

28 Reliabilitätsanalyse

Fortsetzung der Seiten in der 9. Auflage

Die gängigsten Verfahren zur Reliabilitätsprüfung gehen von einer schwachen Definition von Reliabilität, etwa im Sinne hoher Korrelation der Teilmessungen aus. Dem entsprechen die verbreitetsten Testverfahren, wie sie im Buch ausführlich dargestellt wurden. Da in vielen Wissenschaftsbereichen die Voraussetzungen für eine strengere Definition von Reliabilität, etwas im Sinne zusätzlicher Kriterien wie Übereinstimmung der Mittelwerte und/oder der Varianz der Einzelmessungen, Additivität der Einzelmessungen, nicht zu erreichen sind, kommen auch strengere Prüfverfahren selten zur Anwendung. Sie sind deshalb im Buch nur angedeutet. Im vorliegenden Beitrag wird der Abschnitt 28.2.2, der solche Prüfverfahren beschreibt, etwas genauer ausgeführt. 28.2.3 ergänzt die Überlegungen durch Darstellung einiger Strategien, die zur Behebung der Probleme bei nicht gegebenen Reliabilitätsvoraussetzungen angewendet werden können.

28.2.2 Weitere Statistik-Optionen

In der Dialogbox „Reliabilitätsanalyse: Statistik“ können weitere Statistiken angefordert werden.

ANOVA Tabellen.¹ Hier werden drei Arten von Tests angeboten, die dazu dienen zu prüfen, ob sich die Mittelwerte der Items signifikant unterscheiden.

- F-Test.* Es wird eine Varianzanalyse für wiederholte Messung durchgeführt. Dies ist das klassische Vorgehen, setzt aber mindestens intervallskalierte Daten voraus. Ist diese Voraussetzung nicht erfüllt, sollte einer der beiden anderen Tests verwendet werden.
- Friedman Chi-Quadrat.* Chi-Quadrat nach Friedman und Konkordanzkoeffizient nach Kendall. Dieses Angebot ist bei Rangdaten zu verwenden. Friedman Chi-Quadrat ersetzt für Rangdaten den F-Test.
- Cochrans Chi-Quadrat.* Ist für dichotomisierte Daten konzipiert. Cochrans Q ersetzt in der ANOVA-Tabelle das F.

¹ Da im Basismodul von SPSS keine Varianzanalyse bei Messwiederholung angeboten wird, kann man dieses Angebot des Menüs „Reliability“ auch allgemein für Varianzanalyse bei Messwiederholungsdesigns verwenden.

Beispiel. Der F-Test (Varianzanalyse mit wiederholter Messung. DateiA-Skala-Items) ergibt folgende Ausgabe.

ANOVA

	Quadrat-summe	df	Mittel der Quadrate	F	Sig.
Zwischen Personen	402,075	31	12,970		
Innerhalb Personen					
Zwischen Items	180,000	12	15,000	7,895	,000
Nicht standardisierte Residuen	706,769	372	1,900		
Gesamt	886,769	384	2,309		
Gesamt	1288,844	415	3,106		

Gesamtmittelwert = 2,16

Da die Items alle gleichwertige Messungen liefern sollen, müssten die Mittelwerte eigentlich gleich ausfallen. Zwischen den Messungen (Between Measures) sollten keine signifikanten Unterschiede bestehen. Auch hier wird deutlich, dass die Annahme gleicher Mittelwerte nicht zu halten ist. Die Varianz zwischen den Messungen (Between Measures) fällt bei einem F von 7,8951 signifikant aus (Prob.,0000).

Hotellings T-Quadrat und Tukeys Additivitätstest

- ❑ *Hotellings T-Quadrat.* Ein weiterer multivariater Test zur Überprüfung der Hypothese, dass alle Items der Skala den gleichen Mittelwert haben.
- ❑ *Tukeys Additivitätstest.* Testet die Annahme, dass keine multiplikative Wechselwirkung zwischen den Items existiert und schätzt eine Exponenten, mit dem die Skala potenziert werden müsste, um Additivität zu erreichen.

Wird Turkey Additivitätstest aufgerufen, ergibt sich in unserem Beispiel die in Tabelle 28.2. enthaltene Ausgabe. Die Komponente „Nonadditivity“ fällt signifikant aus. Demnach lässt sich die Annahme, dass keine multiplikative Wirkung zwischen den Items bestehe, nicht halten. Als geeigneter Exponent, mit dem die Skala potenziert werden müsste, um Additivität zu erreichen, wird $-0,053$ geschätzt.

Tabelle 28. 2. Ausgabe bei Aufruf des Additivitätstest von Turkey

ANOVA mit Test auf Nicht-Additivität nach Tukey

	Quadrat-summe	df	Mittel der Quadrate	F	Sig.
Zwischen Personen	402,075	31	12,970		
Innerhalb Personen					
Zwischen Items	180,000	12	15,000	7,895	,000
Nicht standardisierte Residuen					
Nicht-Additivität		1	41,522	23,156	,000
Ausgewogenheit	665,247	371	1,793		
Gesamt	706,769	372	1,900		
Gesamt	886,769	384	2,309		
Gesamt	1288,844	415	3,106		

Gesamtmittelwert = 2,16

Tukey-Schätzung des Exponenten, mit dem Beobachtungen potenziert werden müssen, damit gilt: Additivität = $-0,053$.

28.2.3 Varianten

Dargestellt wurde die gebräuchlichste Variante zur Konstruktion und Prüfung einer Summated Rating-Skala. Dabei werden die einzelnen Items mit Hilfe mehrstufiger Ratingskalen gemessen und als Messniveau dieser Einzelmessungen Intervallskalenniveau angenommen. Die Ergebnisse der Prüfung unserer Beispielskala sind eindeutig: Die Zuverlässigkeit – gemessen an den Reliabilitätskoeffizienten – ist befriedigend hoch, kann aber durch Ausschluss des ungeeigneten Statements 2 durchgängig erhöht werden. Weitergehende Ansprüche wie gleich Mittelwerte, gleiche Varianzen, Additivität etc. können aber nicht erfüllt werden.

Wollte man auch diese erfüllen, wären verschiedene Strategien denkbar:

- ❑ *Die Rohwerte in z-Werten transformieren.* Durch diese Standardisierung haben alle Items den Mittelwert 0 und die Streuung 1. Auch die Gesamtskala hat einen Mittelwert von 0. Additivität wird dadurch gewöhnlich gesichert. Die Reliabilitätskoeffizienten müssen dadurch aber nicht erheblich verbessert werden und auch die Übereinstimmung mit dem parallelen bzw. streng parallelen Modell wird dadurch nicht gewährleistet.
- ❑ *Die Rohwerte in Rangwerte transformieren.* Man kann daran zweifeln, ob die Daten Intervallskalenniveau besitzen. Es würde dann nur Ordinalskalenniveau unterstellt. Im Beispiel würde dies die Reliabilität etwas erhöhen und Übereinstimmung der Mittelwerte der Items sowie Additivität garantieren. Anstelle des F-Tests ist Friedmans Chi-Quadrat zu wählen. Ansonsten gilt dasselbe wie bei Verwendung der z-Transformation.
- ❑ *Potenzieren der Skala* mit dem mit Tukeys Additivitätstest ermittelten Koeffizienten. Damit kann u.U. Additivität gesichert werden.
- ❑ *Dichotome Skalen.* Cochrans Chi-Quadrat ersetzt dann das F der Varianzanalyse.