Zur Bestimmung der Güte von Multi-Item-Skalen: Eine Einführung

Beatrice Rammstedt
Zentrum für Umfragen, Methoden und Analysen, Mannheim

Zusammenfassung:

Die vorliegende Einführung vermittelt die Konzepte der drei Hauptkriterien Objektivität, Reliabilität und Validität zur Bestimmung der Güte von Multi-Item-Skalen. Darüber hinaus werden Möglichkeiten zur empirischen Bestimmung dieser drei Gütekriterien aufgezeigt.

Summary:

The present paper gives an introduction how to assess the quality of a multi-item scale. The three main criteria, objectivity, reliability, and validity are presented as well as possibilities for their empirical examination.

ZUMA How-to-Reihe Nr. 12
2004

Um interessierende Merkmale zu erfassen, werden häufig Fragebogen (ob persönlich-mündlich oder telefonisch in Form eines standardisierten Interviews, schriftlich oder online vorgegeben) eingesetzt. Im Zuge der Entwicklung oder der Auswahl eines solchen Fragebogen stellt sich primär die Frage, wie gut dieser Fragebogen für den Untersuchungszweck geeignet ist. Ziel und Zweck dieser Einführung ist, den Blick für die Güte von Fragebogen zu schärfen sowie Verfahren zur Güteüberprüfung zu vermitteln. So sind Güteüberlegungen entscheidend in Situationen, in denen für eine bestimmte Untersuchung ein geeignetes Instrument ausgewählt werden soll. Auch bei der selbständigen Fragebogenentwicklung ist es entscheidend, dessen Qualität zu überprüfen. Hier werden Multi-Item-Skalen betrachtet, also die Teile eines Fragebogens oder gesamte Fragebogen, in denen ein Konstrukt mittels mehrerer Items erfasst wird, deren Beantwortung dann gemittelt oder aufsummiert wird[1].

Von zentraler Bedeutung für die Beurteilung der Qualität von Multi-Item-Skalen sind die sogenannten Hauptgütekriterien, nämlich die Objektivität, die Reliabilität und die Validität des Verfahrens. Jedes der drei Kriterien lässt sich in drei oder vier Aspekte untergliedern, die im Folgenden näher dargestellt werden sollen.

1. Objektivität

Unter Objektivität wird das Ausmaß verstanden, in dem das Untersuchungsresultat unabhängig ist von jeglichen Einflüssen außerhalb der untersuchten Person[2] (vgl. Rost, 1996). Die Objektivität einer Messung nimmt man als gegeben an, wenn das Messergebnis nur von dem zu messenden Merkmal und nicht von dem Befrager (z.B. Untersuchungsleiterverhalten) oder von Situationsvariablen abhängt. Man unterscheidet drei Arten der Objektivität eines Instruments, die Durchführungsobjektivität, die Auswertungsobjektivität und die Interpretationsobjektivität.

1.1 Durchführungsobjektivität

Die Durchführungsobjektivität bezieht sich auf die ***Konstanz der Untersuchungsbedingungen***. Die Durchführungsobjektivität einer Untersuchung kann beeinträchtigt sein, wenn sie anfällig für Störfaktoren ist und es daher nicht gelingt, alle befragten

[1] Ein Teil der dargestellten Analysen lässt sich auch auf Single-Item-Skalen (also Skalen, die ein Konstrukt mittels eines Items erfassen) übertragen.
[2] Im Folgenden wird der Einfachheit halber von Personen als Untersuchungseinheit gesprochen. Diese kann natürlich auch ein Unternehmen, eine Gruppe o.Ä. sein.

Personen unter vergleichbaren und damit in diesem Sinne fairen Untersuchungsbedingungen zu untersuchen. Daher lässt sich die Durchführungsobjektivität am besten gewährleisten durch eine maximale Standardisierung der Untersuchungssituation.

Die Durchführungsobjektivität kann z.B. beeinträchtigt sein durch:

- Interviewereffekte
- Reihenfolgeeffekte der Items
- Anfälligkeit der Itembeantwortungen für momentane individuelle Stimmungen
- Unterbrechungen bei der Fragebogenbearbeitung

Um eine möglichst hohe Durchführungsobjektivität zu erlangen, sollte die Fragebogenerhebung unter möglichst standardisierten Bedingungen durchgeführt werden. Diese sind i.d.R. gegeben bei selbstauszufüllenden Fragebogen. Bei persönlich-mündlichen und telefonischen Verfahren ist auf eindeutige Intervieweranweisungen und deren Einhaltung zu achten.

1.2 Auswertungsobjektivität

Die Auswertungsobjektivität bezieht sich auf die ***Fehler, die bei der Umsetzung der unmittelbaren Reaktionen der befragten Personen in Zahlenwerte auftreten können***. Solche Fehler können insbesondere bei der Codierung offener Antworten auftreten, aber auch bei der einfachen Umwandlung verbaler Antworten der Person in ein Kreuz auf dem Fragebogen. Auch Fehler bei der Dateneingabe, also beim Abtippen oder Einscannen der Fragebogen beeinträchtigen die Auswertungsobjektivität. Diese Form der Objektivität ist demnach umso anfälliger für Beeinträchtigungen, je mehr der Interviewer und/oder Auswerter die unmittelbaren Itembeantwortungen des Befragten in Zahlenwerte transformieren muss.

Quantitative Bestimmung der Auswertungsobjektivität

Die Auswertungsobjektivität kann quantitativ bestimmt werden, indem Interviews oder Fragebogen mindestens 2 verschiedenen Auswertern vorgegeben werden, die unabhängig voneinander die Vercodung für die einzelnen Fälle vornehmen. Die mitt-

lere Korrelation zwischen den Auswertern kann dann als Maß der Auswertungsobjektivität interpretiert werden.

Um eine möglichst hohe Auswertungsobjektivität zu gewährleisten, ist bei der Auswertung geschlossener Fragen wichtig, eindeutige Vorgaben zur Dateneingabe und –transformation zu haben (Umgang mit fehlenden Werten, mit Kreuzen zwischen Kästchen, Recodieranweisungen für Items). Offene Fragen sollten generell vermieden werden. Wenn der Einsatz offener Antwortformate jedoch unvermeidlich ist, sollten eindeutige Klassifikationsanweisungen für die Antworten gegeben sein.

1.3 Interpretationsobjektivität

Die Interpretationsobjektivität bezieht sich auf das ***Ausmaß, in dem die aus den numerischen Befragungsergebnissen gezogenen Schlüsse über verschiedene Interpretatoren vergleichbar sind.***

Demnach ist eine hohe Interpretationsobjektivität dann gegeben, wenn die in einem Instrument gewonnenen Befunde von verschiedenen Diagnostikern in gleicher Weise interpretiert werden. Hierfür ist wichtig, dass die Interpretatoren über vergleichbares Wissen darüber verfügen, was der Fragebogen misst und wie individuelle oder Gruppenwerte quantitativ zu interpretieren sind. Die Interpretation einer eingesetzten Konservatismusskala kann z.B. sehr unobjektiv sein, wenn in der Fragebogendokumentation keine klaren Interpretationshinweise oder keine genaue Beschreibung des erfassten Konstrukts gegeben ist. Ferner sind Normwerte oder Benchmarks wichtig für die quantitative Interpretation. Ohne solche Informationen kann nur ausgesagt werden, dass Person oder Gruppe x einen Wert y auf der Konservatismusskala z hat. Um diesen Wert y als „hoch“ oder „niedrig“ zu interpretieren, sind Vergleichswerte (Mittelwerte und Standardabweichungen) und Konfidenzintervalle notwendig. Zur inhaltlichen Interpretation der Skala z ist eine genaue Konstruktbeschreibung notwendig, da es sonst der Fantasie des Interpretators überlassen ist, das Konstrukt zu definieren. Überprüfbar ist die Interpretationsobjektivität, indem die Schlüsse, die zwei Interpretatoren aus den Werten eines Fragebogens unabhängig voneinander gezogen haben, miteinander verglichen werden.

Um eine hohe Interpretationsobjektivität einer Skala zu gewährleisten, ist es notwendig, dass Vergleichswerte wie Mittelwerte und Standardabweichungen sowie Konfidenzintervalle sowie inhaltliche Interpretationshinweise oder zumindest eine eindeutige inhaltliche Beschreibung der Skala vorliegen.

2. Reliabilität

Die Genauigkeit, mit der eine Skala ein Merkmal misst.

Die Reliabilität eines Fragebogens ist neben dessen Objektivität ein weiteres Kriterium für dessen Güte und i.d.R. auch ein stärkeres. Dies kann man sich am Beispiel einer Waage mit Digitalanzeige verdeutlichen: Diese ist völlig objektiv in dem Sinne, dass zwei Personen genau das gleiche Messergebnis von ihr ablesen, jedoch kann sie sehr unreliabel sein, wenn sie bei einer Person mit stabilem Gewicht deutlich unterschiedliche Angaben macht, wenn diese mehrfach auf die Waage tritt. Die Reliabilität eines Verfahrens kann deshalb als die Replizierbarkeit von Messergebnissen verstanden werden. Diese Replizierbarkeit wird durch Korrelationskoeffizienten ausgedrückt. Im Idealfall ist die Replizierbarkeit gleich 1. Um die Replizierbarkeit von Untersuchungsergebnissen zu überprüfen, müsste man theoretisch eine Person zu einem Zeitpunkt mit einem Verfahren mehrmals testen und diese beiden Ergebnisse (Korrelat 1 und Korrelat 2) miteinander korrelieren. Abbildung 1 veranschaulicht diesen Idealfall.

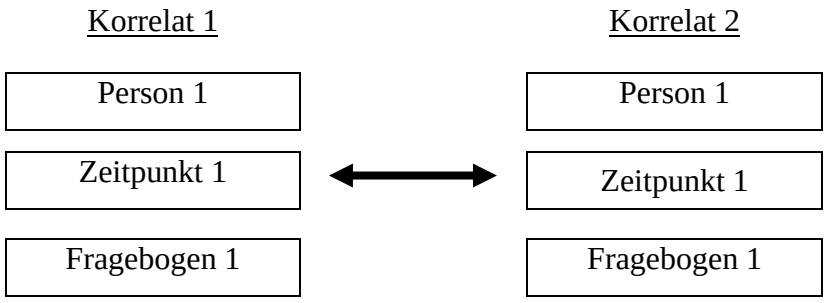


Abbildung 1: Schematische Darstellung des theoretischen Modells der Reliabilitätsbestimmung.

Praktisch ist dieses schwer möglich, da durch die mehrmalige Messung in einem eng umgrenzten Zeitraum keine unabhängige Beantwortung der Items möglich ist. Als mögliche Näherungen an das Ideal werden vier verschiedene Methoden zur Bestim-

mung der Reliabilität unterschieden, die Retest-Reliabilität, die Paralleltest-Reliabilität, die Testhalbierungs-Reliabilität und Konsistenzanalysen.

2.1 Retest-Reliabilität

Bei der Test-Retest-Methode wird der Fragebogen nach einem gewissen Zeitintervall wiederholt vorgegeben. Die Korrelation der Messwerte einer Person zu den beiden Messzeitpunkten wird als Index für die Reliabilität des Verfahrens angesehen (s. Abbildung 2).

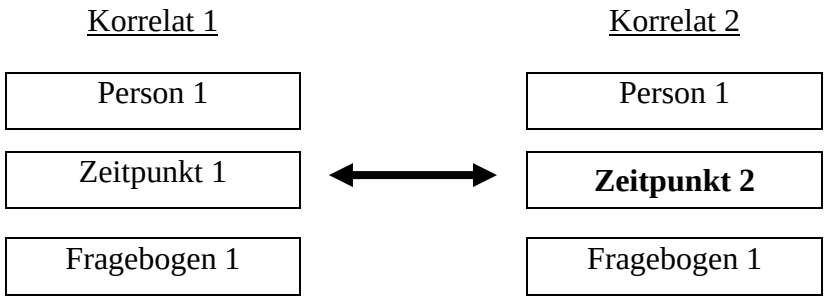


Abbildung 2: Schematische Darstellung der Retest-Reliabilität

Hierbei wird angestrebt, durch ein zeitliches Intervall zwischen den beiden Erhebungen die Erinnerungseffekte zu reduzieren und somit möglichst voneinander unabhängige Messungen zu schaffen. Dieses ist aber nur dann sinnvoll, wenn angenommen werden kann, dass sich die Ausprägung einer Person in dem zu erfassenden Merkmal zwischen den beiden Zeitpunkten nicht (oder nur unwesentlich) ändert. Die Retest-Reliabilität ist daher nur für solche Skalen geeignet, die stabile Merkmale wie z.B. Wertvorstellungen, Einstellungen erfassen, nicht jedoch für Instrumente, die vergleichsweise änderungssensitiv sind (z.B. Stimmungen). Das Ausmaß der Stabilität des Merkmals bestimmt auch die zu wählende Länge des Intervalls zwischen Zeitpunkt 1 und Zeitpunkt 2: je stabiler, desto länger darf das Intervall sein. Bei kurzen Intervallen sollte durch Veränderungen des Fragebogens (z. B. Veränderung der Itemreihenfolge, Einfügen von Füllitems) darauf geachtet werden, dass mögliche Erinnerungseffekte gering gehalten werden. Auch sollte zwischen den beiden Messzeitpunkten kein das Merkmal deutlich beeinflussendes Ereignis stattgefunden haben (z.B. Ereignisse wie Wahlen oder Bestechungsskandale für die Erfassung politischer Einstellungen).

Beispiel zur Bestimmung der Retest-Reliabilität

Zur Bestimmung der Reliabilität des Inventars zur Selbsteingeschätzten Intelligenz (ISI, Rammstedt & Rammsayer, 2003) wurde das Inventar im Abstand von 4 Wochen einer Stichprobe ein zweites Mal vorgegeben. Das ISI erfasst die Dimensionen verbale (V), mathematisch-logische (M), künstlerische (K) und personale Intelligenz (P) mit jeweils zwei, bzw. im Fall der mathematisch-logischen Intelligenz mit fünf Items. Zur Bestimmung der Retest-Reliabilität wurden die individuellen Werte in den vier Skalen zu den Erhebungszeitpunkten miteinander korreliert. Wie aus der in Abbildung 3 wiedergegebenen SPSS-Tabelle ersichtlich, ergaben sich für die Skalen des ISI Retest-Reliabilitäten zwischen 0,608 für personale und 0,787 für verbale Intelligenz.

Correlations

Correlations

		ISIB_V	ISIB_M	ISIB_K	ISIB_P
ISIA_V	Pearson Correlation	.787	.275	.090	.228
	Sig. (2-tailed)	.000	.012	.422	.040
	N	82	82	82	82
ISIA_M	Pearson Correlation	.322	.748	.156	.146
	Sig. (2-tailed)	.003	.000	.162	.191
	N	82	82	82	82
ISIA_K	Pearson Correlation	.133	.280	.752	-.010
	Sig. (2-tailed)	.235	.011	.000	.932
	N	82	82	82	82
ISIA_P	Pearson Correlation	.263	.119	-.043	.608
	Sig. (2-tailed)	.017	.288	.701	.000
	N	82	82	82	82

Abbildung 3: Interkorrelationen der vier ISI-Skalen zu Zeitpunkt A (ISIA) und zu Zeitpunkt B (ISIB).

2.2 Die Paralleltest-Reliabilität

Bei der Paralleltest-Methode wird zu dem in Frage stehenden Fragebogen ein vergleichbarer verwendet. Beide Fragebogen werden dann einer Personengruppe zum gleichen Messzeitpunkt vorgegeben und die Ergebnisse miteinander korreliert.

Angenommen wird, dass beide Fragebogen das selbe Konstrukt erfassen, dass also Fragebogen 2 ein Spiegelbild von Fragebogen 1 darstellt. Durch die Verwendung des

Fragebogen 2 anstatt der wiederholten Vorgabe des Fragebogens 1 werden Erinnerungseffekte und – im Gegensatz zur Retest-Methode – tatsächliche Veränderungen im Merkmal vermieden.

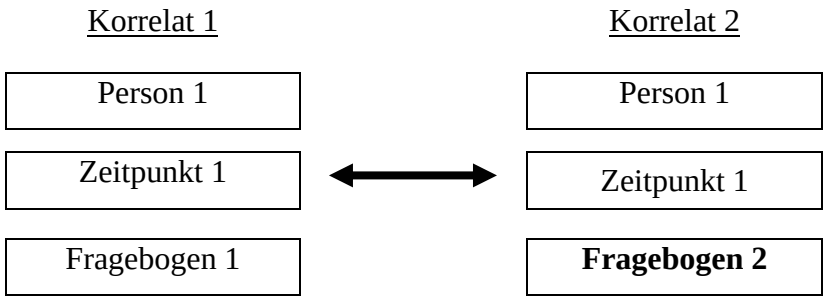


Abbildung 4: Schematische Darstellung der Paralleltest-Reliabilität.

Zur Entwicklung eines parallelen Verfahrens wird für jedes Item eines Fragebogens ein vergleichbares entwickelt. Empirisch vergleichbar sind Items dann, wenn sie hoch miteinander korrelieren und gleiche Mittelwerte und Streuungen aufweisen. In einer abgeschwächten Version der Paralleltest-Reliabilität wird nicht die Vergleichbarkeit auf Itemebene, sondern auf Skalen- oder Indexebene angestrebt. Hierbei kann die Anzahl der Items zwischen den beiden Fragebogenversionen variieren, wichtig ist jedoch eine hohe Korrelation der Skalenwerte. Faktisch ist es sehr schwer, parallele Items für Fragebögen zu entwickeln, so dass diese Reliabilitätsbestimmungsmethode eher in der Leistungsmessung ihre Anwendung findet.

Beispiel zur Bestimmung der Paralleltest-Reliabilität

Zur Bestimmung der Paralleltest-Reliabilität für das oben beschriebene, 11 Items umfassende ISI müsste in einem ersten Schritt eine parallele Fragebogenversion entwickelt werden. Hierzu müssten für die Items parallele formuliert werden. So könnte die parallele Version für das Item „Wortflüssigkeit: Rasches und angemessenes Formulieren von Wörtern“ wie folgt lauten: „Verbale Produktionsgeschwindigkeit: Schnelles und richtiges Produzieren von Wörtern“. Um nun zu überprüfen, ob diese 2. Formulierung als paralleles Item geeignet ist, muss in einer Voruntersuchung an einer Stichprobe, die beide Fragebogenversionen bearbeitet hat, geklärt werden, ob die beiden Itemversionen vergleichbare Mittelwerte und Standardabweichungen aufweisen (hierzu kann ein within subjects t-Test verwendet werden) und ob sie hoch miteinander korrelieren. Ist dies für sämtliche 11 Items und ihre Parallelformen gegeben,

können die beiden Fragebogen in der Untersuchungsstichprobe eingesetzt werden. Hierbei ist zu empfehlen, einige Füllitems zwischen den beiden Versionen vorzugeben und die Itemreihenfolge in der zweiten Version zu verändern. Auf Grundlage der individuellen Itembeantwortungen werden dann separat für die beiden Versionen Skalenwerte ermittelt. Die entsprechenden Skalenwerte der beiden Fragebogenversionen werden miteinander korreliert. In unserem Beispiel könnten die Werte für die Skala „Verbale Intelligenz“ zu 0,75 miteinander korrelieren. Dieser resultierende Koeffizient gibt die Höhe der Paralleltest-Reliabilität der Skala an.

2.3 Die Split-Half-Reliabilität

Bei der Split-Half- (oder Testhalbierungs-)Methode werden die Items eines Fragebogens mit multiplen Indikatoren in zwei äquivalente Hälften geteilt. Die Beantwortung der einen Testhälfte wird dann mit der der zweiten pro Person korreliert. Abbildung 5 veranschaulicht dieses Vorgehen.

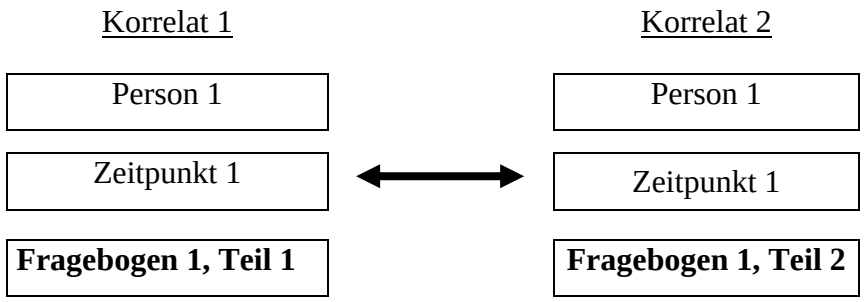


Abbildung 5: Schematische Darstellung der Split-Half-Reliabilität.

Diese Methode ist insofern eine Vereinfachung der Paralleltest-Methode: Anstatt eine neue Skala zu erstellen, wird die bestehende Skala einfach in zwei vergleichbare Hälften geteilt und somit werden zwei Verfahren mit jeweils der Hälfte der Items erstellt. Z.B. könnten zur Bestimmung der Split-Half-Reliabilität eines Instruments, das mittels 10 Items Konservatismus erfasst, die Items in zwei Hälften à 5 Items aufgeteilt werden. Eine wichtige Voraussetzung für die Anwendung der Split-Half-Methode ist die Homogenität der Items, also dass sämtliche Items das selbe Merkmal erfassen: Wenn die Items im 2. Teil des Instruments einen anderen Aspekt des interessierenden Merkmals erfassen als die Items des ersten Teils, wären nur geringe Korrelationen zwischen den Teilen zu erwarten. In unserem Beispiel ist also sicher zu stellen, dass

sämtliche 10 Items der Konservatismusskala ähnliche Aspekte des Konstrukts erfassen. Wenn jedoch z.B. 4 Items inhaltlich eher Konservatismus in politischen Einstellungen und 6 Items Konservatismus in Familien- und Geschlechterrolleneinstellungen erfassen, sollte die Testhalbierung so erfolgen, dass in jeder Hälfte 2 Fragen zu politischem und 3 zu familienorientiertem Konservatismus enthalten sind.

In der Literatur werden verschiedene Verfahren zur Testhalbierung aufgeführt, die hier nur kurz genannt werden: Am Einfachsten bietet sich eine Aufteilung in erste vs. zweite Testhälfte an (z.B. Items 1 – 10 vs. 11 – 20). Dieses Verfahren birgt besonders bei langen Instrumenten die Gefahr, dass Ermüdungseffekte die Itembeantwortung der beiden Testhälften unterschiedlich beeinflussen. Es sollte daher nur bei relativ kurzen Skalen angewandt werden. Alternativ kann die Skala nach gradzahligen und ungradzahligen Itemnummern oder nach Zufall geteilt werden. Idealerweise wird jedoch die Aufteilung nach Itemkennwerten vorgenommen. Bei diesem Vorgehen wird zu jedem Item das auf Grund seiner Itemkennwerte - wie Mittelwert, Streuung, Korrelation mit Gesamtindex (in unserem Beispiel mit der Gesamtskala „Konservatismus“) - am besten passende ausgewählt. Von diesen Itempärchen wird jeweils eines der ersten und das andere der zweiten Testhälfte zugeordnet.

Bei allen Halbierungsverfahren ist natürlich darauf zu achten, dass sämtliche Items in die selbe Richtung des zu erfassenden Merkmals gepolt sind, also zu recodierende Items bereits recodiert wurden, so dass alle Items das Merkmal in positiver Ausprägung erfassen.

Da die Split-Half-Reliabilität im Gegensatz zur verwandten Paralleltestmethode die Reliabilität lediglich auf der Basis der Hälfte der Items bestimmt und da die Reliabilität einer Skala abhängig ist von ihrer Länge, also von der Itemanzahl, wird die Reliabilität mit der Split-Half-Methode geringer ausfallen. Rechnerisch lässt sich diese „Unterschätzung“ mit der Spearman-Brown-Formel für Testverdoppelung (für die allgemeine Form vgl. Lienert & Raatz, 1998) korrigieren:

$$corr\ r_{tt} = \frac{2 \cdot r_{tt}}{1 + r_{tt}}$$

r_{tt} = nach der Split-Half-Methode ermittelte Reliabilität des Tests t

corr r_{tt} = korrigierte Reliabilität des Tests t

Nach dieser Formel lässt sich z.B. für eine nach der Split-Half-Methode bestimmte Reliabilität einer Skala von r_{tt} = 0,70 eine tatsächliche Reliabilität von corr r_{tt} = 0,82 schätzen.

Beispiel zur Bestimmung der Split-Half-Reliabilität

Zur Bestimmung der Split-Half-Reliabilität ist das oben beschriebene ISI nicht gut geeignet, da die meisten Skalen nur zwei Items enthalten und somit die beiden Skalenhälften nur je ein Item enthielten. Ein weitaus umfangreicheres Verfahren ist die Machiavellismusskala von Henning und Sixt (2003). Die Skala erfasst Machiavellismus mittels 18 Items. Die Autoren berichten in ihrer Dokumentation eine Split-Half-Reliabilität von 0,70. Da die verwendete Halbierungsmethode nicht näher beschrieben wird, ist anzunehmen, dass das in SPSS voreingestellte Verfahren in Teilung erste vs. zweite Hälfte verwendet wurde. Der Korrelationskoeffizient wurde anschließend nach der Spearman-Brown-Formel korrigiert und ergab eine korrigierte Reliabilität von 0,82. Zur Berechnung der Split-Half-Reliabilität mittels SPSS muss im Menü „Analyze“ beim Unterpunkt „Scale“ die Option „Reliability Analysis“ ausgewählt werden (s. Abbildung 6).

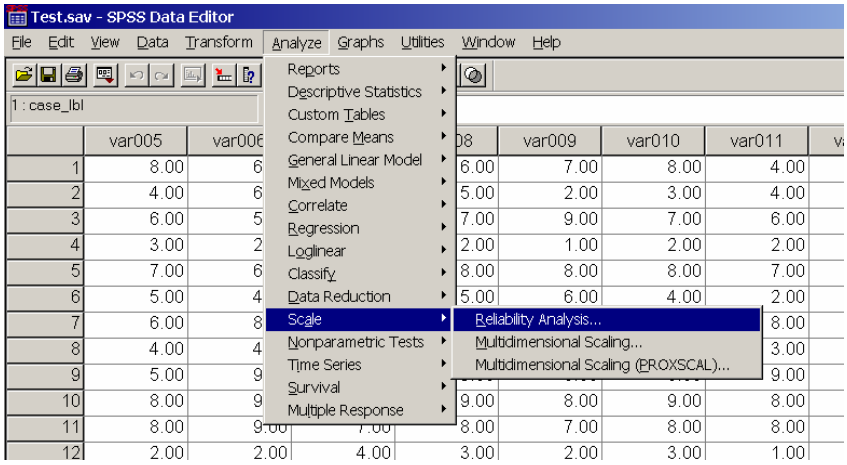


Abbildung 6: Bestimmung der Split-Half-Reliabilität mittels SPSS (Schritt 1).

In dem sich öffnenden Dialogfenster werden dann die Items der Skala ausgewählt und in dem Feld „Model“ „Split-half“ ausgewählt (s. Abbildung 7).

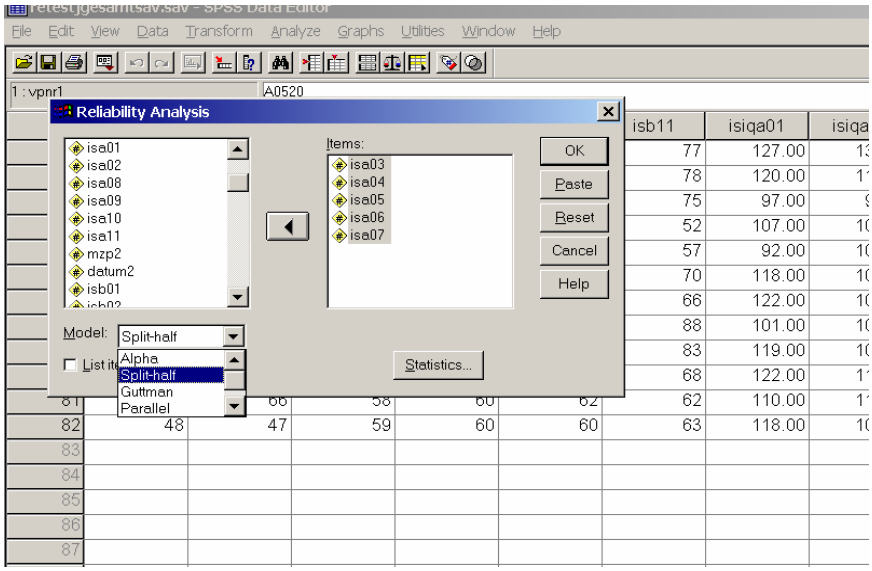


Abbildung 7: Bestimmung der Split-Half-Reliabilität mittels SPSS (Schritt 2).

2.4 Konsistenzanalysen

Die Konsistenzanalyse stellt eine Erweiterung der Split-Half-Methode dar. Da sich bei der Split-Half-Methode das Problem ergibt, dass sich in Abhängigkeit davon, nach welcher Methode man die Skala halbiert, leicht unterschiedliche Reliabilitätskoeffizienten ergeben, wäre es wünschenswert, möglichst viele Splits vorzunehmen und dabei die Skala nicht nur in zwei sondern in vier, acht oder in so viele Teile zu zerlegen, wie Items vorhanden sind. Das Mittel über sämtliche Korrelationen entspräche dann einer „Durchschnittsreliabilität“ der Skala. Dieser Problematik trägt die Konsistenzanalyse Rechnung. Hierbei werden nicht nur zwei Testhälften, sondern sämtliche Items eines Instruments miteinander korreliert (s. Abbildung 8). Wie der Name vermuten lässt, gibt dieser Reliabilitätskoeffizient Auskunft über die Konsistenz, also die Homogenität eines Verfahrens.

Zur Bestimmung der internen Konsistenz existieren verschiedene Formeln, am verbreitetsten ist der Alpha-Koeffizient nach Cronbach (1951), dessen Berechnung auch im Statistikprogramm SPSS als Standardmethode zur Reliabilitätsbestimmung angeboten wird. Dieser Alpha-Koeffizient berechnet sich wie folgt:

$$\alpha = \frac{n\bar{r}}{1+\bar{r}(n-1)}$$

wobei n die Gesamtanzahl der Items und $\bar{r}$ die mittlere Interkorrelation der Items ist.

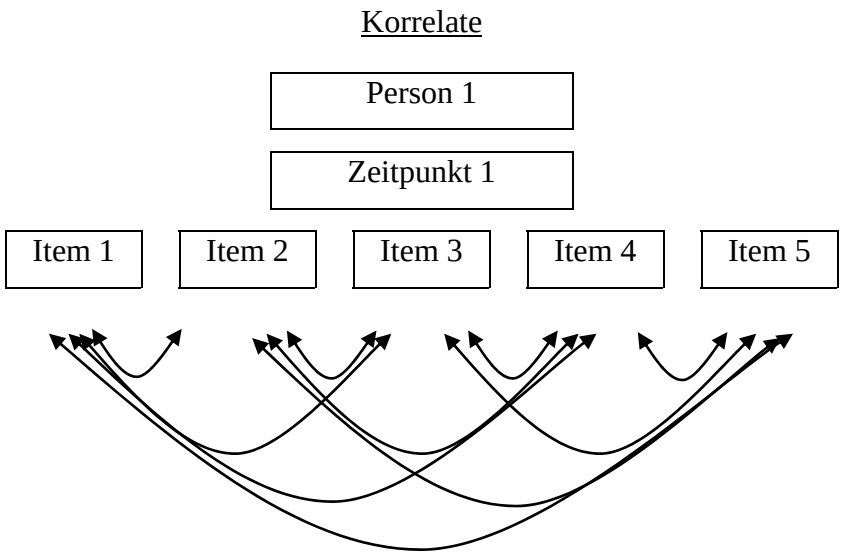
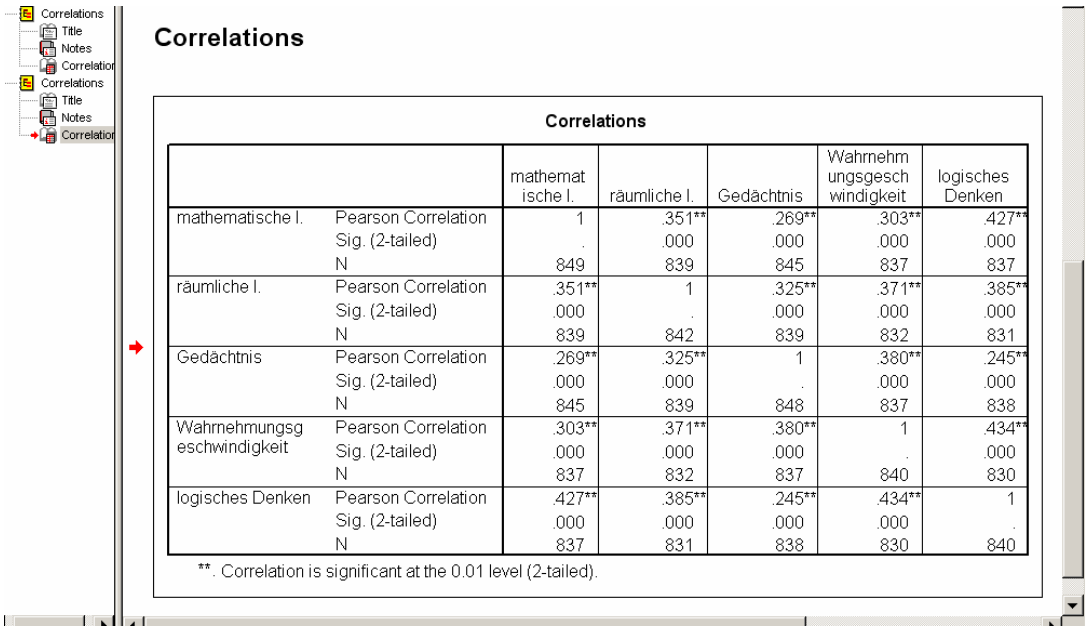


Abbildung 8: Schematische Darstellung der Konsistenzanalysen.

Beispiel zur Bestimmung der internen Konsistenz

Exemplarisch wird die interne Konsistenz der Skala „mathematisch-logische Intelligenz“ des oben beschriebenen ISI dargestellt. Zur Berechnung der mittleren Interkorrelation der Items wurden die fünf Items der Skala miteinander korreliert (s. Abbildung 9). Die zehn Korrelationskoeffizienten ergeben im Mittel eine Korrelation von $\bar{r} = 0{,}344$.



Correlations

		mathematische I.	räumliche I.	Gedächtnis	Wahrnehmungsgeschwindigkeit	logisches Denken
mathematische I.	Pearson Correlation	1	.351**	.269**	.303**	.427**
	Sig. (2-tailed)	.	.000	.000	.000	.000
	N	849	839	845	837	837
räumliche I.	Pearson Correlation	.351**	1	.325**	.371**	.385**
	Sig. (2-tailed)	.000	.	.000	.000	.000
	N	839	842	839	832	831
Gedächtnis	Pearson Correlation	.269**	.325**	1	.380**	.245**
	Sig. (2-tailed)	.000	.000	.	.000	.000
	N	845	839	848	837	838
Wahrnehmungsgeschwindigkeit	Pearson Correlation	.303**	.371**	.380**	1	.434**
	Sig. (2-tailed)	.000	.000	.000	.	.000
	N	837	832	837	840	830
logisches Denken	Pearson Correlation	.427**	.385**	.245**	.434**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.
	N	837	831	838	830	840

**. Correlation is significant at the 0.01 level (2-tailed).

Abbildung 9: Interkorrelation der fünf Items der Skala „mathematisch-logische Intelligenz“.

Eingesetzt in die Formel zur Berechnung der internen Konsistenz ergibt sich:

$$\alpha = \frac{5 \cdot 0{,}344}{1 + 0{,}344(5-1)} = 0{,}724\,.$$

Die Skala „mathematisch-logische Intelligenz“ weist demnach eine interne Konsistenz von 0,724 auf.

Zur Berechnung der internen Konsistenz mittels SPSS muss wiederum im Menü „Analyze“ beim Unterpunkt „Scale“ die Option „Reliability Analysis“ ausgewählt werden. In dem sich öffnendem Dialogfenster können dann die Items des Instruments ausgewählt werden. Im Fenster „Model“ wird diesmal die Voreinstellung „Alpha“ gewählt (Abbildung 10).

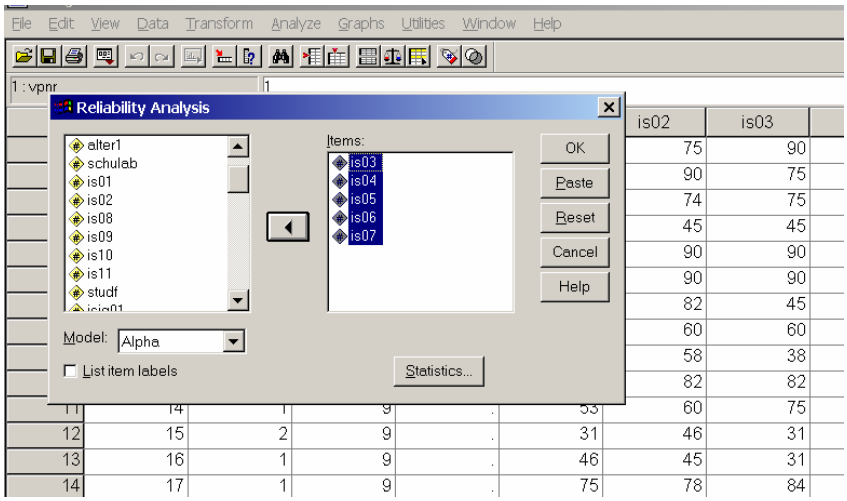


Abbildung 10: Bestimmung der internen Konsistenz mittels SPSS.

Der SPSS-Output enthält in der Standardeinstellung Informationen über die Anzahl der Items, die Anzahl der Fälle sowie den Cronbach-Alpha-Koeffizienten (Abbildung 11).

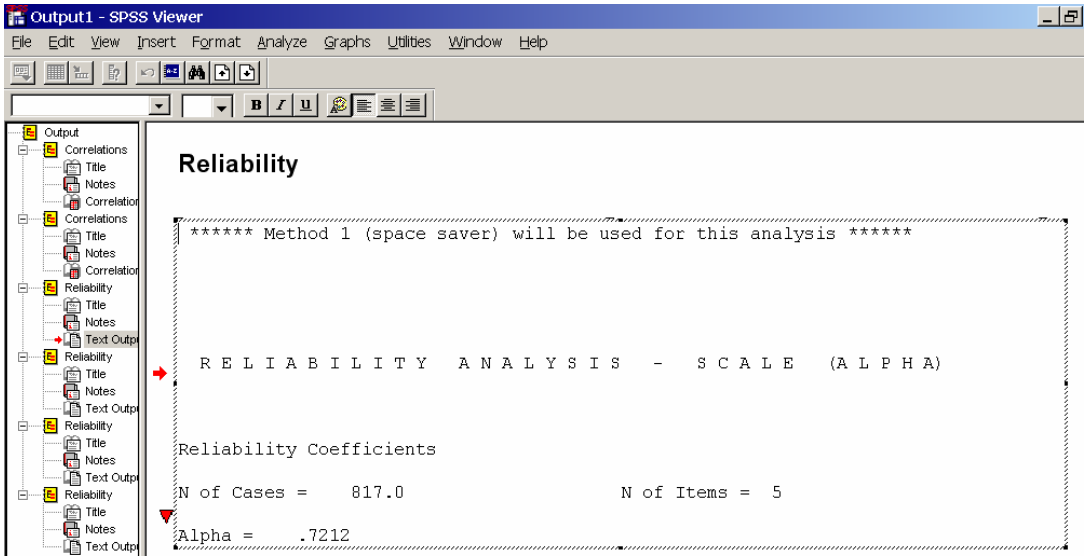


Abbildung 11: SPSS-Output der Berechnung der internen Konsistenz.

Die Beurteilung der Höhe von Reliabilitätskoeffizienten

Wann ist eine Reliabilität als gut zu beurteilen? Diese Frage wird häufig gestellt und ist schwer zu beantworten, da die Höhe des Reliabilitätskoeffizienten neben der eigentlichen Zuverlässigkeit der Skala von einigen Faktoren abhängt:

1. Itemanzahl der Skala

   Je mehr Items eine Skala enthält, desto höhere Reliabilitätskoeffizienten sind zu erwarten.

2. zeitlicher Abstand zwischen den Fragebogenvorgaben beim Retest-Design

   Bei geringerem zeitlichen Abstand (z.B. ein bis zwei Wochen im Vergleich zu sechs Monaten) werden i.d.R. vergleichsweise höhere Koeffizienten erzielt. Daher sollte bei Verwendung der Retest-Methode in der Skalendokumentation immer das zeitliche Intervall berichtet werden.

3. inhaltliche Heterogenität der Items bei Konsistenzanalysen

   Wenn die Items einer Skala recht heterogen sind, ergeben sich vergleichsweise niedrigere Iteminterkorrelationen und somit auch eine niedrigere interne Konsistenz.

Darüber hinaus hängt die Anforderung an die Zuverlässigkeit einer Skala stark vom Untersuchungsziel ab. Während für Individualdiagnosen extrem hohe Reliabilitäten der Messverfahren erforderlich sind, werden für Gruppenvergleiche meist Reliabilitätskoeffizienten über 0,70 als befriedigend angesehen. Als gut gilt eine Reliabilität ab ca. 0,80 (vgl. Nunnally & Bernstein, 1994).

3. Validität

Grad der Genauigkeit, mit der ein Verfahren tatsächlich das misst oder vorhersagt, was es messen oder vorhersagen soll.

Objektive und zuverlässige Verfahren müssen nicht unbedingt valide sein. Dieses lässt sich wiederum an dem oben aufgeführten Beispiel der Waage verdeutlichen: Diese Waage kann objektiv und auch reliabel einen Messwert anzeigen. Jedoch ist dieser Wert nicht das Gewicht der Person (also das zu messende Merkmal) sondern z.B. die Raumtemperatur. Bei der Validität eines Verfahrens geht es also um den Nachweis, dass das Verfahren tatsächlich das zu messende Merkmal erfasst. Hierzu stehen verschiedene, sich ergänzende Validierungskonzepte zur Verfügung. In der Regel werden drei Validitätsarten unterschieden: die Kontentvalidität, die Kriteriumsvalidität und die Konstruktvalidität.

3.1 Die Kontentvalidität

Kontentvalidität (auch Inhaltsvalidität genannt) beruht auf einer inhaltlichen Analyse des Messverfahrens, um festzustellen, ob der Itempool eines Instruments den zu messenden Merkmalsbereich auch tatsächlich hinreichend genau repräsentiert. Voraussetzung für eine kontentvalide Testkonstruktion ist die Definierbarkeit des Itemuniversums für das zu erfassende Merkmal. Diese „Definierbarkeit“ ist oft angezweifelt worden.

Tatsächlich ist es dieser Punkt, der die Verbreitung kontentvalider Testverfahren über die oftmals besonders übersichtlichen klassischen Anwendungsbereiche der Pädagogischen Psychologie (z. B.: „Grundrechnen“) hinaus verhindert hat.

Kontentvalidität setzt daher in der Regel schon zum Zeitpunkt der Fragebogenkonstruktion an. Das Vorgehen zur Erstellung kontentvalider Verfahren besteht aus drei Schritten:

1. Definition des Itemuniversums: Eingrenzung des Merkmals (z.B. Grundrechnen im Zahlenraum bis zehn); Bestimmung des „universe of items“ (vgl. Borg & Shye, 1995), d.h. sämtlicher potenzieller Items (z.B. sämtliche Kombinationen der Zahlen eins bis neun mittels der Grundrechenarten); Definition der Items und Festlegung des Itemformats (z.B. multiple choice mit fünf Antwortalternativen)
2. Ziehung von systematischen Stichproben aus dem Itemuniversum

3. Anwendung.

(Für eine detailliertere Darstellung der Kontentvalidität s. Klauer, 1984)

Um zu überprüfen, inwieweit ein Verfahren kontentvalide ist, wird daher auch primär dessen Herstellungsprozedur überprüft: Wurde ein Itemuniversum definiert? Wurde die Definition des Itemuniversums z.B. von Experten hinsichtlich seiner Gültigkeit eingeschätzt? Wie wurde die Itemauswahl vorgenommen?

Die einzige systematische Methode zur Überprüfung der Kontentvalidität bietet die Facettentheorie (s. z.B. Borg & Shye, 1995), auf die jedoch hier nicht näher eingegangen werden soll.

3.2 Kriteriumsvalidität

Die Kriteriumsvalidität beschreibt den ***Grad der Übereinstimmung des mit einem Fragebogen erzielten Ergebnisses mit den Ergebnissen für ein Außenkriterium wie z.B. Schulerfolg, Wahlverhalten oder Mitgliedschaft in bestimmten Organisationen.***

Bei dem Kriterium handelt es sich um einen Maßstab, der von dem zur Beurteilung eingesetzten Verfahren unabhängig ist und eine häufig im Alltag vorgenommene Beurteilung widerspiegelt (z.B. Ausbildungserfolg, Lehrerurteil). So könnte man z.B. eine Religiositätsskala an der Anzahl der Kirchenbesuche pro Jahr oder eine Skala zum Umweltverhalten an der Spendenbereitschaft für oder Mitgliedschaft in entsprechenden Organisationen (wie z.B. BUND, Greenpeace) validieren. Die Validität wird häufig mit Korrelationsanalysen gemessen.

Je nachdem, wann das Kriterium erhoben wurde, unterscheidet man zwischen der retrograden, konkurrenten und prognostischen (Kriteriums-)Validität. Eine retrograde Validierung wäre z.B., wenn eine Konservatismusskala an dem Wahlverhalten bei der letzten Bundestagswahl validiert würde, während bei der konkurrenten Validierung Verhalten, das zum gleichen Zeitpunkt wie das Verfahren selbst erfasst wurde, als Kriterium dient (z.B. Validierung einer Umwelteinstellungsskala an selbstberichtetem umweltfreundlichen Verhalten). Bei der prognostischen (Kriteriums-)Validität wird geprüft, inwieweit die Befunde eines Verfahrens mit den später tatsächlich eingetretenen Ereignissen übereinstimmen (z.B. Validierung von Zulassungstests für bestimmte Studiengänge am späteren Studien- und Berufserfolg).

Beispiel zur Bestimmung der Kriteriumsvalidität

Schneider und Minkmar (2003) validierten ihren Konservatismusfragebogen an der Einschätzung der eigenen politischen Haltung auf einer Rechts-links-Skala. Hierzu beantwortete eine Stichprobe neben dem Konservatismusfragebogen auch die Frage zur eigenen politischen Einstellung. Die individuellen Werte des Konservatismusfragebogens und der politischen Einstellung wurden miteinander korreliert. Es ergab sich ein Zusammenhang von 0,51 in der Form, dass konservativere Personen ihre eigene politische Einstellung eher als rechts beschrieben.

3.3 Konstruktvalidität

Eine Konstruktvalidierung dient dem Ziel, die Beziehungen zwischen den im Messinstrument berichteten Einstellungen oder Verhaltensweisen und Konstrukten aufzuklären. Es wird also überprüft, inwiefern das Instrument das zu erfassende Merkmal (= Konstrukt) misst. Ein Konstrukt ist ein gedankliches Konzept, das aus Überlegungen und Erfahrungen abgeleitet worden ist, um beobachtbares Verhalten zu erklären, z. B. Konservatismus oder Maskulinität. Es gibt sehr viele unterschiedliche Methoden, um die Konstruktgültigkeit eines Verfahrens zu überprüfen. Eine Methode besteht darin, die Skala mit einem anderem Instrument, das ein stark verwandtes oder das gleiche Konstrukt erfasst, zu vergleichen (z.B. eine neu entwickelte Skala zu Konservatismus mit einem bereits etablierten Konservatismusfragebogen). Hierzu werden beide Instrumente an einer Stichprobe erhoben und die individuellen Werte miteinander korreliert. Eine andere Möglichkeit ist es, Hypothesen über die Dimensionalität des zu erfassenden Merkmals empirisch an dem in Frage stehenden Instrument zu überprüfen.

Konstruktvalidierung mittels Dimensionalitätsüberprüfung

Die Voraussetzung für diese Art der Konstruktvalidierung ist das Vorliegen von Annahmen über die dimensionale Struktur des zu erfassenden Konstrukts. Ist dieses Konstrukt eindimensional (also homogen), oder gliedert es sich in mehrere Teilaspekte? So umfasst das oben beschriebene ISI vier Skalen. Es ist demnach zu erwarten, dass die elf Items des ISI eine vierdimensionale Struktur aufweisen, die die Dimensionen verbale, mathematisch-logische, künstlerische und personale Intelligenz wider-

spiegeln, indem die ersten beiden Items eine Dimension bilden, das 3. bis 7. Item eine zweite, das 8. und 9. Item eine dritte und schließlich das 10. und 11. die vierte Dimension. Um nun zu überprüfen, inwieweit der Fragebogen tatsächlich diese postulierte Struktur des in Frage stehenden Merkmals aufweist, werden die mit dem Instrument erfassten Daten einer Faktorenanalyse unterzogen. Die Faktorenanalyse ist ein Verfahren zur „Gruppierung" von Variablen. Mittels der Faktorenanalyse werden „künstliche" Variablen, nämlich die sogenannten Faktoren erzeugt. Diese Faktoren stellen das Gemeinsame der bivariaten Korrelationen der einzelnen Items dar. Das grundlegende Prinzip der Faktorenanalyse ist, dass so wenige Faktoren wie möglich so viele Gemeinsamkeiten wie möglich abbilden sollen. Es wird somit eine Datenreduktion auf das „Wesentliche" (innerhalb der messfehlerbehafteten Daten) oder ein „Data smoothing" (d.h. eine Glättung der Datenstruktur) angestrebt.

Bei wenigen Variablen kann eine Inspektion der Korrelationsmatrix genügen, um die Dimensionen zusammengehöriger Variablen zu identifizieren. Wie aus dem Beispiel in Tabelle 1 ersichtlich, lassen sich aus der Interkorrelation der vier Variablen deutlich zwei Dimensionen erkennen, nämlich eine Kombination der Items a und b und eine der Items c und d.

Tabelle 1: Intercorrelation der Items a, b, c und d.

	a	b	c	d
a	1,00			
b	0,49	1,00		
c	0,17	0,06	1,00	
d	0,15	0,28	0,55	1,00

Die Zahl der Korrelationen in einer Korrelationsmatrix steigt jedoch mit zunehmender Itemanzahl schnell in unübersichtlichere Ausmaße. Wie aus Tabelle 2 ersichtlich, ist die Korrelationsmatrix im Fall unserer elf ISI-Items schon deutlich weniger übersichtlich.

Tabelle 2. Interkorrelationen der elf ISI-Items (aus Rammstedt & Rammsayer, 2002).

	VV	WF	MaI	RI	GF	WG	LD	MuI	KI	IpI
Verbales Verständnis (VV)	-									
Wortflüssigkeit (WF)	0,59	-								
Mathematische Intelligenz (MaI)	0,23	0,22	-							
Räumliche Intelligenz (RI)	0,18	0,18	0,35	-						
Gedächtnisfähigkeit (GF)	0,18	0,23	0,27	0,33	-					
Wahrnehmungsgeschw. (WG)	0,27	0,29	0,30	0,37	0,38	-				
Logisches Denken (LD)	0,35	0,31	0,43	0,39	0,25	0,43	-			
Musikalische Intelligenz (MuI)	0,17	0,22	0,09	0,15	0,23	0,18	0,14	-		
Körperlich-kinästhetische I. (KI)	0,08	0,18	0,15	0,20	0,15	0,24	0,10	0,32	-	
Interpersonale Intelligenz (IpI)	0,31	0,28	0,09	0,13	0,24	0,32	0,28	0,13	0,20	-
Intrapersonale Intelligenz	0,21	0,27	0,17	0,19	0,32	0,20	0,10	0,17	0,27	0,42

Die Faktorenanalyse berechnet auf Basis der Interkorrelationen der Items die zugrundeliegende Dimensionalität.

In SPSS findet sich die Faktorenanalyse unter „Analyze“ → „Data Reduction“ (s. Abbildung 12).

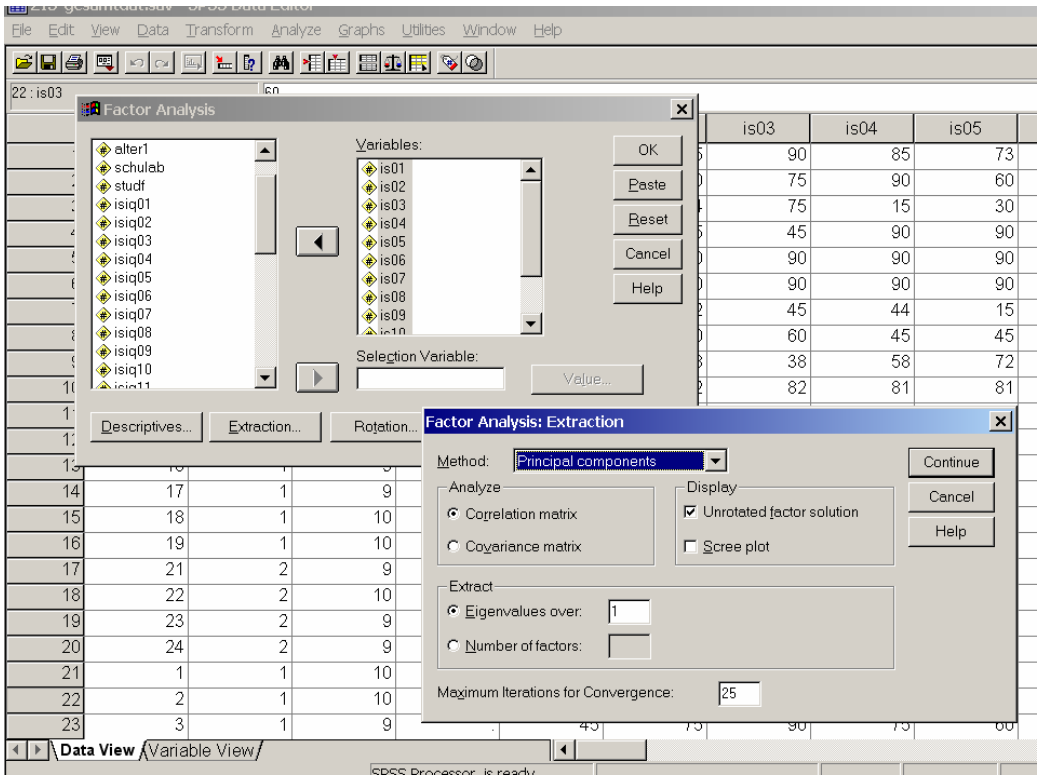


Abbildung 12: Berechnung der Faktorenanalyse mittels SPSS.

Als Verfahren zur Faktorenextraktion ist die Hauptkomponentenanalyse (principal component analysis) voreingestellt und in den meisten Fällen auch die angebrachte Prozedur. Zur Bestimmung der Anzahl der Faktoren ist die durch die einzelnen Faktoren erklärte Varianz der Iteminterkorrelationen entscheidend. Die erklärte Varianz pro Faktor entspricht seinem Eigenwert. Bei SPSS voreingestellt ist das Extraktionsverfahren von Faktoren mit einem Eigenwert > 1 (Kaiser-Guttman-Kriterium). Alternativ kann zur Bestimmung der Anzahl zu extrahierender Faktoren auch der Scree Plot herangezogen werden. Im Scree Plot ist der Eigenwerteverlauf der Faktoren dargestellt. Nach dem Scree-Test (Cattell, 1966) wird der Eigenwerteverlauf auf einen „Knick" hin untersucht und die Anzahl von Faktoren extrahiert, deren Eigenwerte oberhalb des Knicks liegen[3]. Dieses Verfahren bietet sich insbesondere bei einer großen Itemanzahl an, da dann die Bestimmung der Faktorenanzahl nach „Eigenwerten > 1" häufig zu einer Überschätzung der Faktorenzahl führt. Wenn Vorannahmen über die Dimensionalität der Items bestehen, kann alternativ auch in SPSS die Anzahl zu extrahierender Faktoren vorgegeben werden.

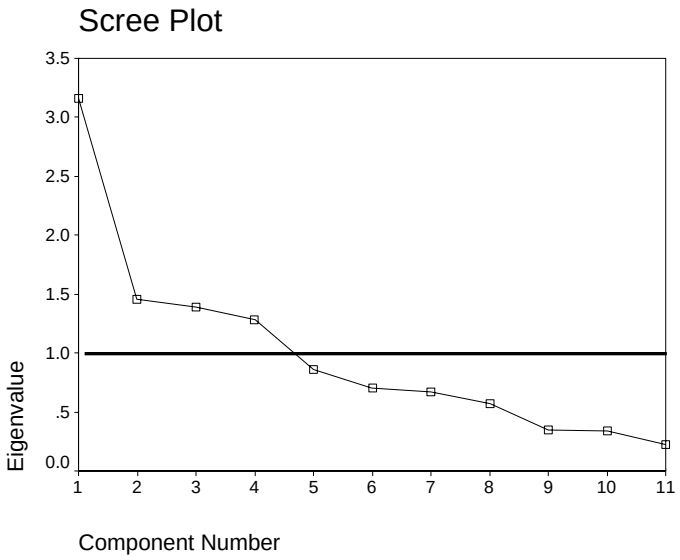


Abbildung 13: Verlauf der Eigenwerte der elf ISI-Items.

[3] Cattell nannte den Test „Scree-Test" (=Geröll-Test), da der Eigenwerteverlauf vorstellbar ist wie Geröll, das einen Berghang hinunter rutscht. Der Knick, an dem der feste Fels beginnt und das Geröllfeld endet, bestimmt die Anzahl zu extrahierender Faktoren.

In unserem Bespiel der elf ISI-Items ergeben sich sowohl nach dem Kaiser-Guttman-Kriterium als auch nach dem Scree-Test vier zu extrahierende Faktoren (siehe Abbildung 13).

Die Voreinstellung von SPSS sieht keine Rotation der resultierenden Faktormatrix vor. Grundsätzlich ist zu empfehlen, von dieser Voreinstellung abzuweichen und die sog. VARIMAX-Rotation zu wählen, eine orthogonale Rotation nach dem Einfachstrukturprinzip (s. Abbildung 14). Das Einfachstrukturprinzip (Thurstone, 1947) besagt, dass die Faktoren so rotiert werden sollen, dass die Items auf jedem Faktor möglichst hoch oder möglichst gering laden. Die Varianz zwischen den Ladungen der Items auf jedem Faktor wird demnach maximiert.

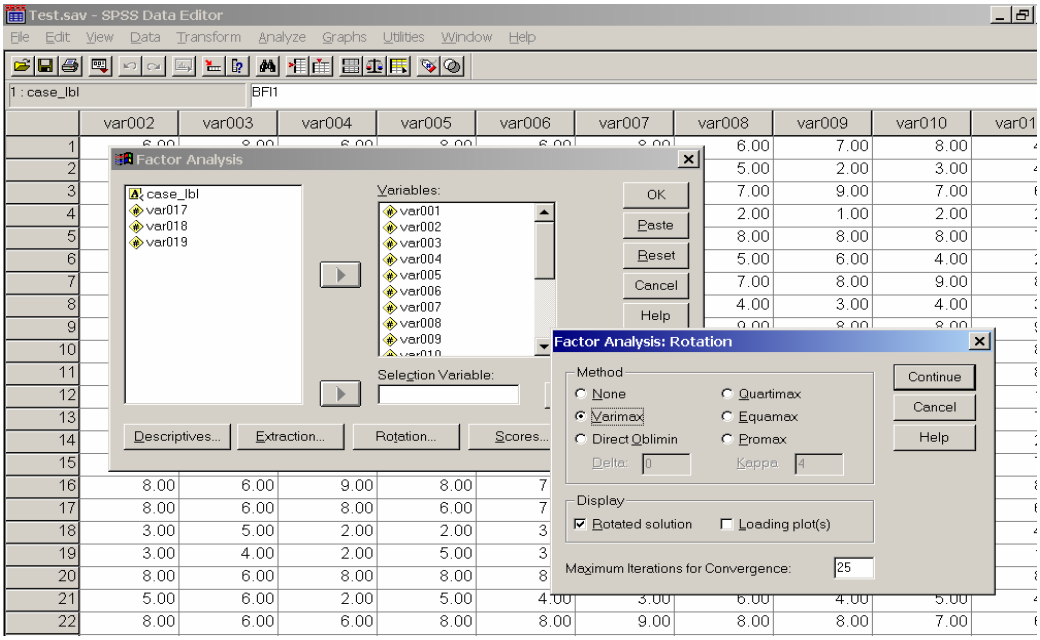
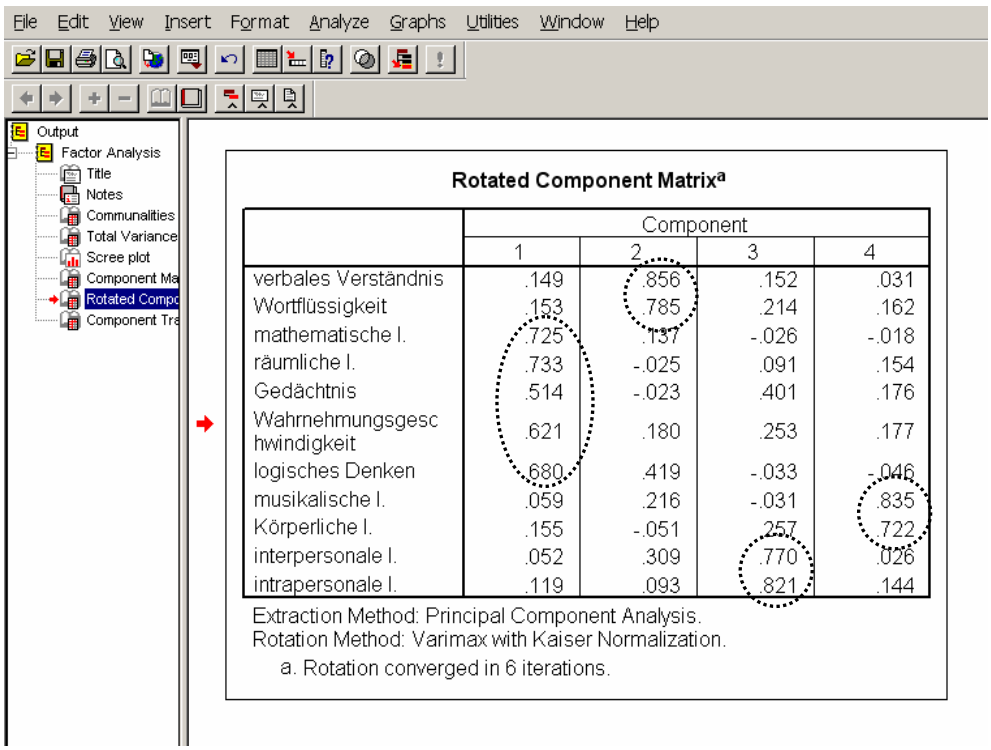


Abbildung 14: Einstellung der Faktorrotation in SPSS.

In der rotierten Komponentenmatrix ist ersichtlich, wie hoch jedes Item auf jedem Faktor lädt (d.h. mit ihm korreliert). Bezogen auf unser Beispiel des ISI müsste überprüft werden, ob in der rotierten Ladungsmatrix tatsächlich die Items 1 und 2, die Items 3 bis 7, 8 und 9 sowie 10 und 11 auf unterschiedlichen Faktoren hoch laden. Erst diese empirische Überprüfung ermöglicht, die für das zugrunde liegende Merkmal postulierte dimensionale Struktur für das verwendete Instrument nachzuweisen, also in diesem Fall die vier Dimensionen verbale, mathematisch-logische, künstlerische und personale Intelligenz.

Wie aus der in Abbildung 15 wiedergegeben rotierten Ladungsmatrix der elf ISI-Items auf den extrahierten vier Faktoren ersichtlich, laden die Items 3 bis 7, die mathematisch-logische Intelligenz erfassen sollen, am höchsten auf dem ersten Faktor und niedrig auf allen anderen. Auf dem zweiten Faktor laden die Items „verbales Verständnis“ und „Wortflüssigkeit“ am höchsten und definieren somit diesen Faktor als „Verbale Intelligenz“. Der dritte Faktor wird markiert von dem 10. und 11. Item und spiegelt damit die Dimension „personale Intelligenz“ wider. Auf dem vierten Faktor schließlich laden die Items „musikalische Intelligenz“ und „körperliche Intelligenz“ am höchsten und ist daher im Sinne der „künstlerischen Intelligenz“ zu interpretieren.



Rotated Component Matrix[a]

	Component			
	1	2	3	4
verbales Verständnis	.149	.856	.152	.031
Wortflüssigkeit	.153	.785	.214	.162
mathematische I.	.725	.137	-.026	-.018
räumliche I.	.733	-.025	.091	.154
Gedächtnis	.514	-.023	.401	.176
Wahrnehmungsgeschwindigkeit	.621	.180	.253	.177
logisches Denken	.680	.419	-.033	-.046
musikalische I.	.059	.216	-.031	.835
Körperliche I.	.155	-.051	.257	.722
interpersonale I.	.052	.309	.770	.026
intrapersonale I.	.119	.093	.821	.144

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 6 iterations.

Abbildung 15: Output der rotierten Ladungsmatrix in SPSS.

4. Vorgehen zur Güteüberprüfung von Skalen

Im Zuge der Dokumentation einer Skala, beispielsweise im Rahmen ihrer Publikation, ist es notwendig, auf die Qualität der Skala einzugehen. In dieser Einführung wurden verschiedene Verfahren zur Bestimmung der Hauptgütekriterien vorgestellt. Häufig stellt sich jedoch die Frage, wie – mit möglichst geringem Aufwand – am besten die Gütemerkmale zu bestimmen sind. Daher soll zum Abschluss hier in Form einer

Checkliste auf das minimale Vorgehen zur Bestimmung der Skalenqualität eingegangen werden.

1. Objektivität
   - ✓ Wird meine Skala standardisiert vorgeben? D.h. gibt es klare Anweisungen zur Durchführung der Befragung? Dann kann die Durchführungsobjektivität i.d.R. als gesichert angesehen werden.
   - ✓ Verwende ich ausschließlich geschlossen Antwortformate? Dann kann die Auswertungsobjektivität als gesichert angesehen werden.
   - ✓ Werden Mittelwerte und Standardabweichungen, eine inhaltliche Beschreibung für die Skala sowie für die Zielpopulation relevante Normen berichtet? Dann ist die Interpretationsobjektivität weitgehend gegeben.
2. Reliabilität
   - ✓ Bei Skalen mit mehr als zwei Items sollte standardmäßig eine Reliabilitätsbestimmung in Form der internen Konsistenz durchgeführt werden.
   - ✓ Wenn ein stabiles Merkmal erfasst wird (oder bei Skalen mit einem oder zwei Items), könnte zusätzlich an einer kleinen (Gelegenheits-)Stichprobe die Retest-Reliabilität bestimmt werden.
3. Validität
   - ✓ Die dimensionale Struktur der Skala sollte mittels Faktorenanalyse überprüft werden.
   - ✓ Wenn Zusammenhänge zu bestimmten Außenkriterien oder mit anderen Skalen, die das gleiche oder ein verwandtes Merkmal erfassen, angenommen werden können, sollte dies überprüft und die Korrelationen berichtet werden. Hierzu reicht i.d.R. eine Gelegenheitsstichprobe aus.

5. Literatur

Borg, I. & Shye, S. (1995). *Facet theory: form and content.* Newbury Park: Sage.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioural Research, 1*, 245 – 276.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Henning, H., & Six, B. (2003). Machiavellismus. In A. Glöckner-Rist (Hrsg.), *ZUMA-Informationssystem. Elektronisches Handbuch sozialwissenschaftlicher Erhebungsinstrumente*. *Version 7.00.* Mannheim: Zentrum für Umfragen, Methoden und Analysen.

Klauer, K.J. (1984). Kontentvalidität. *Diagnostica, 30,* 1-23.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz.

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

Rammstedt, B. & Rammsayer, T. (2002). Die Erfassung der selbsteingeschätzten Intelligenz: Konstruktion, teststatistische Überprüfung und erste Ergebnisse des Inventars zur selbsteingeschätzten Intelligenz (ISI). *Zeitschrift für Differentielle und Diagnostische Psychologie, 23,* 435-446.

Rammstedt, B., & Rammsayer, T. (2003). Fragebogen zur selbsteingeschätzten Intelligenz (ISI). In A. Glöckner-Rist (Hrsg.), *ZUMA-Informationssystem. Elektronisches Handbuch sozialwissenschaftlicher Erhebungsinstrumente. Version 7.00*. Mannheim: Zentrum für Umfragen, Methoden und Analysen.

Rost, J. (1996). Lehrbuch Testtheorie Testkonstruktion. Bern: Huber.

Schneider, J., & Minkmar, H. (2003). Konservatismus. In A. Glöckner-Rist (Hrsg.), *ZUMA-Informationssystem. Elektronisches Handbuch sozialwissenschaftlicher Erhebungsinstrumente*. *Version 7.00.* Mannheim: Zentrum für Umfragen, Methoden und Analysen.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.