I. Testkonstruktion: Planung und Entwurf

1) Testvorbereitung
2) Konstruktionsprinzipien
3) Itemgenerierung
4) Fehlerquellen

1. Testvorbereitung

Bestandteile psychologischer Tests

Testmanual

- Beschreibung, Entstehung und wissenschaftliche Grundlegung des Tests
- Durchführungs-und Interpretationsanweisungen, Normen

Testmaterial

- Stimulus und Reaktionsmaterial

Auswertungshilfen

- Kontrollblätter, Lochfolien, Schablonen

Durchführungsbestandteile

- Testanweisung für den Testleiter (Bedienungen, Durchführung und Auswertung des Test

Vorbereitungsfragen

a) Was soll der Test messen?
   - Definition des Merkmals z.B. in Fachliteratur oder eigene Arbeitsdefinition
   - Konzeptuelle Einengung oder Erweiterung
   - ein-oder mehrdimensional

b) Für wen soll der Test geeignet sein?
   - Itemformulierung (Verständlichkeit)

c) Welchen Verwendungszweck soll der Test haben?
   - Geltungs-oder Anwendungsbereich
   - Einsatzbedingungen
     - Erforderliche Expertise, Durchführungsbedingungen (paper-pencile, Computer, Gruppe, Zeit)

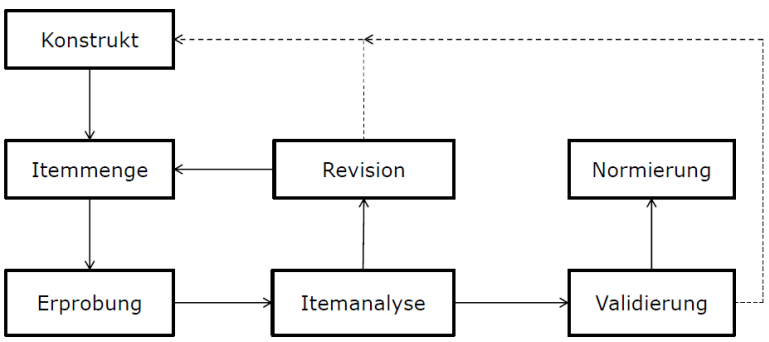
2. Konstruktionsprinzipien

4 verschieden Konstruktionsansätze

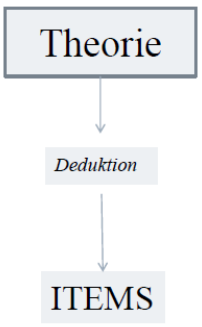
- Rationale (deduktive) Konstruktion
- Induktive Konstruktion
- Externale Konstruktion
- Prototypenansatz

Rationale (deduktive) Konstruktion

= Theoriegeleitet; Deduktion; von allg. Theorie auf Spezifisches schließen



1. Definition und Spezifikation des interessierenden Konstrukts
   - Methode der Deduktion (das Ableiten)
   - Basis der Ableitung -> elaborierte Theorie
   - Definition und Spezifikation des Konstrukts (Eingrenzung oder Erweiterung)
   - Expertengruppe
2. Identifizierung von Verhaltensindikatoren
   - Verhaltensindikatoren = beobachtbare Verhaltensweisen in denen sich das Merkmal manifestiert
3. Testentwurf (vorläufiger Itempool)
   - Formulierung in Statement-oder Fragebogenform
4. Erprobung des Entwurfs
   - Statistische Itemanalyse und (konfirmatorische) Faktorenanalyse
5. (gegebenenfalls) Überarbeitung des Testentwurfs
6. Bestimmung der Testgütekriterien und Normierung



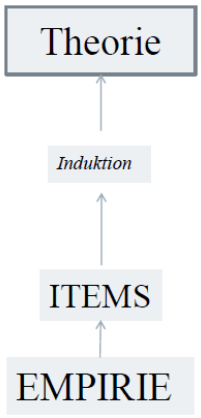
Beispiel: IST-2000

- 2 Theoretische Grundlagen:
  - Primärfaktoren von Thurstone, Catell, Spearman
  - Fluide & kristalline Intelligenz von Horn und Catell
- Zeigt: mehrere Theorien verknüpfbar/reduzierbar

Induktive Konstruktion

= Empirie-geleitet

- keine klare Vorstellung wie Merkmal oder Konstrukt beschaffen
- Generierung von Items auf Basis vager Vorstellung und Literatur
- Gruppierung von Items über die Korrelation
- Durchführung einer exploratorischen Faktorenanalyse
- Interpretation der Faktoren
- (gegebenenfalls) Ableitung einer Theorie



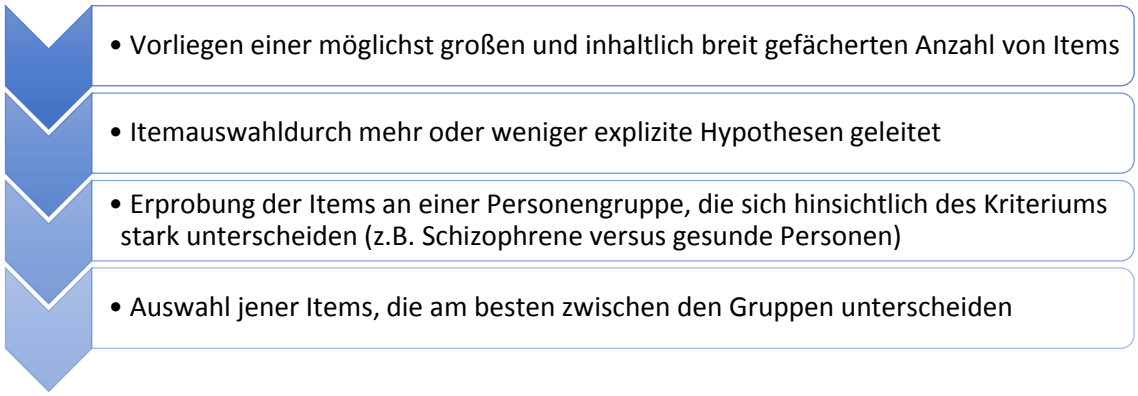
Beispiel: Big Five

- Sedimentationshypothese/Lexikalischer Ansatz
- Allport & Odbert(1936) –18000 Wörter die benutzt werden können, um menschliches Verhalten voneinander zu unterscheiden
- Cattell (1957) extrahierte 4500 Eigenschaftswörter im engeren Sinne
- Personen schätzen andere Personen anhand der ausgewählten Eigenschaften ein
- Faktorenanalyse führte zu 12 (Bekanntenbeurteilungs)-Faktoren von Cattell
- andere Autoren (z.B. Goldberg, 1993, John et al. 1988, John, 1990) fanden mit dem gleichen Variablensatz hingegen nur 5 Faktoren → Big Five
- Entstehung von **5 Dimensionen** mit versch. Inhaltsformaten
  - Über unterschiedlichen Stichproben (Länder, Kinder) und methodische Vorgehensweisen (Selbstbericht, Fremdbericht)
- *Theorieableitung: Five-Factor-Theory*

Externale Konstruktion

= Orientierung an Personengruppen bzw. kriteriumsbezogene Skalenentwicklung

- *Ansatzpunkt*: Vorliegen verschiedener Gruppen
- *Resultat*: Instrument zur Klassifikation oder Diskriminierung von Gruppen (z.B. Schizophrenie, Depression, …)
- *Vorgehen*:



Besonderheiten externaler Tests:

- Kreuzvalidierung nötig
  - Der Psychologe benutzt die Daten der Untergruppe A, um mit ihnen eine Vorhersagegleichung für das Merkmal, das der Test messen soll, zu erstellen. Diese Vorhersagegleichung wendet er nun auf alle Mitglieder der Untergruppe B an und versucht auf deren jeweilige Ausprägung des Merkmals zu schließen. Anschließend vergleicht er die vorhergesagten Ausprägungen mit den tatsächlich vorliegenden. Die Validierung des Tests erfolgt also kreuzweise, deswegen Kreuzvalidierung. Je höher die Übereinstimmung zwischen tatsächlicher und vorhergesagter Ausprägung, umso besser, valider, ist der Test.
- Geringe Augenscheinvalidität
- Heterogene Items
- Interpretation im Sinne von Wahrscheinlichkeiten bezüglich der Gruppenzugehörigkeit

Beispiel externale Testkonstruktion

MMPI-2: Minnesota-Multiphasic-Personality-Inventory

1. Generierung eines Itempools (Originalversion)
   - 1000 Items (allgemeine Gesundheit, familiäre und eheliche Beziehung, sexuelle und religiöse Einstellungen, emotionale Zustände…)
   - Item-Erprobung an a) Kontrollpersonen und b) klinisch auffälligen Personen
   - Auswahl von 566 Items, die signifikant diskriminieren
   - Zusammenstellung von Skalen
2. Überarbeitung (MMPI-2)
   - Überarbeitung der Items (unangemessen/nicht-zeitgemäß/sexuell)
3. Beispielskalen:
   - *L - Lügenskala* (15 Items): „Manchmal möchte ich am liebsten fluchen“
   - *Hd – Hypochondrie* (32 Items) „Ich leide unter Anfällen von Übelkeit und Erbrechen“
   - *D – Depression* (57 Items) „Ich habe einen guten Appetit“ (-)
   - *Pp – Psychopathie, Soziopathie, antisoziale Persönlichkeitsstörungen* (50 Items) „Manchmal habe ich sehr gewünscht, von zu Hause fortzugehen“
   - *Sc – Schizophrenie* (78 Items) „Ich habe manchmal Angst den Verstand zu verlieren“

Prototypenansatz

- Prototypische Vorstellung von Personen mit bestimmten Merkmalen ähneln sich häufig (Cantor & Mischel, 1977)
- Sammlung prototypischer Vorstellungen

Act-Frequency-Approach (Buss & Craik, 1983):

- Stichprobe aus der Zielpopulation
- Sollen an eine Person denken, die eine bestimmte Eigenschaft hat
- Prototypische Verhaltensweisen nennen
- Ableitung von Items
- Beurteilung der Items (auf Basis einer zweiten Stichprobe) bezüglich der Prototypizität für die Eigenschaft

3. Itemgenerierung 1

Konstruktdefinition

Verhaltensindikatoren

Items

Beispiel: Konstruktdefinition: Gewissenhaftigkeit

„Personen mit hoher Merkmalsausprägung in Gewissenhaftigkeit beschreiben sich als eher zielstrebig, willensstark und entschlossen … Hohe Merkmalsausprägungen in G korrespondieren mit schulischem, akademischem und beruflichem Leistungserfolg. Andererseits können auch übertrieben hohes Anspruchsniveau, beinahe zwanghafte Ordentlichkeit und Arbeitssucht als Negativbeispiele hoher Merkmalsausprägung gesehen werden."

Adjektive (hoch): arbeitsam, ausdauernd, beharrlich, besonnen, ehrgeizig, emsig, fleißig, genau, gewissenhaft, kompetent, leistungsfähig, motiviert, ordentlich, ordnungsliebend, perfektionistisch, pflichtbewusst … zielstrebig, zuverlässig.

Adjektive (gering): arbeitsscheu, bequem, chaotisch, ehrgeizlos, faul, flatterhaft, gleichgültig, hedonistisch, inkompetent, inkonsequent, lässig, leichtfertig, leichtsinnig, nachlässig, planlos … unzuverlässig, willensschwach, ziellos.

➔ Kontrolle von Impulsen, Wünschen, Begierden
➔ Selbstkontrolle bezogen auf Planung, Organisation und Ausführung von Aufgaben

Itemtyp

= Was soll gefragt werden?

1. Beschreibung von Reaktionen
   - *beobachtbare Handlungen (Ich räume alles an seine Platz./ Bei mir bleibt schon mal was liegen.)*
   - *physische Reaktionen (Ich schwitze viel.)*
2. Eigenschaftszuschreibungen
   - *Ich bin ordentlich.*
3. Wünsche und Interessen
   - *Ich liebe Ordnung. / Ich wäre gern ordentlich.*
4. Biografische Fakten
   - *In meiner Jugend war ich unordentlich.*
5. Einstellungen und Überzeugungen
   - *Man sollte Ordnung halten.*
6. Reaktionen anderer
   - *Andere halten mich für ordentlich.*

Itemformulierung

Verständlichkeit

- Vermeide Fremdwörtern/Fachbegriffe
  - *Ich fühle mich depressiv. vs. Ich fühle mich desolat.*
- Vermeide negative Formulierungen/ Vermeidnungen
  - *Ich bin nicht oft traurig.*
  - *Niemals würde ich nicht in Notsituationen helfen.*
- Vermeide lange Items.

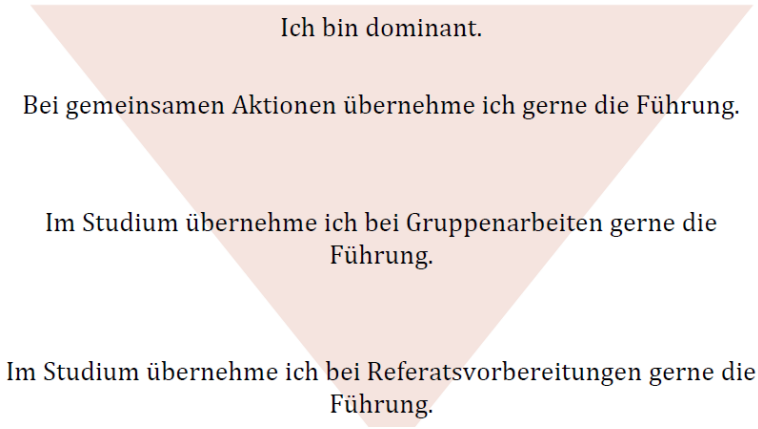
Eindeutigkeit

- Vermeide mehr als einen sachlichen Gedanken
  - *Ich fahre sehr gerne und schnell Auto.*
- Vermeide Konditionalsätze
  - *Ich fühle mich gut, wenn ich Klavier spiele*
- Vermeide Universalausdrücke
  - *Ich bin immer gut gelaunt.*
  - *Ich vergesse nie meine Hausaufgaben.*
  - *Alle Kinder machen Lärm*

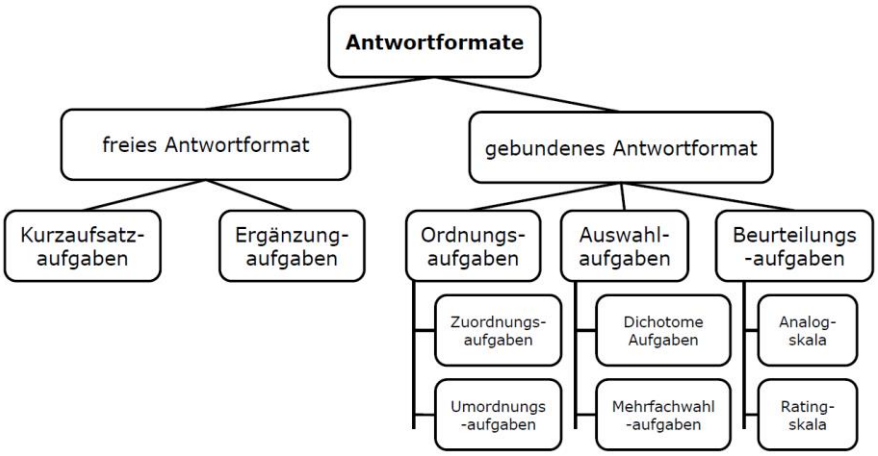
Außerdem zu vermeiden:

- Suggestion
- Gemeinplätze (Ich möchte gerne meine Ziele erreichen.)
- fehlende Passung zwischen Itemstamm und Antwortformat

Allgemeine vs. Spezifische Fragen



3. Itemgenerierung 2 (Antwortformate)



Freies Antworten

- Vorteil: keine Einschränkung der Antwortmöglichkeit, Ratetendenz vernachlässigbar
- Nachteil: Signierung notwendig (d.h. hoher Aufwand); Auswertungsobjektivität

Gebundene Antworten

- Vorteil: hohe Auswertungsökonomie, optimale Auswertungsobjektivität
- Nachteil: eingeschränkte Antwortmöglichkeit, Ratetendenz

1. Freies Antwortformat

Itemformat	Beispiele	Vor- und Nachteile
Völlig freie Antworten	Schriftlich oder mündlich z.B. bei strukturierten klinischen Interviews; projektiven Tests; Kreativitätstests	+ geeignet, wenn komplexes Denken, originelle Lösungen erfasst werden sollen - aufwendige Auswertung - Auswertungsobjektivität eingeschränkt - Abhängigkeit von Ausdrucksfähigkeit
Eingeschränkt freie Antworten	z.B. Wissensfragen oder Rechenaufgaben	+ geeignet, wenn verfügbares Wissen sowie originelle Lösungen erfasst werden sollen - aufwendige Auswertung - Auswertungsobjektivität eingeschränkt

Richtige und falsche Antworten, aber dennoch freies Ausfüllen

2. Gebundenes Antwortformat

Testpersonen wählen aus vorgegebenen Antwortalternativen die korrekte bzw. zutreffendste Alternative aus

- Zuordnungs- und Sortieraufgaben
- Multiple-Choice-Aufgaben (und Forced-Choice-Aufgaben)
- Beurteilungsaufgaben
- Aufgaben mit dichotomen Antworten

2. Gebundenes Antwortformat

Itemformat	Beispiele	Vor- und Nachteile
Zuordnungs- und Sortieraufgaben	z.B. Würfelaufgabe im I-S-T 2000 R; Bilderordnen	+ Erfassung von Wissen und Kenntnissen + objektiv und ökonomisch - nur Wiedererkennen, kein freies Abrufen von Wissen

Itemformat	Beispiele	Vor- und Nachteile
Multiple-Choice-Aufgabe & Forced-Choice-Aufgabe	Mehrere Antwortalternativen (häufig bei Leistungstests) Forced-Choice bei Persönlichkeitstests	+ objektiv und ökonomisch - gute Distraktoren schwer zu finden - nur Wiedererkennen keine freies Abrufen von Wissen/ Gedächtnisinhalten - Ratewahrscheinlichkeit hoch

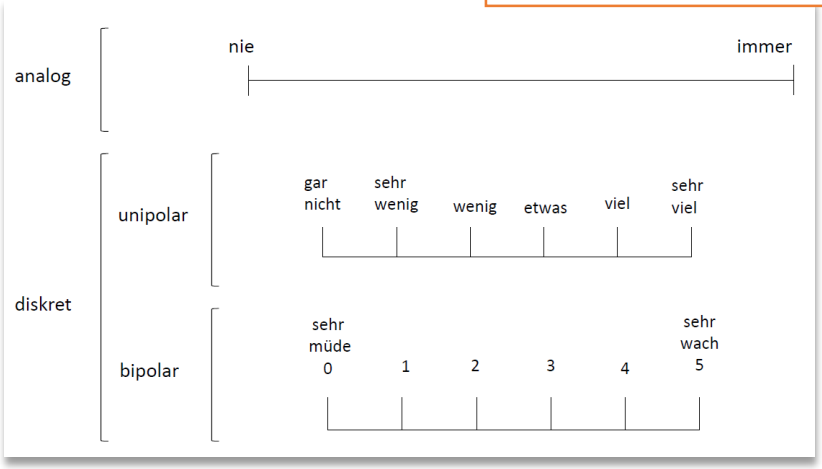
Multiple-Choice
Antworten müssen disjunkt sein

Forced Choice
Ich bin ein Mensch, der
a) ….b) ….c) ….
Achtung: Exhausitivität:
Alle Fälle abdecken

Itemformat	Beispiele	Vor- und Nachteile
Aufgaben mit dichotomen Antworten	„Ja“ oder „Nein“ „Stimmt“ oder „Stimmt nicht“ z.B. FPI-R	+ objektiv und ökonomisch (Schablone) - Entscheidung erzwungen
Beurteilungs-aufgaben	Einstufen, wie gut z.B. eine Aussage zutrifft oder wie häufig ein Verhalten vorkommt	+ objektiv und ökonomisch + differenzierte Informationen

Varianten der Beurteilungsaufgabe

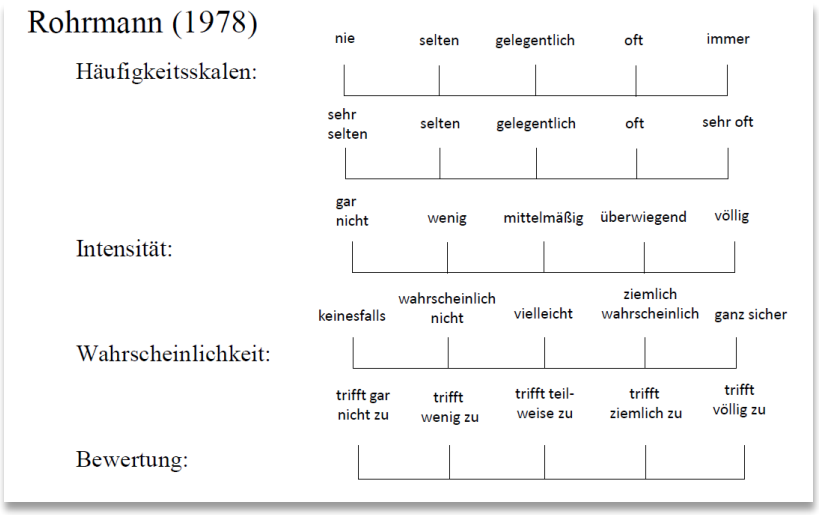
- Analoge vs diskrete Skala
- Benennung der Kategorie
- Anzahl der Antwortstufen
- Existenz einer mittleren Kategorie
- unipolar vs bipolar



+ am ehesten Intervallskalen, gut abstufbar
- Auswertung per paper-pencil

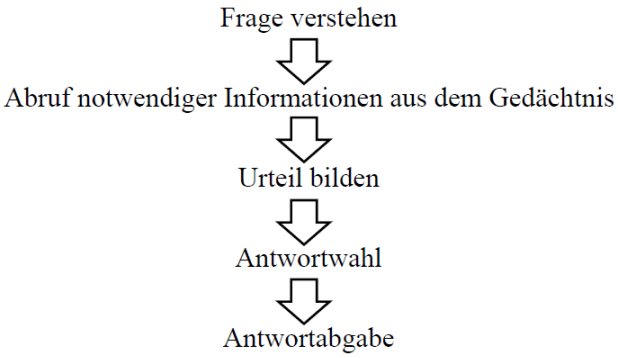
→ eher für computerbasierte Auswertung

- nur Ordinalskalenniveau (durch Aggregation Intervallskala annehmbar)
Richtlinie: 5-7 stufige Skala am besten,
aber Anpassung an Zielperson!



4. Fehlerquellen (5)

Kognitive Stadien bei der Beantwortung eines Items
(vgl. Podsakoff)



Fehlerquellen treten in bzw. zwischen den einzelnen Stadien auf und können so die Antwortabgabe beeinflussen

1. Frage verstehen:

- 1. Fehlerquelle: **Itemmehrdeutigkeit**
- Effekt: Proband versucht Informationen aus dem Kontext zu ziehen oder antwortet willkürlich
- *Beispiel: „Ich stehe unter Spannung“ -> Spannung: dynamisch (positiv) oder abgehetzt (negativ)?*
- 2. Fehlerquelle: durch **Antwortalternativen** sowie Fragekontext
- Effekt: Informationen über Intensität und Deutung aus Antwortalternativen
- *Beispiel: „Wie häufig fühlen sie sich richtig traurig?“ einmal pro Monat-Jahr vs. Woche-Monat?*

2. Abruf notwendiger Informationen aus dem Gedächtnis

- Fehlerquelle: Abruf möglicherweise durch die Stimmung beeinflusst
- Personen urteilen hier nicht nach „recall-and-count“ sondern nach **Heuristiken**
- Mögliche Lösung: offenes Antwortformat, jedoch generell schwierige Fehlerquelle

3. Urteil bilden

- Fehlerquelle: Bewertung der abgerufenen Informationen unter Einfluss/Beantwortung **vorangegangener Items**
- Effekt: globale Meinung/globales Antwortverhalten
- Fehlerquelle**: Reihenfolge von Items**
- Kommen spezifische Items vor allgemeinen, entsteht ein abhängiges Antwortverhalten
- *Beispiel: Zufriedenheit mit Partnerschaft? Zufriedenheit mit Leben?*
- Allgemeine Items sollten spezifischen Items vorangestellt werden

4. Antwortwahl

- Fehlerquelle: **Antworttendenzen**
  - Aquieszenz (Ja-sage-Tendenz)
    - Lösung: zusätzlich gleiche Anzahl an invertierten Items
  - Tendenz zur Mitte oder extremen Urteilen
    - Verzicht auf mittlere Kategorie -> mögl. Frustration
    - Statt mittlere Antwort: überspringbar -> keine Auswertungsmöglichkeit
    - Niedrige Zahl an Antwortalternativen
    - Tendenz zur Mitte kommt häufig bei unverständlichen Items
- **Reihenfolge von Items**
  - Ähnliche Items möglichst getrennt platzieren
  - Systematisch variierte Reihenfolge
  - Randomisiert

Fortsetzung: 5. Antwortabgabe nächste Seite

5. Antwortabgabe

- Fehlerquelle: **Faking** oder **Soziale Erwünschtheit**
- Faking: bewusste Verfälschung
  - Zusicherung von Anonymität
  - Bogus-Pipeline-Technik
  - Verwendung indirekter Testverfahren
- Soziale Erwünschtheit
  - Skalen zur Erfassung von sozialer Erwünschtheit
  - Selbsttäuschung vs. Fremdtäuschung (Musch et al., 2002)
  - Korrekturmaßnahme der sozialen Erwünschtheit mithilfe einer Skala
    - Vorher muss analysiert werden, wie stark soz. Erwünschtheit überhaupt im Test ausgeprägt ist.
    - Korrektur ist immer schlechter als vorherige Reduzierung von sozialer Erwünschtheit bei Itemgenerierung