

*Elektronisches Supplement 8. Detaillierte Ergebnisse zur Konstruktvalidierung.*

Die Hypothesen 1–4 werden auf Basis der Korrelationen untersucht, die in Tabelle 1 für die drei Teilstichproben abgebildet sind. Die Korrelation wurden mit Hilfe der Pakete *lavaan* (Rosseel, 2012) und *lavaan.survey* (Oberski, 2014) in R (R Core Team, 2020) bestimmt. Für die Schätzung der Korrelationen und deren Standardfehler wurde eine Maximum Likelihood (ML) Verfahren verwendet. Das ML-Verfahren ermöglicht die Berücksichtigung der fehlenden Werte (Methode: Full Information Maximum Likelihood [FIML]; Lüdtke, Robitzsch, Trautwein & Köller, 2007) und der hierarchischen Stichprobenstruktur (Oberski, 2014). Fehlende Werte sind in ESM 7 dokumentiert. Die hierarchische Stichprobenstruktur wurde auf Klassenebene berücksichtigt. Alle berichteten Korrelationen sind signifikant unterschiedlich von Null.

Die Hypothesen 1–4 beziehen sich auf die Unterschiedlichkeit von Korrelationen. Die Unterschiedlichkeit wurde mit Williams  $t$  für abhängige Korrelationen, die eine gemeinsame Variable teilen, bestimmt (Williams, 1959). Zusätzlich ermöglichen die 95-prozentigen Konfidenzintervalle der Korrelationen eine Abschätzung der Unterschiedlichkeit zwischen Korrelationen.

Hypothese 5 wird auf Basis des Vergleichs von zwei Regressionsmodellen mit gemischten Effekten zwischen dem Textverständnis gemessen mit dem Itempool (TVIP) und dem ELFE 1-6 als Kriterium sowie den Validierungsmaßen (Maß) als Prädiktoren untersucht. Verglichen werden die Konfidenzintervalle der standardisierten Regressionsgewichte des WLLP-Rs unter Kontrolle der anderen Validierungsmaße (KGFf und KGFv). Überlappen sich die Konfidenzintervalle nicht, kann davon ausgegangen werden, dass WLLP-R unter Kontrolle der anderen Validierungsmaße einen unterschiedlichen Zusammenhang mit TVIP zeigt wie mit dem ELFE 1-6.

*Tabelle 1.* Korrelationen mit ihren Konfidenzintervallen (95 %) zwischen Testverständnis (TVIP und ELFE), Dekodiergeschwindigkeit (WLLP-R), verbale kognitiven Grundfähigkeiten (KGFv) und figuralen kognitiven Grundfähigkeiten (KGFf) nach Erhebung (E1 & E2) und Klassenstufe (K3 & K4).

	Instrument	2	3	4	5
E1 K3	1. TVIP	.61 [.56, .66]	.46 [.40, .51]		.25 [.18, .32]
	2. ELFE 1-6 <sup>1</sup>		.60 [.56, .65]		.22 [.15, .29]
	3. WLLP-R <sub>1</sub>				.18 [.11, .25]
	4. KGFv <sup>2,3</sup>				
	5. KGFf <sup>1</sup>				
	Instrument	2	3	4	5
E2 K3	1. TVIP	.57 [.51, .63]	.49 [.38, .60]		.22 [.10, .35]
	2. ELFE 1-6 <sup>1</sup>		.66 [.60, .71]		.28 [.17, .38]
	3. WLLP-R <sub>1</sub>				.21 [.09, .33]
	4. KGFv <sup>2,3</sup>				
	5. KGFf <sup>1</sup>				
	Instrument	2	3	4	5
E2 K4	1. TVIP	.60 [.53, .66]	.47 [.37, .58]	.60 [.53, .67]	.45 [.39, .51]
	2. ELFE		.54 [.47, .61]	.61 [.54, .67]	.30 [.21, .39]
	3. WLLP			.46 [.37, .55]	.24 [.16, .31]
	4. KGFv				.42 [.35, .48]
	5. KGFf				

*Anmerkungen:* Alle Korrelationen sind signifikant ( $p < .001$ ). Klammern enthalten die untere und obere Schwelle des 95-prozentigen Konfidenzintervalls. <sup>1</sup> Korrelation der einzelnen Testversionen (A, B) weichen nicht signifikant von den berichteten Korrelationen ab. <sup>2</sup> KGF verbal sind aufgrund von geringer interner Konsistenz ( $\alpha = .43$ ) nicht berichtet. <sup>3</sup> KGF verbal wurde in E2 K3 nicht erhoben.

*Tabelle 2.* Tabelle mit Tests der Hypothesen (H) auf Unterschiedlichkeit zwei Korrelationen bzw. zu einem Punktwert in den drei Teilstichproben (E1 K3, E2 K3 & E2 K4).

Hypothesen (H)	E1 K3	E2 K3	E2 K4
<b>H1:</b>			
$r_{\text{TVIP-ELFE}} > r_{\text{TVIP-WLLP}}$	$t = 7.04, p < .001$	$t = 1.98, p = .049$	$t = 4.13, p < .001$
$r_{\text{TVIP-KGFv}} > r_{\text{TVIP-KGFf}}$	-	-	$t = 4.26, p < .001$
$r_{\text{TVIP-KGFf}} > r_{\text{TVIP-WLLP}}$	$t = 6.12, p < .001$	$t = 4.04, p < .001$	$t = 0.47, p = .641$
<b>H2:</b>			
$r_{\text{TVIP-ELFE}} > .50^*$	$t = 4.43, p < .001$	$t = 2.45, p = .015$	$t = 2.95, p < .001$
<b>H3:</b>			
$r_{\text{TVIP-WLLP}} > .30^*$	$t = 5.47, p < .001$	$t = 3.44, p = .001$	$t = 3.19, p = .001$
$r_{\text{TVIP-KGFv}} > .30^*$	-	-	$t = 8.61, p < .001$
<b>H4:</b>			
$r_{\text{TVIP-KGFf}} < .50$	$t = -6.84, p < .001$	$t = -4.36, p < .001$	$t = -1.64, p = .102$

*Anmerkungen:* *t*-Werte durch Williams *t* bestimmt. \*Schwellenwerte basieren auf sprachlichen Konventionen nach Cohen (1988;  $r = 0.1$  für eine geringe Korrelation,  $r = 0.3$  für eine mittlere Korrelation,  $r = 0.5$  für eine hohe Korrelation).

Tabelle 3. Vergleich von zwei Regressionsmodellen mit gemischten mit Textverständnis gemessen mit dem Itempool (TVIP) und dem ELFE 1-6 als Kriterium sowie den Validierungsmaßen (Maß) als Prädiktoren.

Kriterium	Stichprobe	Maß	$\beta$	[UG OG]	$t$	$p$	$mR^2$	$\sigma^2_k$	$N_{sus}$	mis %	$N_k$
TVIP	E1 K3	KGff	0.1	[0.15 0.20]	6.2	<.00	.23	.05	103	6.31	5
		WLLP-R	0.4	[0.39 0.45]	14.9	<.00					7
			2		9	1					7
	E2 K3	KGff	0.1	[0.07 0.17]	2.2	.02	.27	.06	278	0	2
		WLLP-R	0.4	[0.43 0.54]	8.9	<.00					7
			2		3	3					2
	E2 K4	KGff	0.2	[0.19 0.26]	6.7	<.00	.44	.05	566	3.89	3
		KGfV	0.4	[0.36 0.44]	10.5	<.00					6
		WLLP-R	0.2	[0.19 0.27]	5.7	<.00					7
		0		1	1	6					
		3		5	1						
		3		5	1						

  

	Stich- probe	Maß	$\beta$	[UG OG]	$t$	$p$	$mR^2$	$\sigma^2_k$	$N_{sus}$	mis %	$N_k$
ELFE1-6	E1 K3	KGff	0.1	[0.08 0.13]	4.3	<.00	.35	0.0	103	6.58	5
		WLLP-R	0.5	[0.54 0.59]	22.2	<.00					9
			6		5	1					7
	E2 K3	KGff	0.1	[0.07 0.16]	2.5	.01	.47	0.0	274	1.44	2
		WLLP-R	0.6	[0.63 0.72]	14.3	<.00					1
			8		4	1					2
	E2 K4	KGff	0.0	[0.00 0.07]	1.1	.24	.52	0.1	566	3.89	3
		KGfV	0.4	[0.38 0.45]	11.1	<.00					5
		WLLP-R	0.5	[0.47 0.55]	12.5	<.00					9
		4		4	1	6					
		1		3	1						
		1		3	1						

Anmerkungen: Aufgrund des geringen Anteils fehlender Werte wurden Fälle listenweise ausgeschlossen. Regressionsmodelle mit gemischten Effekten (Bates, Mächler, Bolker & Walker, 2014). Teilstichproben: E = Erhebung und K = Klassenstufe.  $\beta$  = standardisierte Regressionsgewicht, [UG OG] = untere und obere Grenze des 95 %-Konfidenzintervalls,  $t$  =  $t$ -Wert,  $p$  =  $p$ -Wert.  $mR^2$  = marginale  $R^2$  der fixed-effects (Edwards, Muller, Wolfinger, Qaqish & Schabenberger, 2008).  $\sigma^2_k$  = Varianz des Klassenzufallseffekts,  $N_{sus}$  = Stichprobengröße, mis% = fehlender Werte, und  $N_k$  = Anzahl der Klassen.

## Referenzen

- Bates, D., Mächler, M., Bolker, B. & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
- Cohen, S. (1988). Perceived stress in a probability sample of the United States. In S. Spacapan & S. Os-kamp (Eds.), *The Claremont Symposium on Applied Social Psychology. The social psychology of health* (p. 31–67). Sage Publications, Inc.
- Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F. & Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in medicine*, 27(29), 6137–6157.
- Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. *Psychologische Rundschau*, 58(2), 103–117.
- Oberski, D. (2014). lavaan. survey: An R package for complex survey analysis of structural equation models. *Journal of statistical software*, 57(1), 1–27.
- R-Core-Team (2020). *R: A Language and Environment for Statistical Computing*. Wien: R Foundation for Statistical Computing.
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12. *Journal of statistical software*, 48(2), 1–36.
- Williams, E. J. (1959) *Regression analysis*. New York: Wiley.