gesis Leibniz-Institut für Sozialwissenschaften

SDM Survey Guidelines

Reliabilität – die Genauigkeit einer Messung

Daniel Danner

Januar 2015, Version 1.0

Zusammenfassung

Die Reliabilität beschreibt die Genauigkeit einer Messung. In diesem Beitrag wird das Konzept Reliabilität definiert und es wird erläutert, warum die Reliabilität einer Messung relevant ist. Danach wird diskutiert, welche Modellannahmen getroffen werden müssen, um die Reliabilität einer Messung zu schätzen und es werden fünf Methoden zum Schätzen der Reliabilität vorgestellt: die Re-Test Korrelation, die Parallel-Test Korrelation, die Split-Half Korrelation, die interne Konsistenz und das Schätzen der Reliabilität mit Strukturgleichungsmodellen. Abschließend wird in knapper Form auf Gemeinsamkeiten und Unterschiede der klassischen Testtheorie und der Item-Response Theorie und deren Bedeutung für die Schätzung der Reliabilität eingegangen.

Zitierung

Danner, Daniel (2015). Reliabilität – die Genauigkeit einer Messung. Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften (SDM Survey Guidelines). DOI: 10.15465/sdm-sg_011

1. Was ist Reliabilität?

In den Sozialwissenschaften werden häufig Unterschiede zwischen Personen gemessen. Beispielweise werden im ALLBUS (z.B. Koch & Wasmer, 2004) und im European Social Survey (European Social Survey, 2014) regelmäßig Einstellungen zu politischen oder gesellschaftlichen Themen gemessen. In einigen Befragungen wie dem International Social Survey Programme (Haller, Jowell, & Smith, 2009) wird darüber hinaus auch die Persönlichkeit der Befragten erfasst. In Studien wie PISA (OECD, 2011), PIAAC (Rammstedt, 2013) oder der NEPS (Blossfeld, von Maurice & Schneider, 2011) werden die kognitiven Fähigkeiten der Teilnehmer gemessen.

Die Reliabilität beschreibt die Genauigkeit einer solchen Messung. Formal betrachtet entspricht der Reliabilitätskoeffizient R dem Verhältnis zwischen wahren Unterschieden τ und beobachteten Unterschieden Y.

$$R = \frac{Varianz\ (\tau)}{Varianz\ (Y)}$$

Wahre Unterschiede sind systematische Unterschiede in der Persönlichkeit, der Einstellung oder der Fähigkeit von Personen. Beobachtete Unterschiede können darüber hinaus von unsystematischen Einflüssen wie situativen Störungen oder zufälligen Messfehlern beeinflusst werden.

Ziel einer Messung ist es, die wahren Unterschiede zwischen Personen abzubilden. Das gelingt, wenn die Reliabilität einer Messung hoch ist. Eine hohe Reliabilität bedeutet, dass ein großer Teil der beobachteten Unterschiede auf wahre Unterschiede zurückgeführt werden kann. Eine geringe Reliabilität bedeutet hingegen, dass die beobachteten Unterschiede substantiell durch Messfehler „verunreinigt" sind. Es gibt keinen verbindlichen Schwellenwert, ab wann die Reliabilitätsschätzung einer Messung ausreichend ist. Für Gruppenuntersuchungen wird eine Reliabilität von 0.70 oft als ausreichend bezeichnet (Rammstedt, 2004), eine Reliabilität von 0.80 wird in der Regel als gut bezeichnet (Nunnally & Bernstein, 1994; Weise, 1975) und eine Reliabilität über 0.90 als hoch (Weise, 1975).

Die geschätzte Reliabilität ist immer die Eigenschaft einer Messung, nicht die eines Instruments. Ein Instrument kann in verschiedenen Stichproben unterschiedlich reliable Messungen hervorbringen. In einer sehr homogenen Stichprobe, in denen es kaum wahre Unterschiede zwischen Personen gibt, kann die Reliabilität kleiner sein als in einer heterogenen Stichprobe mit bedeutsamen Personenunterschieden. Zum Beispiel kann die Messung von politischen Einstellungen in Extremgruppen weniger reliabel sein als die Messung von politischen Einstellungen in einer heterogenen Stichprobe mit demselben Instrument. In der Praxis wird daher häufig versucht, die Reliabilität in einer repräsentativen Stichprobe zu schätzen. Liegt eine Reliabilitätsschätzung aus einer repräsentativen Stichprobe vor, kann angenommen werden, dass die Reliabilität in der Population vergleichbar ist.

2. Warum ist Reliabilität einer Messung relevant?

Die Reliabilität einer Messung ist relevant, wenn Zusammenhänge zwischen verschiedenen Variablen betrachtet werden, aber auch wenn der Messwert einer einzelnen Person betrachtet wird.

Viele Forschungsfragen beschäftigen sich damit, wie verschiedene Konstrukte zusammenhängen. In den Politikwissenschaften wird beispielsweise untersucht, wie Einstellungen mit Wahlverhalten

zusammenhängen (z.B. Wüst, 2002), in der Psychologie wird untersucht, wie kognitive Fähigkeiten mit Berufserfolg zusammenhängen (z.B. Schmitt & Hunter, 2004) oder Persönlichkeit mit Verhalten (z.B. Hossiep & Mühlhaus, 2005). Die Stärke eines solchen Zusammenhangs wird dabei häufig mit Korrelationen untersucht. Die Höhe einer Korrelation wird jedoch von der Reliabilität der Messungen limitiert. Eine hohe Reliabilität ermöglicht eine hohe Korrelation, eine niedrige Reliabilität ermöglicht nur eine geringe Korrelation. Dieser Zusammenhang kann quantifiziert werden: die Korrelation *r* einer Variablen mit einer anderen Variablen kann maximal so hoch sein, wie die Wurzel deren Reliabilität *R*:

$$r_{max} = \sqrt{R}$$

Wird beispielsweise die Einstellung gegenüber Politikern mit einer Reliabilität von $R = 0.90$ erhoben, kann die gemessene Einstellung maximal zu $r_{max} = \sqrt{0.90} = 0.95$ mit einer anderen Variablen korrelieren. Wird die Einstellung mit einer Reliabilität von nur $R = 0.50$ gemessen, kann die gemessene Einstellung maximal zu $r_{max} = \sqrt{0.50} = 0.70$ mit einer anderen Variablen korrelieren.

Darüber hinaus ist die Reliabilität einer Messung relevant, wenn der Messwert einer einzelnen Person betrachtet wird. Eine genaue Messung mit hoher Reliabilität ermöglicht eine präzise Schätzung eines Personenwertes. Der wahre Wert und der beobachtete Wert sind sich dann sehr ähnlich. Eine ungenaue Messung mit geringer Reliabilität ermöglicht nur eine unpräzise Schätzung. Der wahre Wert und der beobachtete Wert können dann bedeutsam voneinander abweichen. Die Abweichung zwischen wahrem Wert und beobachtetem Wert kann anhand des Konfidenzintervalls (oder Vertrauensintervalls) einer Messung quantifiziert werden. Das Vertrauensintervall gibt an, in welchem Bereich der wahre Wert einer Person liegt, wenn der beobachte Wert (*Y*), die Reliabilität der Messung (*R*) und die Standardabweichung (*SD*) des Tests bekannt ist. Das 95% Konfidenzintervall (*KI*) einer Messung kann dann mit folgender Formel geschätzt werden:

$$KI = Y \pm 1.96 * SD * \sqrt{1 - R}$$

Ein Beispiel: Eine Person bearbeitet einen Intelligenztest und die Messung ergibt einen Intelligenzquotient (IQ) von IQ = 111. Die Reliabilität der Messung beträgt $R = 0.90$, die Standardabweichung $SD = 15$. Das 95%-Konfidenzintervall liegt damit zwischen $111 - 1.96 * 15 * \sqrt{1 - 0.90} = 102$ und $111 + 1.96 * 15 * \sqrt{1 - 0.90} = 120$. Würde die Reliabilität des Test nur $R = 0.50$ betragen, wäre die Messung ungenauer und das Konfidenzintervall größer und läge zwischen $111 - 1.96 * 15 * \sqrt{1 - 0.50} = 90$ und $111 + 1.96 * 15 * \sqrt{1 - 0.50} = 132$.

3. Welche Modellannahmen müssen getroffen werden, um die Reliabilität zu schätzen?

Die Reliabilität einer Messung beschreibt das Verhältnis von wahren Merkmalsunterschieden τ zu beobachteten Merkmalsunterschieden *Y*. Der wahre Wert einer Messung kann nicht beobachtet werden. Jedoch kann die Varianz der wahren Werte geschätzt werden, wenn bestimmte Modellannahmen gemacht werden. Ausgangspunkt für verschiedene Messmodelle ist die klassische Testtheorie. Die klassische Testtheorie besagt im Kern, dass sich ein beobachteter Wert *Y* aus einem wahren Wert τ und einem Messfehler ε zusammensetzt (Bühner, 2011; Lord & Novick, 1968; Steyer & Eid, 2001):

$$Y = \tau + \varepsilon$$

Innerhalb der klassischen Testtheorie können verschiedene Messmodelle unterschieden werden:

3.1 Das Parallele Messmodell

Das sparsamste Messmodell ist das parallele Messmodell. Das parallele Messmodell besagt, dass sich ein beobachteter Wert *Y* aus einem wahren Wert τ und einem Messfehler ε zusammen setzt. Weiterhin besagt das Modell, dass mehrere Messungen *i* (von Items oder Tests) denselben wahren Wert τ_i besitzen ($\tau := \tau_i$) und dass die Fehlervarianzen mehrerer Messungen identisch sind ($s_\varepsilon^2 := s_{\varepsilon_i}^2$). Ein Beispiel für ein paralleles Messmodell mit zwei Messungen ist in Abbildung 1 dargestellt. Ein solches Messmodell ist sparsam, macht aber auch die restriktive Annahme, dass die Varianzen mehrerer Messung identisch sind. Diese Annahme muss jedoch nicht erfüllt sein.

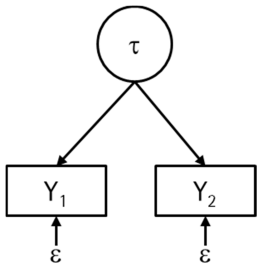

Abbildung 1: Paralleles Messmodell mit zwei Messungen

3.2 Das τ-äquivalente Messmodell

Das τ-äquivalente Messmodell ist weniger restriktiv und besagt, dass mehrere Messungen *i* denselben wahren Wert τ_i besitzen ($\tau := \tau_i$), die Varianzen der Messfehler ($s_{\varepsilon_i}^2$) sich aber unterscheiden dürfen. Ein τ-äquivalentes Messmodell besitzt mehr Parameter, die geschätzt werden müssen. Daher sind mindestens drei Messungen notwendig, um die Modellparameter zu schätzen. Ein Beispiel mit drei Messungen ist in Abbildung 2 dargestellt. Ein solches Modell ist weniger restriktiv, da es zulässt, dass die Messfehler bei unterschiedlichen Messungen unterschiedlich groß sind. Das Modell fordert jedoch, dass der wahre Wert verschiedener Messungen identisch ist. Auch diese Annahme kann verletzt sein.

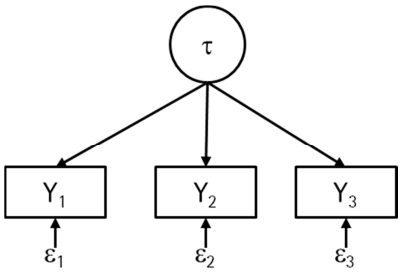

Abbildung 2. τ -äquivalentes Messmodell mit drei Messungen

3.3 Das τ-kongenerische Messmodell

Das τ-kongenerische Messmodell ist weniger restriktiv und besagt, dass sich die Varianzen der Messfehler ($s_{\varepsilon_i}^2$) unterscheiden dürfen und die wahren Werte mehrerer Messungen linear ineinander überführbar sind ($\tau := \lambda_i * \tau_i$). Zwei wahre Werte sind dann ineinander überführbar, wenn ein Wert durch Multiplizieren in den anderen Wert überführt werden kann (z.B. $\tau_2 = 0.75 * \tau_1$). Ein τ-kongenerisches

Messmodell besitzt mehr Parameter. Daher sind mindestens vier Messungen notwendig, um die Modellparameter zu schätzen. Ein Beispiel mit vier Messungen ist in Abbildung 3 dargestellt. Ein solches Modell erlaubt, dass unterschiedliche Messungen den wahren Wert in unterschiedlichem Ausmaß abbilden und dass der Einfluss von Messfehlern bei unterschiedlichen Messungen unterschiedlich groß ist.

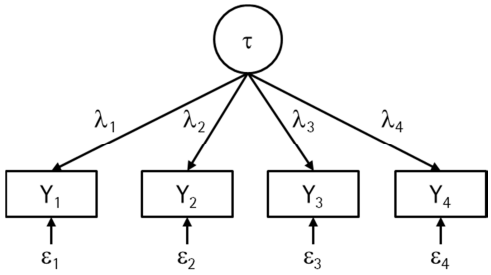

Abbildung 3. τ-kongenerisches Messmodell mit vier Messungen

Die verschiedenen Messmodelle ermöglichen unterschiedliche Methoden, die Reliabilität einer Messung zu schätzen. Welches Messmodell gilt, kann mit Strukturgleichungsmodellen (z.B. Bollen, 1989; Eid, Gollwitzer & Schmitt, 2010; Tabachnick & Fidell, 2013) überprüft werden.

4. Wie kann die Reliabilität einer Messung geschätzt werden?

Die Reliabilität einer Messung kann mit verschiedenen Methoden geschätzt werden. Im Folgenden werden die gebräuchlichsten Schätzmethoden vorgestellt: die Re-Test Korrelation, die Parallel-Test Korrelation, die Split-Half Korrelation, die interne Konsistenz und die Schätzung der Reliabilität mit Strukturgleichungsmodellen.

4.1 Die Re-Test Korrelation

Die Re-Test Korrelation kann verwendet werden, wenn ein Instrument von denselben Personen zu zwei Messzeitpunkten bearbeitet wird. Die Reliabilität R beider Messungen (Y_1 und Y_2) kann dann über die Korrelation r der beiden Messungen geschätzt werden:

$$R = r_{Y_1,Y_2}$$

Die Re-Test Korrelation liefert eine zuverlässige Schätzung der Reliabilität, wenn das parallele Messmodell gilt. Das bedeutet, zu beiden Messzeitpunkten muss derselbe wahre Wert gemessen werden. In der Praxis bedeutet das, der wahre Wert darf sich zwischen den Messungen nicht verändern. Diese Annahme ist bei zeitlich stabilen Konstrukten (wie z.B. Intelligenz oder Extraversion) plausibel. Bei Konstrukten, die sich über die Zeit verändern können (wie z.B. Stimmungen oder Einstellungen) ist diese Annahme weniger plausibel. Wird die Re-Test Korrelation verwendet, muss angenommen werden, dass es keine Übungs- oder Erinnerungseffekte zwischen den Messungen gibt. Die zweite Annahme, die gemacht werden muss ist, dass Varianz des Messfehlers bei beiden Messungen identisch ist. In der Praxis bedeutet das, dass Störeinflüsse wie Lärm bei der Bearbeitung oder Müdigkeit der Probanden bei beiden Messungen gleich groß sind. Diese Annahmen können mit Strukturgleichungsmodellen (z.B. Bollen, 1989; Eid, Gollwitzer & Schmitt, 2010; Tabachnick & Fidell, 2013) überprüft werden. Ein Beispiel für ein

solches Strukturgleichunsgmodell ist in Abbildung 1 dargestellt. In der Praxis wird jedoch häufig auf eine Überprüfung verzichtet.

4.2 Die Parallel-Test Korrelation

Die Parallel-Test Korrelation kann verwendet werden, wenn zwei Parallelformen eines Instruments vorliegen und dieselben Probanden beide Parallelformen bearbeiten. Parallelformen messen exakt denselben wahren Wert und werden in gleicher Weise von Messfehlern beeinflusst. Beispiele für Parallelformen sind die Form A und C des Intelligenz-Struktur-Tests (Liepmann, Beauducel, Brocke & Amthauer, 2007) oder die beiden Formen des Berufseignungstests von Schmale (2001). Liegen zwei Parallelformen eines Instruments vor, kann die Reliabilität R beider Messungen Y_1 und Y_2 über die Korrelation r zwischen beiden Messungen geschätzt werden:

$$R = r_{Y_1,Y_2}$$

Die Parallel-Test Korrelation liefert eine zuverlässige Schätzung der Reliabilität, wenn das parallele Messmodell gilt. Das bedeutet, beide Parallelformen müssen exakt denselben wahren Wert messen und in gleicher Stärke von zufälligen Messfehlern beeinflusst werden. In der Praxis ist es oft schwierig Parallelversionen eines Instruments zu konstruieren, da verschiedene Items oft unterschiedliche Aspekte eines Konstrukts erfassen. Es muss daher gut begründet werden, warum zwei Instrumente parallele Messungen erlauben. Die Parallelität zweier Messungen kann ebenfalls mit einem Strukturgleichungsmodell, wie es in Abbildung 1 dargestellt ist, überprüft werden.

4.3 Die Split-Half Korrelation

Die Split-Half Korrelation kann zum Schätzen der Reliabilität verwendet werden, wenn ein Instrument nur einmal eingesetzt wurde. Dazu wird ein Instrument in zwei Testhälften geteilt und es wird die Annahme gemacht, dass beide Testhälften parallele Messung liefern. Die Split-Half Korrelation wird häufig eingesetzt, wenn ein Instrument aus mehreren Items besteht. Die Items werden dann in zwei gleich große Testhälften eingeteilt. Diese Einteilung kann auf verschiedene Weise vorgenommen werden. Üblich ist eine Einteilung nach geraden und ungeraden Itemnummern (odd-even split), eine Einteilung in eine erste und eine zweite Testhälfte, eine Einteilung nach Itemkennwerten (Methode der statistischen Zwillinge) oder eine zufällige Einteilung der Items. Die Korrelation zwischen den beiden Testhälften schätzt dann die Reliabilität der Testhälften. Diese Schätzung wird mit der Spearman-Brown Formel korrigiert, um eine Schätzung der Reliabilität der Messung des gesamten Instruments zu erhalten (siehe auch Amelang & Schmidt-Atzert, 2006; Bühner, 2011). Die Reliabilität R einer Messung Y, die in zwei Testhälften (Y_1,Y_2) unterteilt wurde, kann dann mit folgender Formel geschätzt werden:

$$R = \frac{2 * r_{Y_1,Y_2}}{1 + r_{Y_1,Y_2}}$$

Die Spearman-Brown Korrektur muss in der Regel nicht manuell vorgenommen werden. Statistikprogramme wie SPSS liefern bereits die korrigierte Reliabilitätsschätzung. Diese wird dann als Spearman-Brown Koeffizient (SPSS) bezeichnet.

Eine weitere Möglichkeit die Split-Half Korrelation zu berechnen bietet die Methode des Maximal Split-Half Coefficients (z.B. Callendar & Osburn, 1979; Hunt & Bentler, 2012). Grundlage für diese Methode ist der Vorschlag von Guttman (1945), einen Test in alle möglichen Testhälften (mit gleicher und ungleicher Itemanzahl) zu zerlegen und alle Korrelationen zu berechnen. Sind diese Testhälften parallel, bietet die höchste Korrelation den besten Schätzer für die Reliabilität des Gesamttests. Ein praktisches Problem bei diesem Vorgehen ist, dass eine Einteilung in alle möglichen Testhälften sehr rechenintensiv ist. So gibt es bei zehn Item $2^{10-1} - 1 = 511$ mögliche Kombinationen, bei 25 Items bereits $2^{25-1} - 1 =$

16,777,215 mögliche Kombinationen. Daher schlagen Hunt und Bentler (2012) vor, eine Stichprobe möglicher Kombinationen (z.B. 10,000 Kombinationen) zu bilden und auf Grundlage dieser Stichprobe die Reliabilität zu schätzen. Die Methode wird bei Hunt und Bentler (2012) ausführlich beschrieben. Eine Schätzung der Reliabilität mit dem Maximal Split-Half Coefficient kann mit dem R-Paket Lambda4 (Hunt, 2013) durchgeführt werden.

Unabhängig davon, wie die Testhälften gebildet werden, setzt die Split-Half Korrelation voraus, dass beide Testhälften parallele Messungen erlauben. Diese Annahme kann mit einem Strukturgleichungsmodell, wie in Abbildung 1 dargestellt, überprüft werden.

4.4 Die interne Konsistenz

Die interne Konsistenz ermöglicht eine Schätzung der Reliabilität, wenn ein Instrument nur einmal eingesetzt wurde. Ausgangspunkt für die Schätzung der Reliabilität mit der internen Konsistenz ist, dass ein Test in einzelne Items zerlegt wird. Die Reliabilität kann dann über die Varianzen (s^2) der Items i und der Varianz des Summenwerts der Items geschätzt werden. Der gebräuchlichste Kennwert zum Schätzen der internen Konsistenz ist Cronbachs α (z.B. Amelang & Schmidt-Atzert, 2006; Bühner, 2011). Die Reliabilität R einer Messung Y, die aus k Items besteht kann dann über folgende Formel geschätzt werden:

$$R = \frac{k}{k-1} * \left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_Y^2}\right)$$

Cronbachs α muss in der Regel nicht manuell berechnet werden, sondern kann mit Statistikprogrammen wie SPSS, SAS oder STATA automatisch berechnet werden.

Cronbachs α ist ein zuverlässiger Schätzer für die Reliabilität einer Messung, wenn die Items τ-äquivalent sind, d.h. alle Items denselben wahren Wert abbilden. Ist diese Voraussetzung verletzt, unterschätzt Cronbachs α die Reliabilität und schätzt nur eine untere Schranke der Reliabilität (Cortina, 1993; Lord & Novick, 1968). Für die Praxis bedeutet dies, dass die Reliabilität einer Messung höher sein kann als Cronbachs α. Weiterhin fordert ein τ-äquivalentes Messmodell, dass die Items nur aufgrund ihrer wahren Werte miteinander kovariieren. Die Messfehler der einzelnen Items dürfen nicht miteinander kovariieren. Anders ausgedrückt, die Items müssen eine eindimensionale Struktur aufweisen: der wahre Wert der Personen ist der einzige Faktor, der die Kovarianz zwischen den Items erklärt. Ist diese Annahme verletzt, ermöglicht Cronbachs α keine zuverlässige Schätzung der Reliabilität. Ob die Items eines Instruments τ-äquivalent sind, kann mit Strukturgleichungsmodellen überprüft werden.

4.5 Strukturgleichungsmodelle

Liegen keine parallelen oder τ-äquivalenten, sondern nur τ-kongenerische Messungen vor, kann die Reliabilität einer Messung mit Strukturgleichungsmodellen geschätzt werden. Diese Methode bietet den Vorteil, dass gleichzeitig die Reliabilität geschätzt wird und die zugrunde gelegten Modellannahmen überprüft werden. Im Folgenden wird die Reliabilitätsschätzung der Composite Reliability nach Raykov (1997) dargestellt. Um die Reliabilität von τ-kongenerischen Messungen mit Strukturgleichungsmodellen zu schätzen, muss ein Instrument mit mindestens vier Items vorliegen. Weiterhin benötigen Strukturgleichungsmodelle verhältnismäßig große Stichproben mit mindestens $N = 200$ Personen (Hoyle, 1995). Sind diese Voraussetzungen erfüllt, kann ein Strukturgleichungsmodell, wie in Abbildung 4 dargestellt, dazu verwendet werden, die Reliabilität zu schätzen. Das abgebildete Strukturgleichungsmodell enthält neben dem Messmodell noch eine Phantomvariable M, die dem

Summen- oder Mittelwert der Items entspricht. Diese Phantomvariable macht das Schätzen der Reliabilität einfacher.

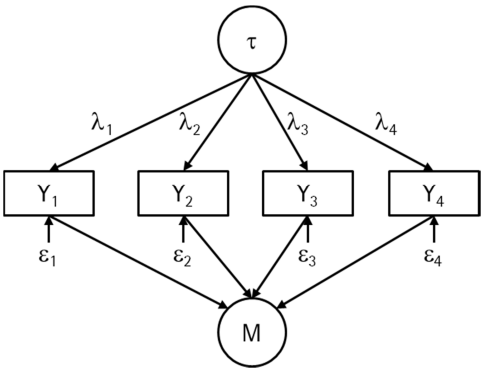

Abbildung 4. τ-kongenerisches Messmodell mit vier Items und einer Phantomvariablen

Die Reliabilität dieses Summen- oder Mittelwerts kann dann mit den Modellparametern geschätzt werden. Dazu wird die Varianz der latenten Variable τ mit den Ladungen λ_i der Items gewichtet und durch die Varianz s^2 der Phantomvariablen *M* geteilt:

$$R = \frac{\sum(\lambda_i^2 * s_\tau^2)}{s_M^2}$$

Die Parameter eines Strukturgleichungsmodells können mit Programmen wie Amos, Mplus, SAS oder R geschätzt werden. Die geschätzten Parameter können dann in die Formel eingesetzt werden. Das Vorgehen wird ausführlich bei Raykov (1997) beschrieben.

Die Composite Reliability (Raykov, 1997) liefert eine zuverlässige Schätzung der Reliabilität, wenn das τ-kongenerische Messmodell gilt (aber auch, wenn das τ-äquivalente oder parallele Messmodell gilt). Das bedeutet, die Zusammenhänge zwischen den Items werden nur durch den wahren Wert der Items erklärt. Die Messfehler der Items dürfen nicht zusammenhängen. Ob ein τ-kongenerisches Messmodell gilt, kann anhand der Passung des Strukturgleichungsmodells überprüft werden. Übliche Kennwerte für die Passung eines Modells sind der Root Mean Square of Approximation (RMSEA) und der Comparative Fit Index (CFI). Ein RMSEA < .06 und ein CFI > .95 zeigen eine gute Passung des Modells an (Hu & Bentler, 1998).

4.6 Vergleich verschiedener Methoden

Die Reliabilität einer Messung kann mit verschiedenen Methoden geschätzt werden. Welche Methode die geeignetste ist, hängt davon ab, welches Messmodell angenommen werden kann. Wird ein Instrument zu zwei Messzeitpunkten eingesetzt und es kann angenommen werden, dass beide Messungen parallel sind, also denselben wahren Wert abbilden und dieselben Fehlervarianzen haben, kann die Reliabilität mit der Re-Test Korrelation geschätzt werden. Liegen zwei parallele Messinstrumente vor, kann die Reliabilität mit der Parallel-Test Korrelation geschätzt werden. Kann ein Instrument in zwei parallele Testhälften zerlegt werden, kann die Reliabilität mit der Split-Half Korrelation geschätzt werden. Wenn die Items eines Instruments die Messung desselben wahren Wertes erlauben (die Items also τ-äquivalent sind) kann die Reliabilität mit der internen Konsistenz der Items geschätzt werden. In Fällen, in denen die Items eines Tests nicht τ-äquivalent, sondern τ-kongenerisch sind, kann die Reliabilität mit Strukturgleichungsmodellen geschätzt werden. Die unterschiedliche

Schätzmethoden machen unterschiedliche Annahmen. Daher können verschiedene Schätzungen voneinander abweichen. Das bedeutet jedoch nicht, dass eine Messung unterschiedliche Reliabilitäten hat. Eine Messung hat immer nur eine Reliabilität. Es soll immer die Schätzmethode verwendet werden, die für die erhobenen Daten am geeignetsten ist. Welches Messmodell gilt und welche Methode die geeignetste ist, kann mit Strukturgleichungsmodellen überprüft werden (z.B. Bollen, 1989; Eid, Gollwitzer & Schmitt, 2010; Tabachnick & Fidell, 2013).

5. Kann die Reliabilität einer Messung geschätzt werden, auch wenn ein Instrument mit der Item-Response Theorie entwickelt wurde?

Ja, das ist möglich. Die klassische Testtheorie und die Item-Response Theorie entstammen unterschiedlichen Forschungstraditionen. Beide Theorien bieten unterschiedliche Sichtweisen auf die Güte einer Messung. Diese Sichtweisen sind aber nicht widersprüchlich, sondern ergänzen sich gegenseitig.

Die Klassische Testtheorie beschreibt einen beobachteten Wert *Y* als Kombination von wahrem Wert τ und Messfehler ε

$$Y = \tau + \varepsilon$$

Der beobachtete Wert ist in der Regel der Summen- oder Mittelwert einer Skala. Dieser Wert wird als intervallskalierte, normalverteilte Variable behandelt. Der wahre Wert τ beschreibt ein Personenmerkmal wie eine Fähigkeit oder eine Einstellung. Die Genauigkeit einer Messung kann dann über den Reliabilitätskoeffizienten bestimmt werden. Der Reliabilitätskoeffizient ist definiert als das Ausmaß wahrer Unterschiede zu beobachteten Unterschieden.

Die Item-Response Theorie beschreibt nicht, wie sich ein beobachteter Wert zusammensetzt, sondern die Wahrscheinlichkeit, einen bestimmen Wert zu beobachten. Im einfachsten Messmodell der Item-Response Theorie, dem Rasch-Modell, hängt die Wahrscheinlichkeit *P* von der Itemschwierigkeit σ und der Personenfähigkeit θ ab:

$$P(Y = 1|\theta, \sigma) = \frac{\exp(\theta - \sigma)}{1 + \exp(\theta - \sigma)}$$

Der beobachtete Wert ist eine Antwortkategorie eines Items (z.B. $Y = 1$). Dieser Wert wird als kategoriale (im einfachsten Fall dichotome) Variable behandelt. Der Itemschwierigkeitsparameter σ beschreibt die Eigenschaft eines Items. Der Personenfähigkeitsparameter θ beschreibt ein Merkmal der Person (z.B. eine Fähigkeit oder eine Einstellung). Der Personenparameter θ kann aus beobachteten Daten geschätzt werden und die Genauigkeit einer Messung kann über den Standardmessfehler dieser Schätzung bestimmt werden. Eine Besonderheit der Item-Response Theorie ist, dass der Standardmessfehler an unterschiedlichen Abschnitten des Fähigkeitskontinuums unterschiedlich groß sein kann. In der Praxis bedeutet dies, dass die Genauigkeit einer Messung in einem mittleren Fähigkeitsbereich genauer sein kann, als die Messung einer Fähigkeit in einem extremen Bereich. Die Schätzung des Personenparameters ist z.B. bei Embretson und Reise (2000) dargestellt.

Mit der Item-Response kann also eine differenzierte Aussage über die Genauigkeit einer Messung gemacht werden. Oft interessiert aber nicht, die Genauigkeit der Messung einer einzelnen Person, sondern die Genauigkeit mehrerer Messungen in einer Stichprobe. Daher berechnen verschiedene Softwarepakete zusätzlich die Average Variance Extracted (AVE). Die AVE beschreibt den durchschnittlichen Varianzanteil der manifesten Variablen, der durch die Personenfähigkeitsvariable erklärt werden kann.

Außerdem ist es möglich, „klassische" Methoden zum Schätzen der Reliabilität zu nutzen. In der Praxis bedeutet dies: auch wenn ein Instrument mit einem Rasch-Modell skaliert wurde, kann die Re-Test Korrelation oder die Split-Half Korrelation genutzt werden, um die Reliabilität des Summen- oder Mittelwerts der Items zu schätzen.

6. Literaturverzeichnis

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Heidelberg: Springer.

Blossfeld, H.-P., von Maurice, J. & Schneider, T. (2011). *Grundidee, Konzeption und Design des Nationalen Bildungspanels für Deutschland* (NEPS Working Paper No. 1). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Bollen, K. A. (1989). *Structural equations with latent variables*. Oxford England: John Wiley & Sons.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson.

Callendar, J. C. & Osburn, H. G. (1979). An empirical comparison of coefficient alpha, Guttman's lambda – 2, and MSPLIT maximized split-half reliability estimates. *Journal of Educational Measurement, 16*, 89-99. doi: 10.1111/j.1745-3984.1979.tb00090.x

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104. doi: 10.1037/0021-9010.78.1.98

Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden*. Weinheim: Belz.

Embretson, S. E. & Reise, S. P. (2000). *Item respose theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

European Social Survey (2014). *ESS Round 6 (2012/2013)* Technical Report. London: Centre for Comparative Social Surveys, City University London.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

Haller, M., Jowell, R., & Smith, T. W. (2009). The International Social Survey Programme. New York: Routledge.

Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA US: Sage Publications, Inc.

Hossiep, R. & Mühlhaus, O. (2005). *Personalauswahl und -entwicklung mit Persönlichkeitstests.* Göttingen: Hogrefe.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3(4)*, 424-453. doi: 10.1037/1082-989x.3.4.424

Hunt, T. D. & Bentler, P. (2012). *Quantile Lower Bounds to Reliability Based on Splits*. Retrieved from the University of California, Los Angeles Website: http://statistics.ucla.edu/preprints/uclastat-preprint-2012:5

Hunt, T. (2013). *Package 'Lambda4'*. Retrieved from http://cran.r-project.org/web/packages/Lambda4/Lambda4.pdf

Koch, A. & Wasmer, M. (2004). Der ALLBUS als Instrument zur Untersuchung sozialen Wandels: Eine Zwischenbilanz nach 20 Jahren. In: R. Schmitt-Beck, M. Wasmer, & A. Koch (Hrsg.): *Sozialer und*

politischer Wandel in Deutschland. Analysen mit ALLBUS-Daten aus zwei Jahrzehnten. 2004, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 13-41, Blickpunkt Gesellschaft, Band 7.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *I-S-T 2000 R - Intelligenz-Struktur-Test 2000 R* (2. Aufl.). Göttingen: Hogrefe.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Welsley.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory* (3. Ed). New York: McGraw-Hill.

OECD. (2011). PISA 2009 Ergebnisse: Potenziale nutzen und Chancengerechtigkeit sichern – Sozialer Hintergrund und Schülerleistungen (Band II). doi: 10.1787/9789264095359-de

Rammstedt, B. (2004). *Zur Bestimmung der Güte von Multi-Item-Skalen: Eine Einführung* (ZUMA How-to-Reihe Nr. 12). Mannheim: ZUMA.

Rammstedt, B. (Ed.). (2014). *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich: Ergebnisse von PIAAC 2012*. Münster: Waxmann.

Raykov, T. (1997). Estimation of Composite Reliability for Congeneric Measures. *Applied Psychological Measurement, 21*, 173-184. doi: 10.1177/01466216970212006

Schmale, H. (2001). *Berufseignungstest (BET). Tabellenband* (4. überarbeitete und ergänzte Aufl.). Bern: Hans Huber.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: occupational attainment and job performance. Journal of Personality and Social Psychology, 86, 162-173. doi: 10.1037/0022-3514.86.1.162

Steyer, R. & Eid, M. (2001). Messen und Testen. Berlin: Springer.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Boston: Allyn and Bacon.

Weise, G. (1975). *Psychologische Leistungstests*. Göttingen: Hogrefe.

Wüst, A. (2002). *Wie wählen Neubürger? Politische Einstellungen und Wahlverhalten eingebürgerter Personen in Deutschland*. Opladen: Leske+Budrich.