tu-dortmund.de

Abstract

This online supplement reports details on the confirmatory factor analysis, the linear mixed models for the Chinese- and German-speaking samples, and item characteristic curves. All data and code is freely available at https://osf.io/xa9gw/.

Online Supplement to: Evidence on the validity and reliability of automatically generated propositional reasoning items: A multilingual study of the challenges of automatic generation of verbal items

Details on the CFA model

Table 1 contains further details on the confirmatory factor analysis model used for the analysis of the tetrachoric correlations, which were computed from the English-speaking sample. Note that the number of pairwise administrations varied, so that the χ^2-value is omitted intentionally. Instead the SRMR is employed, as it is not biased by sample size.

Table E1
Detailed CFA results

		Model			
		Estimate	Std. Err.	z	p
		Factor Loadings			
VR					
	VR.04	0.81	0.03	24.86	.000
	VR.16	0.46	0.04	12.38	.000
	VR.19	0.50	0.04	13.68	.000
	VR.17	0.73	0.03	21.60	.000
R3D					
	R3D.03	0.74	0.03	21.95	.000
	R3D.04	0.76	0.03	22.83	.000
	R3D.06	0.68	0.03	19.69	.000
	R3D.08	0.67	0.03	19.22	.000
LN					
	LN.58	0.75	0.03	23.01	.000
	LN.34	0.75	0.03	23.23	.000
	LN.33	0.77	0.03	23.68	.000
	LN.07	0.76	0.03	23.33	.000
MR					
	MR.45	0.55	0.04	14.92	.000
	MR.46	0.63	0.04	17.41	.000
	MR.47	0.62	0.04	17.13	.000
	MR.55	0.58	0.04	15.85	.000
PR					
	item1	0.50	0.03	14.70	.000
	item2	0.51	0.03	14.91	.000
	item3	0.54	0.03	16.07	.000

item4	0.65	0.03	20.07	.000
item5	0.85	0.03	29.38	.000
item6	0.80	0.03	26.47	.000
item7	0.69	0.03	21.73	.000
item8	0.71	0.03	22.74	.000
item9	0.87	0.03	30.62	.000
item10	0.81	0.03	27.06	.000
item11	0.84	0.03	28.73	.000
item12	0.81	0.03	27.21	.000
item13	0.30	0.04	8.46	.000
item14	0.32	0.04	8.95	.000
item15	0.36	0.04	10.30	.000
item16	0.30	0.04	8.47	.000
item17	0.46	0.03	13.30	.000
item18	0.31	0.04	8.65	.000
item19	0.86	0.03	29.89	.000
item20	0.83	0.03	28.40	.000
item21	0.80	0.03	26.76	.000
item22	0.77	0.03	25.11	.000
item23	0.74	0.03	23.81	.000
item24	0.76	0.03	24.81	.000
item25	0.87	0.03	30.19	.000
item26	0.88	0.03	30.81	.000
item27	0.78	0.03	25.66	.000
item28	0.64	0.03	19.77	.000
item29	0.74	0.03	23.81	.000
item30	0.75	0.03	24.18	.000
	Residual Variances			
VR.04	0.34	0.03	12.18	.000
VR.16	0.79	0.04	18.66	.000
VR.19	0.75	0.04	18.40	.000
VR.17	0.47	0.03	15.35	.000
R3D.03	0.45	0.03	14.82	.000
R3D.04	0.42	0.03	14.12	.000
R3D.06	0.54	0.03	16.22	.000
R3D.08	0.56	0.03	16.46	.000
LN.58	0.44	0.03	15.58	.000
LN.34	0.43	0.03	15.45	.000
LN.33	0.41	0.03	15.16	.000
LN.07	0.43	0.03	15.39	.000
MR.45	0.69	0.04	17.36	.000
MR.46	0.60	0.04	16.16	.000
MR.47	0.61	0.04	16.32	.000
MR.55	0.66	0.04	16.97	.000

item1	0.75	0.04	19.42	.000
item2	0.74	0.04	19.41	.000
item3	0.71	0.04	19.36	.000
item4	0.58	0.03	19.16	.000
item5	0.27	0.02	17.88	.000
item6	0.37	0.02	18.45	.000
item7	0.52	0.03	19.03	.000
item8	0.49	0.03	18.95	.000
item9	0.24	0.01	17.51	.000
item10	0.35	0.02	18.35	.000
item11	0.29	0.02	18.04	.000
item12	0.34	0.02	18.34	.000
item13	0.91	0.05	19.57	.000
item14	0.90	0.05	19.56	.000
item15	0.87	0.04	19.54	.000
item16	0.91	0.05	19.57	.000
item17	0.79	0.04	19.46	.000
item18	0.90	0.05	19.57	.000
item19	0.26	0.01	17.71	.000
item20	0.30	0.02	18.07	.000
item21	0.36	0.02	18.45	.000
item22	0.41	0.02	18.68	.000
item23	0.45	0.02	18.84	.000
item24	0.42	0.02	18.73	.000
item25	0.25	0.01	17.55	.000
item26	0.23	0.01	17.35	.000
item27	0.39	0.02	18.63	.000
item28	0.59	0.03	19.17	.000
item29	0.45	0.02	18.85	.000
item30	0.44	0.02	18.81	.000
	Residual Covariances			
item1 w/item2	0.52	0.03	15.70	.000
item3 w/item4	0.43	0.03	15.31	.000
item5 w/item6	0.10	0.01	7.39	.000
item7 w/item8	0.19	0.02	9.20	.000
item9 w/item10	0.13	0.01	10.05	.000
item11 w/item12	0.10	0.01	7.60	.000
item13 w/item14	0.55	0.04	14.47	.000
item15 w/item16	0.51	0.04	13.75	.000
item17 w/item18	0.43	0.03	12.55	.000
item19 w/item20	0.09	0.01	7.32	.000
item21 w/item22	0.14	0.02	9.12	.000
item23 w/item24	0.21	0.02	11.34	.000
item25 w/item26	0.06	0.01	6.29	.000

item27 w/item28	0.22	0.02	11.04	.000
item29 w/item30	0.25	0.02	12.82	.000
	Latent Variances			
VR	1.00+			
R3D	1.00+			
LN	1.00+			
MR	1.00+			
PR	1.00+			
	Latent Covariances			
VR w/R3D	0.66	0.03	21.55	.000
VR w/LN	0.70	0.03	25.24	.000
VR w/MR	0.74	0.03	23.54	.000
VR w/PR	0.67	0.03	25.94	.000
R3D w/LN	0.53	0.03	15.59	.000
R3D w/MR	0.67	0.03	20.28	.000
R3D w/PR	0.53	0.03	17.43	.000
LN w/MR	0.72	0.03	24.04	.000
LN w/PR	0.55	0.03	18.78	.000
MR w/PR	0.61	0.03	19.36	.000
	Fit Indices			
NPAR	117.00			
DF	964.00			
SRMR	0.08			

+Fixed parameter

Detailed regression results

Table 2 reports all details of the linear mixed models in the Chinese-speaking sample and Table 3 contains the results for the German-speaking sample.

Table E2
Prediction of PR total scores (Chinese-speaking sample).

	model C1	model C2	model C3
fixed effects, β (SE_β)			
intercept	9.78*** (0.38)	9.49*** (0.48)	5.38*** (0.57)
test version (B)		0.60 (0.64)	0.74 (0.38)
ICAR16[1]			0.47*** (0.06)
random effects, SD			
class	1.01	0.95	< 0.01
error	3.15	3.16	2.90
model comparison			
χ^2 (df)		1.07 (1)	52.79*** (1)
AIC	1244.0	1244.9	1194.1
BIC	1254.4	1258.8	1211.5

Note. For binary variables the reference category is given in brackets. Model C2 is contrasted with model C1. Model C3 is contrasted with model C2. χ^2: likelihood ratio test statistic for nested models; [1] total score; * $p < .05$; ** $p < .01$; *** $p < .001$

Table E3
Prediction of PR total scores (German-speaking sample).

	model G1	model G2	model G3	model G4
fixed effects, β (SE_β)				
intercept	11.49*** (0.22)	11.44*** (0.29)	4.20* (1.62)	3.57* (1.40)
test version (B)		-0.09 (0.21)	-0.06 (0.20)	-0.08 (0.20)
session (2)		0.26 (0.21)	0.34 (0.21)	0.36 (0.21)
form (pen and paper)		-0.08 (0.21)	-0.10 (0.20)	-0.05 (0.20)
QCM anxiety[2]			0.08 (0.15)	0.17 (0.13)
QCM challenge[2]			0.29 (0.18)	0.38* (0.17)
QCM interest[2]			0.17 (0.17)	-0.11 (0.16)
QCM success[2]			0.05 (0.18)	0.10 (0.16)
S-C test[1]			0.04*** (0.01)	0.01 (0.01)
ICAR16[1]				0.45*** (0.06)
random effects, SD				
person	2.05	2.05	1.85	1.38
error	1.49	1.50	1.47	1.47
model comparison				
χ^2 (df)		2.00 (3)	26.57*** (5)	43.98*** (1)
AIC	941.4	945.4	928.8	886.8
BIC	951.4	965.5	965.7	927.1

Note. For binary variables the reference category is given in brackets. Model G2 is contrasted with model G1. Model G3 is contrasted with model G2. Model G4 is contrasted with model G3. χ^2: likelihood ratio test statistic for nested models; [1] total score; [2] mean score; * $p < .05$; ** $p < .01$; *** $p < .001$

Item characteristic curves

Figure E1 contains the ICCs of the 15 PR item families from the 2PL models. The English- and Chinese-speaking samples from Studies 1 resp. 2 can be directly compared.

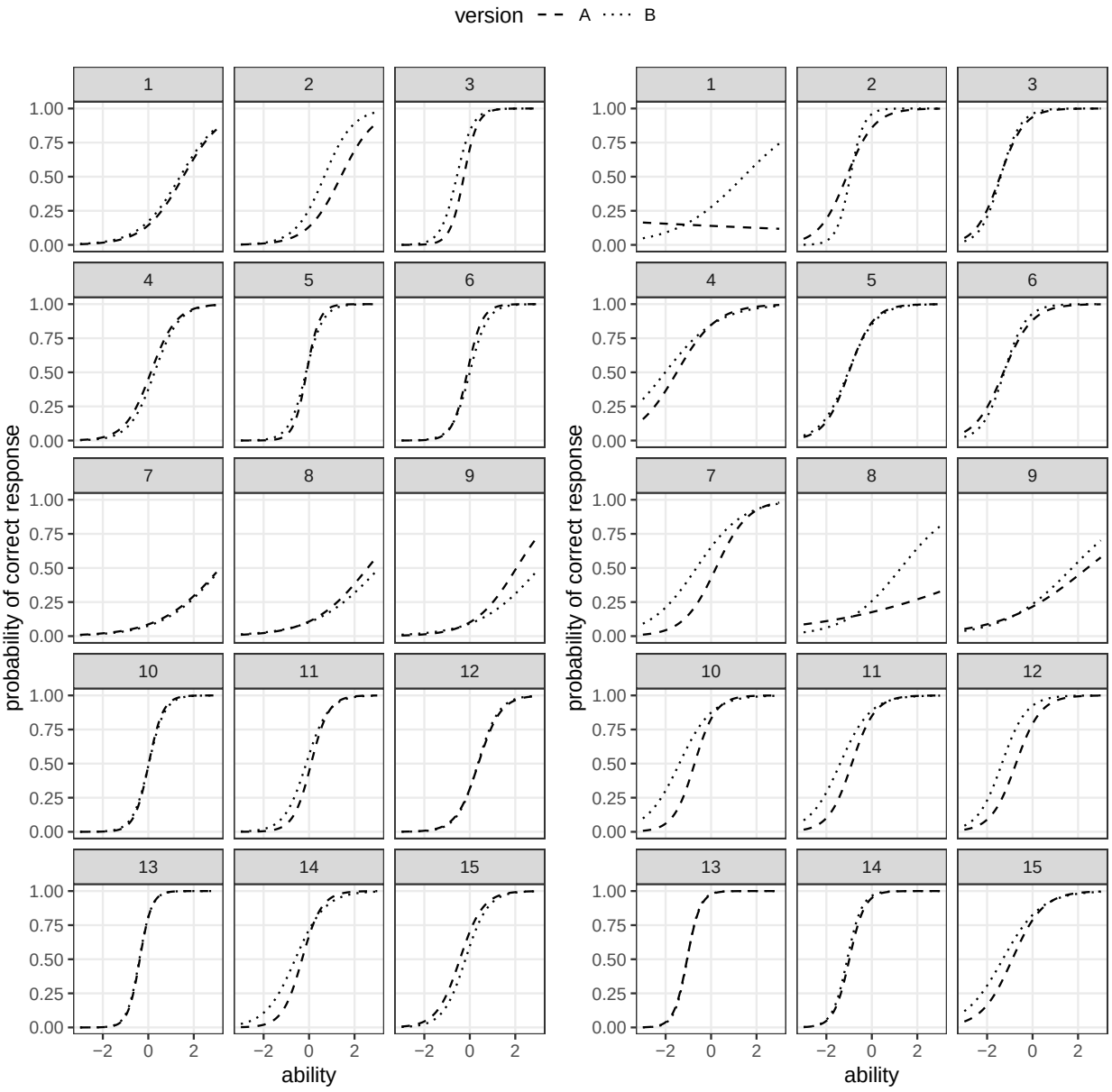

Figure E*1*. ICCs of PR item families 1-15 in 2PL model (left panel: English-speaking sample from Study 1; right panel: Chinese-speaking sample from Study 2).