7. Testkonstruktion

7.1. Phasen der Testentwicklung

1) Testplanung (vgl. Kapitel: Einstellung und Verhalten)

Welche Fragestellung soll der Test beantworten? Stehen eher Fragen der Einstellung (z. B. "Einstellung zur Umwelt") oder des Verhaltens (z. B. "Sporttreiben erhält die Gesundheit") im Blickpunkt?
Welche Literatur lässt sich zu der Fragestellung zusammentragen und gibt es bereits psychologische Theorien oder sogar vorhandene Operationalisierungsvorschläge (vgl. Brickenkamp, 1975) zu der Fragestellung?
Wenn diese Fragen beantwortet sind, müssen die Items ausgewählt werden, die für die Fragestellung repräsentativ sind.

2) Testentwurf (vgl. Kapitel: Wichtigste Skalen zur Messung affektiver Prozesse und Richtlinien zur Formulierung von Einstellungsitems und Antwort-Items zur Messung von Wissen und Verständnis)

Der Test sollte eine klare, einfache Instruktion beinhalten. In der Instruktion sollte die Versuchsperson darüber informiert werden, was untersucht wird, wie sie ankreuzen soll und gegebenenfalls dass ihre Daten anonym behandelt werden.
Der Test soll ökonomisch sein, d. h. es sollte wenig Material verwendet werden, dennoch sollte die Darstellung übersichtlich sein. Die Bearbeitung des Tests darf nicht zu lange dauern.

Die Fragen können nach verschiedenen Prinzipien skaliert werden (s. u.) Innerhalb eines Tests sollte möglichst eine einheitliche Skalierung verwendet werden. Die Polung der Items muss klar sein, so dass aus starker Zustimmung zu einer Frage auch ein hoher Skalenwert resultiert. Manche Items können auch umgepolt werden, um Antworttendenzen vorzubeugen. Dies muss dann allerdings bei der Verrechnung berücksichtigt werden. Schliesslich muss berücksichtigt werden, dass man eine nicht beantwortete Frage (nicht angekreuzt = Missing Data) von einer Frage, die für die Person nicht zutrifft unterscheiden kann.

3) Testdurchführung

Die nun erstellte Testvorform muss sich an einer Stichprobe bewähren. Gibt es Fragen die nicht eindeutig formuliert sind, die nicht zutreffen oder die von allen Personen gleich beantwortet werden?

4) Itemanalyse mit dem Computer (vgl. Kapitel: Durchführung einer Itemanalyse mit dem Computer (SPSS & SYSTAT), Kriterien zur Itemselektion)

Die Kriterien zur Itemselektion wie Trennschärfe und Itemschwierigkeit können berechnet werden. Ein weiteres Kriterium der Güte eines Items ist, wie gut es zu der Skala passt. Diese Frage der Homogenität der Items kann durch die Faktorenanalyse untersucht werden. Schliesslich kann die Reliabilität der Skala berechnet werden.

5) Testendform

Am Ende sollte der bereinigte Test stehen. Items, die sich in der Vorversion als ungeeignet erwiesen haben sind ausgesondert worden. Um den Test zu veröffentlichen muss er an einer möglichst repräsentativen Stichprobe normiert werden. Für verschiedene Teilstichproben (Altersgruppen, Geschlecht) werden die Mittelwerte und die Standardabweichungen auf der Skala berechnet und in die Normalverteilung transformiert (Diehl & Kohr, 1989, S. 137-148). Validitätsuntersuchungen müssen schliesslich zeigen, ob der Test das misst, was er vorgibt zu messen, wo der Test seinen Anwendungsbereich hat, und ob er im Vergleich zu anderen Tests zusätzliche Informationen liefert.

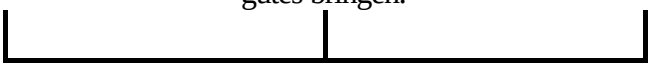
7.2. Einstellung und Verhalten

7.2.1. Historisches

Das psychologische Konzept "Einstellung" spielte eine Hauptrolle in der Geschichte der Sozialpsychologie. In Untersuchungen zeigte sich, dass Einstellungen die Gedanken und schliesslich auch die Handlungen von Menschen beeinflussen können. Thomas & Znaniecki (1918) definierten Einstellungen als individuelle, mentale Prozesse, die die möglichen und tatsächlichen Handlungen einer Person determinieren. Einstellungen wurden zur Erklärung von Verhalten verwendet und damit als Verhaltensdispositionen gesehen.
Damit wurde es notwendig Einstellungen möglichst valide zu erfassen und zu messen. Thurstone (1931) definierte Einstellungen als das Gefühl für oder gegen ein psychologisches Objekt. Er postulierte damit ein Kontinuum zur Einstellungsmessung von "positiv" zu "negativ" oder "stimme zu" bis "stimme nicht zu". Den Wert (Score) einer Person auf dem Kontinuum erhält man, indem man

deren Meinung bzw. Glaube als verbalen Ausdruck der Einstellung erfragt. So unterscheiden sich die folgenden drei Person auf dem Kontinuum der Einstellung zum Krieg:

Person A:	Person B:	Person C:
Es gibt keinen Fortschritt ohne Krieg.	Es ist schwer zu entscheiden, ob Kriege mehr schlechtes als gutes bringen.	Krieg ist ein sinnloser Kampf, der in Selbstzerstörung endet.



Thurstone entwickelte nun verschiedene Methoden der Zuweisung des Scores zu Personen. Die bekannteste ist seine "equal-appearing intervall scale", die in mehreren Schritten abläuft:

a) Sammlung von Einstellungsitems, die mit dem Einstellungsobjekt zusammenhängen.
b) Vorgabe der Itemsammlung (Itempool) an Personen, die repräsentativ für die zu untersuchende Population sind.
c) Diese Personen sollen die Items (auf Kärtchen) entlang einem Kontinuum mit 11 Abstufungen (Kästchen) einordnen. Das Kontinuum geht von positiv zu negativ und soll gleiche Abstände haben.
d) Der Durchschnittswert der Abstufung in die das Item gelegt wurde ergibt dessen Skalenwert.
e) Items, über die die Personen sich uneinig sind (criterion of ambiguity), werden rausgeworfen.
f) Die resultierende Skala besteht aus c.a. 20 Items, die etwa den selben Abstand auf dem Kontinuum haben.
g) Die Items können nun den eigentlichen Versuchspersonen vorgegeben werden. Es wird der Mittelwert der Skalenwerte der Items berechnet, denen eine Versuchsperson zustimmt.

Diese Skalierungsmethode wurde sehr viel verwendet, ist aber mit einem riesigen Konstruktionsaufwandt verbunden. 1932 schlug Likert eine viel einfachere Methode vor:

a) Sammlung von Einstellungsitems (s.o.)
b) Der Untersucher entscheidet nun, welche Items positive bzw. negative Einstellungen gegenüber dem Einstellungsobjekt repräsentieren.
c) Neutrale und unklare Items werden eliminiert.
d) Die restlichen Items werden direkt den Versuchspersonen vorgegeben, und die Personen müssen meisst auf einer 5-Punkte-Skala ankreuzen:

stimme stark zu	stimme zu	unentschieden	stimme nicht zu	stimme gar nicht zu
❑	❑	❑	❑	❑
(5)	(4)	(3)	(2)	(1)

e) Für jede Antwort auf dem Item wird der Score 1 (für starke Ablehnung) bis 5 (für starke Zustimmung) vergeben. Für negative Items werden die Antworten umgepolt.
f) Durch aufsummieren über alle Items erhält man den Skalenwert.
g) Nun können Trennschärfe und Interne Konsistenz berechnet werden, um eine letzte Auswahl der Items zu treffen.

Thurstone konstruierte seine Skala, um auf dem Intervallskalenniveau zu messen (Voraussetzung für die Pearson-Produkt-Momentkorrelation). Likert konnte nun zeigen, dass seine Ordinalskala empirisch zu den selben Ergebnissen führte.

Diese führte zur Entwicklung vieler Einstellungsfragebögen und -untersuchungen (z.B.: Jüdische Studenten in den USA hatten eine positivere Einstellung zur Geburtenkontrolle und zu Kommunismus als katholische und protestantische Studenten). Die eindimensionale Erfassung von Einstellungen wurde allerdings schon bald von Allport (1935) kritisiert. Zudem beantwortet die Frage, ob sich Gruppen von Personen in ihren Einstellungen unterscheiden nicht, welcher Zusammenhang zwischen Einstellung und Verhalten besteht.
Die erste bekannte Studie dazu lieferte LaPiere 1934. Er begleitete ein chinesisches Ehepaar auf ihrer Reise durch die USA. In 251 Restaurants, Hotels und Unterkünften wurden sie nur einmal nicht bedient. Als LaPiere 6 Monate später die Gaststätten per Post befragte, ob sie als Gäste in ihrem Etablissement auch Angehörige der chinesischen Rasse aufnehmen würden, antworteten 90% der 128 Gaststätten mit "Nein"!
Viele ähnliche Forschungen zeigten die Diskrepanz zwischen Einstellung und Verhalten.

Die Erklärung "Einstellung sei ein mehrdimensionales Konstrukt" und die weitere Entwicklung der Einstellungsmessung (Guttman, 1944 und später Osgood, Suici & Tannenbaum 1957) waren Versuche die Diskrepanz zu beheben.
Campbell (1963) argumentierte allerdings anders. Für ihn sind Einstellungsäusserungen und offenes Verhalten gegenüber dem Einstellungsobjekt Manifestationen derselben zugrundeliegenden, latenten Eigenschaft (Disposition), aber mit unterschiedlichen Schwierigkeiten. Es ist natürlich einfacher in einem Brief zu behaupten man würde Chinesen nicht bedienen, als in der realen Situation.
Zudem konnte die Forschung zeigen, dass Einstellung nur ein Faktor unter vielen ist, der Verhalten beeinflussen kann. Die "Theorie of reasoned action" von Ajzen & Fishbein (1980) bezieht die bisherigen Ueberlegungen ein. Einstellungen werden als die Bewertung eines psychologischen Objektes von einer Person definiert. Es wird unterschieden zwischen Glaube, Einstellung, Intention und schliesslich Verhalten. Die Theorie hat in verschiedenen Anwendungsgebieten gezeigt, dass Verhalten tatsächlich durch die postulierten Faktoren vorhergesagt werden kann. Es soll nicht weiter auf die Theorie eingegangen werden (vgl. hierzu Ajzen & Fishbein, 1980), aber für die Testkonstruktion ist die Messung und Definition von Verhalten, wie sie die Autoren vornehmen von grosser Bedeutung.

7.2.2. Definition und Messung von Verhalten

Verhalten steht oft im Blickpunkt von Sozialpsychologen, wie z. B. bei den folgenden Fragen: Warum rauchen Menschen? Wie bringt man jemanden dazu, Geld für eine gute Sache zu spenden? Warum wählt jemand eher christdemokratisch als sozialdemokratisch?
Zunächst muss in der Forschungsfragestellung unterschieden werden, ob man Verhalten, oder das Ergebnis von Verhalten messen will. Ergebnisse von Verhalten, wie z. B. Examenserfolg gehen zum Einen zurück auf Verhaltensweise wie z. B. Vorlesungen besuchen, Bücher lesen oder auch abschreiben in Klausuren. Andererseits können Ergebnisse von Verhalten auch durch andere Faktoren beeinflusst sein, wie z. B. der Schwierigkeitsgrad der Prüfung. Ergebnisse von Verhalten sind daher schwieriger zu erfassen als Verhalten an sich.
Neben der Unterscheidung von Verhalten und Ergebnisse von Verhalten muss man den Unterschied zwischen spezifischem Verhalten (single actions) und Verhaltenskategorien (behavioral category) beachten. Wir können beobachten, wie ein Student ein Buch liest (single act), wir können nicht beobachten wie er "studiert" (behavioral category). Auf die Verhaltenskategorie "studieren" kann nur durch die Beobachtung von single acts geschlossen werden. Um einen möglichst reliablen Schluss auf die Kategorie zu treffen, müssen mehrere single acts gemessen werden. Um einen möglichst validen Schluss auf die Kategorie zu treffen müssen die single acts repräsentativ für die Verhaltenskategorie sein.

Tabelle 9: Beispiel zum Diätverhalten

Single Actions:	**Wertkodierung:**
Naschen zwischen den Mahlzeiten	-1
Eis essen	-1
Kaffee ohne Zucker trinken	+1
Niedrigkaloriengetränke trinken	+1
Ein süsses Dessert nach der Hauptmahlzeit essen	-1
Bier trinken	-1
Diätpillen schlucken	+1
Fettes Fleisch essen	-1

Entsprechend dieser Unterscheidung können auch Einstellungen untergliedert werden. Allgemein sollte zur Konstruktion eines Fragebogens solch eine Aufschlüsselung vorgenommen werden.

7.2.3. Verhaltensbeobachtung und Selbstbeschreibung (self report) von Verhalten

Verhalten kann nicht immer direkt beobachtet werden. So können wir einem Wähler in der Wahlkabine nicht über die Schulter schauen und zusehen, wo er sein Kreuz macht. Wir können eine Person aber befragen, ob bzw. wie oft sie ein bestimmtes Verhalten zeigt.
Solche Selbstbeschreibungsdaten, wie sie in vielen Fragebögen verlangt werden, können falsch sein. Sie sind weniger objektiv als die Beobachtungsdaten. Selbstbeschreibungsdaten haben aber den Vorteil, dass sie viel leichter erhoben werden können. Man braucht weniger Zeit und Geld im Vergleich zu einer Beobachtung. Wenn wir daran interessiert sind, ob eine Person Diätverhalten zeigt, ist es einfacher die single actions abzufragen, als sie bzgl. jeder einzelnen Handlung zu beobachten. Man könnte die Person natürlich auch direkt fragen: "Hältst Du Diät?". Diese Messung der Verhaltenskategorie wäre allerdings nicht sehr objektiv und reliabel, da es sehr stark von der Selbstwahrnehmung der Person abhängt, wie sie darauf antwortet und wir nur eine Messung (eine Frage) für das Verhalten bekommen.

7.3. Wichtigste Skalen zur Messung von affektiven Prozessen

1. Likert Skala:

gebräuchlichstes Verfahren zur Einstellungsmessung
- unaufdringlicher Titel
- Instruktion
- Umpolung, Verrechnung

Beispiel:

Dieses Buch ist gut

ja	eher ja	weder noch	eher nein	nein
❑	❑	❑	❑	❑

2. 2 Punkte Skala:

vermeidet Mittelkategorien
zwingt Vp vielleicht in eine Richtung (forced choice)

Beispiel:

Dieses Buch ist öde

ja	nein
❑	❑

3. Adjektiv Checkliste:

Verrechnung von gewählten "positiven" und "negativen" Adjektiven. Der Stimulus kann wechseln, die Adjektiv-alternativen können gleich bleiben

Beispiel:

Diese Buch ist ...

❑ ... aufregend
❑ ... langweilig
❑ ... informativ
❑ ... wertvoll

4. Bipolare Adjektivskala oder Semantisches Differential:

Drei Dimensionen:
- Bewertung (gut vs. schlecht)
- Potenz (stark vs. schwach)
- Aktivität (aktiv vs. passiv)

(vgl. Osgood, 1957)

Beispiel (fünfstufig):

Dieses Buch ist ...

gut	❑	❑	❑	❑	❑	schlecht
interes.	❑	❑	❑	❑	❑	langweilig
angenehm	❑	❑	❑	❑	❑	unangenehm

...

5. Normierungsprozedur:

Kontext der Normierung muss klar sein. Ermöglicht relative Bewertung von Sachverhalten, Personen, ... (vgl. Soziogramme in der Soziometrie z.B. Identifikation von sozialen Stars).

Beispiel:

Nenne bitte drei der von Dir bisher gelesenen Büchern, die Dir am besten gefielen.

1. _____
2. _____
3. _____

7.4. Richtlinien zur Formulierung von Einstellungsitems

1. Vermeidung von Tatsachenfeststellungen, sie gehören in den Bereich der Messung des kognitiven Bereichs.
2. Vermeidung von Vergangenheitsformulierungen, es wird kein Bericht über vergangenes Verhalten gewünscht (man muss sich in die Situation hineinversetzten können).
3. Die Einstellungsäusserung darf nicht mehrere Interpretationsmöglichkeiten offen lassen, vor allem die Bewertungsdimension muss klar sein ("Ich finde es gut dass ...").
4. Irrelevante Bezüge vermeiden, nur das betreffende Einstellungsobjekt darf angesprochen sein (Person, Ding, Idee, Erlebnis).
5. Keine Items aufnehmen, von denen man annehmen kann, dass sie von allen Leuten gleich beantwortet werden (Es soll eine Differenzierung von Personen hinsichtlich ihrer Einstellungen möglich sein).
6. Die volle Bandbreite der Items soll erfasst werden, deshalb ist eine theoretisch-inhaltliche Aufschlüsselung des Einstellungsbereichs notwendig.
7. Einfach, klar und direkt formulieren, Items sollen nicht Verständnisfähigkeit messen.
8. Einstellungsäusserungen sollten kurze Behauptungen von nicht mehr als 20 Wörtern Länge sein.
9. Es darf nur ein Gedanke pro Behauptung vorkommen.
10. Wörter wie "alle", "immer", "keine", "nie" sollten vermieden werden, sie erzeugen entweder Mehrdeutigkeit oder erzwingen eine bestimmte Antwort aus logischen Gründen.
11. Doppelte Verneinungen vermeiden.

Verfahren 3. & 4. verwenden die Einstellungsobjekte selbst als Stimuli.
Beim Verfahren 5. werden diese erst erfragt.

Bei allen Versuchen der Messung von affektiven Prozessen kommt der Diskussion von ethischen Gesichtspunkten eine besondere Bedeutung zu
=> Privatheit, Freiwilligkeit respektieren.
=> Aufklärung darüber, warum etwas gemessen wird.
=> Diskussion der Ergebnisse einer Messung in einem psychologischen Gespräch.

7.5. Antwort-Items zur Messung von Wissen und Verständnis (kognitiver Bereich)

1. Unstrukturiertes Format: Satz, Phrase oder Zahl wird erfragt:

Beispiel:	Wer war der 17. Präsident der USA? _____ 1/8 + 1/5 = __
Vorteile:	- minimierte Ratewahrscheinlichkeit - keine Hinweisreize wie bei Multiple-Choice - leicht zu konstruieren
Nachteil:	- Schwierigkeiten bei der Auswertung, wenn die Antwort des Probanden nicht genau mit der tatsächlich richtigen übereinstimmt.

Anwendung: Messung spezifischen Wissens (Mathe, Wissenschaft).

Konstruktion:	- Antwort soll in möglichst wenigen Worten wiedergegeben werden können. - einfache, präzise Formulierung. - es soll genau eine richtige Antwort geben.

2. Satzergänzungsformat:

Beispiel:	Eine Verhältnisskala unterscheidet sich von einer Intervallskala durch _____ .
Vorteile:	- bei genauer Formulierung leicht auszuwerten. - weniger Hinweisreize wie bei Multiple-Choice
Nachteil:	- Schwierigkeiten bei der Konstruktion (keine Hinweisreize)
Konstruktion:	- genau so viel Lücken lassen, dass das Item einen mittleren Schwierigkeitsgrad aufweist. - es soll genau eine richtige Antwort existieren (Wort, Phrase) - die Grammatik des Satzes darf das Auffinden der richtigen Lösung nicht erleichtern.

3. Richtig-Falsch/ Ja-Nein - Format:

Beispiel:
- Eine Intervallskala enthält einen absoluten Nullpunkt.
  ❑ richtig ❑ falsch

- Kreise "Ja" ein bei richtiger, "Nein" bei falscher Pluralbildung:

Atlasse	ja	nein
Skalas	ja	nein
Mäuse	ja	nein

Vorteile:
- sehr leicht zu konstruieren.
- schnell zu beantworten.

Nachteil:
- es können Interpretationsschwierigkeiten auftreten.
- hohe Ratewahrscheinlichkeit (=> viele Items verwenden).

Konstruktion:
- Vermeiden absoluter Ausdrücke wie "immer", "niemals", "alle".
- Erst Konstruktion von wahren Items , die dann zur Hälfte umgedreht und in falsche Items verwandelt werden.
- Zufallsreihenfolge der Items.

4. Zwei-Wahl-Klassifikationsformat: Eine Reihe von Reizen sollen klassifiziert werden.

Beispiel: Unterstreiche diejenigen Städte, die in Europa liegen:
a) Brüssel b) Atlanta c) Ulm d) Lyon e) Belfast

Vorteile:
- Ratewahrscheinlichkeit ist relativ niedrig.
- werden als weniger einengend empfunden als z.B.: Multiple Choice

Konstruktion:
- Eindeutige Klassifikation muss möglich sein
- Es müssen nicht unbedingt genauso viele Exemplare wie Nichtexemplare sein.
- zuerst die Klassifikationskategorien aufstellen, und dann Exemplare und Nichtexemplare finden.

5. Multiple-Choice - Format: 3-5 Antwortalternativen, eine davon ist korrekt

Vorteile:	- leichte Verrechenbarkeit. - unkorrekte Antwortmuster sind leicht zu analysieren.
Nachteil:	- richtige Antwort kann evtl. ohne Vorwissen gefunden werden (durch Hinweisreize). - raten kann zu Erfolg führen (=> Verrechnungsprozedur als Ausgleich: z.B.: pro Fehler ein Viertel Punkt Abzug). - man muss plausible Alternativen finden. - Vortestanalysen sind notwendig, um den Kontrast zwischen richtigen und falschen Alternativen zu verstärken.
Konstruktion:	- wahrscheinliche Fehler der Probanden bei der Konstruktion der falschen Alternativen mitberücksichtigen. - unkorrekte Alternativen sollten auch wirklich falsch sein. - unkorrekte Antworten sollten in Länge, Komplexität und grammatikalischer Form der richtigen Antwort vergleichbar sein. - Fragen und Antwortalternativen verständlich formulieren. - Items sollten eindeutig interpretierbar sein. - Bei Antwortalternativen vermeiden von "immer", "niemals", "alle". - Hinweisreize auf richtige Antwort vermeiden. - immer nur einen Aspekt pro Item testen. - korrekte Antworten der Items an zufälliger Stelle (gegen Tendenzen). - Antwortalternativen sollten kurz sein.

6. Paarbildungsformat (matching): Mehrere Fragen und Antworten gleichzeitig in einem Item, zur Unterscheidung zwischen ähnlichen Fakten oder Aspekten.

Beispiel:	1) the argument	a) synthesis
	2) the opposing argument	b) prethesis
	3) the resolving argument	c) thesis
		d) antithesis
Vorteile:	- mit einem Item kann viel abgefragt werden.	
	- Probanden haben Spass.	
Nachteil:	- sehr schwer zu konstruieren.	
	- Tendenz zu Hinweisreizen.	
Konstruktion:	- Item sollte eindimensional sein.	
	- Vermeiden von Hinweisreizen.	
	- kurze Antworten, unabhängig, nicht überlappend.	
	- plausible unkorrekte Antworten sind notwendig.	
	- Antwortpattern dürfen nicht systematisch sein.	
	- räumliche Trennung von Fragen und Antworten.	

7.6. Durchführung einer Itemanalyse mit dem Computer (SPSS & SYSTAT)

Zweck der Itemanalyse ist es, anhand von statistischen Kennwerten unter Berücksichtigung von inhaltlichen Gesichtspunkten zu prüfen, ob die Items zur Messung eines Konstrukts wirklich geeignet sind. Ungeeignete Items sollen ausgelesen werden.

1. Dateneingabe: Man erstellt ein Dateneingabeblatt (Worksheet). Jede Zeile des Blattes repräsentiert eine Vpn. Die Spalten sind die Variablen, wobei die ersten Variablen die Vpn-Nummer und demographische Angaben der Personen beinhalten, dann kommen die einzelnen Items:

Abbildung 18: Organisation der Daten für Computereingabe

	Variablen:						
	VPN	Alter	Geschl.	Item1	Item2	...	$Item_n$
	1	25	1	2	3		2
Personen:	.						
	.						
	.						
	100	30	0	1	3		1

Legende: Es wurden 100 Personen (Vpn) eingegeben. Damit hat das Dateneingabeblatt 100 Zeilen. Nach dem Alter und dem Geschlecht folgen die eigentlichen Fragen. Die Anzahl der Variablen definieren die Spalten der Eingabe.

Die Eintragungen auf diesem Blatt werden nun (oder direkt) in den Computer eingegeben.
Beachte: Die Werte einer Spalte müssen untereinander stehen, am besten durch einen Leerschlag (Space-Taste) getrennt.

Anmerkung: Im folgenden ist der genaue Ablauf für die Computerbenutzung angegeben. Es wurden dabei Programme berücksichtigt, die an der Universität Freiburg/ Schweiz verfügbar sind. Es handelt sich sicher um die weitverbreitetsten Programme. Dennoch muss man einige lokale Gegebenheiten berücksichtigen. Deshalb sind die Befehle, die sich von anderen Standorten unterscheiden können mit einem "*****" gekennzeichnet.

Notation: < > Eingabe durch den Anwender
<BEGIN DATA> Eingabe der Buchstaben auf der Tastatur
<*CTRL*> Drücken spezieller Tasten
<*CTRL*>+<*Z*> Gleichzeitiges drücken zweier Tasten

VAX (SPSSX):	Anmelden:	*****<C UFFAU1>
	User:	*****<PSY>
	Password:	*****<UHR>
	Directory wechseln:	*****<SET DEF [PSY.TESTT]>
	Editieren:	<ED testname.SPS>
	Saven:	<*CTRL*>+<*Z*> <EXIT>
	oder nicht Saven:	<*CTRL*>+<*Z*> <QUIT>

so entsteht der ASCII-Eingabefile

Durch einige SPSSX-Systembefehle entsteht ein SPSS-Programmfile:

Editieren:	<ED testname.SPS>
1. Zeile: VAX-Steu.:	*****<$ SET DEF [PSY.TESTT]>
2. Zeile: VAX-Steu.:	*****<$ SPSSX/OUP=X.X>
3. Zeile: SPSS-Bef.:	<DATA LIST FILE=INLINE
(eine Zeile)	RECORDS=1 NOTABLES>
4. Zeile: Var.-Def.:	< /1 VPN 1-4 ALTER 6-7
(eine Zeile)	GESCHL 9 ITEM1 11> u.s.w.
5. Zeile: Datenanf.:	<BEGIN DATA>
ab 6. Zeile :	Daten
Zeile7 + N:	<END DATA>
Zeile 8 + N:	<weitere SPSSX-Befehle>
letzte Zeile:	<FINISH>
Saven:	<*CTRL*>+<*Z*> <EXIT>

Dieser Programmfile kann nun ausgeführt werden:

Submit:	<SUBNO testname.SPS>

Das Ergebnis steht auf dem SPSS-Ausgabefile:

Anschauen:	<ED X.X>
oder Drucken:	<PRINT X.X>

PC (SYSTAT):	Programmaufruf:	<SYSTAT>
	Data-Modul	<DATA>
	Systemeditor:	<EDIT>
	Oben die Variablen definieren. Dann die Daten eintragen.	
	Saven:	<*ESC*> <SAVE testname.SYS>

Systat arbeitet mit einem Systemeditor, der den Umweg über die ASCII-Daten umgeht.

2. Datenkontrolle: Jeder macht einmal Fehler. Deshalb ist es umso wichtiger die Eingabe der Daten zu kontrollieren. Am schnellsten gelingt dies mit deskriptiven Statistiken:

VAX (SPSSX):	Editieren:	<ED testname.SPS>
	nach END DATA	<DESCRIPTIVES VAR=ALL>
	und	<FREQUENCIES VAR=ALL /BARCHART>
	letzte Zeile:	<FINISH>
	Saven:	<*CTRL*>+<*Z*> <EXIT>
	Submit:	<SUBNO testname.SPS>
PC (SYSTAT):	Stats-Modul:	<STATS>
	Systemfile holen	<USE testname.SYS>
	Berechnungen:	<STATISTICS>
	Graphic-Modul:	<GRAPH>
	Systemfile holen	<USE testname.SYS>
	Barcharts:	<BAR>

3. Faktorenanalyse: Die unrotierte Faktorenladungsmatrix der Faktorenanalyse bietet einige Anhaltspunkte zur Selektion der Items (Trennschärfe; welche Items laden hoch auf einem Faktor, können also zu einer Skala zusammengefasst werden; welche Items haben eine negative Ladung, müssen also umgepolt werden).

VAX (SPSSX):	Editieren:	<ED testname.SPS>
	nach END DATA	<FACTOR VAR=ITEM1 ITEM2>
		usw.
	letzte Zeile:	<FINISH>
	Saven:	<*CTRL*>+<*Z*> <EXIT>
	Submit:	<SUBNO testname.SPS>

Das Ergebnis steht auf dem SPSS-Ausgabefile:

	Anschauen:	<ED X.X>
	oder Drucken:	<PRINT X.X>
PC (SYSTAT):	Faktor-Modul:	<FACTOR>
	Systemfile holen	<USE testname.SYS>
	Faktorenanalyse:	<FACTOR>

4. Analyse nach der klassischen Testtheorie: Um die Programme sinnvoll einzusetzen muss man sich vorher im klaren sein: Sind die Items richtig gepolt? Welche Items will ich zu einer Skala zusammenfassen (inhaltlich und/ oder nach der FA).

VAX (SPSSX):	Editieren:	<ED testname.SPS>
	nach END DATA	<RELIABILITY VAR=
	(eine Zeile)	ITEM1 ITEM2
	(neue Zeile)	/SCALE(Skala1)=ITEM1 ITEM2
		/STAT=ALL
		/SUMMARY=ALL>
	letzte Zeile:	<FINISH>
	Saven:	<*CTRL*>+<*Z*> <EXIT>
	Submit:	<SUBNO testname.SPS>

Das Ergebnis steht auf dem SPSS-Ausgabefile:

	Anschauen:	<ED X.X>
	oder Drucken:	<PRINT X.X>

PC (SYSTAT):	Teststatistik-Modul:	<TESTAT>
	Systemfile holen	<USE testname.SYS>
	KTT:	<MODEL=CLASS>
	lange Ausgabe:	<PRINT=LONG>
	falls Umpolung nötig:	<KEY=+,-> usw.
	Prog. ausführen:	<ESTIMATE ITEM1 ITEM2> usw.

Beachte: Systat schreibt das Ergebnis der Analysen nur auf den Bildschirm. Mit dem Befehl <OUT> kann man die Ausgabe auch umleiten. <OUT> muss nach dem Modulaufruf stehen:

<OUT *>	Ausgabe auf dem Bildschirm
<OUT @>	Ausgabe auf den Drucker
<OUT testname>	Ausgabe auf den File testname.DAT

5. Beenden der Computersitzung:

VAX:	Abmelden:	<LOGOUT>
	Abschalten.	

PC (SYSTAT):	aus dem Programm:	<QUIT>
	Abschalten.	

7.7. Kriterien zur Itemselektion

7.7.1. Trennschärfe

Die Korrelation des Einzelitems mit dem Skalenwert gibt an, ob das Item inhaltlich wirklich zu der Skala passt. Personen, die niedrig auf dem Item angekreuzt haben, sollten auch beim Summenwert relativ niedrig abschneiden. Bei Personen, die hoch angekreuzt haben gilt das entsprechende. Die Frage ist also, wie gut das Einzelitem Personen mit niedrigen und mit hohen Summenwerten trennen kann. Die Trennschärfe sollte >,– .30 sein (siehe unrotierte Faktorenladungsmatrix).

7.7.2. Itemschwierigkeit

Bei dichotomen Items (0/ 1) gibt der Mittelwert die Anzahl der richtigen Antworten im Verhältnis zur Gesamtzahl der Antworten an. Zu schwere (z.B.: <.10) oder zu leichte Items (z.B.: >.90) sollten ausgesondert werden. Ist man allerdings an einer Differenzierung der Personen im Extrembereich interessiert, sollte man diese Items beibehalten. Bei einer mehrfach abgestuften Skala (z.B.: Likert-Skala) kann man den Mittelwert und die Standardabweichung der Stichprobe auf diesem Item betrachten. Die Antworten sollten einigermassen normalverteilt sein (siehe BARCHARTS).

7.8. Uebungsaufgaben zur Testkonstruktion

1. Welche Informationen sollte das Handbuch eines veröffentlichten Tests enthalten? Nenne vier grosse Bereiche.

2. Welche Kriterien zur Itemselektion kennst Du?