Pitt Hild[1]
Christoph Gut[1]
Susanne Metzger[2]
Josiane Tardent[1]

[1]Pädagogische Hochschule Zürich
[2]Pädagogische Hochschule FHNW

Zur Generalisierbarkeit bei Experimentiertests

G-Studien

Neben der klassischen und der probabilistischen Testtheorie (IRT) (Brennan, 2011; Kim & Wilson, 2009), liefern G-Studien (Brennan, 1996; Cronbach, Rajaratnam & Gleser, 1963) wichtige Argumente zur Generalisierbarkeit eines Testinstruments. Bei diesen Studien werden die Einflüsse mehrerer Fehlerquellen, hier als *Facetten* (Cardinet, Tourneur, & Allal, 1976) bezeichnet, auf die Kompetenzmessung gleichzeitig untersucht und mit Hilfe eines Generalisierbarkeitskoeffizienten (ρ^2) Aussagen zur Reproduzierbarkeit und Konsistenz des Messinstrumentes ausformuliert. Hierbei wird die gesamte Varianz, von der Datenmatrix ausgehend, dank unterschiedlicher Schätzer (ANOVA, maximum likelihood, MINQUE, …) in Varianzkomponenten zerlegt, welche gezielte Aussagen über den Einfluss der einzelnen Facetten erlauben (Brennan, 2000):

$$(1) \qquad \hat{\sigma}^2_{X_{pto}} = \hat{\sigma}^2_p + \hat{\sigma}^2_t + \hat{\sigma}^2_o + \hat{\sigma}^2_{pt} + \hat{\sigma}^2_{po} + \hat{\sigma}^2_{to} + \hat{\sigma}^2_{pto,e}$$

Im Beispiel der Gleichung (1) wurde der Einfluss von Personen (*p*), Aufgaben (*t*) und Messzeitpunkten (*o*) auf die gesamte Varianz untersucht. Im letzten Term befindet sich neben der Interaktion $p \times t \times o$ immer noch der (nicht aufgeklärte) Restfehler *e*.
Im Gegenteil zur IRT werden bei G-Studien alle Facetten als *zufällig* gesetzt (Shavelson & Webb, 1981, S.142 & 2006, S.607). Neben Personen, Aufgaben, Testzeitpunkten oder Ratern, können auch die Zuordnung der Personen zu Klassen, Schulen oder Regionen (Verschachtelungen) untersucht werden (vgl. Cronbach, Linn, Brennan & Haertel, 1997). Dank zusätzlicher D-Studien kann des Weiteren vorhergesagt werden, wie sich der G-Koeffizient verändern würde, wenn man die Anzahl der Ausprägungen einzelner Facetten verändern würde (z. B. wenn die Anzahl Aufgaben verdoppelt würde; siehe Brennan, 1996).

Personen x Aufgaben Varianz

Bei Tests, in denen experimentelle Kompetenzen von Schülerinnen und Schülern abgefragt wurden (vgl. Cronbach, Linn, Brennan & Haertel, 1997; Shavelson, Gao & Baxter, 1993; Webb, Schlackman & Sugrue, 2000), wurden sehr gute, jedoch nicht zufriedenstellende G-Koeffizienten ($\rho^2 \geq 0.8$, vgl. Gao, Shavelson, & Baxter, 1994) erzielt. Häufig lag jedoch noch „zu viel“ Varianz im Term *Personen x Aufgaben* (vgl. task-sampling variability bei Brennan, 1996; Shavelson et al., 1999). Dies bedeutet, dass die Leistungen einzelner Probanden über mehrere Aufgaben hinweg zu stark variieren. Ist dies der Fall, kann ein Experimentiertest, obwohl er generalisierbar sein mag, nicht in large-scale Leistungsmessungen eingesetzt werden, da die Anzahl zu lösender Aufgaben nicht reduziert werden kann.

In der Literatur werden unterschiedliche Lösungsvorschläge zur Reduzierung dieser Varianzkomponente vorgeschlagen: Unter anderem sollten die Anzahl Aufgaben erhöht (vgl. Miller, 1998, Shavelson et al., 1993), weitere Facetten wie die Messmethode (Shavelson et al., 1999) oder der Messzeitpunkt (Webb et al., 2000) ins Design integriert, aber auch eine noch stärkere Standardisierung der Aufgabenformate und Kodiermanuale (vgl. Solano-Flores, Jovanovic, Shavelson & Bachman, 1999) gefördert werden.

ExKoNawi – Experimentiertest: Pilot 3

Der Ansatz im Projekt ExKoNawi (Gut, Metzger, Hild & Tardent, 2014), Aufgaben bestimmten experimentellen Problemtypen zuzuordnen, sollte u. a. dazu dienen, die starken Leistungsschwankungen zwischen den unterschiedlichen Aufgaben zu reduzieren. In der dritten Pilotierungsphase (Gut, Hild, Metzger & Tardent, 2017) lösten 190 Schülerinnen (49%) und Schüler der 7. und 9. Jahrgangsstufe aus nicht-gymnasialen Anforderungsniveaus zwei Aufgaben zu vier verschiedenen Problemtypen (*skalenbasiertes Messen, kategoriengeleitetes Beobachten, effektbasiertes Vergleichen* und *fragengeleitetes Untersuchen*). Jede Aufgabe wurde von mindestens zwei Personen geratet und hohe Interrater-Reliabilitäten ($.56 \leq \kappa \leq .97$; $.79 < p_0 \leq .98$) sichergestellt.

Ergebnisse

Die einzelnen Varianzkomponenten wurden mit dem MINQUE Schätzer (Webb, Shavelson, & Haertel, 2006, 35) erzielt. Die G-Koeffizienten wurden dank einer speziellen Synthax (Mushquash & O'Connor, 2006) in SPSS berechnet.

Tabelle 1 zeigt deutlich, dass auch beim ExKoNawi-Experimentiertest sehr viel Varianz in der *Personen* × *Aufgaben* Komponente σ^2_{p*t} steckt. Der letzte Varianzterm $\sigma^2_{p*t*o,e}$ ist in beiden Fällen (Tab. 1 und 2) gleich null weil in unserer Stichprobe keine Person zu zwei unterschiedlichen Messzeitpunkten zweimal die gleiche Aufgabe gelöst hat.

Tabelle 1: Personen (190) x Aufgaben (12) x Messzeitpunkte (2)

Facette	**σ^2-Term**	**Wert**	**% σ^2**
Personen (*p*)	σ^2_p	0.101	14.7
Aufgaben (*t*)	σ^2_t	0.080	11.7
Messzeitpunkte (*o*)	σ^2_o	0.001	0.1
p × *t*	σ^2_{p*t}	0.474	69.1
p × *o*	σ^2_{p*o}	0.009	1.3
t × *o*	σ^2_{t*o}	0.021	3.1
p × *t* × o,*e*	$\sigma^2_{p*t*o,e}$	0.000	0
Rel. Fehlervarianz	σ^2_δ	0.044	
Abs. Fehlervarianz	σ^2_Δ	0.052	
	G-Koeffizient ρ^2	0.697	

Tabelle 2: Personen (190) x [Aufgaben (3) : Problemtypen (4)] x Messzeitpunkte (2)

Facette	**σ^2-Term**	**Wert**	**% σ^2**
Personen (*p*)	σ^2_p	0.096	13.8
Problemtypen (*pt*)	σ^2_{pt}	0.033	4.8
Aufgaben (*t*) in Problemtypen	$\sigma^2_{t:pt}$	0.063	9.1
Messzeitpunkte (*o*)	σ^2_o	0.003	0.4
p × *pt*	σ^2_{p*pt}	0.014	2.0
p × (*t*:*pt*)	$\sigma^2_{p*(t:pt)}$	0.463	66.7
p × *o*	σ^2_{p*o}	0.020	2.9
pt × *o*	σ^2_{pt*o}	0.002	0.3
p × (*t*:*pt*) × *o*,*e*	$\sigma^2_{p*(t:pt)*o,e}$	0.000	0
Rel. Fehlervarianz	σ^2_δ	0.033	
Abs. Fehlervarianz	σ^2_Δ	0.046	
	G-Koeffizient ρ^2	0.743	

Tabelle 2 zeigt, dass das „Nesten" von Aufgaben in Problemtypen, bei gleicher Anzahl Freiheitsgrade, zwar zu einer Verbesserung des G-Koeffizienten führt, die *Personen x Aufgaben* Varianzkomponente $\sigma^2_{p*(t:pt)}$ jedoch nur um 2% verringert werden konnte. Im Vergleich liegt die durch die Personen erzeugte Varianz bei etwa 15% und der Einfluss des Messzeitpunktes (alleine oder in Abhängigkeit mit Personen oder Aufgaben) immer unter 3%.

In Tabelle 3 wird der Einfluss zusätzlicher Facetten (Klasse, Lehrperson, Jahrgang, Schulniveau) auf die Personen untersucht. Hier wird deutlich, dass alle genannten Facetten die Leistungen weniger stark beeinflussen als die Personen selber. Man erkennt, dass der Einfluss der Klassen leicht höher ist als der Einfluss der Lehrpersonen und dass der Einfluss des Schulniveaus deutlich höher liegt als der Einfluss des Jahrgangs.

Tabelle 3: (Personen in Klassen in Lehrpersonen) bzw. (Personen in Schulniveaus in Jahrgängen) × (Aufgaben in Problemtypen)

Facette	**σ^2-Term**	**Wert**	**% σ^2**	**Facette**	**σ^2-Term**	**Wert**	**% σ^2**
Personen (*p*)	σ^2_{p}	0.109	15.6	Personen (*p*)	σ^2_{p}	0.098	14.54
LehrerIn (*te*)	σ^2_{te}	0.026		Jahrgang (*g*)	σ^2_{g}	-0.014	
Klasse (*c*) in *te*	$\sigma^2_{(c:te)}$	0.029		Niveau (*tr*) in *g*	$\sigma^2_{(tr:g)}$	0.064	
(*p*:*c*:*te*)	$\sigma^2_{(p:c:te)}$	0.054		(*p*:*tr*:*g*)	$\sigma^2_{(p:tr:g)}$	0.048	
Aufgaben (*t*)	σ^2_{t}	0.091	13.0	Aufgaben (*t*)	σ^2_{t}	0.091	13.5
Problem-typen (*pt*)	σ^2_{pt}	0.036		Problem-typen (*pt*)	σ^2_{pt}	0.035	
(*t*:*pt*)	$\sigma^2_{(t:pt)}$	0.055		(*t*:*pt*)	$\sigma^2_{(t:pt)}$	0.056	
(*p*:*c*:*te*) × (*t*:*pt*),*e*	$\sigma^2_{(p:c:te)*t}$	0.499	71.4	(*p*:*tr*:*g*) × (*t*:*pt*),*e*	$\sigma^2_{p*(t:pt),e}$	0.485	71.96

Fazit

Obwohl der Problemtypenansatz bei ExKoNawi zu einer deutlichen Steigerung des G-Koeffizienten führt, konnte keine nennenswerte Verringerung (2-3 %) des Varianzterms *Personen* × *Aufgaben* erzielt werden. Auch beim ExKoNawi-Experimentiertest bleibt dieser Term die Archillesferse der Leistungsmessung (Shavelson et al., 1999). Ähnliche Ergebnisse ergaben auch IRT-Analysen (vgl. Gut et al. 2017): Für die getestete Stichprobe, die als Novizen in Bezug auf die getesteten Kompetenzen bezeichnet werden können, fittet ein 1-dimensionales Modell (keine Problemtypen) die Daten am besten.

Literatur

Brennan, R.L. (1996). Generalizability of performance assessments. In G. W. Philips (Ed.), *Technical issues in large-scale performance assessments*. National Center for Education Statistics: Washington DC.

Brennan, R.L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4). 10.1177/01466210022031796

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24, 1–21. 0.1080/08957347.2011.532417

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of Generalizability theory: applications to educational measurement. *Journal of Educational Measurement*, *13*(2), 119-135.

Cronbach, L. J., Rajaratnam, N, & Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology*, *16*(2). 10.1111/j.2044-8317

Cronbach, L. J., Linn, R.L., Brennan, R .L. & Hartel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. Educational and Psychological Measurement, 57(3), S. 373-399.

Gao, X., Shavelson; R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: promises and problems. *Applied Measurement in Education*, *7*(4). 10.1207/s15324818ame0704_4

Gut, C, Metzger, S., Hild, P., & Tardent, J. (2014). Problemtypenbasierte Modellierung und Messung experimenteller Kompetenzen von 12- bis 15-jährigen Jugendlichen. *PhyDid B, Didaktik der Physik*, Beiträge zur DPG-Frühjahrstagung 2014.

Gut, C., Hild, P., Metzger, S., & Tardent, J. (2017). Vorvalidierung des ExKoNawi-Modells. In: C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis* (S. 328-331). Universität Regensburg.

Kim, S.C., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using Generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement*, *4*(10), 408-423.

Miller, M.D. (1998). *Generalizability of performance-based assessments*. Council of the Chief State School Officers: Washington DC.

Mushquash, C., & O'Connor, B. (2006). SPSS and SAS programs for generalizability theory analyses. *Behaviour Research Methods*, *38*(3), 542-547.

Shavelson, R. J., & Webb, N.M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology*, *34*, 133-166. 10.1111/j.2044-8317.1981.tb00625.x

Shavelson, R. J., & Webb, N. (1991). *Generalizability theory: a primer*. SAGE

Shavelson, R. J., Gao, X., & Baxter, G .P. (1993). Sampling variability of performance assessments. CRESST Report 142. University of Los Angeles, California.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E.W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*, *36*(1), 61-71. 10.1111/j.1745-3984.1999.tb00546.x

Solano-Flores, G., Jovanovic, J., Shavelson, R.J., & Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education*, *21*(3), 293–315. 10.1080/095006999290714

Webb, N. M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education*, *13*(3), 277-301. 10.1207/S15324818AME1303_4

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C.R. Rao & S. Sinharay (Eds.), *Handbook of Statistics*, Vol. 26, 1st Edition Psychometrics. Elsevier. 10.1016/S0169-7161(06)26004-8