

Weder die Autoren/innen, noch die Fachschaft Psychologie übernimmt irgendwelche Verantwortung für dieses Skript.

Das Skript soll nicht die Lektüre der Prüfungsliteratur ersetzen.

Verbesserungen und Korrekturen bitte an fs-psycho@uni-koeln.de mailen.

Die Fachschaft dankt den AutorInnen im Namen aller Studierenden!

Klausurfragen und Antworten zu den Themenbereichen „Methoden der Evaluation“ und „Methoden der Entwicklungspsychologie“

zusammengestellt aus den 8 Original-Klausuren vom WS 1999 bis zum WS 2002 von Kristina Siever

I Fragen zur Veränderungsmessung

Literatur:

PETERMANN, FRANZ (1978). Veränderungsmessung. Stuttgart: Kohlhammer. → Kap. 1-5 (S. 11-60).

TRAUTNER, HANNS MARTIN (1992²). Lehrbuch der Entwicklungspsychologie. Bd.1: Grundlagen und Methoden. Göttingen: Hogrefe. → Kap. 4.3.1 (S. 278-286).

Frage 1: Inwiefern ist es ein **Problem**, mit Hilfe von Methoden der **Zustandsbeschreibung** Veränderungen erfassen zu wollen? (2 x gefragt)

Veränderungsmessung

Definition: Veränderungsmessung bedeutet die **Erfassung intraindividuelle Veränderungen über die Zeit**. „Sowohl die Erfassung intraindividuelle Veränderungen als auch die der interindividuellen Unterschiede in diesen Veränderungen erfordert [...] die **wiederholte Erfassung der gleichen Individuen** über den interessierenden Veränderungszeitraum.“ (TRAUTNER, S. 229)

Veränderungsmessung ist also **Prozeßbeschreibung** durch Messung zu mehreren Zeitpunkten und geht von der **Variabilität** psychischer Merkmale über die Zeit aus (bei Wiederholung der Messung werden veränderte Meßwerte erwartet). Dadurch unterscheidet sie sich grundlegend von der Zustandsbeschreibung durch Messung zu nur einem Zeitpunkt, die auf der Annahme der Konstanz psychischer Merkmale basiert (bei Wiederholung der Messung werden gleiche Meßwerte erwartet). Im Rahmen der Zustandsbeschreibung werden Unterschiede zwischen den Meßwerten zu verschiedenen Zeitpunkten auf Meßfehler zurückgeführt; Ziel ist die Zuverlässigkeit der (zu anderen Zeitpunkten wiederholbaren) Messung, die umso höher ist, je kleiner der Meßfehler ist (Reliabilität). Innerhalb der Prozeßbeschreibung werden Unterschiede zwischen den Meßwerten zu verschiedenen Zeitpunkten dagegen auf intraindividuelle Veränderungen von Merkmalen zurückgeführt; Ziel ist die Zuverlässigkeit der Messung des Wandels über die Zeit. Diese **Konstanz-Variabilitäts-Problematisierung** macht deutlich, daß Methoden der Zustandsbeschreibung nicht zur Erfassung von Veränderungen geeignet sind (vgl. PETERMANN, S. 11 f).

Frage 2: Was versteht man unter der „**Regression zur Mitte**“, und unter welchen **Bedingungen** ist dieses Phänomen **besonders stark** ausgeprägt? Geben Sie eine kurze Begrün-

dung! (1 x gefragt) bzw. Was versteht man unter der „**Regression zur Mitte**“, und wie ist dieses statistische Phänomen zu **erklären**? (1 x gefragt)

Definition:

Unter der Regression zur Mitte, auch statistische Regression genannt, versteht man die **scheinbare Veränderung** von individuellen Meßwerten aus einer ersten Messung im Sinn einer **Tendenz zum Mittelwert der Verteilung** bei einer Wiederholungsmessung, die auf die **Meßfehlerbelastetheit** der beobachteten Werte zurückzuführen ist.

Wird bei einer Gruppe von Individuen die gleiche Variable zweimal hintereinander gemessen, ist die Abweichung der individuellen Meßwerte vom Gruppenmittelwert bei der Zweitmessung geringer als die Abweichung der Meßwerte der gleichen Individuen vom Gruppenmittelwert bei der Erstmessung. Es handelt sich hierbei um **statistische Artefakte** aufgrund von **mangelnder Reliabilität** der Meßinstrumente. Die mangelnde Reliabilität eines Merkmals hat zur Folge, daß wiederholte Messungen desselben nicht perfekt miteinander korrelieren (vgl. BORTZ & DÖRING, 1995², S. 517 ff).

Dieser Regressionseffekt ist unter folgenden **zwei Bedingungen** besonders stark ausgeprägt: Er ist um so größer,

- 1) je weiter die individuellen Ausgangswerte vom Gruppenmittelwert der Erstmessung entfernt sind (Extremwerte, unreliable Merkmale)* und
- 2) je niedriger die Korrelation zwischen den beiden Meßwertreihen ist (geringe Retest-Reliabilität → variable Merkmale, Veränderungsmessung!) (vgl. TRAUTNER, S. 279).

*Das bedeutet, daß Individuen, die bei der ersten Messung sehr hohe bzw. sehr niedrige – also extreme – Werte erzielt haben, bei der zweiten Messung weniger extreme Werte erreichen. Es kommt also zu einer Veränderung der Meßwerte über die Zeit im Sinn einer Tendenz zur Mitte.

Erklärung:

Nach der Klassischen Testtheorie setzt sich ein beobachteter Wert x_i zusammen aus dem „wahren“ Variablenwert X_i und dem (zufälligen) Meßfehler e_{xi} : $x_i = X_i + e_{xi}$. Außerdem wird angenommen, daß die Meßfehler zweier Meßzeitpunkte unkorreliert sind. Ferner wird abgeleitet, daß sich extreme Werte aus einem hohen „wahren“ Wert X_i und einem hohen Meßfehler e_{xi} zusammensetzen, d.h., gerade extreme Werte können stark meßfehlerbehaftet sein. „Weichen bei der Erstmessung Meßwerte von Individuen aufgrund zufälliger Meßfehler stärker nach oben oder unten vom allgemeinen Gruppenmittel ab als es ihrem „wahren“ Wert entspricht, ist es wenig wahrscheinlich, daß die gleichen Individuen auch bei der Zweitmessung alle in gleicher Weise vom Gruppenmittel abweichen“ (TRAUTNER, S. 282). Folglich liegen die Meßwerte dieser Individuen in der zweiten Erhebung näher am Mittelwert, da extreme Abweichungen vom Gruppenmittel bzw. extreme Meßfehler seltene Ereignisse sind, die ja gerade durch eine **geringe Wiederholungswahrscheinlichkeit** definiert werden.

Die **Meßfehlerbelastetheit** der Messung in der Psychologie determiniert also die **Richtung der Zufallsänderungen** in den extremen Bereichen der Meßskala (vgl. PETERMANN, S. 28).

Hussy: Die Streuung der Meßwerte ist in den Extrembereichen weniger zufällig als im Durchschnittsbereich.

Der Regressionseffekt stellt nach BALTES & NESSELROADE (1976, zit. n. TRAUTNER, S. 281) ein spezifisches Phänomen der **einmaligen Wiederholungsmessung** (Messung der gleichen Variablen zu *nur zwei Erhebungszeitpunkten*, Prätest-Posttest-Design mit der UV B „Meßzeitpunkt“) dar. Bei der Messung von Veränderungen über eine größere Zahl von Erhebungszeitpunkten, wie es in der Entwicklungspsychologie oft der Fall ist, verliert er zunehmend an Bedeutung. Der Regressionseffekt tritt außerdem kaum auf, wenn eine hohe Korrelation zwischen Erst- und Zweitmessung (r_{xy} , Retest-Reliabilität) besteht. Da eine hohe Retest-Reliabilität, die die Annahme der zeitlichen **Stabilität** von Merkmalen unterstellt, mit der Veränderungsmessung unvereinbar ist, empfiehlt sich zur Kontrolle der statistischen Regression, eine **Kontrollgruppe** in das Untersuchungsdesign einzufügen. Dadurch wird der Regressionseffekt zwar nicht eliminiert, aber er wird kontrollierbar, sofern die Mittelwerte und Streuungen der Experimental- und der Kontrollgruppe in der Erstmessung vergleichbar sind (vgl. PETERMANN, S. 30, S. 57). **Hussy:** Wird bei einem Prätest-Posttest-Kontrollgruppen-

Design die Auswertung mit Hilfe der **Kovarianzanalyse** vorgenommen, müssen die Ausgangsmittelwerte der Experimental- und der Kontrollgruppe nur ungefähr vergleichbar sein; kleinere Abweichungen stören nicht, da die Kovarianzanalyse ja gerade dem Herausrechnen von Mittelwertsunterschieden dient!

Die fünf Axiome der Klassischen Testtheorie (KTT):

- 1) Ein beobachteter Wert x_i setzt sich additiv aus dem wahren Wert X_i und dem Fehlerwert e_{xi} zusammen. / Ein beobachteter Wert x_i setzt sich additiv aus dem wahren Merkmalsanteil X_i und einem Fehleranteil e_{xi} zusammen: $x_i = X_i + e_{xi}$.
- 2) Der Erwartungswert und die Summe der Fehler ist gleich Null / Der durchschnittliche Fehlerwert/Meßfehler in einer beliebigen Population ist gleich Null.

(Die Fehlerwerte gleichen sich über alle Werte aus, so daß sie im Mittel Null ergeben.)

Folge: der Mittelwert der wahren Werte ist gleich dem Mittelwert der beobachteten Werte.

Dieses Axiom setzt voraus, daß der wahre Wert eines Individuums stabil ist und nur der Fehlerwert variiert. Daher läßt sich die KTT nur auf die Messung stabiler Merkmale/Zustände anwenden und nicht auf die Messung von variablen

Merkmale/Veränderung/Prozessen!

- 3) Fehlerwerte zweier verschiedener Tests korrelieren nicht systematisch miteinander. / Fehlerwerte e_{xi} aus verschiedenen Tests sind unkorreliert, sofern die Tests experimentell unabhängig vorgegeben werden.

Aus diesen Axiomen folgt:

- 4) Fehlerwert e_{xi} und wahrer Wert X_i korrelieren nicht systematisch miteinander. / Der Fehlerwert e_{xi} ist mit dem wahren Wert X_i unkorreliert: $r_{X_i, e_{xi}} = 0$.
- 5) Wahrer Wert X_i eines Test und Fehlerwert e_{xi} eines anderen Tests korrelieren nicht systematisch miteinander. / Der wahre Wert X_i eines Test ist mit dem Fehlerwert e_{xi} eines anderen Tests unkorreliert.

x_i = beobachteter/gemessener Wert; X_i = wahrer Wert; e_{xi} = Meßfehler

Frage 3: In welchem Zusammenhang stehen die Begriffe „**Veränderung**“, „**natürliche Regression**“ und „**statistische Regression**“? Illustrieren Sie Ihre Überlegungen an einem Beispiel! (1 x gefragt) bzw. Was versteht man unter „**natürliche Regression**“ und „**statistische Regression**“? Wie wird die statistische Regression **begründet**? (1 x gefragt)

Im Rahmen der Veränderungsmessung stellt sich die Frage, ob eine erfaßte **Veränderung** aufgrund einer wahren, tatsächlichen Veränderung des Merkmals zustande kommt oder nur als ein statistisches Artefakt anzusehen ist. Man unterscheidet daher die natürliche Regression und die statistische Regression.

Natürliche Regression: Man bezeichnet eine Veränderung als natürliche Regression, wenn der beobachtete Effekt unabhängig von den Meßfehlern ist und die beobachtete Veränderung eine **wahre Veränderung** darstellt (vgl. PETERMANN, S. 27).

Statistische Regression: Im Gegensatz dazu versteht man unter der statistischen Regression die Tendenz zur Mitte, die auf die (zufälligen) Meßfehler zurückzuführen ist (vgl. PETERMANN, S. 28); es handelt sich also nur um eine **scheinbare Veränderung** aufgrund mangelnder Reliabilität der Meßinstrumente.

Es ist zu bedenken, daß zumeist sowohl die natürliche als auch die statistische Regression die Meßwerte beeinflussen.

Frage 4: Nennen Sie **drei Variablen**, die die Stärke der **statistischen Regression** beeinflussen! **Begründen** Sie die Wirksamkeit dieser Variablen! (1 x gefragt)

Folgende **drei Variablen** beeinflussen die Stärke der statistischen Regression:

- 1) **Art der Stichprobe**: bei vielen extremen Werten in der Erstmessung/Extremgruppen/selegierten Stichproben starker Effekt (vgl. BORTZ & DÖRING, 1995², S. 519);
- 2) **Höhe der Korrelation zwischen Erst- und Zweitmessung**: bei geringer Korrelation starker Effekt;
- 3) **Anzahl der Wiederholungsmessungen**: bei nur einmaliger Wiederholungsmessung (ein Posttest) starker Effekt.

Begründung zu 1) extreme Werte in der Erstmessung:

Nach der Klassischen Testtheorie setzt sich ein beobachteter Wert x_i zusammen aus dem „wahren“ Variablenwert X_i und dem (zufälligen) Meßfehler e_{xi} : $x_i = X_i + e_{xi}$. Außerdem wird angenommen, daß die Meßfehler zweier Meßzeitpunkte unkorreliert sind. Ferner wird abgeleitet, daß sich extreme Werte aus einem hohen „wahren“ Wert X_i und einem hohen Meßfehler e_{xi} zusammensetzen, d.h., gerade extreme Werte können stark meßfehlerbehaftet sein. „Weichen bei der Erstmessung Meßwerte von Individuen aufgrund zufälliger Meßfehler stärker nach oben oder unten vom allgemeinen Gruppenmittel ab als es ihrem „wahren“ Wert entspricht, ist es wenig wahrscheinlich, daß die gleichen Individuen auch bei der Zweitmessung alle in gleicher Weise vom Gruppenmittel abweichen“ (TRAUTNER, S. 282). Folglich liegen die Meßwerte dieser Individuen in der zweiten Erhebung näher am Mittelwert, da extreme Abweichungen vom Gruppenmittel bzw. extreme Meßfehler seltene Ereignisse sind, die ja gerade durch eine **geringe Wiederholungswahrscheinlichkeit** definiert werden.

Die **Meßfehlerbelastetheit** der Messung in der Psychologie determiniert also die **Richtung der Zufallsänderungen** in den extremen Bereichen der Meßskala (vgl. PETERMANN, S. 28).

Hussy: Die Streuung der Meßwerte ist in den Extrembereichen weniger zufällig als im Durchschnittsbereich.

Begründung zu 2) geringe Korrelation zwischen Erst- und Zweitmessung/den beiden Meßwertreihen (r_{xy} , Retest-Reliabilität):

Nach FURBY (1973, zit. n. TRAUTNER, S. 282) ist die statistische Regression aufgrund von zufälligen Meßfehlern nur ein Sonderfall eines allgemeineren Phänomens, das zwangsläufig auftritt, wenn zwei Meßwertreihen weniger als $r = 1.0$ korreliert sind, auch im hypothetischen Fall, daß Meßfehler auszuschließen sind. Er geht in seiner Interpretation dieses Phänomens von der Bedeutung eines Korrelationskoeffizienten (r) aus: Daß zwei Meßwertreihen niedriger korrelieren als $r = 1.0$ – was in der Psychologie der Regelfall ist – heißt ja, daß die miteinander korrelierten Variablen bzw. die miteinander korrelierten Meßwerte einer Variable zu zwei Zeitpunkten außer gemeinsamen Varianzanteilen auch noch **spezifische Varianzanteile** aufweisen. (Die gemeinsame Varianz zweier Meßwertreihen ist gleich dem Quadrat des Korrelationskoeffizienten (r^2).) Genauso wie keine überzeugende Wahrscheinlichkeit dafür gegeben ist, daß sich zufällige Meßfehler beim gleichen Individuum wiederholt in gleicher Richtung auswirken, ist es **wenig wahrscheinlich, daß die jeweils spezifischen Varianzquellen (Faktoren) zweier Meßwerte diese Meßwerte jeweils in der gleichen Richtung beeinflussen**. Je höher die spezifischen Varianzanteile zweier Meßwerte sind (und damit je geringer die Korrelation zwischen den beiden Meßwerten ist), desto unwahrscheinlicher wird also eine gleichsinnige Beeinflussung der beiden Meßwerte, d.h., desto stärker ist die statistische Regression.

Begründung zu 3) nur einmalige Wiederholungsmessung (ein Posttest):

Bei der Messung von Veränderungen in einer Variablen über eine größere Zahl von Erhebungszeitpunkten verliert die statistische Regression zunehmend an Bedeutung. Läßt sich das mit dem zweiten Axiom der Klassischen Testtheorie „Der durchschnittliche Fehlerwert/Meßfehler in einer beliebigen Population ist Null.“ begründen? In dem Sinn, daß sich die Meßfehler über eine zunehmende Zahl von Wiederholungsmessungen ausgleichen und damit die beobachteten Werte immer weniger meßfehlerbelastet sind, so daß die statistische Regression immer schwächer wird?

„Mit wachsender Anzahl der Meßzeitpunkte wird der Einfluß eines fehlerhaften bzw. wenig reliablen Meßinstruments auf die Reliabilität der Veränderungsmaße zunehmend kompensiert.“ (BORTZ & DÖRING, S. 517). → Die mangelnde Reliabilität des Meßinstruments, die ja für die statistische Regression verantwortlich ist, wird also immer unbedeutender, der Regressionsseffekt immer schwächer. „Die Reliabilität der Veränderungsmaße läßt sich drastisch

verbessern, wenn die Anzahl der Meßzeitpunkte erhöht wird, wobei der **Reliabilitätsgewinn am größten** ist, wenn der Untersuchungsplan statt zwei Meßzeitpunkten (z.B. Prä- und Posttest) **drei Meßzeitpunkte** vorsieht. WILLETT (1989) berichtet, daß die Reliabilität allein durch das Hinzufügen eines dritten Meßzeitpunktes um 250% und mehr erhöht werden kann.“ (ebda.).

Frage 5: Mit Hilfe welcher Veränderungsindizes, versuchplanerischen Maßnahmen und Auswertungsverfahren kann der **Regressionseffekt eliminiert** werden? Geben Sie jeweils eine kurze Begründung! (3 x gefragt)

1. Veränderungsindizes

1) Regressionsmaße: Eine Möglichkeit, den Regressionseffekt zu kontrollieren (nicht zu eliminieren!), besteht in der Korrektur der Prätestwerte x und der Posttestwerte y hinsichtlich ihrer Meßfehlerbehaftetheit über die Regressionsschätzung. Eine solche Korrektur leistet die **Kovarianzanalyse**, die eine **Kombination von Regression und Varianzanalyse** darstellt. Sie untersucht die Unterschiede zwischen zwei Gruppen (Experimental- und Kontrollgruppe) hinsichtlich eines Kriteriums, nachdem die Unterschiede der Gruppen in den Ausgangswerten ausgeschaltet wurden (vgl. PETERMANN, S. 56). Man bestimmt dabei die Regression vom Nachttest auf den Vortest und schätzt aus den Vortestwerten die Nachttestwerte. Die **Differenz zwischen den geschätzten und den beobachteten Werten** gibt die Veränderung an. Dieses Auswertungsverfahren trennt den individuumspezifischen Trend (Veränderung) von dem allgemeinen Trend, der allen Versuchspersonen gemeinsam ist; man korrigiert damit auch den Regressionseffekt und ermittelt, welche Veränderung ein Individuum über den Trend hinaus erfährt (vgl. PETERMANN, S. 34). Diese Vorhersage der Posttestwerte aus den Prätestwerten in Form einer Schätzung zeigt jedoch keinen kausalen Zusammenhang auf! Nach PETERMANN (S. 34, S. 57) ist die Voraussetzung für die Anwendung der Kovarianzanalyse als Auswertungsverfahren innerhalb eines Prätest-Posttest-Kontrollgruppen-Designs (Meßwiederholungsplan → die Meßwerte einer Stichprobe hängen von Messung zu Messung voneinander ab/korrelieren miteinander) daß die Ausgangswerte der Experimental- und der Kontrollgruppe vergleichbar sind. **Hussy:** Wird bei einem Prätest-Posttest-Kontrollgruppen-Design die Auswertung mit Hilfe der Kovarianzanalyse vorgenommen, müssen die Ausgangsmittelwerte der Experimental- und der Kontrollgruppe **nur ungefähr vergleichbar** sein; kleinere Abweichungen stören nicht, da die Kovarianzanalyse ja gerade dem Herausrechnen von Mittelwertsunterschieden dient!

LORD (1958, 1963) schlug zur Behebung dieser Schwäche der Regressionsschätzung über die Kovarianzanalyse bei voneinander abhängigen Messungen (vergleichbare Ausgangswerte als Voraussetzung) ein Verfahren zur Regressionsschätzung vor, das x und y über Partialkorrelationen bereinigt und in die Schätzgleichung miteinbezieht (**Lord-Verfahren**). CRONBACH & FURBY (1970) sahen dieses Korrekturverfahren als das brauchbarste an und erweiterten es um eine Korrelation der Veränderungsbeträge mit Außenvariablen (Drittvariablen). Der **CRONBACH & FURBY-Ansatz** ist bei weitem der exakteste zur Schätzung der Veränderungseffekte. Zudem kann man bei ihm noch die Effekte von Drittvariablen auf die Veränderung bestimmen. Allerdings bleibt ungeklärt, warum gerade eine bestimmte Drittvariable einbezogen wird. (vgl. PETERMANN, S. 34 ff).

2) Residualmaße

Der einfachste und häufigste Weg zur Korrektur von Differenzwerten besteht nach T RAUTNER in der Verwendung von residualen Veränderungswerten/Residualmaßen, die Regressionseffekte eliminieren. Zu ihrer Berechnung geht man von den Werten aus, die entsprechend der Regressionsgeraden, d.h. der Regression von y (Nachttestwert) auf x (Vortestwert), bei der Zweitmessung zu erwarten wären. „Der Residualwert eines Individuums ist dann gleich der **Abweichung des zweiten Meßwerts von dem Wert, der aufgrund der Regressionsanalyse nach Kenntnis des ersten Meßwerts vorhergesagt wird**“ (T RAUTNER, S. 284). Nach PETERMANN (S. 35 ff) ist das Maß für die Veränderung das **Residuum, das entsteht, wenn man von dem Nachttestwert y jenen Anteil abzieht, der auf den Einfluß des Vortestwerts x zurückzuführen ist**; der Vortestwert x wird aus dem Nachttestwert y auspartialisiert.

Die Berechnung von Residualmaßen beruht also auf der Eliminierung des Einflusses des Ausgangswerts auf die Höhe des Zuwachswerts (vgl. TRAUTNER, S. 284). Der **Residualansatz** geht ebenfalls auf LORD (1958, 1963) zurück. Residualwerte eliminieren zwar Regressionseffekte, vernachlässigen aber die Meßfehler der beiden Einzelmessungen; ferner treten auch bei der Verwendung des Residualmaßes Verzerrungen der Korrelationen des Veränderungsmaßes mit theoretisch bedeutsamen Drittvariablen auf.

Zur Ausschaltung des Einflusses von Meßfehlern auf die Berechnung von Residualwerten haben CRONBACH & FURBY (1970) den Residualansatz in gleicher Weise wie den Regressionsansatz modifiziert und Drittvariablen einbezogen. Die Abhängigkeit der Beobachtungen wird über Partialkorrelationen berücksichtigt. Die Schätzung der wahren Veränderung ergibt sich aus der sukzessiven Auspartialisierung der fehlerkorrigierten Variablen. Auch hier bleibt allerdings ungeklärt, warum gerade eine bestimmte Drittvariable einbezogen wird. (vgl. PETERMANN, S. 35 ff).

2. Versuchsplanerische Maßnahmen und Auswertungsverfahren

1) Quasi-experimentelle Untersuchungspläne mit mehr als zwei Erhebungszeitpunkten

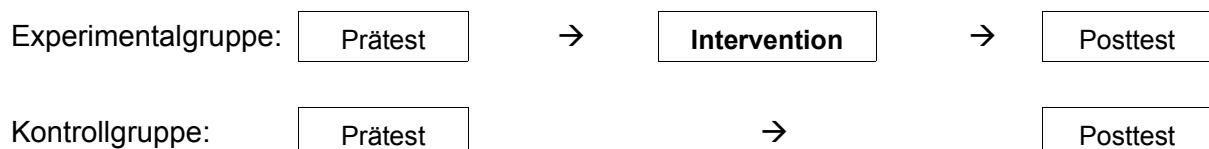
Der Regressionseffekt stellt nach BALTES & NESSELROADE (1976, zit. n. TRAUTNER, S. 281) ein spezifisches Phänomen der *einmaligen Wiederholungsmessung* (Messung der gleichen Variablen zu *nur zwei Erhebungszeitpunkten*, Prätest-Posttest-Design mit der UV B „Meßzeitpunkt“) dar. Bei der Messung von Veränderungen über eine größere Zahl von Erhebungszeitpunkten, wie es in der Entwicklungspsychologie oft der Fall ist, verliert er zunehmend an Bedeutung. „Mit wachsender Anzahl der Meßzeitpunkte wird der Einfluß eines fehlerhaften bzw. wenig reliablen Meßinstruments auf die Reliabilität der Veränderungsmaße zunehmend kompensiert.“ (BORTZ & DÖRING, S. 517). → Die mangelnde Reliabilität des Meßinstruments, die ja für die statistische Regression verantwortlich ist, wird also immer unbedeutender, der Regressionseffekt immer schwächer. „Die Reliabilität der Veränderungsmaße läßt sich drastisch verbessern, wenn die **Anzahl der Meßzeitpunkte erhöht** wird, wobei der **Reliabilitätszugewinn am größten** ist, wenn der Untersuchungsplan statt zwei Meßzeitpunkten (z.B. Prä- und Posttest) **drei Meßzeitpunkte** vorsieht. WILLETT (1989) berichtet, daß die Reliabilität allein durch das Hinzufügen eines dritten Meßzeitpunktes um 250% und mehr erhöht werden kann.“ (ebda.).

2) Quasi-experimentelle Untersuchungspläne mit einer Kontrollgruppe

Es empfiehlt sich, zur Kontrolle der statistischen Regression eine **Kontrollgruppe** in das Untersuchungsdesign einzufügen. Dadurch wird der Regressionseffekt zwar nicht eliminiert, aber er wird kontrollierbar, sofern die Mittelwerte und Streuungen der Experimental- und der Kontrollgruppe in der Erstmessung vergleichbar sind (vgl. PETERMANN, S. 30, S. 57). **Hussy:** Wird bei einem Prätest-Posttest-Kontrollgruppen-Design die Auswertung mit Hilfe der Kovarianzanalyse vorgenommen, müssen die Ausgangsmittelwerte der Experimental- und der Kontrollgruppe nur ungefähr vergleichbar sein; kleinere Abweichungen stören nicht, da die Kovarianzanalyse ja gerade dem Herausrechnen von Mittelwertsunterschieden dient!

Prätest-Posttest-Kontrollgruppen-Design

Die **interne Validität** von quasi-experimentellen Untersuchungen (ohne Randomisierung) mit Meßwiederholung läßt sich dadurch erhöhen, daß neben der Experimentalgruppe (UV A₁: mit Intervention) auch eine **Kontrollgruppe** (UV A₂: ohne Intervention) geprüft wird (Prätest-Posttest-Kontrollgruppen-Design, Zwei-Gruppen-Prätest-Posttest-Plan). Dieses Untersuchungsdesign ermöglicht die **Kontrolle von zeitabhängigen Störeinflüssen** (externe zeitliche Einflüsse, Reifungsprozesse, Testübung, vgl. BORTZ & DÖRING, S. 522). So läßt sich die „wahre“ Veränderung zwischen zwei Meßzeitpunkten direkter untersuchen als durch eine nachträgliche statistische Korrektur der beobachteten Differenzwerte in einem einfachen Prätest-Posttest-Design/Ein-Gruppen-Prätest-Posttest-Plan (durch Regressions- oder Residualmaße, s.o.). Der einfachste Plan umfaßt eine Experimentalgruppe (mit Bedingungsfaktor/Intervention) und eine Kontrollgruppe (ohne Bedingungsfaktor/Intervention), die beide jeweils zweimal untersucht werden. Die beiden Gruppen/Stichproben sollen sich lediglich darin unterscheiden, daß in der Experimentalgruppe zwischen Vortest und Nachtest ein experimenteller Bedingungsfaktor/eine Intervention eingeführt wird, während die Kontrollgruppe ausschließlich dem Vortest und dem Nachtest unterzogen wird. Sowohl die Prätestwerte der beiden Versuchsgruppen als auch alle denkbaren Einflüsse zwischen Erst- und Zweitmessung, abgesehen von der Intervention, müssen dabei gleich sein. Schematisch (vgl. TRAUTNER, S. 285):



Der große **Vorteil** bei einem solchen Versuchsplan ist, daß man zur Abschätzung der „wahren“ Veränderung aufgrund der experimentellen Variablen/Intervention auf die Berechnung von Differenz- oder Veränderungswerten zwischen den beiden Meßzeitpunkten verzichten kann und stattdessen die **Nachtestwerte** der beiden Versuchsgruppen **direkt miteinander**

vergleichen kann. Aus dem Unterschied der beiden Nachtestwerte läßt sich der Effekt des experimentellen Faktors/der Intervention unmittelbar ablesen (vgl. TRAUTNER, S. 285). So läßt sich die „wahre“ Veränderung zwischen zwei Meßzeitpunkten direkter untersuchen als durch eine nachträgliche statistische Korrektur der beobachteten Differenzwerte in einem einfachen Prätest-Posttest-Design/Ein-Gruppen-Prätest-Posttest-Plan (durch Regressions- oder Residualmaße, s.o.). Der Vergleich der Nachtestwerte der beiden Gruppen kann mit einem t-Test für unabhängige Stichproben vorgenommen werden; dabei geht jede Vp nur mit einer Messung (Nachtest) in die Prüfung ein, so daß das Problem der statistischen Regression bei abhängigen Messungen vermieden wird.

Nach BORTZ & DÖRING (S. 521 f) eignet sich für die statistische Auswertung des Zwei-Gruppen-Prätest-Posttest-Plans die zweifaktorielle Varianzanalyse mit Meßwiederholungen. Um den „Netto-Effekt“ der Intervention zu ermitteln, berechnet man die Differenz der Veränderung in der Experimental- und der Kontrollgruppe.

Schema zur Ermittlung des Effekts einer Intervention:

	Prätest	Posttest	Differenz
Experimentalgruppe	E ₁	E ₂	E = E ₁ - E ₂
Kontrollgruppe	K ₁	K ₂	K = K ₁ - K ₂ Netto-Effekt = E - K

Die Buchstaben E und K stehen hier für Durchschnittswerte in der Experimental- bzw. Kontrollgruppe. Ein statistisch signifikanter „Netto-Effekt“ wird durch eine signifikante Interaktion zwischen dem Gruppenfaktor und dem Meßwiederholungsfaktor nachgewiesen.

Da bei quasi-experimentellen Kontrollgruppen-Designs keine Randomisierung durchgeführt werden kann, ist die Durchführung von Vortests der abhängigen Variable unerlässlich, um Aussagen über die Veränderung in der Experimentalgruppe machen zu können. Die Vortests (z.B. t-Test für unabhängige Stichproben) haben hier die Funktion, Ausgangsunterschiede zwischen der Experimental- und der Kontrollgruppe in der abhängigen Variable zu Beginn der Untersuchung festzustellen. Die stichprobenspezifischen „Startbedingungen“ sind die Referenzdaten, auf die sich interventionsbedingte Veränderungen beziehen (vgl. BORTZ & DÖRING, S. 515). Nur wenn die beiden Gruppen sich im Vortest in ihren Ausgangswerten (Mittelwerte und Streuungen) nicht deutlich unterscheiden (Vergleich von A₁B₁ mit A₂B₁ z. B. mit einem t-Test für unabhängige Stichproben), kann der statistische Regressionseffekt kontrolliert (nach PETERMANN, S. 30 aber nicht eliminiert) werden und ist die interne Validität so akzeptabel, daß ein möglicher Gruppenunterschied im Posttest (Vergleich von A₁B₂ mit A₂B₂) eindeutig kausal auf die Intervention zurückgeführt werden kann (vgl. PETERMANN, S. 30).

Versuchsplan:

		UV B: Meßzeitpunkt	
		B ₁ vorher	B ₂ nachher
UV A: Intervention	A ₁ mit	AV: A ₁ B ₁	AV: A ₁ B ₂
	A ₂ ohne	AV: A ₂ B ₁	AV: A ₂ B ₂

Kombination von Prätest-Posttest-Kontrollgruppen-Design und Kovarianzanalyse

Besteht ein Unterschied zwischen der Experimental- und der Kontrollgruppe hinsichtlich ihrer Ausgangswerte im Prätest, kann die Veränderung der Werte der Experimentalgruppe beim Posttest (auch) auf die Ausgangsunterschiede und nicht (allein) auf die Intervention zurückgeführt werden. Ferner besteht bei vorliegendem Ausgangsunterschied die Gefahr eines statistischen Regressionseffekts, der sich darin äußern würde, daß sich eine hohe Differenz im Prätest im Posttest verringert. Zur **Kontrolle des Ausgangsunterschieds** bzw. zur **Eliminierung des Regressionseffekts** sollte daher bei vorliegendem Ausgangsunterschied eine **Kovarianzanalyse** durchgeführt werden.

Die Kovarianzanalyse ist eine **Kombination** von **Regression** und **Varianzanalyse** zur Überprüfung des Unterschieds von zwei Gruppen hinsichtlich eines Kriteriums (y, AV), nachdem die Unterschiede in den Ausgangswerten ausgeschaltet wurden (vgl. PETERMANN, S. 56). Die Kovarianzanalyse bestimmt hierzu eine Regression vom Nachtest auf den Vortest und schätzt die Nachtestwerte aus den beobachteten Vortestwerten. Die **Differenz zwischen den geschätzten und den beobachteten Nachtestwerten** gibt die Veränderung an. Dieses Auswertungsverfahren trennt den individuumspezifischen Trend (Veränderung) von dem allgemeinen Trend, der allen Versuchspersonen gemeinsam ist; es beseitigt den allgemeinen Trend aus den Veränderungswerten, so daß die individuumspezifischen Veränderungen übrig bleiben. Damit wird auch der Regressionseffekt eliminiert und ermittelt, welche Veränderung ein Individuum über den Trend hinaus erfährt. (vgl. PETERMANN, S. 30, S. 38).

Merksatz: „Eine **Kontrollvariable** ist eine (personengebundene) Störvariable, deren Einfluß mittels Kovarianzanalyse aus der abhängigen Variablen herausgerechnet (herauspartialisiert) wird.“ (BORTZ & DÖRING, S. 509)

Die vor der Untersuchung angetroffenen a-priori-Unterschiede zwischen den Gruppen in bezug auf die abhängige Variable werden kovarianzanalytisch aus den Messungen herauspartialisiert, damit sie das Untersuchungsergebnis nicht beeinflussen. In der Kovarianzanalyse werden varianzanalytische Techniken mit regressionsanalytischen Techniken kombiniert. Mit Hilfe der Regressionsrechnung erfolgt – vereinfacht dargestellt – die Bestimmung einer Regressionsgleichung zwischen der abhängigen Variable und der Kontrollvariable, die eingesetzt wird, um die abhängige Variable aufgrund der Kontrollvariable vorherzusagen. Die vorhergesagten Werte der abhängigen Variable sind dann vollständig durch die Kontrollvariable determiniert. Werden die **Differenzen** zwischen den tatsächlichen Werten der abhängigen Variable und den aus der Kontrollvariable vorhergesagten Werten berechnet, so resultieren **Regressionsresiduen**, die von der Kontrollvariable unbeeinflusst sind (vgl. BORTZ, 1999⁵, S. 349 f).

Merksatz: Eine **Kovarianzanalyse** ist eine Varianzanalyse über Regressionsresiduen, mit deren Hilfe der Einfluß einer Kontrollvariablen auf die abhängige Variable neutralisiert wird. (BORTZ, S. 350)

Die regressionstechnische Eliminierung des Einflusses einer Kontrollvariablen auf die abhängige Variable entspricht dem Prinzip nach einer Partialkorrelation zwischen der unabhängigen Variable und der abhängigen Variable unter Ausschaltung der Kontrollvariable (vgl. BORTZ & DÖRING, S. 508).

In der kovarianzanalytischen Auswertung der Posttestwerte der beiden Gruppen werden die Prätestwerte der beiden Gruppen als Kontrollvariable einbezogen. **Aus der abhängigen Variable (AV) im Prätest wird dadurch eine Kontrollvariable/Kovariate (KV) im Posttest** (identischer Wert). Der Einfluß der Prätestwerte (KV) auf die Posttestwerte (AV) ist somit statistisch kontrollierbar. Aus dem eigentlich zweifaktoriellen Design (UV A „Intervention“ mit den beiden Stufen „mit“ und „ohne“, UV B „Meßzeitpunkt“ mit den beiden Stufen „vorher“ und „nachher“) wird so ein einfaktorielles (Auswertungs-)Design: man beschränkt sich auf die wesentliche UV A (mit oder ohne Intervention). Außerdem ist der statistische Regressionseffekt vermieden, da Vor- und Nachtestwerte nicht direkt miteinander verglichen werden, sondern nur die Nachtestwerte der Experimental- und der Kontrollgruppe (jede Vp geht nur mit einer Messung der AV in die (Posttest-) Prüfung ein, die AV wird nur einfach verwertet). Der Unterschied von Experimental- und Kontrollgruppe im Posttest ist dann von möglichen Ausgangsunterschieden bereinigt, so daß die **interne Validität höher** ist und eine **eindeutige Aussage über eine tatsächliche Veränderung aufgrund der Intervention** möglich wird.

Auswertungsplan:

	UV B: Meßzeitpunkt		Haupteffekt (HE A)
	B ₁ vorher	B ₂ nachher	

UV A: Intervention	A ₁ mit	KV: A ₁ B ₁	AV: A ₁ B ₂	
	A ₂ ohne	KV: A ₂ B ₁	AV: A ₂ B ₂	

Frage 6: Wie und weshalb kann die **Kombination** eines **Prätest-Posttest-Designs** mit einer **Kovarianzanalyse** einen eventuell vorhandenen **Regressionseffekt eliminieren** (mit kurzer Begründung)? (1 x gefragt)

siehe Frage 5:

Kombination von Prätest-Posttest-Kontrollgruppen-Designs und Kovarianzanalyse!

Frage 7: Welche **Vorteile** hat ein **Prätest-Posttest-Kontrollgruppen-Plan** im Vergleich zu einem reinen **Prätest-Posttest-Plan**? (1 x gefragt)

Wenn Veränderung aufgefaßt wird als der Unterschied in einer Variablen (AV) zu zwei Zeitpunkten, da eine einmalige Beobachtung einer Variablen keine Aussage über ihre Veränderung zuläßt, dann entspricht diesem Verständnis der **Prätest-Posttest-Plan** mit mindestens zwei Meßzeitpunkten. Man kann weder allein aus dem Prätest noch allein aus dem Posttest Aussagen zu Veränderungen machen, sondern erst durch die gleichzeitige Berücksichtigung beider Meßzeitpunkte ist Veränderung zu erschließen. Damit wird eine UV (B, Meßzeitpunkt) mit zwei Stufen eingeführt (UV B₁: vorher; UV B₂: nachher), zwischen welche eine Intervention tritt. Hervorgerufen wird die Veränderung durch eine Intervention (UV A), die im Fall der Evaluationsforschung durch eine Maßnahme und in der Entwicklungspsychologie bei Hypothesen über Veränderungen in Abhängigkeit vom Alter durch das Verstreichen eines Zeitraums repräsentiert wird.

Bei diesem Design wird eine repräsentative Stichprobe der interessierenden Zielpopulation einmal vor und einmal nach der Intervention untersucht. Die durchschnittliche Differenz auf der abhängigen Variablen (**Vergleich von Vortestwerten und Nachtestwerten der Experimentalgruppe**) gilt behelfsweise als Indikator für die Wirkung der Intervention. Die *interne Validität* (Eindeutigkeit der Ergebnisse) dieses Designs ist jedoch *gering*, da alle möglichen zeitabhängigen Störeinflüsse die Veränderung (bzw. Nichtveränderung) (mit-)bewirkt haben können. Ferner treten durch die Meßwiederholung (abhängige Messungen) *statistische Regressionseffekte* auf. Daher sollte dieser Versuchsplan nur bei Fragestellungen eingesetzt werden, bei denen eine Intervention interessiert, von der praktisch alle Personen betroffen sind, so daß keine Kontrollgruppe gebildet werden kann oder bei Fragestellungen, bei denen aus ethischen Gründen die Bildung einer Kontrollgruppe nicht möglich ist. Als Signifikanztest verwendet man bei zwei Messungen z.B. den t-Test für abhängige Stichproben und bei mehr als zwei Messungen die einfaktorielle Varianzanalyse mit Meßwiederholungen.

Der große **Vorteil** des **Prätest-Posttest-Kontrollgruppen-Plans** gegenüber dem reinen Prätest-Posttest-Plan ist, daß man zur Abschätzung der „wahren“ Veränderung aufgrund der experimentellen Variablen/Intervention auf die Berechnung von Differenz- oder Veränderungswerten zwischen den beiden Meßzeitpunkten verzichten kann und stattdessen die **Nachtestwerte der beiden Versuchsgruppen direkt miteinander vergleichen** kann. Aus dem Unterschied der beiden Nachtestwerte läßt sich der Effekt des experimentellen Faktors/der Intervention unmittelbar ablesen (vgl. TRAUTNER, S. 285). So läßt sich die „wahre“ Veränderung zwischen zwei Meßzeitpunkten direkter untersuchen als durch eine nachträgliche statistische Korrektur der beobachteten Differenzwerte in einem einfachen Prätest-Posttest-Design/Ein-Gruppen-Prätest-Posttest-Plan (durch Regressions- oder Residualmaße, s.o.). Der Vergleich der Nachtestwerte der beiden Gruppen kann mit einem t-Test für unabhängige Stichproben vorgenommen werden; dabei geht jede Vp nur mit einer Messung (Nachtest) in die Prüfung ein, so daß das **Problem der statistischen Regression** bei abhängigen Messungen **vermieden** wird. Die **interne Validität** ist hier durch die Kontrolle von zeitabhängigen Störeinflüssen **höher**.

Schematische vergleichende Darstellung:

Prätest-Posttest-Plan:

	Prätest (t_1)	Posttest (t_2)
Experimentalgruppe (E)	AV_{E1} ←	→ AV_{E2}

Prätest-Posttest-Kontrollgruppen-Plan:

	Prätest (t_1)	Posttest (t_2)
Experimentalgruppe (E)	AV_{E1}	AV_{E2}
	↑	↑
Kontrollgruppe (K)	AV_{K1}	AV_{K2}
	↓	↓

Frage 8: Was versteht man unter dem **Reliabilitäts-Validitäts-Dilemma**, und inwiefern ist es in der **klassischen Testtheorie** begründet? (3 x gefragt)

Eine Veränderung bezieht sich immer auf den Unterschied einer beobachteten/gemessenen Variablen bezogen auf mindestens zwei Beobachtungs-/Meßzeitpunkte. Diesen Unterschied (g_i) erfaßt man in der Regel durch die Differenz der Variablenwerte x_i (Zeitpunkt 1) und y_i (Zeitpunkt 2): $g_i = y_i - x_i$. Das Ziel der Veränderungsmessung besteht darin, von den beobachteten Differenzen ($g_i = x_i - y_i$) auf die wahren Differenzen ($G_i = Y_i - X_i$) bzw. tatsächlichen Veränderungen zu schließen. Damit ein solcher Schluß durchgeführt werden kann, ist es notwendig, die **Reliabilität der Differenzen** zu bestimmen, da festgestellt werden muß, ob die beobachtete Differenz bzw. Veränderung (g) eine Verhaltensfluktuation (wahre Veränderung) abbildet oder auf eine Zufallsfluktuation zurückzuführen ist. Handelt es sich um eine Zufallsfluktuation, so verwendet man als Erklärung für den beobachteten Effekt den Meßfehler ($e_{xi} = x_i - X_i$) (vgl. PETERMANN, S. 30).

Wichtig ist bei der Bestimmung der Reliabilität von Differenzen, daß der Meßfehler der Differenz zweier Werte (des Differenzwerts) größer ist als der jeweilige Meßfehler der Einzelwerte, da sich Meßfehler summieren, so daß die Differenzwert-Reliabilität kleiner ist als die Reliabilitäten der Einzelwerte (vgl. TRAUTNER, S. 283).

Das Problem, das nun entsteht, wenn man von den beobachteten Differenzen auf die wahren Differenzen schließen will, wird als **Reliabilitäts-Validitäts-Dilemma** bezeichnet. Es äußert sich folgendermaßen:

„Je höher die Korrelation zwischen Erst- und Zweitmessung ist, desto niedriger ist die Meßgenauigkeit (Reliabilität) der Differenzwerte. Je niedriger die Korrelation zwischen Erst- und Zweitmessung ist, desto niedriger ist die Validität der Werte.“ (BEREITER, 1963, zit. n. TRAUTNER, S. 283)

Je höher die Korrelation zwischen Erst- und Zweitmessung (Retest-Reliabilität) und damit die Validität der Einzelwerte ist (es ist bei Erst- und Zweitmessung das Gleiche gemessen worden bzw. es sind die gleichen Faktoren für das Zustandekommen der Meßwerte von x und von y verantwortlich), desto niedriger ist die Reliabilität der Differenzwerte, d.h., die Differenzwerte geben kein zuverlässiges Abbild der wahren Veränderung. Je niedriger die Korrelation zwischen Erst- und Zweitmessung (Retest-Reliabilität) und damit je höher die Reliabilität der Differenzwerte ist (die Differenzwerte bilden die wahre Veränderung zuverlässig ab), desto niedriger ist die Validität der Einzelwerte, d.h., es ist bei Erst- und Zweitmessung nicht das Gleiche gemessen worden bzw. es sind verschiedene Faktoren für das Zustandekommen der Meßwerte von x und von y verantwortlich. Die Erhöhung der Differenzwert-Reliabilität mit dem Ziel, ein zuverlässiges Abbild der wahren Veränderung zu erhalten, wird also mit dem Absinken der Validität der Einzelwerte erkaufte (vgl. TRAUTNER, S. 283).

→ Hohe Validität der Einzelwerte = niedrige Reliabilität der Differenzwerte; hohe Reliabilität der Differenzwerte = niedrige Validität der Einzelwerte.

→ Entweder hohe Validität der Einzelwerte, oder hohe Reliabilität der Differenzwerte; beides zugleich geht nicht.

→ Da eine *hohe Retest-Reliabilität* (basierend auf der Annahme zeitlicher *Stabilität* von Merkmalen) für die Veränderungsmessung uninteressant ist, ist eine **hohe Differenzwert-Reliabilität** (i.S. von Paralleltest-Reliabilität, vgl. PETERMANN, S. 32) anzustreben.

Mathematisch läßt sich dieses Dilemma mit Hilfe der **Formel zur Berechnung der Reliabilität der Differenzen ($r_{gg'}$)** nach GUILFORD (1964, zit. n. PETERMANN, S. 30) verdeutlichen:

$$r_{gg'} = (r_{xx'} - 2r_{xy} + r_{yy'}) / (2(1 - r_{xy}))$$

$r_{xx'}$ = Reliabilität der 1. Messung, die sich ergibt, wenn der Quotient aus wahrer Varianz (s^2_x) und der Gesamtvarianz (S^2_x) gebildet wird (s^2_x / S^2_x)

$r_{yy'}$ = Reliabilität der 2. Messung (Berechnung analog $r_{xx'}$)

r_{xy} = Korrelation zwischen 1. und 2. Messung (Retest-Reliabilität)

Fall 1)

Ist die **Korrelation** (r_{xy}) zwischen erster und zweiter Messung sehr **niedrig (niedrige Retest-Reliabilität)** und weisen die Einzelmessungen hohe Reliabilitäten auf, dann erhält der Zähler in der Formel einen relativ hohen positiven Wert, und $r_{gg'}$ wird groß. Es liegt also eine **hohe Differenzwert-Reliabilität** vor.

Bei einer niedrigen Korrelation zwischen zwei Messungen kann man davon ausgehen, daß entweder eine Veränderung über die Zeit stattgefunden hat oder sich das Meßinstrument verändert hat (oder beides), d.h., zum zweiten Meßzeitpunkt liegt ein quantitativ und/oder qualitativ anderer Zustand vor als zum ersten Meßzeitpunkt. Dies bedeutet, daß zahlenmäßig gleiche Meßwerte bei der ersten und der zweiten Messung nicht mehr in derselben Weise interpretiert werden können. Man kann folglich unter diesen Bedingungen keine Aussage darüber treffen, was der Test mißt; es liegt **geringe Validität** vor.

Fall 2)

Ist die **Korrelation** (r_{xy}) zwischen erster und zweiter Messung **mittel bis hoch (mittlere bis hohe Retest-Reliabilität)** und weisen die Einzelmessungen mittlere Reliabilitäten auf, dann wird der Zähler in der Formel klein. Da der Nenner relativ groß bleibt, wird $r_{gg'}$ klein, d.h. die Differenzwerte sind völlig unreliabel. Es liegt also eine **niedrige Differenzwert-Reliabilität** vor. Eine hohe Korrelation zwischen den Messungen einer Variable zu zwei Meßzeitpunkten bedeutet nun nach den Überlegungen aus Fall 1), daß die **Validität hoch** ist.

Das Reliabilitäts-Validitäts-Dilemma der Veränderungsmessung kann im Rahmen der klassischen Testtheorie nicht gelöst werden, sofern man verlangt, daß die Messungen x und y identische Inhalte messen (hohe Validität) und die Korrelation r_{xy} (Retest-Reliabilität) in die Differenzwert-Reliabilitäts-Formel eingeht. Die diskutierten Extremfälle (Fall 1 und Fall 2) zeigen deutlich die Gegenläufigkeit von Retest-Reliabilität und Differenzwert-Reliabilität (vgl. PETERMANN, S. 32). **Die Bestimmung der Retest-Reliabilität, der die Annahme der zeitlichen Stabilität von Ereignissen zugrunde liegt, ist mit dem Anliegen der Veränderungsmessung unvereinbar.** HELMREICH (1977, S. 47, zit. n. PETERMANN, S. 32) folgert, „daß *Testwiederholungs-Reliabilität* in Untersuchungen, in denen Veränderung im Sinne von Variabilität interessiert, auf keinen Fall ein geeignetes Maß der Reliabilitätsbestimmung ist. Die Korrelation der Meßreihen von erster und zweiter Testung ist in diesem Fall nicht das adäquate Mittel, etwas über Tests auszusagen, die zur Messung der Zustände zum ersten und zweiten Meßzeitpunkt eingesetzt werden.“

Frage 9: Das **Reliabilität-Validitäts-Dilemma** läßt sich mit der Formel zur Berechnung der **Differenzwertreliabilität** erläutern: $r_{gg'} = (r_{xx'} - 2r_{xy} + r_{yy'}) / (2 * (1 - r_{xy}))$. Verwenden Sie zur Veranschaulichung ein Beispiel! (1 x gefragt)

siehe Frage 8.

Beispiel: Intelligenzentwicklung bei zwei Extremfällen (vgl. PETERMANN, S. 31):

Bei denselben hohen Reliabilitäten der Einzelmessungen ($r_{xx'} = .83$ und $r_{yy'} = .85$) bei beiden Fällen erhält man:

- im Fall 1) mit geringer Retest-Reliabilität ($r_{xy} = .30$) eine hohe Differenzwert-Reliabilität von 0.77 (bei niedriger Validität) und
- im Fall 2) mit hoher Retest-Reliabilität ($r_{xy} = .30$) eine geringe Differenzwert-Reliabilität nahe Null von .06 (bei hoher Validität).

II Fragen zu Methoden der Evaluation

Literatur:

BORTZ, JÜRGEN & DÖRING, NICOLA (1995²). Forschungsmethoden und Evaluation. Berlin: Springer. → Kap. 3.1 (S. 95-106) und 9.4 (S. 589-607).

WOTTAWA, HEINRICH & THIERAU, HEIKE (1998²). Lehrbuch Evaluation. Bern: Huber. → Kap. 1-6 (S. 13-164).

Frage 10: Welche Besonderheiten zeichnen die **Evaluationsforschung** im **Vergleich** zur **Grundlagenforschung** aus? (2 x gefragt) bzw. Erläutern Sie knapp die **Unterschiede** zwischen **Grundlagen-** und **Evaluationsforschung!** (1 x gefragt)

Evaluationsforschung befaßt sich – als Teilbereich der empirischen Forschung – mit der **Bewertung von Maßnahmen, Programmen, Interventionen, Personen oder Strukturen**. Sie ist keine eigenständige Disziplin, sondern eine **Anwendungsvariante wissenschaftlicher Forschungsmethoden auf eine spezielle Gruppe von Fragestellungen** (vgl. BORTZ & DÖRING, S. 95 f).

Nach ROSSI & FREEMAN (1993) beinhaltet Evaluationsforschung in diesem Sinn die **systematische Anwendung empirischer Forschungsmethoden zur Bewertung des Konzepts, des Untersuchungsplanes, der Implementierung und der Wirksamkeit sozialer Interventionsprogramme** (vgl. BORTZ & DÖRING, S. 96).

Mit WOTTAWA & THIERAU (S. 14) kann man zusammenfassend **drei allgemeine Merkmale** der Evaluationsforschung nennen:

- 1) Evaluation meint „**Bewerten**“, dient als Planungs- und Entscheidungshilfe und hat somit etwas mit der Bewertung von Handlungsalternativen zu tun.
- 2) Evaluation ist **ziel- und zweckorientiert**. Sie hat primär das Ziel, praktische Maßnahmen zu überprüfen, zu verbessern oder über sie zu entscheiden.
- 3) Evaluationsforschung soll dem **aktuellen Stand** wissenschaftlicher Techniken und Forschungsmethoden angepaßt sein.

Evaluationsforschung trägt somit zur **Handlungsoptimierung in komplexen Situationen** bei. Wie aus der Denkpsychologie bekannt, sind offene komplexe Probleme nicht endgültig optimierbar, so daß das Ziel darin bestehen muß, innerhalb eines wissenschaftsexternen, vorläufigen und in gewissen Grenzen willkürlichen Rahmens die Wahrscheinlichkeit für die Auswahl einer besonders guten Verhaltensalternative zu erhöhen und analog dazu die Wahl einer besonders schlechten Alternative zu verringern (Ideallösung vs. Übelminimierung). Evaluationsvorhaben rechtfertigen sich nicht aufgrund des Findens von absoluten Wahrheiten, sondern aufgrund ihres Beitrages zu einem Entscheidungsprozeß bezüglich der Auswahl von Verhaltensalternativen, der in jedem Fall ein Ergebnis erbringen muß (vgl. WOTTAWA & THIERAU, S. 21). Die **Methodik** in Evaluationsprojekten ist deshalb stets eine **Gratwanderung** zwischen dem wünschenswerten, dem prinzipiell möglichen, dem aufgrund von Ressourcenbegrenzungen tatsächlich möglichen und dem in Anbetracht der verschiedenen Interessenseinflüsse gewünschten Vorgehen (vgl. WOTTAWA & THIERAU, S. 553).

Hilfreich für die Begriffsklärung ist eine Unterscheidung zwischen Evaluation und Evaluationsforschung von SUCHMAN (1967, zit. n. WOTTAWA & THIERAU, S. 13):

Evaluation (Bewertung): Prozeß der Beurteilung des Wertes eines Produktes, Prozesses oder Programms, was nicht notwendigerweise systematische Verfahren oder datengestützte Beweise zur Untermauerung einer Beurteilung erfordert.

Evaluationsforschung: explizite Verwendung **wissenschaftlicher** Forschungsmethoden und -techniken für den Zweck der Durchführung einer Bewertung. Evaluationsforschung betont die Möglichkeit des **Beweises** anstelle der reinen Behauptung bzgl. des Wertes und Nutzens einer bestimmten sozialen Aktivität.

Nach BORTZ & DÖRING (S. 96 ff) ergibt sich für die Frage nach den **Unterschieden** zwischen **Grundlagen-** und **Evaluationsforschung** folgender **Vergleich:**

Kriterien	Grundlagenforschung	Evaluationsforschung
Erkenntnisinteresse	Generierung von Hintergrundwissen → offene Forschungsziele Überprüfung wissenschaftlicher Theorien (Beschreibung, Erklärung und Prognose eines Sachverhaltes oder Phänomens)	Begleitung und Bewertung (Erfolg/Mißerfolg) eines Evaluationsobjekts → begrenzt/zielgerichtet, gebundene Forschungsziele Überprüfung technologischer Theorien (praktische Umsetzung wissenschaftlicher Erkenntnisse in konkrete Handlungsanweisungen)
Forschungsauftrag	internes Forschungsinteresse	externer Auftraggeber
Anwendbarkeit / funktionaler Wert der Ergebnisse	von untergeordnetem Interesse	hohe Priorität, es sollen Entscheidungen bzw. Handlungsanweisungen ermöglicht werden
Publikation der Ergebnisse	Voraussetzung für die kritische Würdigung durch die „scientific community“ (Kontrolle)	selten (kein Interesse daran: a) negative Ergebnisse: schlechter Ruf; b) positive Ergebnisse: Verlust des Wettbewerbsvorteils)
Ergebnisinterpretation	mit gebotener wissenschaftlicher Zurückhaltung: vorsichtig und selbstkritisch hinsichtlich der Frage, ob der Geltungsbereich der geprüften Hypothese gefestigt oder ausgeweitet werden kann	unter Entscheidungszwang: eindeutige und verständliche Beantwortung der Evaluationsfrage, Erfolgs-/Mißerfolgsrückmeldung bzw. Handlungsanweisung (Ratgeberpflicht)

Frage 11: Erläutern Sie den **Zusammenhang** von **Interventions-** und **Evaluationsforschung!** Welches **Evaluationsmodell** ist in diesem Zusammenhang von besonderer Bedeutung? Geben Sie eine kurze Begründung! (1 x gefragt) bzw. **Unterscheiden** Sie die **Evaluationsforschung** von der **Interventionsforschung!** Welche **Berührungspunkte** weisen sie auf? (1 x gefragt)

Auch der Vergleich von Evaluations- und Interventionsforschung ist aufschlußreich (BORTZ & DÖRING, S.100 f).

Wenn die Grundlagenforschung bsw. erkannt hat, daß bestimmte Verhaltensstörungen auf traumatische Kindheitserlebnisse zurückgeführt werden können, wäre es Aufgabe psychologischer Experten, diese Erkenntnisse in eine Therapie/Intervention umzusetzen, also eine **technologischer Theorie zu entwickeln, die die zielgerichtete Ableitung einer Therapie ermöglicht.** Diesen Vorgang bezeichnet man mit **Interventionsforschung.**

Die Aufgabe der **Evaluationsforschung** besteht darin, solche **Interventionen zu bewerten.** Dabei kann es auch vorkommen daß Interventions- und Evaluationsforschung nicht sequentiell ablaufen (summative Evaluation), sondern parallel (formative Evaluation).

Frage 12: Welche **Funktion** hat die **Evaluationsforschung** im Zusammenspiel von wissenschaftlichen und technologischen Theorien? (1 x gefragt) bzw. **Beschreiben** Sie den Begriff der **Evaluationsforschung!** Ordnen Sie den Begriff in die **Systematik** von **technologischer** und **wissenschaftlichen Theorien** (HERRMANN, 1979) ein! (1 x gefragt)

Zur weiteren begrifflichen Klärung, was „Evaluationsforschung“ bedeutet, trägt nach BORTZ & DÖRING (S. 99 f) auch die Unterscheidung zwischen technologischen und wissenschaftlichen Theorien im Sinn von HERRMANN (1979) bei.

Wissenschaftliche Theorien dienen der Beschreibung, Erklärung und Vorhersage von Sachverhalten; sie werden in der Grundlagenforschung entwickelt.

Technologische Theorien geben konkrete Handlungsanweisungen zur **praktischen Umsetzung wissenschaftlicher Theorien;** sie fallen in den Aufgabenbereich der angewandten Forschung sowie Interventions- bzw. **Evaluationsforschung.**

Beide Theoriearten sind für eine Wissenschaft gleichermaßen wichtig. Die empirische Überprüfung wissenschaftlicher Theorien zählt zu den Aufgaben der Grundlagenwissenschaften, während die Überprüfung technologischer Theorien vorrangig Evaluationsforschung ist. Eine gute wissenschaftliche Theorie ist durch eine präzise Terminologie, einen logisch konsistenten Informationsgehalt (Widerspruchsfreiheit), eine möglichst breite inhaltliche Tragweite sowie eine begrenzte Anzahl von Annahmen (Sparsamkeit) gekennzeichnet. Eine gute technologische Theorie sollte wissenschaftliche Erkenntnisse in effiziente, routinisierte Handlungsanweisungen umsetzen und Wege ihrer praktischen Nutzbarmachung aufzeigen.

Frage 13: Mit welchen **ethischen Problemen** kann ein Evaluator im Rahmen eines **Evaluationsprojekts** konfrontiert werden? (1 x gefragt)

1) Ethisch nicht akzeptable Interventionsziele

Ein Ziel der Evaluationsforschung besteht in der Verbesserung von Entscheidungen, die ihrerseits **Auswirkungen auf menschliche Schicksale** haben können. Verbesserungen von innerbetrieblichen Arbeitsabläufen können (zunächst versteckt für den Evaluator) den Abbau von Arbeitsplätzen mit sich bringen (oder sogar zum Ziel haben). Ein Evaluator muß also in verstärktem Maße die möglichen Konsequenzen seines Handelns durchdenken und im Zweifel – aus ethisch-moralischer Verantwortung - auch bereit sein, die Übernahme eines Evaluationsprojektes abzulehnen.

Die Einbettung in den unmittelbaren Entscheidungsbezug bringt es mit sich, daß mit erheblichem **Druck in Richtung eines erwünschten Ergebnisses** durch Auftraggeber zu rechnen ist (vgl. WOTTAWA & THIERAU, S. 551 f). So ist es etwa denkbar, daß in einem schulischen Modellversuch das Evaluationsergebnis in Richtung der Auffassung des Bildungsministeriums beeinflußt wird. Dazu werden oft getarnte Maßnahmen eingesetzt, wie beschränkter Zugang zu bestimmten Gesprächspartnern, Informationen usw. Der Evaluator muß in der Lage sein, solche Verzerrungen zu erkennen und anzusprechen, gegebenenfalls auch persönliche Konsequenzen zu ziehen.

2) Ethisch bedenkliche Bedingungen der Untersuchungsdurchführung

Ethische Probleme entstehen aber auch dann, wenn

- zur Bewertung einer Maßnahme Informationen benötigt werden, die die **Intimsphäre** der Betroffenen verletzen;
- die Verweigerung, an der Evaluationsstudie teilzunehmen, an **Sanktionen** geknüpft ist;
- die Mitwirkung an der Evaluationsstudie mit **psychischen oder physischen Beeinträchtigungen** verbunden ist (vgl. BORTZ & DÖRING, S. 103 f).

Die voranstehenden Überlegungen belegen, daß Evaluationsprojekte für den Evaluator nicht nur an ethisch nicht akzeptablen Interventionszielen, sondern auch an ethisch bedenklichen Bedingungen der Untersuchungsdurchführung scheitern können.

WOTTAWA & THIERAU sehen neben einem ausgeprägten **ethisch-moralischen Verantwortungsbewußtsein** und den erforderlichen **fachlichen Kenntnissen** (abgeschlossenes empirisch orientiertes psychologisches bzw. sozialwissenschaftliches Studium, gute Methodenkenntnisse, gute Allgemeinbildung im jeweiligen Evaluationsfeld) folgende **weitere persönliche Voraussetzungen** für einen erfolgreichen Evaluator (vgl. S. 51 f):

- **hohe Leistungsmotivation:** der Evaluator hat kaum unmittelbare Macht, und auch seine Anschlußmotivation sollte (wegen der Neutralität gegenüber konkurrierenden Gruppen) nicht allzu stark ausgeprägt sein;
- **gute Kommunikationsfähigkeit:** dies betrifft sowohl eine schwer veränderbare Disposition, am Kontakt mit anderen Menschen Spaß zu haben, als auch die Beherrschung entsprechender Gesprächs- und Gruppenmoderationstechniken;
- **gute Fähigkeit zur Perspektivenübernahme** („role-taking“): die gedankliche Übernahme der Welt- und Problemsicht verschiedenster Gruppen ist oft die Voraussetzung, um eine allen Beteiligten bzw. Betroffenen annähernd gerecht werdende

Projektplanung durchführen zu können und (unbewußte) Einseitigkeiten, etwa bei der Auswahl von Bewertungskriterien, zu vermeiden.

Frage 14: Beschreiben Sie knapp eine **Systematik** von **Evaluationsprojekten!** (1 x gefragt)
Nach WOTTAWA & THIERAU kann man folgende **sechs Gestaltungsaspekte** von Evaluationsstudien unterscheiden, nach denen man sie systematisieren kann (vgl. S. 55 ff):

1) Evaluationsziele: Erarbeitung der dem Projekt zugrunde liegenden Ziele zusammen mit dem Auftraggeber, Interventionsleiter und den Beteiligten bzw. Betroffenen. (→ Warum/ Wozu wird evaluiert?)

2) Evaluationsbereich: Das Praxisfeld, in welchem die Evaluation stattfinden soll (häufiger im Bildungssektor, Wirtschaft, Agrar- und Verkehrspolitik, Familien- und Sozialpolitik, Justizvollzug, Gesundheitswesen usw.). (→ In welchem gesellschaftlichen Bereich wird evaluiert?)

3) Evaluationsobjekt: Oberbegriff für die zu bewertenden Alternativen. Inzwischen gibt es zahlreiche Objekte wie Personen, Umwelt- und Umgebungsfaktoren, Produkte, Techniken/ Methoden, Programme usw. (→ Wer/Was wird evaluiert?)

4) Ort der Evaluierung: die raum-zeitliche Spezifizierung der Studie. In der Regel finden Evaluationsstudien im Feld statt und können sowohl experimentell als auch quasi-experimentell angelegt sein. (→ Wo wird evaluiert?)

5) Evaluationsmodell: Die **formative** Evaluation stellt vor allem Informationen für noch in der Vorbereitungs- oder Implementierungsphase befindliche oder laufende Programme bereit, die verbessert werden sollen. Die **summative** Evaluation soll die Qualität und den Einfluß bereits stattgefundenen Programme feststellen und abschließend bewerten. Sie ist dann sinnvoll, wenn mehrere disjunkte Handlungsformen vorliegen, deren Konsequenzen miteinander verglichen und somit bewertet werden können. (→ Wie wird evaluiert?)

6) Evaluationsnutzung: Betrifft die Weise, in der die Ergebnisse in praktisches Handeln umgesetzt werden (→ Wie werden die Ergebnisse aufbereitet und verwendet?):

- geschlossene Selbstevaluation,
- Ergebnisse dienen Machtentscheidung,
- Ergebnisse für die Fachöffentlichkeit,
- Ergebnisse werden für eine politische Entscheidung verwendet und veröffentlicht usw.

Frage 15: Erläutern Sie die **Zielgruppenbestimmung** im Rahmen der **Zielexplication** bei Evaluationsstudien! (1 x gefragt)

Zielgruppenbestimmung: Aufgrund der häufig anzutreffenden eingeschränkten Vorstellungen der Auftraggeber darüber, um welche Zielgruppe es bei den zu evaluierenden Alternativen geht, gilt es, deren diesbezügliche Ideen zu erweitern. Dazu dienen **Kreativitätstechniken** wie z.B. die hierarchisch gesteuerte Assoziationskette, die in Gruppensitzungen durchgeführt wird und den Beteiligten so ermöglicht, die notwendigen Erkenntnisse selbst zu erlangen – was sinnvoller ist als von außen kommende Vorschläge (Abwehr!).

Beispiel: **Hierarchisch gesteuerte Assoziationsketten**

Auftraggeber: ein Schulbuchverlag

Fragestellung: „Prüfen Sie, ob durch „advanced organizer“* die Verständlichkeit von Texten (und damit die Verbreitung eines Schulbuches) verbessert wird!“

* „advanced organizer“: vor jedem Kapitel bzw. Abschnitt wird eine Übersicht über die folgenden Ausführungen geboten, um dem Leser den schnellen Aufbau einer entsprechenden kognitiven Struktur zu ermöglichen.

Die so formulierte Fragestellung könnte man direkt aufgreifen und das Projekt auf ein *Einfach-Design* beschränken (*laborexperimentell; randomisierte Schülergruppen werden verschiedenen Materialvariationen ausgesetzt und hinsichtlich ihres Textverständnisses geprüft*). Dieses Vorgehen wäre zwar einfach, überschaubar und kostengünstig, aber nicht un-

bedingt wirklich relevant für die zu treffende praktische Entscheidung des Schulbuchverlags. Um die Zielgruppe des Verlags zu konkretisieren und damit dessen Zielsetzung zu präzisieren (was die Praxisrelevanz des Vorgehens erhöht), werden statt dessen in Gruppensitzungen gemeinsam hierarchische Assoziationsketten entwickelt. Ablauf:

Überlegung: Welche Personengruppen sind von den zu evaluierenden Alternativen betroffen?

z.B. Schulbuch: Welche Personengruppen werden mit dem Buch unmittelbar Kontakt haben? → z.B. erste Assoziationskette:

- Autor- Lehrer- Schüler.

Erweiterung: Zusammenstellung von **Personen-Obermengen**, z.B. weitere Ketten:

- Autor - Verlag - Lehrer - Schüler - Eltern
- Autor - Verlag - Schulbehörden - Händler - Lehrer - Schüler – Eltern.

Ausdifferenzierung: Einteilung jeder Personen-Obermenge nach Relevanz für das zu evaluierende Problem;

z.B. Kette für die Personen-Obermenge „Lehrer“:

- Alter – beruflicher Status – Fach – Unterrichtsmethodik – Schulform usw.

z.B. Kette für die Personen-Obermenge „Schüler“:

- Klassenstufe – Schulform - Intelligenz - Vorkenntnisse - Arbeitsmotivation - Geschlecht - Hausaufgabenbetreuung usw.

Kombination von Teilgruppen einer Personen-Obermenge: wenn Wechselwirkungen zwischen den einzelnen Definitionsteilen hinsichtlich der Fragestellung plausibel sind, sind solche Ketten notwendig;

z.B. Kette für die Kombination aus der Personen-Obermenge „Schüler“: Schüler der Klasse 6 an einem Gymnasium ohne Hausaufgabenbetreuung.

Kombination von Teilgruppen verschiedener Personen-Obermengen: soweit dies sachlich sinnvoll ist, werden so im Sinn einer weiteren Verfeinerung relevante Subgruppen gebildet;

z.B. Kette für die Kombination aus den Personen-Obermengen „Schüler“, „Lehrer“ und „Eltern“: Schüler der Klasse 6 an einem Gymnasium ohne Hausaufgabenbetreuung, die bei einem Lehrer mit besonderer Vorliebe für Frontalunterricht das Fach Englisch lernen, deren Eltern keine Kenntnisse in Englisch haben und wenig am Schulerfolg ihrer Kinder interessiert sind.

Mit dieser kreativen Methode läßt sich eine nahezu unübersehbare Vielfalt von potentiellen Zielgruppen herausarbeiten. Die entscheidende Aufgabe der Zielgruppenbestimmung liegt daher in der **Reduktion** der Zielgruppen auf die wirklich **wesentlich erscheinenden Teilgruppen**. Im Beispiel des „advanced organizer“ wird man zumindest überlegen, das ursprüngliche Einfach-Design um Aspekte des Entwicklungsgrads der Schüler (Klassenstufe), des Faches, der Unterrichtsmethodik des Lehrers und evtl. des Ausmaßes an Unterstützung bei den Hausaufgaben durch die Eltern zu erweitern. Außerdem wird man durch die hierarchischen Assoziationsketten darauf hingewiesen, daß nicht nur der Lernerfolg, sondern auch Aspekte wie Akzeptanz des Schulbuchs durch die Lehrer (ohne sie kann sich kein Schulbuch durchsetzen!), der durch solche didaktischen Hilfen vergrößerte Umfang des Buches und der Preis als Bewertungskriterien berücksichtigt werden müssen (vgl. dazu die Bewertungskriterienexplikation (WOTTAWA & THIERAU, S. 86) → Bewertungsprozeß).

Frage 16: Was versteht man unter „**antizipatorischer Ergebnisverwertung**“ bei der **Zielexplikation**? Verwenden Sie zur Erläuterung eine passende Technik! (2 x gefragt)

Antizipatorische Ergebnisverwertung: Vor allem bei großen sozialwissenschaftlichen Evaluationsprojekten besteht das **Problem der praktischen Verwendbarkeit**, die aufgrund von zwei strukturellen Gegebenheiten eingeschränkt sein kann:

- 1) **projektunabhängige** Veränderungen der **Rahmenbedingungen** und, daraus resultierend, auch der Zielsetzung des Auftraggebers durch die lange Dauer der Projekte (meist mehrere Jahre);
- 2) nachträgliche, **projektergebnisabhängige** Verschiebungen der Zielsetzung vor dem Hintergrund der gewonnenen neuen Problemsicht.

Als vorbeugende Maßnahme gegen eine mangelnde praktische Verwendbarkeit des Evaluationsprojekts sollten die **späteren Verwertungssituationen antizipiert** werden.

Um mögliche Veränderungen der **Rahmenbedingungen** abzuschätzen, kann die Szenario-Technik eingesetzt werden und helfen, Fehler bei der Zielgruppenbestimmung oder der Konkretisierung des Evaluationsobjekts zu vermeiden.

Szenario-Technik

Dies ist eine systematische Methodik zur Entwicklung und Beschreibung möglicher zukünftiger Situationen (Szenarien) sowie zum Aufzeigen des Entwicklungsverlaufs, der zu diesen Situationen geführt hat.

Charakteristika:

- sorgfältige Analyse von gegenwärtigen Situationen;
- Einbeziehung von quantitativen und qualitativen Aspekten;
- Ermittlung von Annahmen über die Haupteinflussfaktoren;
- verfahrensmäßig relativ problemlose Verarbeitung von Störereignissen;
- Entwicklung von alternativen, in sich konsistenten (stimmigen) Zukunftsbildern.

Für eine grobe Abschätzung der durch die **Projektergebnisse** möglichen Situationsveränderungen und sich daraus ergebenden zusätzlichen Untersuchungsziele kann auch die Planspiel-Methode zum Einsatz kommen.

Planspiel

Dies ist eine Unterweisungsmethode, die speziell dem Entscheidungshilfetraining dient. Dem Planspiel liegt immer eine reale Situation zugrunde, die in einem Modell simuliert wird. Auf diese Weise soll die wechselseitige Abhängigkeit der einzelnen Systemelemente verdeutlicht werden und die Wirkung einzelner Entscheidungen auf das Gesamtsystem transparent gemacht werden. Beim Planspiel übernehmen die Teilnehmer die Rolle von Entscheidungsinstanzen. Aufgrund der modellartigen Simulation des Gesamtsystems können die Folgen der Entscheidungen ermittelt und bewertet werden.

Die wichtigsten **Elemente** des Planspiels sind:

- Nachahmung der Realität im Modell bzw. Reduktion der Realität auf zielrelevante Faktoren des Planspiels;
- aktives Handeln in Form abstrakter Denktätigkeit bzw. Interaktion der Spieler in der simulierten Realität;
- hohe Motivationskraft durch Ausnutzung des dem Menschen innewohnenden Spieltriebs und damit Wirkung und Verstärkung des sachbezogenen Interesses;
- rollenspielartige Übernahme bestimmter Verhaltensweisen innerhalb der Simulationssituation;
- Konflikttraining verursacht durch abweichende Zielvorstellungen zwischen den Spielgruppen sowie innerhalb der einzelnen Gruppen;
- Training der Kommunikationsfähigkeit, da die gestellte Aufgabe einen Informationsaustausch innerhalb der Gruppen, zwischen den Gruppen und zum Spielleiter erfordert.

Frage 17: Skizzieren Sie die **Szenario-Technik** im Rahmen der **antizipatorischen Ergebniserwartung!** (1 x gefragt)

Antizipatorische Ergebnisverwertung: Vor allem bei großen sozialwissenschaftlichen Evaluationsprojekten besteht das **Problem der praktischen Verwendbarkeit**, die aufgrund von zwei strukturellen Gegebenheiten eingeschränkt sein kann:

- 1) **projektunabhängige** Veränderungen der Rahmenbedingungen und, daraus resultierend, auch der Zielsetzung des Auftraggebers durch die lange Dauer der Projekte (meist mehrere Jahre);
- 2) nachträgliche, **projektergebnisabhängige** Verschiebungen der Zielsetzung vor dem Hintergrund der gewonnenen neuen Problemsicht.

Als vorbeugende Maßnahme gegen eine mangelnde praktische Verwendbarkeit des Evaluationsprojekts sollten die **späteren Verwertungssituationen antizipiert** werden.

Um mögliche **Veränderungen der Rahmenbedingungen** abzuschätzen, kann die Szenario-Technik eingesetzt werden und helfen, Fehler bei der Zielgruppenbestimmung oder der Konkretisierung des Evaluationsobjekts zu vermeiden.

Szenario-Technik

Definition: Eine systematische Methodik zur Entwicklung und Beschreibung möglicher zukünftiger Situationen (Szenarien) sowie zum Aufzeigen des Entwicklungsverlaufs, der zu diesen Situationen geführt hat. Sie besteht aus acht logisch aufeinander aufbauenden Schritten, die den gesamten Prozeß transparent und in allen Phasen nachvollziehbar machen.

Charakteristika:

- sorgfältige Analyse von gegenwärtigen Situationen;
- Einbeziehung von quantitativen und qualitativen Aspekten;
- Ermittlung von Annahmen über die Haupteinflussfaktoren;
- verfahrensmäßig relativ problemlose Verarbeitung von Störereignissen;
- Entwicklung von alternativen, in sich konsistenten (stimmigen) Zukunftsbildern.

Zugrundeliegendes Denkmodell:

Aus der Fülle der plausiblen, in sich stimmigen Szenarien wird das plausibelste ausgewählt, das sog. „**Trend Szenario**“. Zusätzlich werden mindestens zwei „**Extrem Szenarien**“ ausgewählt; eins mit einer besonders positiven, das andere mit einer extrem negativen Entwicklungstendenz. Man geht davon aus, daß man mit der Auswahl von mindestens drei Szenarien ein „**Trichtermodell**“ der zukünftigen Entwicklung hat (Entwicklung als ein umgekehrter Trichter): Das Trend Szenario entspricht der Hauptachse des Trichters, die Extrem Szenarien definieren die äußere Hülle, der Trichter hat seine punktförmige Spitze in der Gegenwart (in der ja alle Szenarien zusammenfallen) und erweitert sich im Lauf der Zeit immer mehr, so daß dann die verschiedenen Szenarien immer stärker auseinanderklaffen. Je mehr Zeit vergeht, umso unsicherer wird auch die Prognose, da immer mehr unkontrollierte und nicht vorhergesehene Störereignisse die Entwicklung verändert haben.

Ziel solcher Szenarien-Studien ist es vor allem, durch rechtzeitig eingeleitete Maßnahmen dafür zu sorgen, daß prognostizierte unerwünschte Szenarien nicht Realität werden. Ihre Ergebnisse sind selbstverständlich nicht unfehlbar, bieten aber für die sinnvolle Steuerung zukünftiger Entwicklungen im Rahmen von antizipatorischer oder prognostischer Evaluation eine rationalere Grundlage als persönliche Zukunftserwartungen.

Ablauf in acht Schritten/Phasen:

1. Strukturierung und Definition des Untersuchungsfeldes
2. Identifizierung und Strukturierung der wichtigsten Einflußbereiche auf das Untersuchungsfeld
3. Ermittlung von Entwicklungstendenzen und kritische Beschreibung der Umfeldler
4. Bildung und Auswahl konstanter Annahmebündel
5. Interpretation der ausgewählten Umfeldszenarien
6. Einführung und Auswirkungsanalyse signifikanter Störereignisse
7. Ausarbeitung der Szenarien bzw. Ableiten von Konsequenzen für das Untersuchungsfeld
8. Konzeption von Maßnahmen und Planungen.

Es wird aufgrund des hohen Aufwandes selten möglich sein, eine ‚ideale‘ Szenario-Technik für die Zielsetzung von Evaluationsprojekten durchzuführen. Doch schon eine relativ grobe Abschätzung der erwartbaren projektunabhängigen Veränderungen kann Fehler vermeiden helfen. Das Ergebnis kann durchaus der Verzicht auf das Evaluationsprojekt selbst sein, wenn gravierende bzw. schnelle Veränderungen der Rahmenbedingungen im Lauf des Projekts zu erwarten sind. In den meisten Fällen werden dadurch jedoch Teile des Projekts akzentuiert.

Frage 18: Was versteht man im Rahmen eines **Evaluationsprojekts** unter dem Prozeß der **Zielkriterienerstellung**, und wie geht man dabei vor? (1 x gefragt)

Der Begriff der „Zielkriterienerstellung“ kommt weder im Text von WOTTAWA & THIERAU noch im Text von BORTZ & DÖRING vor! Entweder - unwahrscheinlich - meint HUSSY die „Zielexplication“ (Oberbegriff für 1. „Zielgruppenbestimmung“, 2. „Konkretisierung des Evaluationsobjekts“, 3. Antizipatorische Ergebnisverwertung) oder - wahrscheinlicher - den **„Bewertungsprozeß“** (Oberbegriff für 1. „Auswahl der Bewertungskriterien“, 2. „Nebenfolgenabschätzung“, 3. „Operationalisierungsfragen“) bzw. hiervon nur 1. „Auswahl der Bewertungskriterien“ (vgl. WOTTAWA & THIERAU, Kap. 4: Zielexplication und Bewertungskriterien → 4.1 Zielexplication, → 4.2 Bewertungsprozeß und 4.3 Bewertungs- und Entscheidungshilfen). WOTTAWA & THIERAU verweisen auf S. 86 unten „auf die Kriterienexplikationen in Abschnitt 4.2“. Da der „Bewertungsprozeß“ auf der „Zielexplication“ aufbaut und bisher keine Klausurfrage zur „Konkretisierung des Evaluationsprozesses“ im Rahmen der „Zielexplication“ gestellt wurde, wird hier eine ausführliche Antwort gegeben, die sowohl auf die „Zielexplication“ als auch auf den „Bewertungsprozeß“ im Sinn einer **Bewertungskriterienexplikation** eingeht.

Evaluation dient der Bewertung verschiedener Maßnahmen, Organisationsformen etc., und ihre Ergebnisse sollen praktische Konsequenzen ermöglichen, z.B. im Sinn der Auswahl der ‚besseren‘ Alternative. Diese Bewertung setzt ein (subjektiv bewertetes) Ziel voraus, das durch die einzelnen zu evaluierenden Alternativen besser oder schlechter erreicht wird. Anders formuliert dient Evaluation der Optimierung der Grundlage für nutzenmaximierendes Verhalten. **Nutzenmaximierendes Verhalten** kann also als **übergeordnetes Ziel** der zu evaluierenden Alternativen angesehen werden, das jedoch auf konkreter Ebene, in Abhängigkeit des zu evaluierenden Gegenstandes bzw. des Verwertungszusammenhangs der Ergebnisse, völlig unterschiedlich aussehen kann.

Für erfolgreiche Evaluationsstudien (Verlauf und Ergebnisse) ist es daher unverzichtbar, zu wissen

- in welchem Verwertungszusammenhang die Ergebnisse zu sehen sind (Zielexplication)
- welche Bewertungskriterien dafür herangezogen werden sollen (Bewertungsprozeß)
- wie der Nutzen der aufgetretenen Ausprägungsgrade der Bewertungskriterien einzuschätzen ist und wie auf dieser Basis eine globale Alternativenbewertung erfolgen kann (Bewertungs- und Entscheidungshilfemethoden) (vgl. WOTTAWA & THIERAU, S. 83).

Die Fragen der (konkreten) Zielsetzung und der Nutzenaspekte der zu evaluierenden Alternativen sind in vielen Evaluationsprojekten die größte – und basale - Schwachstelle, die die Praxisrelevanz des Projekts beeinträchtigen. Gerade bei sozialwissenschaftlich interessanten Themen ist es nicht einfach, überhaupt zu Beginn des Projekts zwischen allen Beteiligten einen **Konsens über Zielsetzung und Nutzenaspekte** herzustellen. Noch schwieriger ist es, einen solchen Konsens auch als Grundlage für die nachträgliche Bewertung von Evaluationsprojekten beizubehalten, wenn ‚unerwünschte‘ Ergebnisse aufgetreten sind oder wenn sich inzwischen Rahmenbedingungen des Projektes so geändert haben, daß sie – im nachhinein - eine andere Projektausrichtung als sinnvoll erscheinen lassen.

Zielexplication

Die möglichen Zielvorgaben, die ein Evaluator von seinen Auftraggebern bekommt, beschreiben ein Kontinuum von konkret formulierten Zielen, bei denen der Evaluator lediglich Datensammler ist, bis hin zu völlig vage formulierten Zielen, bei denen er die faktische Verantwortung für die Verwertbarkeit der Ergebnisse übernehmen muß. In der Regel sind die Zielvorgaben der Auftraggeber eher in der Mitte dieses Kontinuums anzusiedeln, was den Evaluator vor die Aufgabe der **Verbesserung der Zielsetzung** im Sinn von rationalen und konkretisierten Zielen stellt. Gründe dafür liegen darin, daß Auftraggeber häufig nur eine eingeschränkte Vorstellung vom möglichen Umfang des Projektes haben, sowohl in bezug auf die zu evaluierenden Alternativen (Evaluationsobjekte) als auch auf die betroffenen Personen (Zielgruppe der Evaluationsobjekte). Grundlegende Schwierigkeiten hinsichtlich des zu bildenden Konsenses stellen **Konflikte innerhalb des Auftraggebers** dar. Diese können **institutionalisiert** sein; dann gilt es, ausdrücklich allen Beteiligten die Projektstrategie offen zu legen und die konträren Standpunkte anzusprechen

und zu integrieren. Oder sie können **verdeckt** sein, was es dem Evaluator nahezu unmöglich macht, adäquat zu reagieren.

Bei der Zielexplication sind **drei zentrale Aspekte** zu berücksichtigen:

1. Zielgruppenbestimmung
2. Konkretisierung des Evaluationsobjekts
3. Antizipatorische Ergebnisverwertung.

1. Zielgruppenbestimmung

Aufgrund der häufig anzutreffenden eingeschränkten Vorstellungen der Auftraggeber darüber, um welche Zielgruppe es bei den zu evaluierenden Alternativen geht, gilt es, deren diesbezügliche Ideen zu erweitern. Dazu dienen **Kreativitätstechniken** wie z.B. die hierarchisch gesteuerte Assoziationskette, die in Gruppensitzungen durchgeführt wird und den Beteiligten so ermöglicht, die notwendigen Erkenntnisse selbst zu erlangen – was sinnvoller ist als von außen kommende Vorschläge (Abwehr!) (siehe Frage 15!).

2. Konkretisierung des Evaluationsobjektes

Die Kenntnis der möglichst genauen Zielsetzung ist die Basis für den Evaluator, um das Evaluationsobjekt im Sinn einer **Operationalisierung** der theoretischen Begriffe (Konstrukte) in **empirisch erfaßbare Indikatoren** zu konkretisieren. Da diese Konkretisierungen die spätere Verwendung der Ergebnisse beeinflussen, ist es wichtig, daß der Evaluator seine diesbezüglichen Vorstellungen dem Auftraggeber als Vorschläge vorlegt, um sie von ihm prüfen und möglichst formell festlegen zu lassen.

Die Beeinflussung der Ergebnisse durch die konkrete Festlegung der Begriffe wird schon an einem so einfachen Problem wie dem „advanced organizer“ deutlich: Die Realisierungsmöglichkeiten eines „advanced organizers“ sind vielfältig (optische Aspekte: kleiner Kasten im Kleindruck mit ausschließlicher Angabe der kommenden Zwischenüberschriften bis hin zu mehrseitigen Darstellungen; inhaltliche Aspekte: bloße Aufzählung der folgenden Hauptpunkte bis hin zu einer umfassenden, evtl. noch Sekundäraspekte mit beinhaltenden Begründung gerade dieser Auswahl und Reihenfolge). Diese **Gestaltungsdetails** eines „advanced organizer“ wirken sich wahrscheinlich stärker auf relevante Bewertungskriterien wie Verständlichkeit, Akzeptanz und Kosten aus als die bloße Unterscheidung zwischen dem Vorhandensein oder Nichtvorhandensein irgendeines „advanced organizer“. Je komplexer eine Maßnahme ist, umso vielfältiger wird der Gestaltungsspielraum des Evaluators. Es empfiehlt sich, daß er zunächst eine Vielzahl möglicher Gestaltungsdimensionen erarbeitet.

Ablauf der Konkretisierung in **zwei Schritten**:

1. Ausarbeitung einer Vielzahl von Gestaltungsdimensionen

Techniken (je nach Problemstellung):

- **Analyse der bereits vorhandenen unterschiedlichen Ausprägungen** der zu evaluierenden Maßnahme auf den relevanten Dimensionen (Literaturstudium, Hospitationen, Experteninterviews)
- **Gruppendiskussionen** (mit verschiedenen Betroffenen, Ideenstiftern, Auftraggeber, Experten)
- **Brain-Storming-Techniken** und ähnliche kreativitätsfördernde Gruppenverfahren (mit Mitarbeitern des Auftraggebers und des Projektteams).

2. Systematische Kombination verschiedener Ausprägungsgrade der Gestaltungsdimensionen (z.B. in Anlehnung an die Facettentheorie). Die viel zu große Vielfalt potentieller Konkretisierungen muß für ein durchführbares Projekt begrenzt werden. Die Begrenzung auf wenige Varianten sollte unter Berücksichtigung der späteren praktischen Verwendung der Ergebnisse und der zukünftig zu erwartenden Rahmenbedingungen durchgeführt werden (→ antizipatorische Ergebnisverwertung!).

3. Antizipatorische Ergebnisverwertung

Vor allem bei großen sozialwissenschaftlichen Evaluationsprojekten besteht das **Problem der praktischen Verwendbarkeit**, die aus zwei Gründen eingeschränkt sein kann:

- 1) durch Veränderung der Rahmenbedingungen und einer daraus resultierenden Verschiebung der Zielsetzung des Auftraggebers (projektunabhängig);
- 2) durch nachträgliche Verschiebung der Zielsetzung vor dem Hintergrund der gewonnenen Ergebnisse (projektergebnisabhängig).

Als vorbeugende Maßnahme können **Verwertungssituationen antizipiert** werden. Hierfür geeignete Methoden sind z.B. die **Szenario-Technik** und die **Planspiel-Methode** (siehe Frage 16 und Frage 17!).

Bewertungsprozeß = Bewertungskriterienexplikation (≈ Zielkriterienerstellung)

Nach abgeschlossener Zielexplication (Festlegung der (Teil-) Ziele der zu evaluierenden Maßnahme) muß ein **Konsens über die Bewertungskriterien** hergestellt werden, also über die empirischen Beobachtungen, anhand derer das Ausmaß der Zielerreichung beurteilt werden soll. Der empirisch-wissenschaftliche Informationsgewinn durch die Evaluation baut auf der Menge der erhobenen Ausprägungsgrade der ausgewählten Indikatoren (Kriterien) auf, so daß das gesamte Ergebnis des Projekts entscheidend von der konkreten Auswahl dieser Kriterien abhängt. Bsp.: Es kann einen großen Unterschied im Bewertungsergebnis machen, ob man den „Lernerfolg“ von Schülern in verschiedenen Organisationsformen an den von den Lehrern vergebenen Noten, objektiven Testverfahren, Einschätzungen der Eltern oder der späteren leistungsmäßigen Entwicklung der Schüler in folgenden Klassen mißt (verschiedene Operationalisierungen/Kriterien für „Lernerfolg“).

Folgende **drei Punkte** sind für den Bewertungsprozeß im einzelnen zu klären:

1. Auswahl der Bewertungskriterien
2. Nebenfolgenabschätzung
3. Operationalisierungsfragen.

1. Auswahl von Bewertungskriterien

Hierzu muß zunächst die Ausgangssituation geklärt werden. Dann muß die Zielsetzung der zu evaluierenden Maßnahme im Sinn von Richtzielen (Grobziele) möglichst detailliert werden. Danach müssen für die gefundenen Grobziele ideenreich geeignete Vorschläge im Sinn von Feinzielen gefunden werden, wobei alle auch nur annähernd sinnvollen Vorschläge in die Projektplanung aufgenommen werden sollten, damit sich jeder Beteiligte im Projektplan wiederfinden kann und so keine offenen oder verdeckten Konflikte geschürt werden. Die Feinziele werden dann nach (subjektiven) Kriterien hierarchisiert, um im Konsens der potentiellen Konfliktpartner eine Reduktion der Feinziele mit den dazu passenden Bewertungsdimensionen (theoretische Bewertungskriterien) auf eine bewältigbare Arbeitsmenge vorzunehmen.

Vorgehen in **vier Schritten**:

1. Analyse der Ist-Situation
2. Festlegung der Richtziele (Grobziele)
3. Bestimmung der Feinziele
4. Hierarchisierung der Feinziele nach (subjektiven) Kriterien und Reduktion.

Beispiel: „advanced organizer“ aus Sicht der Zielsetzung des Lehrers:

1. Analyse der Ist-Situation:

Erfassung

- des durchschnittlichen Leistungsniveaus in der unterrichteten Klasse
- der Motivation der Schüler, mit dem alten Schulbuch zu arbeiten
- der auftretenden Probleme, die sich bei der Arbeit mit dem alten Schulbuch ergeben.

2. Festlegung der Richtziele (Grobziele):

- a) schulische Ziele
- b) persönliche Ziele

3. Bestimmung der Feinziele:

mögliche Feinziele von a):

- didaktische Verbesserung des Unterrichts
- Straffung des Unterrichts
- Verbesserung des Klassendurchschnitts
- zeitökonomische Aspekte
- Steigerung der Motivation der Schüler
- Erleichterung der Informationsaufnahme
- usw.

mögliche Feinziele von b):

- Demonstration von Innovationsfreudigkeit
- Erhöhung der eigenen Motivation
- Hoffnung auf höheres Ansehen/Status
- Zeitersparnis bei der Themenauswahl und Vorbereitung
- Durchsetzung im Kollegium
- wissenschaftliche Orientierung
- usw.

4. Hierarchisierung der Feinziele nach (subjektiven) Kriterien und Reduktion.

Das Finden der Zielhierarchie und der dazu passenden Bewertungskriterien sollte im Team erfolgen, wobei eine interessens- und vorbildungsmäßig heterogene Gruppe empfehlenswert sein kann. Geeignete Methoden dazu sind die **Brain-Storming-Technik** oder/und die **Metaplan-Methode**.

Brain-Storming-Technik

Dies ist eine **Technik zur kreativen Problemlösung** nach dem **Prinzip der freien Assoziation**, die darauf abzielt, die negativen Erscheinungen von Diskussionsrunden und Konferenzen wie destruktive Kritik, Rivalität unter den Teilnehmern und Verzettelung in unwichtige Einzelheiten zu überwinden.

Das klassische Brain-Storming beinhaltet **zwei Phasen**:

1. **Ideenfindung**: Der Moderator fordert die Teilnehmer auf, zu einem spezifischen Problem möglichst viele Ideen zu produzieren (Quantität, nicht *Qualität!*) Wichtig ist der **Ideenfluß**, d.h., die Teilnehmer sollen alle aufkommenden Ideen, auch ungewöhnliche oder unrealistisch erscheinende, aussprechen, und es reicht die Andeutung des Gedankengangs, ohne ausführliche Erläuterung. Hierbei ist jegliche Kritik – positive wie negative („Killerphrasen“) – an den einzelnen Vorschlägen untersagt, auch nonverbale Äußerungen. Alle Ideen werden protokolliert.

2. **Ideenbewertung**: Die einzelnen Ideen werden anhand von **drei Kriterien** bewertet, um die Auswahl sinnvoller Ideen zu erleichtern:

- **Einfachheit**
- **Realisierbarkeit**
- **Schwierigkeitsgrad.**

Der Grad dieser drei Bewertungsdimensionen wird dazu auf einer Punkte-Skala eingetragen.

Metaplan-Methode

Dies ist eine **Gesprächs- bzw. Diskussionstechnik**, die durch **hierarchiefreies Arbeiten** die Teilnehmer motiviert und deren Kreativität fördert.

Aufgaben der Teilnehmer:

- Sammlung von Beiträgen zu einer bestimmten Problematik durch Kartenabfrage
- Gewichtung dieser Probleme
- Zusammenfassung der gewichteten Probleme zu Problembündeln.

Aufgaben des Moderators:

- Organisation des Ablaufs der Moderation
- Visualisierung der Sach- und Beziehungsprobleme in der Gruppe
- Sorge für Gleichberechtigung der Teilnehmer.

Ablauf der Moderation in **drei Phasen**:

1. **Einstieg:**

- Warming-up
- Entwicklung eines Problembewußtseins der Teilnehmer
- Sichtbarmachung von Interessen.

2. **Bearbeitung der Problematik:**

- Formulierung von Problemfragen
- Problemspeicherung
- Kleingruppenarbeit
- Vorstellen der Ergebnisse in der (Groß-)Gruppe
- Feedback durch die Teilnehmer selbst oder durch den Moderator.

3. Finale:

- Erstellen eines Tätigkeitskatalogs in der Gruppe oder in Kleingruppen
- Feststellen der Zufriedenheit und des Gruppenklimas durch den Moderator.

Anwendungsgebiete:

- in konflikträchtigen Situationen (z.B. Bildungsbedarfsanalyse)
- zur Erarbeitung neuer Problemstellungen.

Vorteile:

- Selbstverantwortlichkeit der Teilnehmer
- Anhäufung verschiedener Informationen, Meinungen, Ideen zu einer bestimmten Problematik.

Nachteile:

- hoher personeller und finanzieller Aufwand
- Zweifel, ob die reale Ungleichheit der Teilnehmer überhaupt ausgleichbar ist
- Einengung des Entscheidungsraums durch nicht-veränderliche Strukturen.

2. Nebenfolgenabschätzung

Bei jedem Evaluationsprojekt ist prinzipiell neben den gewünschten Effekten auch mit unintendierten Effekten zu rechnen. Diese betreffen sowohl die zu evaluierenden Maßnahmen als auch die Evaluationsstudie selbst, die ihrerseits unerwartete Konsequenzen haben kann, die nichts mit der eigentlichen Zielsetzung zu tun haben. Zu einer umfassenden Bewertung gehört, auch solche Nebenfolgen schon bei der Projektplanung mit zu berücksichtigen. Das rechtzeitige Entdecken potentieller Nebenwirkungen ist besonders schwierig, weil diese ja eben nicht zu den ursprünglich intendierten Maßnahmezielen gehören.

Eine nützliche Hilfe zur Identifizierung von möglichst vielen denkbaren Nebenfolgen im Sinn einer hypothetischen Formulierung ist die Berücksichtigung von **Handlungsplänen**. Für jede irgendwie von den Evaluationsobjekten oder der Evaluationsstudie selbst betroffene Personengruppe wird überlegt, in welcher Weise sich die Maßnahmen in diesen Handlungsplänen auswirken könnten. Dabei ist danach zu fragen, ob die Maßnahmen für die jeweiligen Personen

- ein (neues oder zusätzliches) Problem darstellen bzw. zur Folge haben
- die Mittel für die Bearbeitung bestehender Probleme verändern bzw. erweitern
- die Handlungsziele beeinflussen
- die Bewertung der Konsequenzen von Zielerreichungen verändern.

Beispiele:

Lehrer, die einen methodisch schlecht gestalteten Unterricht abhalten, könnten in der Verfügbarkeit didaktisch gut aufbereiteter Schulbücher ein Mittel zur partiellen Problemlösung sehen und daher darauf verzichten, ihren Unterricht zu verbessern → erwartbare Nebenfolge: Reduktion des Bestrebens der Lehrer, sich selbst optimal zu verhalten bzw. weiterzubilden.

Eine politische Partei hat sich seit Jahren vehement für eine bestimmte Schulorganisation eingesetzt; ein gegenteiliges Evaluationsergebnis könnte u.a. die Glaubwürdigkeit ihrer Aussagen und ihr Prestige herabsetzen, so daß dieses Ergebnis zu einem ‚Problem‘ für sie würde → erwartbare Nebenfolge: Maßnahmen der Partei zur Vermeidung ‚unerwünschter‘ Resultate.

Bei rechtzeitiger Berücksichtigung möglicher Nebeneffekte ist es möglich, die Leistungsfähigkeit des Projekts durch Erweiterung zu erhöhen. Mögliche Erweiterungen sind:

- **Ergänzung des Bewertungskriterienkatalogs**
- **methodische Vorkehrungen** (z.B. Vermeidung von direkter oder indirekter Selbstevaluation)
- **Empfehlung vorbereitender Maßnahmen**, insbesondere zur Vermeidung einer Einflußnahme auf die Ausgestaltung und Berichtlegung des Projekts (z.B. im Partei-Schulsystem-Beispiel).

Um zu vermeiden, daß die Ausarbeitung etwaiger Nebenfolgen durch die Beteiligten als „zynisch“ bezeichnet wird und die benannten Nebenwirkungen als unsachgemäße Unterstellun-

gen zurückgewiesen werden (z.B. fehlende Innovationsfreudigkeit bei Lehrern, irrationales Verhalten politischer Parteien), weil die erwartbaren Nebenfolgen gegen die sozial akzeptierten Normen der jeweiligen Gruppe verstoßen, sollte in solchen Fällen die Ideensammlung mit **anonymisierenden Methoden** (Metaplan oder auch vertrauliches Interview) durchgeführt werden. Der Evaluator sollte dann deutlich machen, daß diese Ideen nicht vom Projektteam selbst entwickelt wurden, sondern von praxiserfahrenen Außenstehenden.

3. Operationalisierungsfragen

Nach der Auswahl der theoretischen Bewertungskriterien der zu evaluierenden Maßnahmen müssen diese abstrakten Bewertungskriterien (und auch die hypothetisch formulierten Nebenfolgen) in konkreter Weise faßbar gemacht werden. Hier ist eine **Konsensfindung** mit dem Auftraggeber bzw. mit den relevanten Teilgruppen innerhalb des Auftraggebers wichtig, da ansonsten mit einer nachträglichen Abwertung der Meßinstrumente bei ‚unerwünschten‘ Ergebnissen gerechnet werden muß. Diese Operationalisierung bringt sowohl inhaltliche als auch methodische Probleme mit sich.

Inhaltliche Operationalisierungsprobleme:

Die Antwort auf die Frage, an welchen Beobachtungen das Ausmaß einer Verbesserung (z.B. „Lernerfolg“, „Therapieerfolg“) erfaßt werden kann, ist nicht Gegenstand einer empirischen Wissenschaft, sondern erfordert eine **geisteswissenschaftlich begründete Setzung**, etwa anhand von subjektiver Plausibilität, Verträglichkeit mit etablierten Ansätzen oder Nutzen-Überlegungen. Dies macht eine intensive Abstimmung mit dem Auftraggeber bezüglich dieser Setzung erforderlich, selbst wenn man auf ‚bewährte‘ Tests zurückgreift.

Methodische Operationalisierungsprobleme:

Die methodischen Probleme sind zwar fast ebenso schwierig, aber wissenschaftlich leichter zu bearbeiten. Die Wahl des methodischen Ansatzes sollte in jedem Evaluationsprojekt möglichst nach Sachaspekten getroffen werden, auch wenn vielfach die Berücksichtigung eines Mangels an Ressourcen nicht zu vermeiden ist. WOTTAWA & THIERAU gehen auf folgende Ansätze ein (vgl. S. 95 ff):

- ideographische Ansätze;
- nomothetische Ansätze, die Itemmengen sind definiert durch
 - Stoffgebiete
 - Konstruktionsregeln
 - eindimensionale probabilistische Modelle.

Ideographische Ansätze

Ideographische Ansätze der Operationalisierung der theoretischen Bewertungskriterien in Itemmengen werden vor allem dann verwendet, wenn die Evaluationsergebnisse stark in Abhängigkeit von Individuen bewertet werden müssen; es handelt sich also um eine **individuumorientierte Operationalisierung** und Datenerhebung.

Beispiel: Erfolgskontrolle von Psychotherapien → für jeden Klienten sind andere Interventionsziele bedeutsam, so daß die konkreten Bewertungskriterien für das Erreichen bestimmter individueller Interventionsziele auf jeden einzelnen Klienten zugeschnitten sein müssen.

Methoden:

- **Interviews** und **Fallbeispiele**: sind ‚weiche‘ Methoden, meist von hohem heuristischen Wert; ihre Ergebnisse können aber kaum verallgemeinert werden und hängen in erheblichem Maß von der subjektiven Voreinstellung des Untersuchers ab.
- **Struktur-lege-Technik** (SLT) (GROEBEN & SCHEELE, 1984) und **HYPAG/Structure** (WOTTAWA & ECHTERHOFF, 1982): sind stärker strukturierte Methoden, bei denen die (unbeabsichtigte) Einflußnahme des Untersuchers auf die Ergebnisse geringer ist; Ziel ist die Erhebung der für die jeweilige Fragestellung relevanten kognitiven Strukturen („subjektive Theorien“, „Entscheidungsregeln“) der Gesprächspartner.
- **Goal-Attainment-Scaling** (GAS) (WITTMANN, 1981,1985): Methode, die eine eindeutige Zusammenfassung der nur individuell operationalisierbaren Einzelbewertungen zu einer Gesamtbewertung ermöglicht. Diese Methode ist vor allem im Bereich der Therapie-Evaluation notwendig, wo es nicht nur um eine Betrachtung des Interventionserfolgs im Einzelfall geht, sondern auch um eine vergleichend-verallgemeinernde Aussage über die relative Bewährung verschiedener Therapiemethoden für spezielle Indikationsstellungen.

Nomothetische Ansätze

Bei nomothetischen Ansätzen der Operationalisierung der theoretischen Bewertungskriterien in Itemmengen geht es um eine **gesetzmäßige Operationalisierung** und Datenerhebung für alle betroffenen Personen in der gleichen Form. Daher stellt sich hier die Frage nach der Rechtfertigung einer bestimmten Indikatorenauswahl, denn das Evaluationsergebnis hängt wesentlich von der genauen Ausformung des Meßinstruments (der Itemmenge) ab.

Für die **Auswahl von Indikatoren** gibt es grundsätzlich **drei Vorgehensweisen**:

1. **Unsystematische Auswahl** einer größeren Anzahl von Einzelindikatoren (Items, Beobachtungen usw.), Konsensbildung über diese Einzelindikatoren und nachträgliche Zusammenfassung der Informationsmenge in Richtung auf einige besonders ‚wesentliche‘ konkrete Bewertungskriterien mit deskriptiven Verfahren (z.B. Faktorenanalyse).
2. **Anwendung bereits vorhandener möglichst gut konstruierter Test- oder Erhebungsverfahren**. Dieses Vorgehen spart Entwicklungsarbeit, erleichtert den Vergleich mit anderen Studien und ermöglicht das Delegieren der Verantwortung für eventuelle Schwächen der Messung an die Autoren. Wichtig ist es zu prüfen, ob die gegebene konkrete Operationalisierung auch tatsächlich einer Evaluationsfragestellung, die meist Veränderungsaspekte zum Gegenstand hat, gerecht wird.
3. **Spezifische Neukonstruktion der Meßinstrumente** entsprechend der besonderen Zielrichtung des Evaluationsprojekts. Dieses zeit- und kostenintensive Vorgehen sollte zumindest bei großen Projekten mit Verwendung der Meßinstrumente im Längsschnitt gewählt werden. Hierzu können folgende drei Ansätze wichtig werden: 1) Festlegung von Stoffgebieten mit Umsetzungsregeln, 2) Systematische Itemkonstruktion und 3) Eindimensionale probabilistische Modelle.

1) Festlegung von Stoffgebieten mit Umsetzungsregeln

Bei diesem **inhaltlich** orientierten Ansatz werden Stoffgebiete als operationalisierte Bewertungskriterien festgelegt, die dann in konkrete Meßinstrumente umgesetzt werden. Vor allem im pädagogisch-psychologischen Bereich liegen für manche Teilgebiete komplette Operationalisierungen der Bewertungskriterien (Stoffgebiete) vor, insbesondere in bezug auf kognitive Lerninhalte. Z.B. ergibt sich der Wissensstoff eines Schulfaches aus der Zusammenfassung aller dafür zugelassenen Lehrbücher oder die Kriterien für Schulreife aus der Sammlung aller für eine erfolgreiche Einschulung notwendigen Verhaltensweisen. Daher basiert die Testkonstruktion in diesem Bereich stark auf der **kriteriumsorientierten Messung** im Sinn einer sachgerechten Zusammenstellung von Itemsätzen aus einer das Kriterium definierenden Gesamtmenge. Ein solcherart gegebenes Stoffgebiet muß dann anhand eines objektiven Verfahrens mit festen Regeln in ein konkretes Meßinstrument umgesetzt werden.

2) Systematische Itemkonstruktion

Innerhalb dieses **inhaltlich** orientierten Ansatzes findet die Itemerstellung anhand expliziter Konstruktionsregeln statt. Hierfür sind zwei verwandte Denkansätze verbreitet:

Facettentheorie: Kombination verschiedener Aspekte der Aufgaben zu einzelnen „Facetten“.

Rationale/regelgeleitete Itemkonstruktion: Systematische Kombination kognitiver Prozesse, die für die Aufgabenlösung benötigt werden.

3) Eindimensionale probabilistische Modelle

Dieser an den **methodischen** Aspekten von Messung orientierte Ansatz bezieht sich auf die im Hinblick auf die Meßeigenschaften optimale Zusammenstellung der inhaltlich festgelegten Indikatoren. Da „eindimensional“ als eine besondere Definition von „ähnlich“ aufgefaßt werden kann, ist es im Prinzip möglich, für jedes einzelne ausgewählte Item durch Hinzufügen von entsprechend gleich strukturierten Items eine ganze Itemdimension zu erstellen. Die verbesserten Meßeigenschaften für Einzelpersonen (Anpassung der Itemschwierigkeit an den Leistungsstand verschiedener Subgruppen) sind besonders geeignet für die Messung von Entwicklungsverläufen einzelner Personen im Längsschnitt.

Frage 19: Beschreiben Sie anhand eines Beispiels die **MAUT-Technik** zum Zweck der **Nutzenbestimmung!** (1 x gefragt)

Bewertungs- und Entscheidungshilfen

Mit dem Erheben aller relevanten empirisch erfaßbaren Informationen, die die Bewertungskriterien darstellen, endet der wissenschaftliche Teil der Arbeit an Evaluationsobjekten. Es beginnt die **Bewertung** dieser sinnvoll gewählten empirisch-objektiven Fakten in Form einer Einschätzung des Nutzens der gefundenen Ausprägungsgrade der Bewertungskriterien und einer globalen Alternativenbewertung. Ziel ist im Fall einer summativen Evaluation bereits stattgefundener Programme eine Auswahl-Entscheidung über mehrere disjunkte Handlungsalternativen. Bei einer formativen Evaluation von noch in der Vorbereitungs- oder Implementierungsphase befindlichen oder laufenden Programmen ist das Ziel eine Verbesserungs-Entscheidung (Gestaltungsvorschlag). Für eine Bewertung ist die **Übersetzung der objektiven Fakten in subjektive Nutzenwerte** erforderlich. Da diese den Abnehmern der Evaluationsergebnisse oft sehr schwer fällt, gehört es auch zur Aufgabe des Evaluators, diesen Übersetzungsprozeß durch geeignete **Sozialtechniken** zu unterstützen. Es liegt eine Vielzahl solcher Techniken vor, die in **drei Bereiche** eingeteilt werden können: 1) Expizite Verfahren der Nutzenbestimmung, 2) Entscheidung durch Experten und 3) Entscheidung durch Betroffene.

Explizite Verfahren der Nutzenbestimmung

Für eine objektive, formalisierte Nutzenbestimmung sind folgende **drei Teilschritte** erforderlich:

1. Für jedes Evaluationsobjekt muß der Ausprägungsgrad auf den festgelegten Bewertungskriterien erhoben werden.
2. Für jeden empirisch gefundenen Ausprägungsgrad eines jeden Bewertungskriteriums muß der ‚Nutzen‘ festgestellt werden (Nutzenmessung).
3. Liegen mehrere Bewertungskriterien vor, muß bestimmt werden, wie die einzelnen Ergebnisse zu einem ‚Gesamtnutzen‘ zusammengefaßt werden können; dieser Vorgang erfordert subjektive Setzungen (Nutzenverrechnung).

Beispiel: Evaluationsstudie zu verschiedenen Schulsystemen

1. Feststellung der Ergebnisse (Ausprägungsgrade) der Schüler eines Schulsystems (ein Evaluationsobjekt) in einem Vokabeltest im Fach Englisch (eines der Bewertungskriterien).
2. Festlegung des Nutzens von verschiedenen Ergebnissen (Anzahl der Lösungen) im Vokabeltest, z.B. des Nutzens eines Ergebnisses von 20 richtigen Vokabeln im Vergleich zu einem Ergebnis von nur 15 richtigen Vokabeln.
3. Verrechnung der gefundenen Nutzenwerte auf z.B. den Bewertungskriterien „Vokabeltest Englisch“, „Test Mathematik“, „Wohlbefinden in der Klassengemeinschaft“ etc.

Schwierigkeiten treten nicht nur bei der Nutzenverrechnung auf (subjektive Setzungen), sondern auch dadurch, daß der ‚Nutzen‘ einer zu evaluierenden Alternative für verschiedene Gruppen von Betroffenen bzw. Entscheidern sehr unterschiedlich gesehen werden kann.

Nutzenmessung

Die Übersetzung einzelner Kriteriumsausprägungen in zugeordnete Nutzenwerte erfüllt zwei Funktionen:

- 1) Sie ermöglicht die Anwendung formalisierter Bewertungs- bzw. Entscheidungsverfahren vor allem bei Vorliegen ‚harter‘ Daten bzw. Skalen (Intervall- oder Rational-Skalen).
- 2) Sie trägt auch bei schwachen (Ordinal-)Skalen zur Problemexplikation bei und kann damit eine rationale, konsensbezogene Entscheidungsfindung auch ohne formalisierte Verrechnungsmethoden erleichtern.

Die Vorteile einer auch nur auf Ranginformationen aufbauenden Nutzenmessung werden deutlich, wenn man sich vor Augen führt, daß ein monotoner Zusammenhang zwischen Kriteriumsausprägung und Nutzen (monotone Nutzenfunktion) in keiner Weise selbstverständlich ist. So kann z.B. der Nutzen bei mittlerer Kriteriumsausprägung höher sein als bei hoher Ausprägung (Bsp. Bewertungskriterium „intellektuelle Leistungsfähigkeit“: ob ein Mensch mit einer Spitzenintelligenz für a) einfachste Berufstätigkeiten oder für b) Konzentration erfordernde Tätigkeiten wie Autofahren wirklich besser geeignet ist als ein Mensch mit durchschnittlicher Intelligenz, ist zumindest fraglich; im Fall a) wegen des subjektiven Anspruchsniveaus des Hochbegabten, das im Gegensatz zur Tätigkeit steht; im Fall b)

wegen der intensiven gedanklichen Beschäftigung mit einem Problem während der Tätigkeit).

Man unterscheidet 1) eindimensionale und 2) mehrdimensionale Verfahren zur Nutzenmessung. Mehrdimensionale Ansätze weisen für Evaluationsstudien größere praktische Relevanz auf, da sie die **Mehrdimensionalität des Nutzens von Handlungsalternativen** berücksichtigen. Ein für die mehrdimensionale Nutzenmessung besonders wichtiges Verfahren ist die **Multi-Attributive Nutzen-Technik** (MAUT-Technik), bei der auch direkt eine Nutzenverrechnung stattfindet.

Vorgehen der MAUT-Technik – Nutzenmessung und -verrechnung:

1. Identifizierung der Personen oder Organisationen, deren Nutzen zu maximieren ist.
2. Erarbeitung des Problembereichs, das heißt der Entscheidung, für wen die Nutzenmaximierung relevant ist.
3. Identifizierung der Alternativen, die in die Bewertung eingehen sollen.
4. Zusammentragen der relevanten Bewertungskriterien, anhand derer die Alternativen bewertet werden sollen.
5. Einordnung der relevanten Bewertungskriterien in eine Zielhierarchie.
6. Gewichtung der Bewertungskriterien.
7. Erstellung von Nutzenfunktionen für jedes einzelne Bewertungskriterium.
8. Feststellung der Ausprägungsgrade jeder zu bewertenden Alternative auf jeder Bewertungsdimension (gleicher Skalenbereich für alle Kriterien wichtig!).
9. Bestimmung des Gesamtnutzens jeder zu bewertenden Alternative nach folgender Formel: $u_i(A_j) = \text{Summe}(w_i \cdot u_i(x_{ij}))$.
 $u(A_j) = \text{Gesamtnutzen der Alternative } A_j$, $u(A_{ij}) = \text{Teilnutzen von } x_{ij}$, $x_{ij} = \text{Ausprägung der Alternative } A \text{ auf dem } i\text{-ten Kriterium}$, $w_i = \text{Gewicht des } i\text{-ten Kriteriums}$
(Vorher müssen alle Werte normiert werden.)
10. Entscheidung: Wahl derjenigen Alternative mit dem höchsten Nützlichkeitswert u_i .

Trotz eines numerischen Nutzenwertes darf nicht vergessen werden, daß es sich immer um **subjektive Setzungen** handelt.

Bei der MAUT-Technik findet bereits eine einfache **additive Nutzenverrechnung** statt (*lineare* Zusammenfassung von Teilnutzenwerten zu einer gewichteten Summe). Diese ist jedoch nicht immer sinnvoll, denn sie impliziert z.B., daß ein Evaluationsobjekt mit ausschließlich durchschnittlichen Nutzenwerten auf allen Bewertungskriterien den gleichen Gesamtnutzen erhält wie ein Evaluationsobjekt mit sehr hohen Nutzenwerten auf einigen Kriterien und sehr niedrigen auf anderen. Wenn die einzelnen Teilaspekte/Bewertungskriterien für verschiedene Betroffene unterschiedlich bedeutsam sind, ist diese *rechnerische Gleichheit* bei inhaltlich verschiedenen Sachverhalten problematisch. Ferner setzt die Bildung gewichteter Summen mindestens *Intervallskalenniveau* voraus, was meistens nicht sinnvoll angenommen werden kann.

Frage 20: Beschreiben Sie kurz **drei Methoden**, mit deren Hilfe der **Evaluator** zu Entscheidungen (Beurteilungen) über die untersuchte Maßnahme gelangt (**summatives Modell**)! (3 x gefragt)

Bewertungs- und Entscheidungshilfen

Mit dem Erheben aller relevanten empirisch erfaßbaren Informationen, die die Bewertungskriterien darstellen, endet der wissenschaftliche Teil der Arbeit an Evaluationsobjekten. Es beginnt die **Bewertung** dieser sinnvoll gewählten empirisch-objektiven Fakten in Form einer Einschätzung des Nutzens der gefundenen Ausprägungsgrade der Bewertungskriterien und einer globalen Alternativenbewertung. Ziel ist im Fall einer summativen Evaluation bereits stattgefundener Programme eine **Auswahl-Entscheidung** über mehrere disjunkte Handlungsalternativen. Für eine Bewertung ist die **Übersetzung der objektiven Fakten in subjektive Nutzenwerte** erforderlich. Da diese den Abnehmern der Evaluationsergebnisse oft sehr schwer fällt, gehört es auch zur Aufgabe des Evaluators, diesen Übersetzungsprozeß durch geeignete **Sozialtechniken** zu unterstützen. Es liegt eine Vielzahl solcher Techniken vor, die in **drei Bereiche** eingeteilt werden können: 1) Explizite Verfahren der Nutzenbestimmung, 2) Entscheidung durch Experten und 3) Entscheidung durch Betroffene.

1) Explizite Verfahren der Nutzenbestimmung

Nutzenmessung und -verrechnung:

Mehrdimensionale Ansätze weisen für Evaluationsstudien größere praktische Relevanz auf, da sie die **Mehrdimensionalität des Nutzens von Handlungsalternativen** berücksichtigen. Ein besonders wichtiges Verfahren ist die **Multi-Attributive Nutzen-Technik** (MAUT-Technik), bei der auch direkt eine Nutzenverrechnung stattfindet (siehe Frage 19!).

Bei der MAUT-Technik findet bereits eine einfache **additive Nutzenverrechnung** statt (*lineare* Zusammenfassung von Teilnutzenwerten zu einer gewichteten Summe). Diese ist jedoch nicht immer sinnvoll, denn sie impliziert z.B., daß ein Evaluationsobjekt mit ausschließlich durchschnittlichen Nutzenwerten auf allen Bewertungskriterien den gleichen Gesamtnutzen erhält wie ein Evaluationsobjekt mit sehr hohen Nutzenwerten auf einigen Kriterien und sehr niedrigen auf anderen. Wenn die einzelnen Teilaspekte/Bewertungskriterien für verschiedene Betroffene unterschiedlich bedeutsam sind, ist diese *rechnerische Gleichheit* bei inhaltlich verschiedenen Sachverhalten problematisch. Ferner setzt die Bildung gewichteter Summen mindestens *Intervallskalenniveau* voraus, was meistens nicht sinnvoll angenommen werden kann.

Screening

Da die Nutzenmessung, von Ausnahmen abgesehen, nur auf Ordinalskalen (Rangfolgen) erfolgen kann und insbesondere die Gewichtung der einzelnen Teilaspekte über Personen oder Personengruppen hinweg unterschiedlich ist, ist auf der Basis einer formalen Nutzenverrechnung oft nur ein „**screening**“ der Evaluationsobjekte möglich. Es dient einer möglichst einfachen **formalen Vorauswahl letztlich nicht relevanter Alternativen**. Zwei mögliche Screening-Verfahrensweisen sind 1) die Bildung der pareto-optimalen Teilmenge und 2) die Anwendung multipler cut-off-Strategien.

Bildung der pareto-optimalen Teilmenge:

Die Idee bei diesem Vorgehen ist, daß eine Alternative unabhängig von der speziellen Gewichtung oder Verrechnung immer dann einer anderen unterlegen ist und daher ausgeschlossen werden kann, wenn sie gleichzeitig in allen Bewertungskriterien niedrigere Nutzenwerte hat. Eine Gesamtmenge von alternativen Objekten wird nach diesem Prinzip vergleichend betrachtet und alle Objekte ausgeschlossen, die unterlegen sind. Es verbleiben mehrere alternative Objekte, die nicht weiter selektiert werden können und die pareto-optimale Teilmenge der Alternativen darstellen [pareto = teilweise schwach].

Dieses Vorgehen ist vor allem dann nützlich, wenn viele ‚Objekte‘ zu evaluieren sind und klare Nutzenverrechnungsregeln fehlen (z.B. Entscheidungen über Personal oder über Anbieter von Weiterbildungsseminaren). Es liefert eine formale Vorselektion der Objekte, die zu Recht in die engere Wahl kommen.

Anwendung multipler cut-off-Strategien:

Bei diesem Vorgehen werden auf den einzelnen Teilnutzenaspekten bestimmte Mindestwerte festgelegt, die überschritten werden müssen, damit ein ‚Objekt‘ in die engere Auswahl kommt. Es ist auch dann plausibel anzuwenden, wenn keine klare Zusammenfassung der einzelnen Nutzendimensionen vorliegt oder wenn nur Ordinalskalenniveau für die Nutzenmessung vorliegt. (Bsp.: Auswahl von Anbietern von Weiterbildungsprogrammen: es kommen nur solche in Frage, die für einzelne Teildimensionen wie Lernerfolg, Seminarklima oder Übertragbarkeit der Inhalte bestimmte Mindestwerte überschritten haben; Bsp. Auswahl von Personen in der Berufseignungsdiagnostik). Auch dieses Verfahren ermöglicht eine formale Vorselektion.

Um zu einer endgültigen Auswahl-Entscheidung zwischen den verbliebenen Alternativen zu kommen, muß im Anschluß an eine solche formale Vorselektion mit einem der beiden Screening-Verfahren eine weniger formalisierte Vorgehensweise angewandt werden, z.B. Expertengestützte Entscheidungsfindung oder Entscheidungsfindung durch Betroffene.

2) Expertengestützte Entscheidungsfindung

Die einfachste Art der Zusammenfassung unterschiedlicher Aspekte ist eine persönliche Einzelentscheidung oder eine Entscheidung in Form eines Gruppenkonsenses durch die für die Entscheidung zuständigen oder von ihr betroffenen Personen. Für kompliziertere, spezielle

Kenntnisse erfordernde Problemlagen ist es aber sinnvoll, zusätzlich das Wissen von Experten (z.B. Evaluatoren) einzubeziehen. Institutionalisierte Vorgehensweisen sind in der Praxis z.B. Enquête-Kommissionen, Anhörungen oder Gutachten. Sie sind aber nur bei der Darstellung wissenschaftlich unumstrittener Fakten und für unemotionale politische Fragestellungen unproblematisch.

Im sozialwissenschaftlichen Bereich sind die empirischen ‚Fakten‘ fast nie eindeutig im Hinblick auf ihre Bedeutung und Verursachung beschreibbar und Bewertungen immer subjektiv, so daß hier der Einsatz von Techniken wichtig ist, die Rollenverhalten einschränken und konsensbildend wirken – vor allem, wenn es um politisch-emotional sehr umstrittene Evaluationsfragestellungen geht. Für die Arbeit mit Experten hat sich die DELPHI-Methode sehr bewährt.

DELPHI-Methode

Definition: Eine spezielle Form der schriftlichen Befragung, mittels derer ein Kreis von Experten zu einem ausgewählten Problembereich in einem mehrstufigen Prozeß individuell und anonym befragt wird, so daß gruppendynamische Effekte ausgeschaltet werden können. Ein Leitungsgremium übernimmt dabei eine Koordinationsfunktion, indem es einen Katalog von Ausgangsfragen und Zielen entwickelt, die Antworten der Experten auswertet sowie einen ständig verbesserten Fragenkatalog ausarbeitet. Der Name bezieht sich auf das berühmte griechische „Orakel von Delphi“, das besonders weise Ratschläge gegeben haben soll.

Ziele:

- Gewinnung von neuen Ideen von Experten;
- Annäherung der Standpunkte durch ständige anonyme Rückmeldung über die Angaben der Expertenkollegen, so daß ein übereinstimmender Lösungsvorschlag für das behandelte Problem entwickelt werden kann.

Ablauf:

- Das Leitungsgremium erarbeitet für die anstehende Problematik einen speziellen Fragebogen.
- Ein ausgewähltes Expertenteam wird mit Hilfe des vorbereiteten Fragebogens um seine Meinung gebeten. Die schriftliche Befragung findet individuell und anonym statt.
- Die Ergebnisse der Umfrage werden durch das Leitungsteam ausgewertet (qualitativ und quantitativ).
- Auf der Basis der Resultate dieser ersten Befragung wird durch das Leitungsgremium ein neuer Fragenkatalog entworfen.
- In einer zweiten Befragungsrunde erhalten die einzelnen Experten den neuen Fragebogen zusammen mit den Ergebnissen der ersten Umfrage sowie zusätzlich anonymisierte Informationen über die Standpunkte und Lösungsbeiträge der anderen Experten. Die Experten werden dabei um eine Kommentierung ihrer Antwort im Vergleich zu den Gruppenergebnissen gebeten, wobei eine gewisse Angleichung der Ansichten erwartet wird.
- Weitere Auswertungen und Umfragen (meistens werden drei bis fünf Wiederholungsrunden nach diesem Ablaufschema durchlaufen). Dabei werden extreme Meinungen eliminiert und das Schwergewicht auf strittige Punkte gelegt.
- Liegt eine Ideensammlung von größerer Aussagefähigkeit vor, erarbeitet das Leitungsteam schließlich einen umfassenden Lösungsvorschlag für das relevante Problem.

Vorteil:

Durch die Vermeidung von gruppendynamischen Effekten fällt es den Experten leichter, die von anderen vorgebrachten Argumente ohne Emotionen zu prüfen und die eigene ursprüngliche Meinung als Ergebnis solcher zusätzlicher Überlegungen ohne Gesichtsverlust zu ändern.

Nachteile:

- Das Verfahren ist sehr kostenaufwendig.
- Das Verfahren widerspricht dem Selbstverständnis vieler Wissenschaftler (wer von ihnen gibt schon gern zu, daß er auch in der Rolle als Sachverständiger für bestimmte Themen dazu neigt oder neigen könnte, auch sachfremde Einflüsse wie Emotionen in seine Aussagen einfließen zu lassen?).

3) Entscheidungsfindung durch Betroffene

Eigentlich sind die Betroffenen (aus wissenschaftlicher Sicht meist Laien) am ehesten befugt, bei Evaluationsfragestellungen eine Entscheidung zu treffen. Jedoch verfügen sie in der Regel nicht über die erforderlichen relativ weitgehenden Sachkenntnisse über die Grundlagen und Nebenfolgen für eine begründete Auswahl-Entscheidung (bzw. im Fall der formativen Evaluation für eine begründete Gestaltungsentscheidung). Außerdem neigen Betroffene dazu, ihre persönlichen Nutzenaspekte besonders hoch zu veranschlagen und evtl. sogar massive Nachteile für andere als weniger gravierend einzuschätzen. Um diese Störungen auszuschalten, sind folgende **vier Anforderungen** zu gewährleisten:

- 1) intensiver Kontakt zwischen verschiedenen Betroffenen mit unterschiedlichen Interessen;
- 2) sorgfältige Information dieser ‚Entscheider‘;
- 3) konsensfördernde Diskussionsgestaltung;
- 4) repräsentative Auswahl der ‚Entscheider‘ aus der Grundgesamtheit der Betroffenen.

Ein für die Lösung dieser Probleme hervorragend geeigneter Ansatz ist die Planungszelle.

Planungszelle

Prinzip: Das Prinzip liegt in der Zusammenfassung einer größeren Anzahl von Betroffenen in mehreren Kleingruppen. Diese erhalten sorgfältige und umfassende Informationen durch Fachexperten. In einer Diskussion der Informationen in der Kleingruppe wird abschließend eine Bewertung des Problems bzw. eine Beschlussfassung festgelegt.

Definition: Eine Gruppe von Bürgern, die nach einem Zufallsverfahren ausgewählt und für begrenzte Zeit von ihren arbeitstäglichen Verpflichtungen vergütet freigestellt worden sind, um, assistiert von Prozeßbegleitern, Lösungen für vorgegebene, lösbare Planungsprobleme zu erarbeiten.

Merkmale:

1. unerläßliche konstruktive Merkmale

- Gruppenentscheid
- akzeptable Rollenzuordnung für alle Teilnehmer
- Freistellung der Teilnehmer von Arbeits- und Familienverpflichtungen
- vergütete Teilnahme
- befristete Teilnahme
- Zufallsauswahl der Teilnehmer
- Laien als Teilnehmer
- fachliche Begleitung
- vorgegebene Aufgabenstellung
- Freizügigkeit des Einsatzes
- simultane Anwendbarkeit durch andere Gruppen

2. variable Merkmale

- Teilnehmerzahl (meistens 25 Personen)
- Dauer (meistens drei Wochen)
- Programmdichte (Einflußmöglichkeiten auf den Programmablauf)

Das Verfahren ist besonders dann angemessen, wenn die ‚wissenschaftlichen‘ Grundlagen für die Entscheidung entweder weniger wichtig oder leicht verständlich sind (z.B. antizipatorische Evaluation von Stadtplanungsmaßnahmen). Bei entsprechender Modifikation wäre es sicher auch sehr gut geeignet, um in anderen Bereichen eine wirkliche Beteiligung der Betroffenen an (politischen) Entscheidungen zu ermöglichen, deren Qualität weit über die in manchen Bereichen gesetzlich vorgesehenen Anhörungen hinausgeht.

Vorteil:

Ermöglicht demokratische Entscheidungen durch direkte Beteiligung von informierten Betroffenen (Entscheidungsdelegation an „mündige Bürger“).

Nachteile:

- Das Verfahren ist sehr kostenaufwendig.

- Die Entscheidungsdelegation bedeutet subjektiv und objektiv eine Entmachtung der (politischen Mandats- und) Entscheidungsträger, was deren Selbstverständnis widerspricht (und daher werden Planungszellen so selten angewendet!).

Frage 21: Skizzieren Sie kurz die **Netzplantechnik** bei der **Planung** von **Evaluationsstudien!** (2 x gefragt)

Bei der **Planung** von Evaluationsstudien ist eine **Zeit- und Kostenabschätzung** erforderlich.

Das Angebot eines Evaluators an den potentiellen Auftraggeber schließt eine Kalkulation der benötigten Zeit für das Projekt und der anfallenden Kosten mit ein. Um Aussagen über diese beiden Punkte machen zu können, wird das Projektvorhaben zunächst in Teilschritte (Vorgänge/Arbeitsschritte: Zeit erfordernde Geschehnisse mit definiertem Anfang und Ende; Ereignisse: Eintreten eines definierten Zustands im Ablauf des Projekts; Anordnungsbeziehungen: quantifizierbare Abhängigkeiten zwischen den Vorgängen, deren Gesamtheit die Ablaufstruktur des Projektes bildet) zergliedert (Strukturanalyse). Anschließend wird ermittelt, wieviel Zeit für die unterschiedlichen Teilschritte benötigt wird (Zeitanalyse) und welche Kosten für die einzelnen Schritte zu erwarten sind (Kostenabschätzung).

Zeitanalyse

Es gibt zwei Möglichkeiten, die erforderliche Dauer der einzelnen Vorgänge zu ermitteln. Die erste Möglichkeit besteht in der Abschätzung möglichst plausibel erscheinender Zeitintervalle („**stochastisches Konzept**“ mit wahrscheinlichen Zeiten). Die zweite Möglichkeit besteht im Rückgriff auf die Erfahrungswerte aus früheren Projekten („**deterministisches Konzept**“ mit fest angenommenen Zeiten).

Auf dieser groben Abschätzung bauen dann die spezifischen **Zeitplanungstechniken** auf. Zwei Techniken, die eine übersichtliche graphische Darstellung der Teilschritte des Projektes und deren Dauer in Form von Diagrammen ermöglichen, sind 1) die Balkenplantechnik und 2) die Netzplantechnik.

Netzplantechnik

Die Netzplantechnik ist auch für komplizierte Projekte mit vielfältig abhängigen Vorgängen geeignet (wenn z.B. die Erledigung von drei verschiedenen Vorarbeiten mit jeweils unterschiedlicher Zeitdauer die Voraussetzung für den Beginn des vierten Projektschritts ist). Mit dieser Technik werden alle Vorgänge und Ereignisse des Projektes logisch miteinander verknüpft.

Fünf Arbeitsschritte/Phasen:

Phase 1: Strukturanalyse: Ereignisse, Vorgänge/Arbeitsschritte

Phase 2: Zeitanalyse: Dauer (Kosten) der Vorgänge/Arbeitsschritte

- a) bei bekannter Zeit-Kosten-Variable Bestimmung des „kritischen Weges“ aufgrund von Erfahrungswerten (deterministisch)
- b) bei unbekannter Zeit-Kosten-Variable Netzplanberechnung aufgrund von Schätzwerten (stochastisch)

Phase 3: Erstellung des Netzplans

Bestandteile zur Konstruktion eines Netzplans:

a) Elemente zur Darstellung strukturanalytischer Ergebnisse:

- Ereignisse = „Knoten“ = ٢
- (reale) Vorgänge/Arbeitsschritte = Pfeile = →
- fiktive Vorgänge/Arbeitsschritte, die eingeführt werden, um den Nebenbedingungen zu genügen = Scheinvorgänge = ----→

b) Darstellung zeitanalytischer Werte:

- Zeitangaben (deterministisch und/oder stochastisch)
- die Zeitvariablen werden nur den realen Vorgängen zugeordnet; Scheinvorgänge haben grundsätzlich die Zeitdauer null.

Grundsätze/Nebenbedingungen:

1. Alle Vorgänge und Ereignisse müssen genau definiert sein.

2. Das Netzwerk hat nur einen Start- und einen Endpunkt.
3. Das Netz ist lückenlos verknüpft, d.h., jedes Ereignis ist über eine Kette von Vorgängen mit dem Start und dem Ende verbunden.
4. Zwei Ereignisse dürfen nur durch einen einzigen Vorgang miteinander verbunden sein.
5. Jedem Vorgang ist eine Zeitvariable zugeordnet (bei dieser Methode sagt die Länge des Pfeils nichts über den Zeitbedarf aus!).
6. Das Netz muß schleifenfrei verlaufen.
7. Parallel verlaufende Vorgänge werden durch „Scheinvorgänge“ miteinander verbunden (damit Bedingung 4 nicht verletzt wird).

Phase 4: Graphische Darstellung des Netzplans

Phase 5: Netzplanberechnung: Unter verschiedenen Wegen vom Anfangs- bis zum Endpunkt des Netzes gibt es einen Weg von kürzester Zeitdauer und einen Weg von längster Zeitdauer. Diese beiden Wege (sog. „kritische Wege“) bestimmen den frühestmöglichen bzw. den ungünstigsten (spätesten) Zeitpunkt für das Projektende.

Berechnung des frühestmöglichen Projektendes: Wenn man alle (deterministischen) Zeitvariablen, die den Vorgängen des „kritischen Weges“ (hier der Weg der kürzesten Zeitdauer) zugeordnet sind, addiert, repräsentiert die Summe der Zeitvariablen den frühestmöglichen Endzeitpunkt. Genauer: Zunächst wird die Dauer aller möglichen Wege vom Startpunkt zum Endpunkt des Netzes berechnet, indem jeweils alle (deterministischen) Zeitvariablen, die den Vorgängen jedes Weges zugeordnet sind, addiert werden. Aus den verschiedenen Summenwerten für jeden möglichen Weg wird dann der Wert von kleinster numerischer Größe ausgewählt, der die Dauer des „kritischen Weges“ im Sinn des Weges von kürzester Zeitdauer re-präsentiert.

Berechnung des ungünstigsten (spätesten) Projektendes: Man addiert jeweils den maximalen stochastischen Wert aller Vorgänge, die auf dem „kritischen Weg“ (hier der Weg von längster Zeitdauer) liegen. Der sich ergebende Summenwert repräsentiert die Dauer des „kritischen Weges“ im Sinn des Weges von längster Zeitdauer.

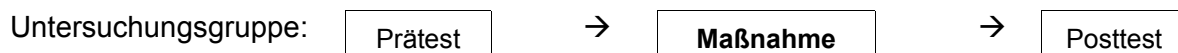
Vorteile/besondere Stärken:

- Übersichtliche Darstellung termingebundener Projekte.
- Geringer Rechenaufwand.
- Der Planer wird gezwungen, alle Projekte zusammenhängend gründlich zu durchdenken.
- Realistische Festlegung von Terminen.
- Potentielle Engpässe/Störungen können klar erkannt werden, da der Netzplan eine systematische und lückenlose Darstellung der zwischen den Vorgängen bestehenden Zusammenhänge ermöglicht.
- Zugänglich für EDV-gestützte Optimierungsabschätzungen (vor allem notwendig bei wirklich komplexen langfristigen formativen Evaluationen).

Frage 22: Welche **Vor-** und **Nachteile** bringt der **Prä-Posttest-Kontrollgruppen-Plan** im Rahmen eines Evaluationsprojekts mit sich? (1 x gefragt)

Wenn Veränderung aufgefaßt wird als der Unterschied in einer Variablen (AV) zu zwei Zeitpunkten, da eine einmalige Beobachtung einer Variablen keine Aussage über ihre Veränderung zuläßt, dann entspricht diesem Verständnis der **Prätest-Posttest-Plan** mit mindestens zwei Meßzeitpunkten. Man kann weder allein aus dem Prätest noch allein aus dem Posttest Aussagen zu Veränderungen machen, sondern erst durch die gleichzeitige Berücksichtigung beider Meßzeitpunkte ist Veränderung zu erschließen. Damit wird eine UV (B, Meßzeitpunkt) mit zwei Stufen eingeführt (UV B₁: vorher; UV B₂: nachher), zwischen welche eine Intervention tritt. Hervorgerufen wird die Veränderung durch eine Intervention (UV A), die im Fall der Evaluationsforschung durch eine Maßnahme repräsentiert wird.

Schema:



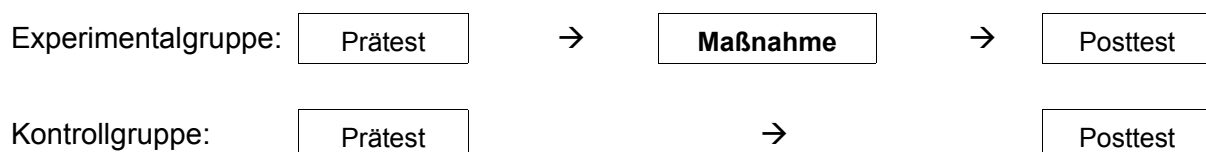
Bei diesem Design wird eine repräsentative Stichprobe der interessierenden Zielpopulation einmal vor und einmal nach der Intervention untersucht. Die durchschnittliche Differenz auf der abhängigen Variablen (**Vergleich von Vortestwerten und Nachtestwerten der Experimentalgruppe**) gilt behelfsweise als Indikator für die Wirkung der Intervention. Die *interne Validität* (Eindeutigkeit der Ergebnisse) dieses Designs ist jedoch *gering*, da alle möglichen zeitabhängigen Störeinflüsse die Veränderung (bzw. Nichtveränderung) (mit-)bewirkt haben können. Ferner treten durch die Meßwiederholung (abhängige Messungen) *statistische Regressionseffekte* auf. Daher sollte dieser Versuchsplan nur bei Evaluationsfragestellungen eingesetzt werden, bei denen eine Maßnahme evaluiert werden soll, von der praktisch alle Personen betroffen sind, so daß keine Kontrollgruppe gebildet werden kann oder bei Fragestellungen, bei denen aus ethischen Gründen die Bildung einer Kontrollgruppe nicht möglich ist. Als Signifikanztest verwendet man bei zwei Messungen z.B. den t-Test für abhängige Stichproben und bei mehr als zwei Messungen die einfaktorielle Varianzanalyse mit Meßwiederholungen.

Beispiel für ein Evaluationsprojekt:

EIH: Personen, die an einem Zeitmanagementtraining teilnehmen, kommen danach (MZP₂) mit ihrer Arbeitszeit besser zurecht als zuvor (MZP₁). SV: $\mu_1 < \mu_2$. TH: $H_1: \mu_1 < \mu_2$; $H_0: \mu_1 = \mu_2$. Es wird deutlich, daß die Hypothese als Kausalaussage formuliert ist (die Intervention verursacht die Veränderung); außerdem ist sie gerichtet (Veränderung in Richtung einer Verbesserung). Folglich ist die H_0 mit einem t-Test oder einem a priori-Kontrast zu prüfen. Schließlich liegen aufgrund der Meßwiederholung abhängige Stichproben vor. Es handelt sich also um ein **Quasi-Experiment durch Meßwiederholung (VPL1Q(W))**. Der Versuchsplan ermöglicht jedoch *keine Kausalaussage* in dem Sinn, daß im Fall einer festgestellten Verbesserung im Umgang mit der Arbeitszeit zum zweiten Meßzeitpunkt diese Verbesserung eindeutig auf die Teilnahme an dem Zeitmanagementtraining zurückzuführen ist. Die Veränderung könnte auch durch zeitabhängige Störeinflüsse wie Reifungsprozesse (vgl. WOTTAWA & THIERAU, S. 125) und Sequenzeffekte (Positionseffekte, Übertragungseffekte, Effekte durch das zwischenzeitliche Geschehen) oder statistische Regressionseffekte bewirkt worden sein. Die Kontrolle der Sequenzeffekte durch vollständiges interindividuelles Ausbalancieren ist hier aufgrund des Aufbaus des Versuchsplans nicht möglich. Deshalb muß nach Möglichkeiten gesucht werden, andere Kontrollmaßnahmen zu installieren. Eine Möglichkeit besteht in der **Einführung einer Kontrollgruppe**.

Prätest-Posttest-Kontrollgruppen-Plan

Die **interne Validität** von quasi-experimentellen Evaluationsuntersuchungen (ohne Randomisierung) mit Meßwiederholung läßt sich dadurch erhöhen, daß neben der Experimentalgruppe auch eine **Kontrollgruppe** geprüft wird (Prätest-Posttest-Kontrollgruppen-Design, Zwei-Gruppen-Prätest-Posttest-Plan). Dieses Untersuchungsdesign ermöglicht die **Kontrolle von zeitabhängigen Störeinflüssen** (externe zeitliche Einflüsse, Reifungsprozesse, Testübung, vgl. BORTZ & DÖRING, S. 522). So läßt sich die „wahre“ Veränderung zwischen zwei Meßzeitpunkten direkter untersuchen als durch eine nachträgliche statistische Korrektur der beobachteten Differenzwerte in einem einfachen Prätest-Posttest-Design/Ein-Gruppen-Prätest-Posttest-Plan (durch Regressions- oder Residualmaße, s.o.). Der einfachste Plan umfaßt eine Experimentalgruppe (UV A₁: mit Maßnahme) und eine Kontrollgruppe (UV A₂: ohne Maßnahme), die beide jeweils zweimal untersucht werden. Die beiden Gruppen/Stichproben sollen sich lediglich darin unterscheiden, daß in der Experimentalgruppe zwischen Vortest und Nachtest eine Maßnahme durchgeführt wird, während die Kontrollgruppe ausschließlich dem Vortest und dem Nachtest unterzogen wird. Sowohl die Prätestwerte der beiden Versuchsgruppen als auch alle denkbaren Einflüsse zwischen Erst- und Zweitmessung, abgesehen von der Intervention, müssen dabei gleich sein. Schematisch (vgl. TRAUTNER, S. 285):



Der große **Vorteil** des Prätest-Posttest-Kontrollgruppen-Plans gegenüber dem reinen Prätest-Posttest-Plan ist, daß man zur Abschätzung der „wahren“ Veränderung aufgrund der Maßnahme auf die Berechnung von Differenz- oder Veränderungswerten zwischen den beiden Meßzeitpunkten verzichten kann und stattdessen die Nachtestwerte der beiden Versuchsgruppen direkt miteinander vergleichen kann. **Aus dem Unterschied der Nachtestwerte der beiden Versuchsgruppen läßt sich der Effekt der Maßnahme unmittelbar ablesen** (vgl. TRAUTNER, S. 285). So läßt sich die „wahre“ Veränderung zwischen zwei Meßzeitpunkten direkter untersuchen als durch eine nachträgliche statistische Korrektur der beobachteten Differenzwerte in einem einfachen Prätest-Posttest-Design/Ein-Gruppen-Prätest-Posttest-Plan (durch Regressions- oder Residualmaße, s.o.). Der Vergleich der Nachtestwerte der beiden Gruppen kann mit einem t-Test für unabhängige Stichproben vorgenommen werden; dabei geht jede Vp nur mit einer Messung (Nachtest) in die Prüfung ein, so daß das Problem der statistischen Regression bei abhängigen Messungen vermieden wird. Die **interne Validität** ist hier durch die Kontrolle von zeitabhängigen Störeinflüssen **höher**.

Schematische vergleichende Darstellung:

Prätest-Posttest-Plan:

	Prätest (t_1)	Posttest (t_2)
Experimentalgruppe (E)	AV _{E1} ←	→ AV _{E2}

Prätest-Posttest-Kontrollgruppen-Plan:

	Prätest (t_1)	Posttest (t_2)
Experimentalgruppe (E)	AV _{E1}	AV _{E2}
Kontrollgruppe (K)	AV _{K1}	AV _{K2}

Ein **Nachteil** des Prätest-Posttest-Kontrollgruppen-Plans bei quasi-experimentellen Evaluationsuntersuchungen (ohne Randomisierung) ist, daß der **Regressionseffekt nur dann kontrollierbar** ist, wenn die **Prätestwerte** (Mittelwerte und Streuungen) der Experimental- und der Kontrollgruppe **vergleichbar** sind (vgl. PETERMANN, S. 30, S. 57). Die Durchführung von Vortests der abhängigen Variable ist unerlässlich, um Aussagen über die Veränderung in der Experimentalgruppe machen zu können. Die Vortests haben hier die Funktion, Ausgangsunterschiede zwischen der Experimental- und der Kontrollgruppe in der abhängigen Variable zu Beginn der Evaluationsuntersuchung festzustellen. Die stichprobenspezifischen „Startbedingungen“ sind die Referenzdaten, auf die sich maßnahmebedingte Veränderungen beziehen (vgl. BORTZ & DÖRING, S. 515). Nur wenn die beiden Gruppen sich im Vortest in ihren Ausgangswerten (Mittelwerte und Streuungen) nicht deutlich unterscheiden (Vergleich von A₁B₁ mit A₂B₁ z. B. mit einem t-Test für unabhängige Stichproben), kann der statistische Regressionseffekt kontrolliert (nach PETERMANN, S. 30 aber nicht eliminiert) werden und ist die **interne Validität** so akzeptabel, daß ein möglicher Gruppenunterschied im Posttest (Vergleich von A₁B₂ mit A₂B₂) eindeutig kausal auf die Maßnahme zurückgeführt werden kann (vgl. PETERMANN, S. 30).

Versuchsplan:

		UV B: Meßzeitpunkt	
		B ₁ vorher	B ₂ nachher
UV A: Maßnahme	A ₁ mit	AV: A ₁ B ₁	AV: A ₁ B ₂
	A ₂ ohne	AV: A ₂ B ₁	AV: A ₂ B ₂

Hussy: Wird bei einem Prätest-Posttest-Kontrollgruppen-Design die Auswertung mit Hilfe der **Kovarianzanalyse** vorgenommen, müssen die Ausgangsmittelwerte der Experimental- und der Kontrollgruppe **nur ungefähr vergleichbar** sein; kleinere Abweichungen stören nicht, da die Kovarianzanalyse ja gerade dem Herausrechnen von Mittelwertsunterschieden dient!

Kombination von Prätest-Posttest-Kontrollgruppen-Design und Kovarianzanalyse

Besteht ein Unterschied zwischen der Experimental- und der Kontrollgruppe hinsichtlich ihrer Ausgangswerte im Prätest, kann die Veränderung der Werte der Experimentalgruppe beim Posttest (auch) auf die Ausgangsunterschiede und nicht (allein) auf die Maßnahme zurückgeführt werden. Ferner besteht bei vorliegendem Ausgangsunterschied die Gefahr eines statistischen Regressionseffekts, der sich darin äußern würde, daß sich eine hohe Differenz im Prätest im Posttest verringert. Zur **Kontrolle des Ausgangsunterschieds** bzw. zur **Eliminierung des Regressionseffekts** sollte daher bei vorliegendem Ausgangsunterschied eine **Kovarianzanalyse** durchgeführt werden.

In der kovarianzanalytischen Auswertung der Posttestwerte der beiden Gruppen werden die Prätestwerte der beiden Gruppen als Kontrollvariable einbezogen. **Aus der abhängigen Variable (AV) im Prätest wird dadurch eine Kontrollvariable/Kovariate (KV) im Posttest** (identischer Wert). Der Einfluß der Prätestwerte (KV) auf die Posttestwerte (AV) ist somit statistisch kontrollierbar. Aus dem eigentlich zweifaktoriellen Design (UV A Maßnahme mit den beiden Stufen „mit“ und „ohne“, UV B Meßzeitpunkt mit den beiden Stufen „vorher“ und „nachher“) wird so ein einfaktorielles (Auswertungs-)Design: man beschränkt sich auf die wesentliche UV A (mit oder ohne Maßnahme). Jede Vp geht nur mit einer Messung der abhängigen Variable in die (Posttest-) Prüfung ein, da die AV nur einfach verwertet wird, so daß ein statistischer Regressionseffekt vermieden wird. Der Unterschied von Experimental- und Kontrollgruppe im Posttest ist dann von möglichen Ausgangsunterschieden bereinigt, so daß die **interne Validität höher** ist und eine **eindeutige Aussage über eine tatsächliche Veränderung aufgrund der Maßnahme** möglich wird.

Auswertungsplan:

		UV B: Meßzeitpunkt		Haupteffekt (HE A)
		B ₁ vorher	B ₂ nachher	
UV A: Maßnahme	A ₁ mit	KV: A ₁ B ₁	AV: A ₁ B ₂	
	A ₂ ohne	KV: A ₂ B ₁	AV: A ₂ B ₂	

Frage 23: Geben Sie ein Beispiel für die **a priori- inhaltliche Bestimmung** der **Effektgröße** aus dem Projekt „Schule zum Anfassen“ mit den **Konsequenzen** für den Versuchsplan! (1 x gefragt)

Neben der statistischen Bestimmung der Effektgröße kann man diese auch im voraus (a priori) inhaltlich im Sinn eines **Nutzenkriteriums** definieren. Bei dem Projekt „Schule zum Anfassen“ (SZA) ist dies z.B. für die abhängige Variable „Planungstiefe“ (PT, Antizipation der Handlungsschritte: wie weit denken die Kinder voraus?) durch folgende Forderung realisiert worden:

„Die mittlere Ausprägung der Planungstiefe in der Experimentalgruppe ($\mu_{PT EG}$) soll der Leistung von gleichaltrigen Schülern entsprechen, die nicht aus sozialen Brennpunktgebieten stammen.“

Vom Untersuchungsdesign aus gesehen wäre diese Forderung durch eine **weitere Kontrollgruppe** zu überprüfen: eine Stichprobe gleichaltriger Schüler ohne Förderunterricht aus einer Schule in ‚normaler‘ Umgebung. (vgl. Handout **Hussy** „Ergänzungen zum Evaluationsprojekt „Schule zum Anfassen“ (SZA))

Frage 24: Nennen Sie **folgeschwere Fehler** bei der **Durchführung** von **Evaluationsprojekten**, und erläutern Sie drei davon anhand eines Beispiels! (1 x gefragt)

Die **Durchführung größerer Evaluationsprojekte mit längerer Laufzeit** ist nur möglich mit einer regelmäßigen Rückmeldung des Evaluators über die Einhaltung 1) der Zeitplanung, 2) des Kostenrahmens und 3) der vereinbarten Qualitätsstandards. Für einen ‚nur sozialwissenschaftlich‘ ausgebildeten Projektleiter stellt insbesondere die sachgerechte **Kontrolle des Kostenrahmens** eine Schwierigkeit dar. Eine regelmäßige (evtl. wöchentliche) Gegenüberstellung der bis zum jeweiligen Arbeitsschritt geplanten Ausgaben, der durch Verträge

eingegangenen Zahlungsverpflichtungen und den bereits tatsächlich verausgabten Beträgen sollte selbstverständlich sein.

Jedoch kommt es gar nicht selten sogar innerhalb des jeweiligen Projektbudgets, also ohne Berücksichtigung der zugeschlüsselten Gemeinkosten, zu folgenden **Fehlern**:

- **Keine Berücksichtigung der Lohnnebenkosten**
- **Kalkulation der Projektarbeit auf der Basis eines 52-Wochen-Jahres**: Der deutsche Arbeitnehmer ist wegen Urlaubs- und anderen Ausfallzeiten durchschnittlich nur 43 Wochen im Jahr produktiv tätig. Werden z.B. Urlaubszeiten bei der Planung nicht berücksichtigt, so kann dies zu einem ganz erheblichen Kostenfaktor werden (Ablösung des Urlaubsanspruchs bzw., sofern überhaupt möglich, die Bezahlung von Ersatzkräften).
- **Keine Reserven für die Überbrückung von Ausfällen**: Erhebliche Probleme entstehen z.B. bei mangelnden Reserven im Fall von vorzeitigen Kündigungen, Mutterschaftsurlaub oder längeren Erkrankungen.
- **Keine rechtzeitige Einplanung von Lohn- und Preissteigerungen**: Auch die mangelnde Einplanung von Lohn- und Preissteigerungen, z.B. bedingt durch die Inflation, kann die Kontrolle des Kostenrahmens schwierig machen.

Besondere Probleme können auftreten, wenn mit dem Auftraggeber kein Festpreis, sondern zumindest in Teilen der Ersatz der tatsächlichen Aufwendungen vereinbart wurde. Ein folgenschwerer Fehler kann hier die **Mißachtung der Aufwandsentschädigungsbestimmungen des Auftraggebers** sein (z.B. Höchstbeträge für Übernachtungs- und Verpflegungsspesen, Kilometergeld, anzurechnendes Stundenhonorar). Diese Bestimmungen müssen zusätzlich zu der eigenen Kalkulation berücksichtigt werden. Die Vereinbarung von Aufwandsentschädigungen kann z.B. erforderlich sein, wenn sich die Kosten mancher Arbeitsschritte in der Planungsphase nicht realistisch einschätzen lassen (z.B. für das teilweise mühevoll und mit Dienstreisen verbundene Einholen von Zustimmungen durch Eltern, Lehrer und Schüler bei Schuluntersuchungen) oder wesentlich von dem späteren Verhalten des Auftraggebers selbst abhängen.

Frage 25: Was versteht man unter „Meta-Analyse“, und nach welchen **Kriterien** vollzieht sich die **Auswahl** der verwendeten Untersuchungen (jeweils kurze Beschreibung)? (2 x gefragt)

Definition:

Unter Meta-Analyse wird eine Gruppe von Verfahren verstanden, mit denen die **Ergebnisse verschiedener Untersuchungen mit gemeinsamer Thematik zusammengefaßt werden**, um so einen **Überblick über den aktuellen Stand der Forschung** zu gewinnen.

Es gibt bisher keine einheitliche metaanalytische Methodik, sondern die Meta-Analyse befindet sich in einem Entwicklungsprozeß, dessen Ende noch nicht absehbar ist. Für eine Vereinheitlichung im Sinn eines allgemein akzeptierten methodischen Vorgehens ist eine Norm für eine verbindliche Art der Ergebnisdarstellung in Publikationen erforderlich (Umfang und Genauigkeit der Informationen, statistische Ergebnisindikatoren).

Zielsetzung:

Das vorrangige Ziel aller metaanalytischen Techniken besteht in der **statistischen Aggregation der Einzelergebnisse von inhaltlich homogenen Primäruntersuchungen**. Von besonderer Bedeutung ist hierbei die Frage nach der **Wirksamkeit** einer Intervention (Maßnahme oder Behandlung), deren Beantwortung alle einschlägigen Forschungsergebnisse berücksichtigen soll. Damit ist die **Effektgröße** bzw. der durch viele einzelne Untersuchungen geschätzte „**wahre**“ **Effekt** einer Intervention ein zentraler Begriff der Meta-Analyse.

Nach BORTZ & DÖRING sind Meta-Analysen auch zur **Vorbereitung größerer (summativer) Evaluationsprojekte** wichtig. Bereits vorhandene bzw. vom Evaluator selbst durchgeführte Meta-Analysen erleichtern die Formulierung einer spezifischen Hypothese unter Vorgabe einer bestimmten, unter Kosten-Nutzen-Aspekten festgelegten Effektgröße für die Wirkung der zu evaluierenden Maßnahme (praktisch bedeutsamer Effekt) (vgl. S. 590). BORTZ & DÖRING betonen, daß die summative Evaluation idealerweise am Ende eines

Forschungsprozesses steht, der mit der Grundlagenforschung beginnt und über die Interventionsforschung zu einer konkreten Maßnahme führt, die einer abschließenden Evaluation unterzogen wird. Daher müßten die Themen der summativen Evaluationsforschung genügend elaboriert sein, um spezifische Hypothesen mit Effektgrößen formulieren zu können. Ferner ist es auch im Interesse des Auftraggebers zu erfahren, ob die von ihm finanzierte Maßnahme einen praktisch bedeutsamen Effekt erzielt (statt nur die Information zu erhalten, daß die Maßnahme ‚irgendwie‘ wirkt) (vgl. S. 111).

Kriterien für die Selektion der metaanalytisch zusammenfassenden Untersuchungen:

Das **Ergebnis** einer Meta-Analyse hängt von der **Auswahl** der einbezogenen Primäruntersuchungen ab, so daß diese mit großer Sorgfalt erfolgen muß. Akribische Bemühungen um eine möglichst vollständige Erfassung aller thematisch einschlägigen Arbeiten (Monographien, Zeitschriftenartikel, Dissertationen, institutsinterne Berichte) sind hierzu unerlässlich. Diese Aufgabe wird durch fachspezifische Bibliotheken, Sammelreferate, Literaturdatenbanken, Informationsvermittlungsstellen der Universitäten und regelmäßigen Informationsaustausch mit Wissenschaftlern mit vergleichbaren Arbeitsschwerpunkten erheblich erleichtert. So wird das Ergebnis einer Meta-Analyse auch von der Fähigkeit des Meta-Analytikers zur Beschaffung von interessantem Material, das möglicherweise schwer zugänglich ist, beeinflusst.

Drei Selektionskriterien:

Es hat sich die Auffassung durchgesetzt, daß bei der Auswahl von Studien eher **liberale Kriterien** nützlich sind.

1) Methodische Qualität (interne Validität) der Primäruntersuchungen:

Es sollen alle Untersuchungen einbezogen werden, die methodischen Mindeststandards genügen. Wichtig ist hierbei, daß die Auswahl der Qualitätskriterien (z.B. Verwendung eines Kontrollgruppen-Designs, Größe, Art und Auswahl der Stichprobe, Reliabilität der Meßinstrumente, Fehlerkontrolle, Genauigkeit der Implementierung der Maßnahme/Behandlung, Angemessenheit der statistischen Verfahren) begründet wird und daß die Bewertung der Untersuchungen anhand dieser Kriterien möglichst objektiv erfolgt (mehrere Urteiler mit Nennung der Urteilerübereinstimmung, so daß die Meta-Analyse prinzipiell replizierbar ist).

2) Inhaltliche Kohärenz (Homogenität der UV und der AV) der Primäruntersuchungen:

Bei der Frage nach der inhaltlichen Kohärenz der Untersuchungen müssen zwei Aspekte unterschieden werden: die Homogenität in bezug auf die unabhängige Variable und hinsichtlich der abhängigen Variable.

Unabhängige Variable: Bei der Auswahl der Operationalisierungsvarianten der unabhängigen Variable können liberale Kriterien verwendet werden. Dies ist im umfassenderen Forschungsinteresse der Meta-Analyse im Vergleich zur Primärforschung begründet, das häufig bewußt Abstraktionen von konkreten Ausprägungen (Operationalisierungen) der unabhängigen Variable (und damit größere Heterogenität) anstrebt. Damit ist eine genaue Definition der UV mit objektivierbarer Ausgrenzung von nicht zulässigen Operationalisierungsvarianten hier schwierig. Die (liberale) Abgrenzung sollte den spezifischen Besonderheiten der inhaltlichen Fragestellung bzw. dem Erkenntnisinteresse des Meta-Analytikers überlassen werden. Sehr heterogene Varianten der Operationalisierung einer unabhängigen Variablen können im Nachhinein in homogene Gruppen aufgeteilt und getrennt analysiert werden.

Abhängige Variable: Für die Auswahl der Operationalisierungsvarianten der abhängigen Variable sollten hingegen strengere Kriterien gelten, da eine Zusammenfassung von Untersuchungsergebnissen nur dann sinnvoll ist, wenn die verschiedenen abhängigen Variablen Indikatoren eines gemeinsamen inhaltlichen Konstruktes und damit hoch korreliert sind. Daher ist hier besonderer Wert auf die genaue Definition der abhängigen Variable einschließlich der Ausgrenzung unzulässiger Operationalisierungsvarianten zu legen.

3) Unabhängigkeit der Ergebnisse (bzw. Stichproben) der Primäruntersuchungen:

Metaanalytische Aussagen basieren auf einer Gesamtstichprobe, die sich additiv aus den Stichprobenumfängen der Einzelstudien zusammensetzt, bei der also keine Teilstichprobe doppelt oder gar mehrfach gezählt wird. Daher sollten nur voneinander unabhängige Einzelergebnisse, die aus unabhängigen Stichproben stammen, in eine Meta-Analyse eingehen. Werden in einer Primärstudie mehrere (abhängige) Teilergebnisse derselben Stichprobe be-

richtet, geht in die Meta-Analyse entweder nur ein Teilergebnis ein, oder man faßt die Teilergebnisse zu einem Gesamtwert zusammen.

Merksatz: „Die Untersuchungen bzw. Daten für eine Meta-Analyse sind so auszuwählen, daß möglichst alle einschlägigen Arbeiten berücksichtigt werden, die 1) methodische Mindestkriterien erfüllen und 2) vergleichbare abhängige Variablen untersuchen. Die in die Meta-Analyse eingehenden Einzelergebnisse sollten 3) von unabhängigen Stichproben stammen (Teilergebnisse derselben Stichprobe sind ggf. zu einem Gesamtwert zu aggregieren).“ (BORTZ & DÖRING, S. 592)

Frage 26: In welcher **Beziehung** stehen **Meta-Analyse** und **Evaluationsforschung**? Illustrieren Sie Ihre Ausführungen knapp anhand eines Beispiels! (2 x gefragt)

Nach BORTZ & DÖRING sind **Meta-Analysen** zur **Vorbereitung größerer (summativer) Evaluationsprojekte** wichtig. Bereits vorhandene bzw. vom Evaluator selbst durchgeführte Meta-Analysen erleichtern die Formulierung einer **spezifischen** Hypothese unter Vorgabe einer bestimmten, unter Kosten-Nutzen-Aspekten festgelegten **Effektgröße** für die Wirkung der zu evaluierenden Maßnahme (praktisch bedeutsamer Effekt) (vgl. S. 590). BORTZ & DÖRING betonen, daß die summative Evaluation idealerweise am Ende eines Forschungsprozesses steht, der mit der Grundlagenforschung beginnt und über die Interventionsforschung zu einer konkreten Maßnahme führt, die einer abschließenden Evaluation unterzogen wird. Daher müßten die Themen der summativen Evaluationsforschung genügend elaboriert sein, um spezifische Hypothesen mit Effektgrößen formulieren zu können. Ferner ist es auch im Interesse des Auftraggebers zu erfahren, ob die von ihm finanzierte Maßnahme einen **praktisch bedeutsamen Effekt** erzielt (statt nur die Information zu erhalten, daß die Maßnahme ‚irgendwie‘ wirkt) (vgl. S. 111).

WOTTAWA & THIERAU führen im Zusammenhang mit der Schwierigkeit, wirklich aussagekräftige Evaluationen durchzuführen, aus, daß die **metaanalytische Zusammenfassung** der Ergebnisse in Form der Effektgrößen von **möglichst vielen verschiedenen Evaluationsstudien** eine wertvolle und unverzichtbare Grundlage für **allgemeine Empfehlungen** darstellt. Sie fordern daher, daß alle publizierten Evaluationsprojekte die für eine sachgerechte Aufarbeitung in Meta-Analysen erforderlichen Angaben in zusammengefaßter, übersichtlicher Form geben sollen, um die spätere metaanalytische Zusammenfassung zu erleichtern (vgl. S. 138).

Frage 27: Welche Funktion hat das **Δ -Maß** im Rahmen der **Meta-Analyse**? (1 x gefragt)

Das **Δ -Maß (Delta-Maß)** hat im Rahmen der Meta-Analyse die Funktion, **die spezifischen Effektgrößen der verschiedenen Primäruntersuchungen zu vereinheitlichen** und damit direkt vergleichbar und statistisch aggregierbar zu machen. Das Delta-Maß ist ein einheitliches/universelles Effektgrößen-Maß, das der bivariaten Produkt-Moment-Korrelation (r) entspricht, d.h. es ist ein Korrelations-Äquivalent. Jede signifikanztestspezifische Effektgröße läßt sich in einen Delta-Wert transformieren.

Die metaanalytische Zusammenfassung verschiedener Effektgrößen aus unabhängigen Primäruntersuchungen ist nur dann sinnvoll, wenn die einzelnen Effektgrößen Schätzungen einer gemeinsamen Populationseffektgröße δ (Delta) darstellen, was Homogenität der untersuchungsspezifischen Effektgrößen impliziert, d.h., die gefundenen Zusammenhänge müssen inhaltlich die gleiche Richtung aufweisen. Bevor die Effektgrößen mehrerer Einzeluntersuchungen metaanalytisch aggregiert werden, muß daher geprüft werden, ob die betreffenden Effektgrößen auch homogen sind. Dazu kann ein Homogenitätstest und/oder die 75%-Regel eingesetzt werden.

Frage 28: Was versteht man unter „**Meta-Analyse**“, und welche **Prüfschritte** müssen vor der **Aggregation** der einzelnen **Δ -Maße** unternommen werden? (1 x gefragt)

Definition:

Unter Meta-Analyse wird eine Gruppe von Verfahren verstanden, mit denen die **Ergebnisse verschiedener Untersuchungen mit gemeinsamer Thematik zusammengefaßt werden**, um so einen **Überblick über den aktuellen Stand der Forschung** zu gewinnen.

Es gibt bisher keine einheitliche metaanalytische Methodik, sondern die Meta-Analyse befindet sich in einem Entwicklungsprozeß, dessen Ende noch nicht absehbar ist. Für eine Vereinheitlichung im Sinn eines allgemein akzeptierten methodischen Vorgehens ist eine Norm für eine verbindliche Art der Ergebnisdarstellung in Publikationen erforderlich (Umfang und Genauigkeit der Informationen, statistische Ergebnisindikatoren).

Zielsetzung:

Das vorrangige Ziel aller metaanalytischen Techniken besteht in der **statistischen Aggregation der Einzelergebnisse von inhaltlich homogenen Primäruntersuchungen**. Von besonderer Bedeutung ist hierbei die Frage nach der **Wirksamkeit** einer Intervention (Maßnahme oder Behandlung), deren Beantwortung alle einschlägigen Forschungsergebnisse berücksichtigen soll. Damit ist die **Effektgröße** bzw. der durch viele einzelne Untersuchungen geschätzte „**wahre**“ **Effekt** einer Intervention ein zentraler Begriff der Meta-Analyse.

Das **Δ -Maß (Delta-Maß)** hat im Rahmen der Meta-Analyse die Funktion, **die spezifischen Effektgrößen der verschiedenen Primäruntersuchungen zu vereinheitlichen** und damit direkt vergleichbar und statistisch aggregierbar zu machen. Das Delta-Maß ist ein einheitliches/universelles Effektgrößen-Maß, das der bivariaten Produkt-Moment-Korrelation (r) entspricht, d.h. es ist ein Korrelations-Äquivalent. Jede signifikanztestspezifische Effektgröße läßt sich in einen Delta-Wert transformieren.

Die metaanalytische Zusammenfassung verschiedener Effektgrößen aus unabhängigen Primäruntersuchungen ist nur dann sinnvoll, wenn die einzelnen Effektgrößen Schätzungen einer gemeinsamen Populationseffektgröße δ (Delta) darstellen, was Homogenität der untersuchungsspezifischen Effektgrößen impliziert, d.h., die gefundenen Zusammenhänge müssen inhaltlich die gleiche Richtung aufweisen. Bevor die Effektgrößen mehrerer Einzeluntersuchungen metaanalytisch aggregiert werden, muß daher geprüft werden, ob die betreffenden Effektgrößen auch homogen sind. Dazu kann ein Homogenitätstest und/oder die 75%-Regel eingesetzt werden.

Prüfung der Ergebnishomogenität vor der Aggregation der einzelnen Δ -Maße

1. Homogenitätstest

Nach der Transformation der verschiedenen unabhängigen Effektgrößen der Primäruntersuchungen in Delta-Maße (Δ -Maße) ist vor einer Zusammenfassung der Delta-Maße zu überprüfen, ob die untersuchungsspezifischen Effektgrößen als Schätzungen eines gemeinsamen Populationsparameters δ (Delta) anzusehen sind. Diese Überprüfung geschieht mit Hilfe eines X^2 -Homogenitätstests nach HUNTER ET AL. (1982). Ein signifikanter X^2 -Wert weist darauf hin, daß die Varianz der durch die einzelnen Untersuchungen geschätzten wahren Effektgrößen ungleich Null ist, so daß von einem Modell heterogener Effektgrößen auszugehen ist. Dieses Ergebnis würde die Suche nach Moderatorvariablen bzw. mehrere Meta-Analysen über homogene Teilgruppen von Untersuchungen erforderlich machen.

Allerdings muß berücksichtigt werden, daß die **Varianz der Delta-Maße** auch durch die **Verschiedenartigkeit der Untersuchungen** (stichprobenbedingte Ergebnisunterschiede, mangelhafte bzw. fehlende Reliabilität der Messungen, verschiedene Meßbereiche der in Beziehung gesetzten Merkmale, Operationalisierungsunterschiede bei der Erfassung der untersuchten Konstrukte, Unterschiede hinsichtlich der Störvariablenkontrolle, Fehler in den Ergebnisberichten etc.) beeinflusst wird. Somit ist die (ausschließliche) Schätzung der Streuung der wahren Effektgrößen aufgrund der Streuung der Delta-Maße wenig geeignet.

2. 75%-Regel

Für eine zuverlässigere Überprüfung der Homogenität der Delta-Maße wird daher die Heranziehung eines deskriptiven Maßes, der sog. 75%-Regel, empfohlen. Dieses Maß bezieht sich auf denjenigen Anteil der Varianz der Delta-Maße, der auf Stichprobenfehler und sonstige Artefakte zurückgeht. Die 75%-Regel lautet: Wenn mindestens 75% der Varianz

der Delta-Maße durch die Verschiedenartigkeit der Untersuchungen erklärbar sind, kann auf Homogenität der Delta-Maße geschlossen werden. → höhere Teststärke (BARBARA REHSE)

3. Signifikanztest für den Gesamteffekt

Ist die Homogenitätshypothese, überprüft durch den X^2 -Homogenitätstest und/oder die 75%-Regel, beizubehalten, dann stellt das **durchschnittliche Δ -Maß (Gesamteffektgröße)** einen akzeptablen Schätzwert der wahren Effektgröße δ dar. Um zu überprüfen, ob dieser Schätzwert signifikant von Null abweicht, wird ein Konfidenzintervall berechnet. Umschließt dieses Konfidenzintervall den Wert Null, muß davon ausgegangen werden, daß die metaanalytisch ermittelte Gesamteffektgröße nicht signifikant ist. Liegt der Wert Null außerhalb des Konfidenzintervalls, ist die Gesamteffektgröße statistisch signifikant.

Merksatz: „Die Meta-Analyse faßt homogene Effektgrößen bzw. Δ -Maße aus Einzeluntersuchungen zu einem durchschnittlichen Δ -Maß (Gesamteffektgröße) zusammen, mit dem die Populationseffektgröße δ geschätzt wird. Umschließt das 95%-ige Konfidenzintervall für die Populationseffektgröße den Wert Null, so ist der Gesamteffekt nicht signifikant; umschließt es nicht den Wert Null, ist von einem signifikanten Populationseffekt auszugehen.“ (BORTZ & DÖRING, S. 597)

Frage 29: Was versteht man unter „**Moderatorvariablen**“, und welche **Funktion** haben sie in der **Meta-Analyse**? (1 x gefragt)

Definition:

Eine **Moderatorvariable** (moderierende Variable) ist eine Variable, die den **Einfluß** einer unabhängigen Variable auf die abhängige Variable **verändert**.

Funktion von Moderatorvariablen in der Meta-Analyse:

Im Rahmen einer Meta-Analyse können Moderatorvariablen **Unterschiede in den Δ -Maßen** verursachen, da sie die Effekte in den einzelnen Primäruntersuchungen beeinflusst und damit zu heterogenen Δ -Maßen geführt haben. Ergibt die Homogenitätsprüfung heterogene Δ -Maße, sollte daher immer der potentielle Einfluß von Moderatorvariablen geprüft werden, die die Unterschiede in den Δ -Maßen erklären könnten. Für eine Meta-Analyse relevante Moderatorvariablen umfassen Besonderheiten der aggregierten Studien wie z.B. den Designtyp, Operationalisierungsvarianten, Kontrolltechniken, Art der Publikation, Jahr der Veröffentlichung etc. oder weitere Merkmale, die sich aus dem inhaltlichen oder methodischen Vergleich der metaanalytisch zusammengefaßten Studien ergeben.

Die Analyse potentieller Moderatorvariablen sollte varianzanalytisch erfolgen. Hat man keine Hypothese zum Einfluß spezieller Moderatorvariablen, kann man zwischen den potentiellen Moderatorvariablen und den studienspezifischen Δ -Maßen eine (multiple) Korrelation berechnen, deren Höhe über die Bedeutung der Studienmerkmale für die Heterogenität der Δ -Maße informiert. Signifikante bzw. praktisch bedeutsame Korrelationen sollten zum Anlaß genommen werden, die Primärstudien in homogene Subgruppen aufzuteilen, für die dann getrennte Meta-Analysen berechnet werden müssen. Die Subgruppenbildung sollte insbesondere bei großen Meta-Analysen mit vielen Einzelstudien clusteranalytisch abgesichert werden. Dabei werden die Studien so gruppiert, daß die Unterschiede in möglichst vielen Studienmerkmalen innerhalb der einzelnen Gruppen möglichst gering und zwischen den Gruppen möglichst groß sind.

Läßt sich bei einem signifikanten Homogenitätstest die Bildung homogener Subgruppen für getrennte Meta-Analysen nicht sinnvoll begründen, sollte auf eine Meta-Analyse verzichtet werden, denn metaanalytische Aussagen, die auf heterogenen Studien basieren, sind eher irreführend als klärend.

III Fragen zu Methoden der Entwicklungspsychologie

Literatur:

TRAUTNER, H.M. (1992²). Lehrbuch der Entwicklungspsychologie. Bd. 1. → Kap. 4.1 und 4.2 (S. 227-278).

Frage 30: Welche Probleme bringt die Variable „Alter“ im Sinne einer UV für die **entwicklungspsychologische Forschung** mit sich? (3 x gefragt) bzw. Welche Probleme bestehen im Zusammenhang mit der **Altersvariable** in der **entwicklungspsychologischen Forschung**? (1 x gefragt)

Die über die Zeit zu beobachtenden Entwicklungsveränderungen werden häufig in Form von Veränderungen über das Lebensalter dargestellt, d.h. es wird deskriptiv eine korrelative Beziehung zwischen dem Eintreten bestimmter Veränderungen und dem Lebensalter dargestellt. Viele Veränderungen treten in Zusammenhang mit einem bestimmten Lebensalters auf, so daß eine derartige Darstellung durchaus sinnvoll sein kann (z.B. Schulreife mit sechs Jahren, schulischer Leistungsabfall in der Pubertät).

Unter methodischen Aspekten besitzt die Variable „Alter“ allerdings nicht den Status einer echten unabhängigen Variable. Gegen ihre Verwendung als UV sprechen **zwei Gründe**:

- 1) Die **fehlende experimentelle Manipulierbarkeit** des Alters eines Individuums: Es ist nicht möglich, die Variable „Alter“ unter Konstanthaltung aller übrigen möglichen Entwicklungsfaktoren zu kontrollieren und zu variieren (keine Randomisierung möglich), da sie eine **organismische Variable** ist. Alter als Einflußgröße ist ja gerade durch seine **korrelative Beziehung** zu den verschiedensten Entwicklungsfaktoren definiert (vgl. auch TRAUTNER, S. 33 f), und dies individuell verschieden. Außerdem ist keine Replikation einer Untersuchung an der gleichen Stichprobe unter Konstanthaltung des Alters möglich.
- 2) Von seiner inhaltlichen Definition her hat das Alter selbst **keinen Erklärungswert**, es ist nur eine **Trägervariable** für verschiedene dahinter wirksame Entwicklungsfaktoren. Die Variable „Alter“ beinhaltet nur eine (jeweils individuell zu definierende) **zeitliche Dimension** (chronologische Zeit nach der Geburt, vgl. TRAUTNER, S. 245), in der sich unter Gegebenheit bestimmter Bedingungen (unabhängige Variablen) bestimmte Veränderungen (in abhängigen Variablen) vollziehen. Vorgefundene Altersverläufe für eine (abhängige) Variable sind immer an die für die jeweils untersuchten Individuen gegebenen früheren und gegenwärtigen Entwicklungsbedingungen (unabhängige Variablen) gebunden. (vgl. TRAUTNER, S. 229 f)

Frage 31: Was versteht man unter einer **Entwicklungsfunktion**? Welche **Informationen** liefert sie für Individuen und Gruppen? (2 x gefragt) bzw. Was versteht man unter einer **Entwicklungsfunktion**, und wie kann man sie **analysieren**? (1 x gefragt)

Definition: Unter einer **Entwicklungsfunktion** versteht man die Form oder Art der Beziehung zwischen dem chronologischen Alter eines Individuums und den im Verlauf der Entwicklung auftretenden Veränderungen in einer bestimmten Variablen. Nach KESSEN ist Entwicklung bzw. Veränderung (V) eine Funktion (f) der Zeit (t): $V = f(t)$ und kann folgendermaßen abgebildet werden: Veränderung = Ordinate, Zeit = Abzisse, Funktion = Kurve. Wird der Begriff der Entwicklungsfunktion weit gefaßt, so lassen sich auch qualitative Veränderungen auf diese Art darstellen: die Entwicklungsfunktion ist dann als regelhafte Abfolge einzelner Variablen auf dem Zeitkontinuum definiert (vgl. TRAUTNER, S. 232) → **Entwicklungssequenz:** die regelhafte/gesetzmäßige Abfolge qualitativer Veränderungen in einer Variable bzw. die **regelhafte Abfolge mehrerer aufeinander bezogener Variablen** (vgl. TRAUTNER, S. 230).

Überall, wo sich gesetzmäßige ontogenetische Veränderungen in einer Variablen zeigen, besteht die Möglichkeit, diese Veränderungen in Form von Entwicklungsfunktionen auszudrücken, sofern in einer solchen Entwicklungsfunktion tatsächlich die systematische intraindividuelle Variation gemessen wird und nicht ein (zufälliger oder systematischer) Meßfehler.

Die Erstellung von Entwicklungsfunktionen bzw. –kurven bezieht sich auf das einzelne Individuum und kann entsprechend nur durch die wiederholte Messung des gleichen Individuums gewonnen werden. Die **individuelle Entwicklungsfunktion** liefert also die Information, wie sich ein Merkmal bzw. mehrere aufeinander bezogene Merkmale bei **einem Individuum** (intraindividuell) über die Zeit verändern.

Unter der Voraussetzung weitgehend ähnlicher Kurvenverläufe verschiedener Individuen lassen sich **prototypische Entwicklungsfunktionen für Gruppen** von Individuen erstellen. Zu

beachten ist dabei, daß bei der Zusammenfassung (Mittelung) der individuellen Kurven zu einer Gruppenkurve ein Maßstab gewählt wird, der die gemeinsamen Verlaufsmerkmale der individuellen Kurven adäquat wiedergibt (vgl. TRAUTNER, S. 233). Prototypische überindividuelle Entwicklungsfunktionen abstrahieren von den individuellen Unterschieden in den Verlaufskurven und liefern damit generalisierende Informationen darüber, wie sich ein Merkmal bzw. mehrere aufeinander bezogene Merkmale bei einer **Gruppe von Individuen derselben Altersgruppe** (intraindividuell) über die Zeit verändern.

Für die Entwicklungspsychologie ist wie für andere Wissenschaften auch die Generalisierbarkeit von Aussagen ein wesentliches Ziel; daher kommt der Erstellung überindividueller Entwicklungsfunktionen große Bedeutung zu. Um trotz des Vorhandenseins interindividueller Entwicklungsunterschiede zu einem möglichst hohen Grad von Generalisierbarkeit zu kommen, empfiehlt es sich, Entwicklungsfunktionen für Gruppen von Individuen zu erstellen, die in einer Reihe von Merkmalen, welche die betreffende Entwicklungsfunktion beeinflussen, möglichst vergleichbar sind (z.B. hinsichtlich Intelligenz, Geschlecht etc.).

Die Erstellung von Entwicklungsfunktionen bezieht sich also zunächst auf die primäre Aufgabe der Entwicklungspsychologie, intraindividuelle Veränderungen von Individuen oder Gruppen von Individuen zu beschreiben und zu erklären. Innerhalb dieser Aufgabenstellung erfolgt auch die Analyse der einzelnen individuellen Entwicklungsfunktionen (intraindividuelle Analyse).

Es gibt **vier Niveaus der Analyse** von Entwicklungsfunktionen:

1. Analyse nach der allgemeinen **Richtung** der Veränderung (z.B. ansteigend oder abfallend)
2. Analyse nach der allgemeinen **Form** des Kurvenverlaufs (z.B. linear ansteigend, nach anfänglich positiver Beschleunigung zunehmend negativ beschleunigt)
3. Analyse unter Berechnung der speziellen **mathematischen Funktion**, die dem Kurvenverlauf am besten entspricht (z.B. Kurven höherer Ordnung/ Exponentialfunktion, Polygon, logistische Kurve) zur Anpassung empirischer Kurven an theoretische Verläufe und zur Trendbestimmung
→ Ebenen 1. bis 3.: Gesamtcharakterisierung von Entwicklungskurven
4. Analyse unter Berechnung von **Parametern für Teilcharakteristika** einer Kurve wie Änderungsgeschwindigkeit innerhalb eines bestimmten Zeitraumes, Zeitpunkte des Erreichens von bestimmten Charakteristika der Kurve (z.B. Minimum, Maximum, Asymptote).

Die Analyse der Unterschiede in den individuellen Entwicklungsfunktionen dient der zweiten Aufgabe der Entwicklungspsychologie, die interindividuellen Unterschiede in den intraindividuellen Veränderungen zu beschreiben und zu erklären (interindividuelle Analyse). Auch diese Analyse kann auf den vier genannten Niveaus erfolgen.

Frage 32: Welches Problem ist bei der **Mittelung individueller Entwicklungsverläufe** zu beachten? Welche Lösungsvorschläge gibt es? (1 x gefragt)

Die Erstellung von Entwicklungskurven bezieht sich auf das einzelne Individuum und kann entsprechend nur durch die wiederholte Messung des gleichen Individuums gewonnen werden. Unter der Voraussetzung weitgehend ähnlicher Kurvenverläufe verschiedener Individuen lassen sich prototypische Funktionen für Gruppen von Individuen erstellen. Allerdings ergibt sich das Problem, daß sich aus der entstandenen überindividuellen Durchschnittskurve (gemittelte Kurve) nicht unmittelbar auf individuelle Entwicklungsverläufe schließen läßt, weil sich die Entwicklungsveränderungen bei verschiedenen Individuen im **Zeitpunkt des Eintretens**, der **Geschwindigkeit**, dem **Niveau** und/oder der **Verlaufsform** unterscheiden können (vgl. TRAUTNER, S. 17 f).

Beispiel: Geschwindigkeit des Größenwachstums von fünf Mädchen in Kindheit und Jugend.

Gruppenkurve a): Durchschnittliche jährliche Größenzunahme (Körperhöhe) bei Mädchen als Funktion des **Lebensalters**: Die Durchschnittskurve individueller Verläufe aufgrund der

gemittelten Gruppenwerte **sieht anders aus** als jede einzelne individuelle Kurve aufgrund der individuellen Veränderungswerte, obwohl die Individualkurven in ihrer Verlaufsform untereinander sehr ähnlich sind. Die Abweichung der Individualkurven und der Durchschnittskurve kommt hier dadurch zustande, daß das Maximum der Wachstumsgeschwindigkeit bei den einzelnen Mädchen zu verschiedenen Zeitpunkten (Alterswerten) eintritt. Dies führt zu einer **Verkleinerung** der in Jahresabständen berechneten **mittleren Zuwachsrates** und damit zu einer erheblichen **Glättung** der Durchschnittskurve (vgl. Abbildung, vgl. TRAUTNER, S. 17).

!!!!hier Kopie aus TRAUTNER, S. 18, Abb. 1.1 aufkleben!!!!
7 cm

—— = Individualkurven von fünf Mädchen - - - - - = Gemittelte Gruppenkurve

Gruppenkurve b): Durchschnittliche jährliche Größenzunahme (Körperhöhe) bei Mädchen als Funktion des **Zeitpunkts der größten Wachstumsgeschwindigkeit**: Die Durchschnittskurve individueller Verläufe **gibt die Abfolge der Zuwachswerte adäquat wieder**, da hier die zeitliche Entfernung vom Punkt der maximalen Zunahme des Größenwachstums zum Maßstab genommen wird (vgl. Abbildung, vgl. TRAUTNER, S. 17).

!!!!hier Kopie aus TRAUTNER, S. 18, Abb. 1.1 aufkleben!!!!
7 cm

—— = Individualkurven von fünf Mädchen - - - - - = Gemittelte Gruppenkurve

Die Lösung des Problems einer geglätteten Durchschnittskurve, die von allen individuellen Kurven abweicht, liegt in der **Wahl eines adäquaten Maßstabs** bei der Zusammenfassung (Mittelung) der individuellen Kurven zu einer Gruppenkurve, der die gemeinsamen Verlaufsmerkmale der einzelnen individuellen Kurven adäquat wiedergibt (vgl. TRAUTNER, S. 17, S. 233).

Frage 33: Beschreiben Sie kurz **drei prototypische Fälle** der **Messung** von **ontogenetischen Veränderungen!** (1 x gefragt)

Unter dem Aspekt der wechselseitigen Zuordnung von a) einem „wahren“ Variablenwert eines Individuums, b) einem gemessenen Skalenwert bei dem Individuum und c) einem Zeitpunkt bzw. gemessenen Lebensalter sowie unter Berücksichtigung des Skalenniveaus der gemessenen Variable lassen sich **drei prototypische Fälle der Messung ontogenetischer Veränderungen** in abhängigen Variablen unterscheiden (vgl. TRAUTNER, S. 240 f, S. 238 f):

Veränderung einer kontinuierlichen Variable, intervallskalierte Messung	Veränderung einer kontinuierlichen Variable, ordinalskalierte Messung	regelhafte Abfolge von mehreren Variablen (Entwicklungssequenz*), nominalskalierte Messung
Veränderungen werden mit einer unbeschränkten Anzahl von Ausprägungsgraden des Merkmals gemessen; zwischen je zwei Punkten mit zahlenmäßig gleichem Unterschied besteht in allen Skalenbereichen der gleiche zahlenmäßige Abstand (Intervall) der Variablenwerte → quantitative Veränderungen; ihre Messung erlaubt die Erstellung von Entwicklungsfunktionen	Veränderungen werden nur in einer beschränkten Anzahl von Ausprägungsgraden/Abstufungen des Merkmals angeordnet (z.B. nicht vorhanden, etwas ausgeprägt, stark ausgeprägt, sehr stark ausgeprägt); gleiche Rangabstände (z.B. zwischen den Stufen 1 u. 3 und den Stufen 4 u. 6) müssen nicht zahlenmäßig gleiche Unterschiede in der Ausprägung des Merkmals widerspiegeln → quantitative Veränderungen; ihre Messung erlaubt die Erstellung von Entwicklungsfunktionen	Veränderungsreihen werden durch die Zuordnung von Personen zu mehreren aufeinanderfolgenden Variablen gemessen; jeweils nominale Klassifikation in diskrete Klassen („Merkmal ist vorhanden“ versus „Merkmal ist nicht vorhanden“, diskontinuierlicher Wechsel von einem Zustand zum anderen); mehrere qualitative/diskontinuierliche Variablen, deren Ausprägungen sich nicht in numerischen Werten einer kontinuierlichen Skala ausdrücken lassen → qualitative Veränderungsreihen; ihre Messung erlaubt keine Erstellung von Entwicklungsfunktionen
die Beziehung zwischen den Veränderungen und dem Lebensalter kann unterschiedlicher Art sein (Skalierung der Variable ist unabhängig von der Art der Beziehung zum Alter)	die Beziehung zwischen den Veränderungen und dem Lebensalter kann unterschiedlicher Art sein (Skalierung der Variable ist unabhängig von der Art der Beziehung zum Alter)	die Beziehung zwischen der Variablenabfolge und dem Lebensalter kann nur monoton sein (Skalierung der Variablenabfolge basiert auf der Altersvariation)
eindimensional quantifizierbare (= kontinuierliche) Variable	eindimensional quantifizierbare (= kontinuierliche) Variable	mehrere Variablen, deren Abfolge sich nicht auf einer Dimension anordnen läßt (außer auf Zeitdimension), daher nur nominalskalierte Messung möglich
Beispiele: 1) visuell-motorische Koordination (Angabe der Zeitdauer zur Bewältigung der Koordinationsaufgaben) 2) zeitliche Veränderungen der Variable Intelligenz	Beispiele: 1) visuell-motorische Koordination (Grade der Koordinationsfähigkeit, z.B. keine, Ansätze, gute, sehr gute) 2) Entwicklung des moralischen Urteils (Stufen)	Beispiel: regelhafte motorische Sequenzen im Kleinkindalter (z.B. Abfolge der aufeinander bezogenen Variablen Sitzen-Krabbeln-Stehen-Laufen)

***Definition Entwicklungssequenz:** „Die regelhafte/gesetzmäßige Abfolge qualitativer Veränderungen in einer Variable bzw. die regelhafte/**gesetzmäßige Abfolge mehrerer aufeinander bezogener Variablen**“ (TRAUTNER, S. 230).

Frage 34: Wie können in der **Entwicklungspsychologie qualitative Veränderungen** in **quantitative Werte** umgewandelt werden? (1 x gefragt)

In der Entwicklungspsychologie interessiert sowohl die Untersuchung von **quantitativen** Veränderungen (z.B. Zunahme des Wortschatzes, Abnahme der Häufigkeit aggressiven Verhaltens) als auch von **qualitativen** Veränderungen (z.B. Wandel im Gebrauch von Satzstrukturen, Ausbildung sekundärer Geschlechtsmerkmale in der Pubertät).

Nur quantitative Veränderungen können in Form von Entwicklungskurven nach dem Muster $V = f(t)$ abgebildet werden (Veränderung (V, Ordinate) als Funktion (f) der Zeit (t, Abzisse)) → eindimensionale Darstellung der beobachteten Veränderungen in kontinuierlichen (= eindimensional quantifizierbaren) Variablen.

Wird der Begriff der Entwicklungsfunktion weit gefaßt, so lassen sich auch qualitative Veränderungen auf diese Art darstellen: die Entwicklungsfunktion ist dann als regelhafte Abfolge einzelner Variablen auf dem Zeitkontinuum definiert (vgl. TRAUTNER, S. 232) → **Entwicklungssequenz**: die regelhafte/gesetzmäßige Abfolge qualitativer Veränderungen in einer Variable bzw. die **regelhafte Abfolge mehrerer aufeinander bezogener Variablen** (vgl. TRAUTNER, S. 230). Um qualitative Veränderungen in Form von Entwicklungskurven darstellen zu können, muß eine Umwandlung der Qualitäten in quantifizierbare Indizes wie Zeit (Alter) des erstmaligen Auftretens der einzelnen Qualitäten oder Prozentsatz von Individuen auf den einzelnen Altersstufen, die an einem bestimmten Punkt der Entwicklungssequenz ange­langt sind, erfolgen (vgl. TRAUTNER, S. 230, S. 241 ff).

Drei Arten der Umwandlung qualitativer Veränderungen in quantitative Werte

1) Angabe der Anzahl der VP einer Stichprobe, die ein Kriterium erreichen.

Um aus diskontinuierlichen individuellen qualitativen Veränderungen ein quantifizierbares Maß abzuleiten, muß man als Bezugspunkt anstelle des einzelnen Individuums die Gruppe wählen und für jeden Zeitpunkt des interessierenden Altersbereichs angeben, wieviele Individuen der einzelnen Altersstufen ein bestimmtes Kriterium erreichen (Angabe meist in Prozentwerten). Damit werden diskontinuierliche individuelle Entwicklungsverläufe in eine **kontinuierliche Entwicklungsverlaufskurve der Altersgruppe** aufgrund der sich kontinuierlich verändernden Gruppenwerte umgewandelt. Sie kann allerdings nicht als typischer Entwicklungsverlauf angesehen werden, denn es ist nicht zu erkennen, was für den Kurvenverlauf verantwortlich ist: die individuellen Unterschiede in der Veränderungsgeschwindigkeit oder im Altersbereich, in dem die Veränderung eintritt (vgl. TRAUTNER, S. 241).

Beispiel: Motorische Entwicklung im frühen Kindesalter (qualitative Veränderungsreihe); Variable: Allein-laufen-Können (ohne fremde Hilfe) vs. Nicht-allein-laufen-Können.

Für jedes Kind läßt sich feststellen, ob es allein laufen kann oder nicht (diskrete Variable). Außerdem läßt sich für jedes Kind, das allein laufen kann, feststellen, seit welchem Alter es dies kann. Für jedes Kind bedeutet der Wechsel von einem Zustand (nicht allein laufen) in den anderen (allein laufen) eine diskontinuierliche, qualitative Veränderung. Betrachtet man nicht das einzelne Kind (Frage: Kann das Kind allein laufen oder nicht?), sondern die ganze Gruppe des interessierenden Altersbereichs, ist es möglich, unter Beibehaltung des diskreten Charakters der Variable ein quantifizierbares Maß zu erhalten, indem man für jeden Zeitpunkt des Altersbereichs angibt, wie viele Kinder der einzelnen Altersstufen allein laufen können (Frage: Wieviel Prozent der Kinder einer Altersgruppe können allein laufen?). Dieser Gruppenwert verändert sich kontinuierlich.

!!!!hier Kopie aus TRAUTNER, S. 224, Abb. 4.4 aufkleben!!!!
8,5 cm

2) Bildung eines Gesamtwertes auf der Basis von Items heterogener Schwierigkeit

Um eine qualitative Veränderung zu quantifizieren, werden eine Reihe unterschiedlich schwieriger Aufgaben zur Feststellung des Ausprägungsgrades der betreffenden Variable konstruiert und die Summe der gelösten Aufgaben als quantitatives Maß verwendet. Die Veränderungen in der Anzahl der gelösten Aufgaben im Verlauf der individuellen Entwicklung können als **individuelle Entwicklungskurve** nach der Formel $V = f(t)$ dargestellt werden.

Beispiel: Allein-laufen-Können. Die Variable „Allein-laufen-Können“ wird mit verschiedenen motorischen Aufgaben verknüpft, die unterschiedlich schwer zu bewältigen sind, z.B. über Strecken mit unterschiedlichen Steigungsgraden laufen, Hindernisse umgehen etc. Man kann für jedes Kind auszählen, wieviele Aufgaben bewältigt wurden. Die Summe der gelösten Aufgaben dient als quantitativ abstufbares Maß der Fähigkeit, allein laufen zu können.

Zu dieser Art der Umwandlung sind **drei Voraussetzungen** nötig (vgl. TRAUTNER, S. 242 f):

1. Die Variable muß quantitative Abstufungen erlauben, d.h., es muß eine latente **kontinuierliche Dimension** gegeben sein, und es können zur Messung der Variable eine Reihe von Aufgaben gefunden werden, die in ihrer **Schwierigkeit abgestuft** sind.
2. Alle Items sollen zwar die gleiche Variable (Dimension) messen, gleichzeitig soll jedoch die **Wahrscheinlichkeit der Lösung** eines Items möglichst **unabhängig** von der Lösung anderer Items sein (relative stochastische Itemunabhängigkeit: die Wahrscheinlichkeit für die Lösung eines Items/einer Aufgabe ist unabhängig von der Lösung anderer Items/Aufgaben; sie hängt nur von dem Ausprägungsgrad der latenten Dimension und der Schwierigkeit des Items/der Aufgabe ab, vgl. Probabilistische Testtheorie). Ist diese Voraussetzung der relativen Itemunabhängigkeit nicht erfüllt, dann resultiert eine bimodale Werteverteilung, die ein Hinweis darauf ist, daß die (erste) Voraussetzung einer latenten kontinuierlichen Dimension nicht aufrechterhalten werden kann und stattdessen von einer diskontinuierlichen Dimension/Variable auszugehen ist. (Damit auch niedrige Konstruktvalidität, z.B. durch Abstufung von Lösungen unabhängig von der Schwierigkeit der Aufgaben oder durch mehrere Dimensionen von Schwierigkeit → Nachweis Autor?).
3. **Äquivalenz aller Items:** Jedes Item geht mit dem gleichen Gewicht in den Gesamtwert ein; der gleiche Gesamtwert kann sich daher aus verschiedensten Items zusammensetzen. (Dies spricht für Kontinuität der Dimension/Variable → Nachweis Autor?).

3) Betrachtung qualitativer Variablen unter rein quantitativem Aspekt

Die Messung einer qualitativen Variable wird auf deren leicht quantifizierbare (Neben-)Aspekte beschränkt, z.B. die Dauer zur Lösung einer Aufgabe oder die Effizienz eines Verhaltens. Diese Form der Umwandlung ist eher als ein **zusätzliches Maß** einzusetzen und weniger als *Ersatz* für eine direkte Art der Messung (vgl. TRAUTNER, S. 243 f).

Beispiel: Motorische Entwicklung im Kleinkindalter. Quantitative Messung der Zeit, die jedes einzelne Kind benötigt, um sich zu einem Ziel fortzubewegen (von Punkt 1 zu Punkt 2), oder der Effektivität der individuellen Art der Fortbewegung zur Erreichung eines Zieles, unabhängig von der Art (Qualität) der Fortbewegung.

Frage 35: Inwiefern hilft der Begriff der „Kontrollgruppe“ bei der Unterscheidung zwischen Sequenzplänen und den klassischen Längs- bzw. Querschnittplänen? (1 x gefragt)

Aus der angegebenen Literatur zum Thema/Referat „Abbildung von Entwicklungsvorgängen“ (TRAUTNER, 1992, Bd. 1, Kap. 4.2, S. 244-278) kann man die Antwort auf diese Frage nicht direkt entnehmen – der Begriff „Kontrollgruppe“ kommt im Text nicht vor! Auf S. 260 f befindet sich der Punkt „Die Einordnung der konventionellen Stichprobenpläne in SCHAIES dreifaktorielles Entwicklungsmodell“ und auf S. 271 f der Punkt „Die Einordnung der konventionellen und der sequentiellen Untersuchungspläne in das zweifaktorielle Entwicklungsmodell von BALTES“ → aus diesen Abschnitten kann man den Begriff der „Kontrollgruppe“ zur Kennzeichnung der Sequenzpläne von SCHAIES und von BALTES ableiten.

Aus dem Abschnitt „Die Einordnung der konventionellen Stichprobenpläne in SCHAIES dreifaktorielles Entwicklungsmodell“ (S. 260 f): Die Tabelle 4.4 zeigt, „daß die konventionellen Methoden der Querschnitt- und der Längsschnittuntersuchung als Sonderfälle von SCHAIES Allgemeinem Entwicklungsmodell angesehen werden können. **Jede Spalte entspricht einer einzelnen Querschnittuntersuchung, jede Zeile einer einzelnen Längsschnittuntersuchung.**(..) Darüber hinaus beinhaltet das Modell noch einen dritten Plan, die sog. Zeitwandelmethode. Hier stellt jede Diagonale des Plans eine eigene Zeitwandeluntersuchung dar. Die Zeitwandelmethode eignet sich zur Überprüfung des Einflusses sich über die historische Zeit wandelnder Entwicklungsbedingungen auf das Verhalten bestimmter Altersgruppen.“

!!!!hier Kopie aus TRAUTNER, S. 259, Tab. 4.4 aufkleben!!!!

5,5 cm

Durch Erweiterung der drei aus dem Allgemeinen Entwicklungsmodell abgeleiteten konventionellen Stichprobenpläne (Längsschnitt-, Querschnitt- und Zeitwandeluntersuchung) gelangt SCHAIE ZU seinen drei Sequenzmodellen/sequentiellen Untersuchungsplänen (vgl. TRAUTNER, S. 261 ff):

!!!!hier Kopie aus TRAUTNER, S. 263, Tab. 4.5 aufkleben!!!!

7,5 cm

Aus dem Abschnitt „Die Einordnung der konventionellen und der sequentiellen Untersuchungspläne in das zweifaktorielle Entwicklungsmodell von BALTES“ (S. 271 f): BALTES sieht die klassische Längsschnittuntersuchung und die konventionelle Zeitwandeluntersuchung als „adäquate einfaktorische Versuchspläne (Teilpläne) an, bei denen der Effekt des einen (mehrstufigen) Faktors festgestellt werden soll, ohne über den zweiten (konstant gehaltenen) Faktor zu verallgemeinern“. Bei der Längsschnittuntersuchung wird das Alter variiert und der Faktor Kohorte konstant gehalten; Stichprobenunterschiede werden dann als reine Alterseffekte interpretiert. Bei der Zeitwandeluntersuchung wird der Faktor Kohorte mehrstufig definiert und das Alter konstant gehalten; Stichprobenunterschiede werden dann als reine Kohorteneffekte interpretiert.

Die konventionelle Querschnittuntersuchung stellt für ihn dagegen einen inadäquaten einfaktorischen Versuchsplan dar, da hier die beiden Faktoren Alter und Kohorte unausweichlich miteinander konfundiert sind (systematische Variation des Faktors Kohorte mit den verschiedenen Altersstufen). „Lediglich in ihrer sukzessiven Anwendung als Querschnittsequenz ist die Querschnittmethode für BALTES ein legitimer Untersuchungsplan“.

Durch die einfache Erweiterung der beiden aus seinem auf zwei Faktoren reduzierten Allgemeinen Entwicklungsmodell abgeleiteten konventionellen Pläne (Längsschnitt- und Quer-

schnittuntersuchung) gelangt BALTES zu seinen zwei Sequenzmodellen/sequentiellen Untersuchungsplänen (vgl. TRAUTNER, S. 271 f):

!!!!hier Kopie aus TRAUTNER, S. 272, Tab. 4.10 aufkleben!!!!
8,5 cm

Die klassischen Längs- und Querschnittpläne sind quasiexperimentelle einfaktorielle Pläne mit der UV Alter; der konventionelle Längsschnittplan ist ein VPL 1 Q(W)-Plan, der konventionelle Querschnittplan ein VPL 1 Q-Plan. Die Sequenzpläne von SCHAIE und von BALTES kann man als quasi-experimentelle **zweifaktorielle Kontrollgruppenpläne** bezeichnen, da die Sequenzen von Längs- bzw. Querschnittuntersuchungen als Untersuchungen mit Experimental- und Kontrollgruppen aufgefaßt werden können:

Bei einem **Längsschnittsequenzplan** (bei SCHAIE Kohorten-Sequenz-Methode, bei BALTES Längsschnitt-Sequenz) werden mehrere Kohorten über mehrere aufeinanderfolgende Altersstufen untersucht, d.h., Kohorte und Alter der Vpn werden systematisch variiert. Es handelt sich also um einen Meßwiederholungsplan **VPL 2 QQ(W)** mit meßwiederholter hypothesenrelevanter UV B „Alter“ und Kontrollfaktor UV A „Kohorte“, dessen Stufen als systematische Replikationsstudien angesehen werden können: Für den gleichen Altersbereich werden für verschiedene Kohorten Daten im Längsschnitt erhoben (sukzessive Durchführung mehrerer Längsschnittuntersuchungen → Sequenzen von Längsschnittuntersuchungen), wobei der Vergleich der Daten der verschiedenen Kohorten (Kohortendifferenzen) der Kontrolle der Altersdifferenzen dient (vgl. TRAUTNER, S. 262, S. 271 f); **die verschiedenen Kohorten stellen also gegenseitig Kontrollgruppen dar.**

Dieser Plan erlaubt Aussagen über a) die durchschnittlichen **Altersdifferenzen** für die untersuchten Kohorten und b) die durchschnittlichen **Kohortendifferenzen** für die untersuchten Altersstufen. Altersdifferenzen und Kohortendifferenzen sind hier also **gegenseitig kontrolliert** (vgl. TRAUTNER, S. 262) (vgl. Vorteil von zweifaktoriellen Versuchsplänen durch Funktion der beiden Faktoren als wechselseitige Kontrollfaktoren, s.u.).

Bei einem **Querschnittsequenzplan** (bei SCHAIE Testzeit-Sequenz-Methode, bei BALTES Querschnitt-Sequenz) werden mehrere Altersstufen zu mehreren aufeinanderfolgenden Testzeiten (SCHAIE) bzw. über mehrere Kohorten (BALTES) untersucht, d.h., Testzeit bzw. Kohorte und Alter der Vpn werden systematisch variiert. Es handelt sich also um einen Plan **VPL 2 QQ** mit hypothesenrelevanter UV B „Alter“ und Kontrollfaktor UV A „Testzeit“ bzw. „Kohorte“, dessen Stufen als systematische Replikationsstudien angesehen werden können: Für den gleichen Altersbereich werden für verschiedene Testzeitpunkte bzw. Kohorten Daten im Querschnitt erhoben (sukzessive Durchführung mehrerer Querschnittuntersuchungen → Sequenzen von Querschnittuntersuchungen), wobei der Vergleich der Daten der verschiedenen Testzeiten bzw. Kohorten (Testzeitdifferenzen bzw. Kohortendifferenzen) der Kontrolle der Altersdifferenzen dient (vgl. TRAUTNER, S. 262, S. 271 f); **die verschiedenen Testzeiten bzw. Kohorten stellen also gegenseitig Kontrollgruppen dar.**

Dieser Plan erlaubt Aussagen über a) die durchschnittlichen **Altersdifferenzen** für die untersuchten Testzeiten bzw. Kohorten und b) die durchschnittlichen **Testzeit- bzw. Kohortendifferenzen** für die untersuchten Altersstufen. Altersdifferenzen und Testzeit- bzw. Kohortendifferenzen sind hier also **gegenseitig kontrolliert** (vgl. TRAUTNER, S. 262) (vgl.

Vorteil von zweifaktoriellen Versuchsplänen durch Funktion der beiden Faktoren als wechselseitige Kontrollfaktoren, s.u.).

In den Sequenzplänen ist also der hypothesenrelevante Faktor (UV B) immer, wie bei den konventionellen Längs- bzw. Querschnittplänen auch, das Alter. Als zweiter Faktor (UV A) kommt ein **Kontrollfaktor** dazu, dessen Stufen als systematische Replikationsstudien angesehen werden können.

Merksatz: Ein **Kontrollfaktor** ist eine personengebundene Störvariable, deren Bedeutung für die abhängige Variable durch Berücksichtigung als gesonderter Faktor (UV A) neben dem eigentlich interessierenden gruppenkonstituierenden Faktor (hypothesenrelevante UV B) in mehrfaktoriellen Plänen kontrolliert wird. (vgl. BORTZ & DÖRING, S. 492, S. 499 ff)

Systematische Variation: Eine potentielle Störvariable wird dadurch kontrolliert, daß sie zu einer weiteren UV gemacht wird (Kontrollfaktor). Diese Technik wird für alle allgemeinen Störeffekte – also auch für Probandenmerkmale – eingesetzt (vgl. Script **Hussy** zur Versuchsplanung, 1999, S. 48).

Die Kontrolle einer potentiellen Störvariable durch systematische Variation hat **vier Vorteile**:

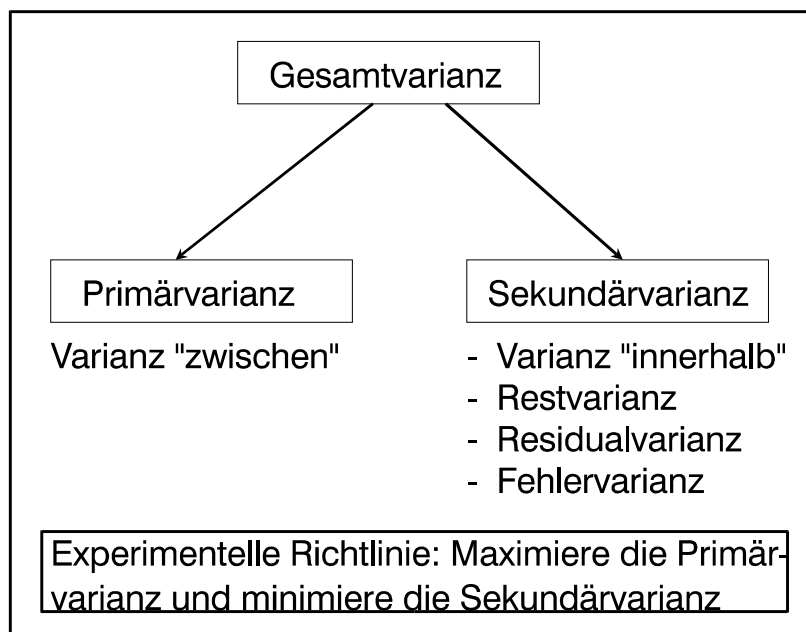
1. Die interne Validität wird erhöht, da eine Einflußgröße kontrolliert wird.
2. Der Anteil der Primärvarianz vergrößert sich, da die Variable zu einer UV wird.
3. Die Sekundärvarianz verringert sich, da der entsprechende Fehleranteil wegfällt.
4. Man erhält Informationen über die Wirksamkeit der kontrollierten Variable.

(vgl. Handout **Hussy** zu „Störvariablen und ihre Kontrolle“ zur Vorlesung Versuchsplanung im SS 2003)

Zweifaktorielle Untersuchungspläne überprüfen simultan drei verschiedene Unterschiedshypothesen: zwei Haupteffekthypothesen und eine Interaktionshypothese. Diese drei Hypothesen müssen jedoch nicht immer explizit formuliert sein. Häufig steht nur eine Hypothese im Vordergrund (z.B. die Wirkung einer Intervention), und der zweite Faktor wird nur zu Kontrollzwecken eingeführt. Die Einführung eines Kontrollfaktors ist immer dann sinnvoll, wenn die gruppenkonstituierende unabhängige Variable sehr wahrscheinlich durch ein anderes personengebundenes Merkmal überlagert ist (Konfundierung), das ebenfalls als Erklärung für den gefundenen Gruppenunterschied in der abhängigen Variable in Frage kommt. Die Forschungshypothese bezieht sich dabei auf den anderen Faktor, die eigentlich interessierende unabhängige Variable (hypothesenrelevanter Faktor).

Mit der Berücksichtigung eines (zweiten) **Kontrollfaktors**, der die Untersuchungsteilnehmer gemäß ihrer Ausprägung in diesem Merkmal in **homogene Teilgruppen (Blöcke)** einteilt, wird derjenige Varianzanteil der abhängigen Variable, der auf den Kontrollfaktor bzw. auf die Interaktion von hypothesenrelevantem Faktor und Kontrollfaktor zurückgeht (Sekundärvarianz/Varianz „innerhalb“), varianzanalytisch bestimmbar, und die zwischen den Gruppen registrierten Unterschiede in der AV (Primärvarianz/Varianz „zwischen“) sind unabhängig vom Kontrollfaktor. (vgl. BORTZ & DÖRING, S. 499 ff)

Die folgende Abbildung verdeutlicht das **Prinzip der Varianzaufteilung** in einem **einfaktoriellen Plan** (vgl. Script **Hussy** zur Versuchsplanung, S. 34):

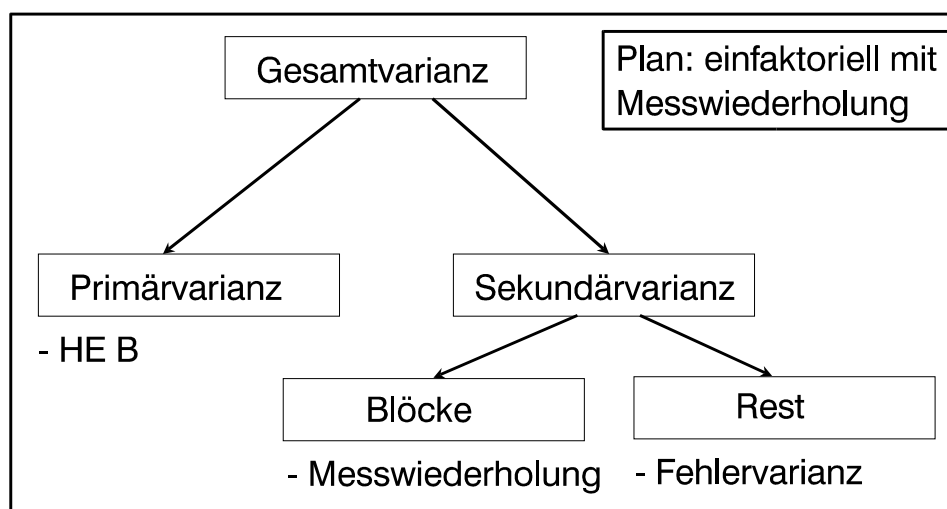


Die **Gesamtvarianz** ergibt sich aus den unterschiedlichen Werten in der AV aller Vpn, die an der Untersuchung teilgenommen haben. Ein Teil davon ergibt sich aus den verschiedenen Stufen der UV; dieser Teil der Varianz wird **Primärvarianz** genannt, da er auf die experimentellen Bedingungen zurückzuführen und somit erwünscht ist. Die alternative Bezeichnung „Varianz zwischen“ bezieht sich auf die Stufen der UV: je größer der Unterschied zwischen den experimentellen Bedingungen ist, desto größer ist diese Varianz. Die **Sekundärvarianz** kommt durch weitere Einflußgrößen (Störvariablen) innerhalb der beiden Gruppen von Vpn zustande (Varianz „innerhalb“). Es handelt sich dabei also um den nicht erklärten Rest der Gesamtvarianz, der deshalb auch Restvarianz, Residualvarianz oder Fehlervarianz genannt wird.

Durch unterschiedliche Arten der Operationalisierung der UV und AV nimmt man Einfluß auf die Primärvarianz. Mit der Kontrolle von Störvariablen und der Auswahl und Zuteilung der Vpn nimmt man Einfluß auf die Sekundärvarianz. Die experimentelle Richtlinie muß deshalb lauten: Maximiere die Primärvarianz und minimiere die Sekundärvarianz.

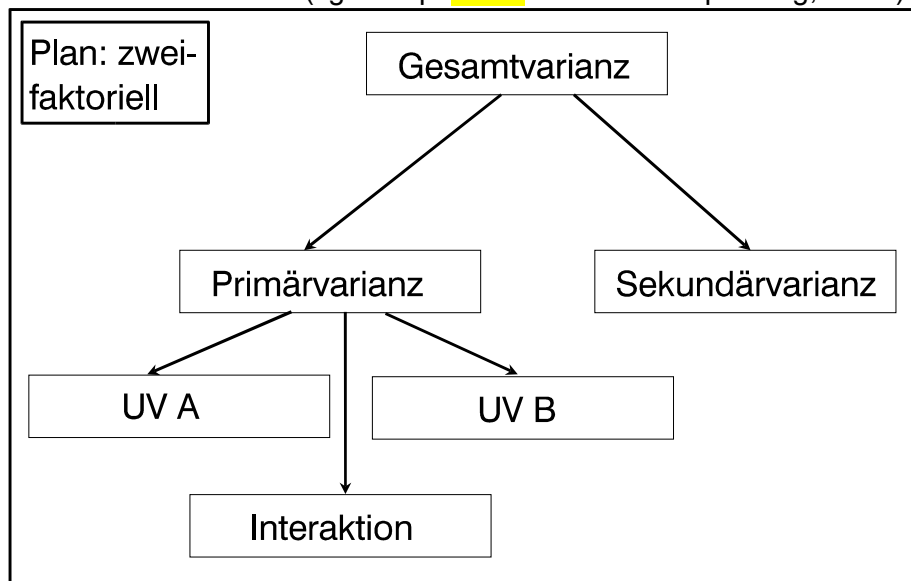
Die Ergebnistabelle der Varianzanalyse spiegelt auch das erläuterte Prinzip der Varianzaufteilung wider. Der Haupteffekt der UV entspricht der Primärvarianz, der Fehleranteil der Sekundärvarianz. Dabei wird auch deutlich, daß der F-Wert durch eine Division der MQS (mittlere Quadratsumme) der Primärvarianz mit der MQS der Sekundärvarianz ermittelt wird: je größer die Primär- und je kleiner die Sekundärvarianz, desto größer der F-Wert. Es ist die Höhe des F-Werts, die darüber entscheidet, ob ein gefundener Mittelwertsunterschied signifikant (überzufällig) ist oder nicht. Je höher der F-Wert, desto größer die Wahrscheinlichkeit für ein signifikantes Ergebnis. Ist der F-Wert signifikant, dann wird die H_0 abgelehnt, und die H_1 gilt als statistisch nachgewiesen. Ist der Mittelwertsunterschied nicht signifikant, dann wird die H_0 beibehalten, und die H_1 gilt als nicht statistisch nachgewiesen.

Die folgende Abbildung verdeutlicht das **Prinzip der Varianzaufteilung** in einem **einfaktoriellen Plan mit Meßwiederholung** (vgl. Script **Hussy** zur Versuchsplanung, S. 61):



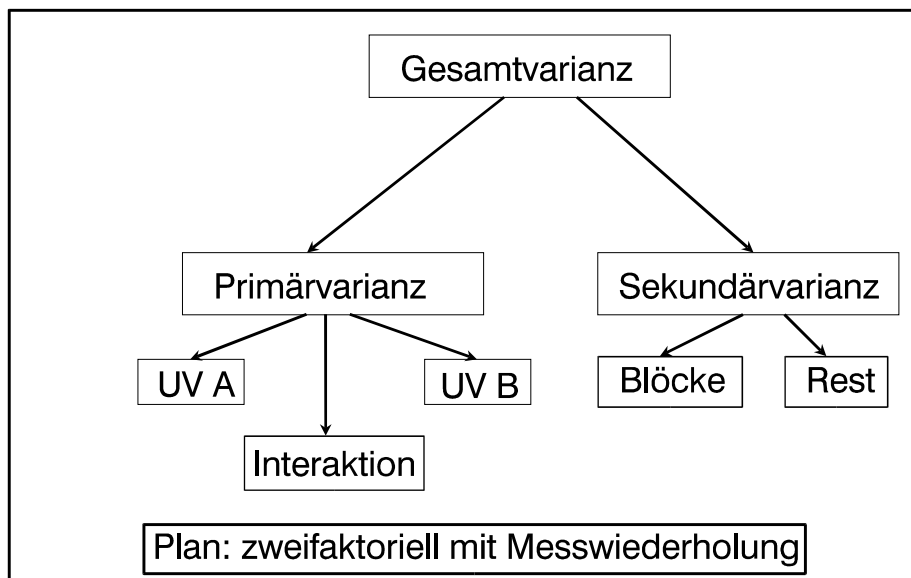
Die Unterschiede zwischen einem einfaktoriellen Plan und einem einfaktoriellen Plan mit Meßwiederholung beziehen sich lediglich auf die Fehlervarianz. Im einfaktoriellen Plan mit Meßwiederholung wird die **Fehlervarianz in zwei Quellen zerlegt (Blöcke und Rest)**. Die Varianzquelle „**Blöcke**“ bezieht sich dabei auf die **wiederholt beobachteten Vpn**, und nur mit der MQS der **Restvarianz** wird der **F-Wert des Haupteffekts der UV B (HE B)** bestimmt. Im Vergleich zum einfaktoriellen Plan ist der **F-Wert höher**, wodurch sich eine **höhere Wahrscheinlichkeit für einen überzufälligen Mittelwertsunterschied** ergibt → Minimierung der Sekundärvarianz und damit erhöhte Präzision der Untersuchung (vgl. Script **Hussy** zur Versuchsplanung, S. 74).

Die folgende Abbildung verdeutlicht das **Prinzip der Varianzaufteilung in einem zweifaktoriellen Plan** (vgl. Script **Hussy** zur Versuchsplanung, S. 65):



In zweifaktoriellen Plänen setzt sich die Primärvarianz aus dem Haupteffekt der UV A (HE A), dem Haupteffekt der UV B (HE B) und der Interaktion zusammen.

Die folgende Abbildung verdeutlicht das **Prinzip der Varianzaufteilung in einem zweifaktoriellen Plan mit Meßwiederholung** (vgl. Script **Hussy** zur Versuchsplanung, S. 66):

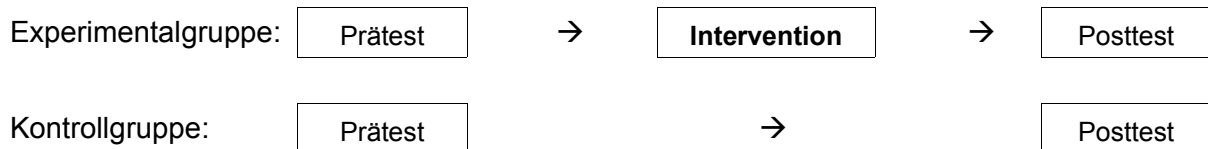


Aus der varianzanalytischen Ergebnistabelle, die die Meßwiederholung auf der UV B berücksichtigt, ist zu entnehmen, daß die Sekundärvarianz zusätzlich in die Anteile „Blöcke“ und „Rest“ zerlegt wird. Auch die Abbildung verdeutlicht diesen Vorgang anschaulich. Mit der **Sekundärvarianz, die auf die Blöcke zurückgeht**, wird sowohl der **Effekt der meßwiederholten UV B** als auch die **Interaktion** auf Signifikanz überprüft. Die **restliche Sekundärvarianz** dient der Überprüfung der **UV A**. Aus der Tabelle geht auch klar hervor, daß sich an den absoluten Werten der QS und FG für die Primär-, Sekundär- und Gesamtvarianz durch die Meßwiederholung (im Vergleich zum zweifaktoriellen Plan) nichts verändert hat, da der Berechnung der gleiche Datensatz zugrunde liegt. → Minimierung der Sekundärvarianz durch Blöcke und damit erhöhte Präzision der Untersuchung (vgl. Script **Hussy** zur Versuchsplanung, S. 74).

Drei Vorteile der Verwendung einer **zweifaktoriellen Versuchsplananlage VPL-A2** bei der Untersuchung von mehreren UVn anstelle der mehrfachen Benutzung der einfaktoriellen Versuchsplananlage VPL-A1:

- 1) Bei einfaktorieller Vorgehensweise können keine Bedingungskombinationen untersucht werden, sondern lediglich Haupteffekte (HE). Bedingungskombinationen geben dabei an, wie es sich auswirkt, wenn je eine Stufe der betrachteten Faktoren gleichzeitig an denselben Pbn realisiert wird, und diese Information ist sehr häufig von großem Interesse und/ oder großer Bedeutung (Interaktion).
- 2) Bei einer zweifaktoriellen Versuchsplananlage **können die beiden Faktoren wechselseitig als Kontrollfaktoren fungieren**: Betrachtet man bei zwei Faktoren zunächst Faktor B als hypothesenrelevant, fungiert aus dieser Perspektive der Faktor A als Kontrollfaktor; aus der Perspektive von Faktor A fungiert dagegen der Faktor B als Kontrollfaktor. Die UV B hat also gleichsam eine Doppelfunktion. Die Funktion der beiden Faktoren als Kontrollfaktor für den jeweils anderen Faktor ist unabhängig davon, ob sie beide der Prüfung einer (oder mehrerer) Hypothesen dienen oder ob nur einer direkt auf eine Hypothese bezogen ist.
- 3) Bei einer zwei-/mehrfaktoriellen Versuchsplananlage werden relativ weniger Pbn benötigt (Ökonomieaspekt), denn jeder Pb erbringt einen Meßwert sowohl unter einer Stufe von B als auch gleichzeitig einen unter einer Stufe von A usw. Es werden dann - in der zweifaktoriellen Versuchsplananlage - pro Kombination von Bedingungen (AB_{jk}) $n_{jk} = n$ Pbn untersucht, und damit resultieren pro Spalte jeweils $J \cdot n$ Pbn, während pro Zeile $K \cdot n$ Pbn zur Verfügung stehen. (vgl. Script **Hussy** zur Versuchsplanung, S. 63)

Im Sinn von HAGER (1987, in LÜER (Hrsg.), zit. n. Script **Hussy** zur Versuchsplanung, S. 74) repräsentiert z.B. der Plan **VPL 2 QQ(W)** – also ein **Längsschnittsequenzplan**, s.o. - einen **Prätest-Posttest-Kontrollgruppen-Plan/nichtäquivalenten Kontrollgruppenplan**. Prätest-Posttest-Kontrollgruppen-Pläne implizieren die Existenz einer Experimental- und einer Kontrollgruppe, die beide vor und nach der experimentellen Behandlung (Intervention) hinsichtlich der AV beobachtet werden. Schematisch:



In diesem zweifaktoriellen quasiexperimentellen Plan könnte die **UV A** für die **Experimental- und Kontrollgruppe stehen, denen die Vpn nicht zufällig zugeteilt werden**, während die **UV B** durch die **Vor- und Nachtestmessung** entsteht (**Meßwiederholung**). Zwischen Vor- und Nachtest liegt für die Experimentalgruppe die Intervention, die Kontrollgruppe erfährt keine gezielte Beeinflussung. In der **Entwicklungspsychologie** wird bei Hypothesen über Veränderung in Abhängigkeit vom **Alter** (UV B, hypothesenrelevanter Faktor) unter einer Intervention das Verstreichen eines **Zeitraums** (ohne spezifische Maßnahme) verstanden. Die Experimentalgruppe besteht hier also aus einer Stichprobe, die vor und nach dem Verstreichen eines bestimmten Zeitraums – in unterschiedlichen Altersstufen – untersucht wird. Im Fall eines Längsschnittsequenzplans besteht die Experimentalgruppe aus einer Stichprobe von Personen, die alle derselben Kohorte (Zeitraum, in dem eine Population geboren wurde) angehören, während die Kontrollgruppe aus Personen besteht, die alle derselben anderen Kohorte angehören. Beide Gruppen werden in denselben Altersstufen – vor und nach dem Verstreichen desselben Zeitraums - untersucht und ihre Ergebnisse verglichen, um Aussagen über von Kohortenunterschieden bereinigte altersbedingte Veränderungen zu machen.

Versuchsplan VPL 2 QQ(W):

		UV B: Meßzeitpunkt	
		B ₁ vorher	B ₂ nachher
UV A: Kohorte	A ₁	AV: A ₁ B ₁	AV: A ₁ B ₂
	A ₂	AV: A ₂ B ₁	AV: A ₂ B ₂

Frage 36: Welches sind die **prinzipiellen Unterschiede** von **Sequenzplänen** zu den **konventionellen entwicklungspsychologischen Plänen**? (2 x gefragt)

- 1.
- 2.
- 3.
- 4.

siehe Frage 35.

Frage 37: Welchem **Ziel** dient die **Hinzunahme** eines **Kontrollfaktors** bei **SCHAIÉ** bzw. bei **BALTES** (mit kurzer Begründung)? (1 x gefragt)

Aus der angegebenen Literatur zum Thema/Referat „Abbildung von Entwicklungsvorgängen“ (TRAUTNER, 1992, Bd. 1, Kap. 4.2, S. 244-278) kann man die Antwort auf diese Frage nicht direkt entnehmen – der Begriff „Kontrollfaktor“ kommt im Text nicht vor! Hier der Versuch, aus dem Text eine Antwort abzuleiten.

Ausgangspunkt für die Erarbeitung ihrer Allgemeinen Entwicklungsmodelle und der daraus abgeleiteten Sequenzpläne waren für SCHAIÉ und für BALTES die Mängel der konventionellen Untersuchungspläne, insbesondere die Konfundierung von Altersunterschieden und Kohortenunterschieden bei der Querschnittmethode und die Konfundierung von Altersunterschieden und Testzeitunterschieden bei der Längsschnittmethode (vgl. TRAUTNER, S. 258). Die Sequenzpläne sollen also mit der Hinzunahme von jeweils einem Kontrollfaktor neben dem hypothesenrelevanten Faktor „Alter“ diese **Konfundierungen** und damit **Störungen der internen Validität** der Untersuchungen **vermeiden**.

Um die einfache Beziehung zwischen dem Lebensalter und den über das Alter beobachtbaren Veränderungen mit dem Ziel der Vermeidung von Konfundierungen zu erweitern und damit eine Präzisierung des Begriffs „Lebensalter“ zu erreichen, betrachtet SCHAIÉ Entwicklung/Veränderung (V) als eine Funktion von Alter (A), Kohorte (K) und Testzeit (T): $V = f(A, K, T)$. Diese drei Komponenten seines Entwicklungsmodells definiert er dabei folgendermaßen:

- **Alter** = Anzahl der Zeitintervalle (Jahre, Monate etc.) zwischen der Geburt eines Individuums und dem Zeitpunkt der Messung;
- **Kohorte** = alle Individuen, die in einem bestimmten Zeitraum (z.B. in einem kalendarischen Jahr) geboren sind;
- **Testzeit** = Zeitpunkt, an dem die Individuen untersucht werden (Tag/Monat/Jahr).

SCHAIÉ betrachtet die Beziehungen zwischen Alters-, Kohorten- und Testzeitunterschieden in einem **erklärenden** Sinn als **Komponenteneffekte** und versucht daher, die Bedeutung der drei Komponenten im Sinn von **Entwicklungsbedingungen/Entwicklungsfaktoren** inhaltlich zu interpretieren:

- **Alterseffekte** = Ausdruck individueller neurophysiologischer Reifungsprozesse der Organismen, die während des Untersuchungszeitraums, d.h. in der untersuchten Altersspanne, aufgetreten sind;
- **Kohorteneffekte** = Ausdruck unterschiedlicher allgemeiner Umweltbedingungen vor dem ersten Testzeitpunkt und/oder genetischer Unterschiede zwischen den Kohorten;
- **Testzeiteffekte** = Ausdruck von für alle Organismen gemeinsamen Umweltbedingungen oder von allgemeinen Veränderungen der Umwelt der Organismen: Ausdruck kultureller Wandlungsprozesse in dem betreffenden historischen Zeitabschnitt, von denen alle Individuen gleichermaßen betroffen werden (vgl. TRAUTNER, S. 260, S. 269).

Dementsprechend hat SCHAIÉ die aus seinem dreifaktoriellen Entwicklungsmodell abgeleiteten Sequenzpläne nicht als Untersuchungspläne für die *Beschreibung von Entwicklungsver-*

läufen konzipiert, sondern als Versuchspläne mit dem **primären Ziel der Überprüfung seiner spezifischen Erklärungsmodelle** (vgl. TRAUTNER, S. 260, S. 274).

BALTES geht dagegen davon aus, daß sich Alter und Testzeit für jede einzelne Kohorte meßtechnisch auf den gleichen Abschnitt des Zeitkontinuums beziehen und damit keine inhaltlich verschiedene Bedeutung haben (vgl. PETERMANN, 1978, S. 49), so daß eine der beiden Variablen zu dessen Kennzeichnung ausreicht. Da in der Entwicklungspsychologie dem Alter als Dimension der zeitlichen Einteilung ontogenetischer Veränderungen im Vergleich zur Testzeit die größere Bedeutung zukommt, entscheidet BALTES sich für die Beibehaltung der Komponente Alter. Allerdings sind auch in seinem allgemeinen Entwicklungsmodell alle drei Komponenten Alter, Kohorte und Testzeit enthalten; nur wird die Testzeit nicht als eigenständiger Faktor verwendet, sondern mit der Kohorte zusammengefaßt (K'). Er präzisiert also den Begriff „Lebensalter“, indem er Entwicklung/Veränderung (V) als eine Funktion von Alter (A) und Kohorte (K') betrachtet: $V = f(A, K')$ (vgl. PETERMANN, 1978, S. 49).

BALTES sieht die inhaltliche Interpretation der drei Komponenten Alter, Kohorte und Testzeit durch SCHAIE im Sinn von Entwicklungsfaktoren (Erklärungskonstrukten) als unangemessen und spekulativ an und spricht den Sequenzplänen lediglich einen **deskriptiven** Charakter zu. Dementsprechend dienen die aus seinem zweifaktoriellen Entwicklungsmodell abgeleiteten Sequenzpläne als Untersuchungspläne mit dem **Ziel der Beschreibung von Entwicklungsverläufen** (vgl. TRAUTNER, S. 269, S. 274).

„Allerdings können die Ergebnisse einer sequentiellen Untersuchung Hinweise auf die Art der wirksamen Entwicklungsfaktoren geben und zur Bildung gezielter Hypothesen führen.“ (TRAUTNER, S. 269). Die Sequenzpläne von SCHAIE und von BALTES stellen einen „wesentlichen methodischen Fortschritt bei der Erforschung von Entwicklungsvorgängen dar:

- Mit ihrer Hilfe lassen sich unter definierten Voraussetzungen bis zu einem bestimmten Grad die sonst unausweichlich auftretenden Konfundierungen zwischen Alters-, Kohorten- und Testzeitdifferenzen vermeiden oder zumindest kontrollieren.
- Gleichzeitig erlauben sie eine gemeinsame Analyse von Alterseffekten und Kohorten – bzw. Testzeiteffekten und damit eine ansatzweise ‚reine‘ Erfassung (Deskription) von Altersdifferenzen.“ (TRAUTNER, S. 276).

Die verschiedenen Sequenzpläne sind daher geeignet zur **Beschreibung von Entwicklung** und zur **Hypothesengenerierung über mögliche Entwicklungsdeterminanten** (vgl. TRAUTNER, S. 276).

Hier folgt eine Antwort auf die Frage nach der Zielsetzung bei der Hinzunahme von Kontrollfaktoren anhand des auf der Homepage des Lehrstuhls von Prof. HUSSY unter „Veranstaltungsunterlagen“ veröffentlichten Handouts „Entwicklungspsychologische Untersuchungspläne: Konventionelle und sequentielle Designs“ aus dem Sommersemester 2002 bzw. des im Handapparat in der Bibliothek als Kopiervorlage abgehefteten Handouts vom SS 2000. In beiden kommt der Begriff „Kontrollfaktor“ explizit vor, allerdings fehlt jeglicher Literaturnachweis; die für das Referat zum Thema „Abbildung von Entwicklungsvorgängen“ angegebene Literatur war, wie auch im SS 1998, TRAUTNER, 1992, Bd. 1, Kap. 4.2, S. 244-278.

Merksatz: Ein **Kontrollfaktor** ist eine personengebundene Störvariable, deren Bedeutung für die abhängige Variable durch Berücksichtigung als gesonderter Faktor (UV A) neben dem eigentlich interessierenden gruppenkonstituierenden Faktor (hypothesenrelevante UV B) in mehrfaktoriellen Plänen kontrolliert wird. (vgl. BORTZ & DÖRING, S. 492, S. 499 ff)

Bei SCHAIE dient die Hinzunahme des Kontrollfaktors der **Prüfung des Anwendungsbereichs der Hypothese**, also der **externen Validität**, während sie bei BALTES die **Erhöhung der internen Validität der Ergebnisse** zum Ziel hat. Die Begründung erfolgt hinsichtlich der auf die konventionellen Längs- und Querschnittpläne bezogenen sequentiellen Varianten von SCHAIE und von BALTES.

Konventioneller Längsschnittplan VPL 1 Q (W): quasi-experimenteller Plan aufgrund der Meßwiederholung in der UV Alter (Sequenzeffekte sind nicht kontrollierbar); Störung der in-

ternen Validität der Ergebnisse möglich (z.B. durch Testzeiteffekte); Minimierung der Sekundärvarianz durch Blöcke (Meßwiederholung) und dadurch erhöhte Präzision

UV B Alter (hypothesenrelevanter Faktor)			
B ₁	B ₂	B _k
1	1	1
2	2	2
.....
12	12	12

In einer **Längsschnittstudie** wird eine Stichprobe zu verschiedenen Zeitpunkten mit demselben oder einem vergleichbaren Meßinstrument untersucht. Die gleiche Stichprobe wird also mehrfach beobachtet.

Konventioneller Querschnittplan VPL 1 Q: quasi-experimenteller Plan aufgrund der organismischen UV Alter; Störung der internen Validität der Ergebnisse möglich (z.B. durch Kohorteneffekte); keine Minimierung der Sekundärvarianz durch Blöcke und dadurch verringerte Präzision

UV B Alter (hypothesenrelevanter Faktor)			
B ₁	B ₂	B _k
1	13	40
2	14	41
.....
12	25	50

In einer **Querschnittsstudie** werden Stichproben aus verschiedenen Altersgruppen mit demselben oder einem vergleichbaren Meßinstrument an einem Zeitpunkt einmal untersucht.

1. Zu SCHAIES Sequenzplänen

1) Kohortensequenzplan VPL2QQ (W)

Bei einem Kohortensequenzplan (Kohorten-Sequenz-Methode, **Längsschnittsequenzplan**) werden mehrere Kohorten über mehrere aufeinanderfolgende Altersstufen untersucht, d.h., Kohorte und Alter der Vpn werden systematisch variiert. Es handelt sich also um einen Meßwiederholungsplan **VPL 2 QQ(W)** mit meßwiederholter hypothesenrelevanter UV B „Alter“ und Kontrollfaktor UV A „Kohorte“, dessen Stufen als systematische Replikationsstudien angesehen werden können: Für den gleichen Altersbereich werden für verschiedene Kohorten Daten im Längsschnitt erhoben (sukzessive Durchführung mehrerer Längsschnittuntersuchungen → Sequenzen von Längsschnittuntersuchungen), wobei der Vergleich der Daten der verschiedenen Kohorten (Kohortendifferenzen) der Kontrolle der Altersdifferenzen dient (vgl. TRAUTNER, S. 262, S. 271 f); **die verschiedenen Kohorten stellen also gegenseitig Kontrollgruppen dar.**

Dieser Plan erlaubt Aussagen über a) die durchschnittlichen **Altersdifferenzen** für die untersuchten Kohorten und b) die durchschnittlichen **Kohortendifferenzen** für die untersuchten Altersstufen. Altersdifferenzen und Kohortendifferenzen sind hier also **gegenseitig kontrolliert** (vgl. TRAUTNER, S. 262) (vgl. Vorteil von zweifaktoriellen Versuchsplänen durch Funktion der beiden Faktoren als wechselseitige Kontrollfaktoren, s.o.). Die Testzeitdifferenzen lassen sich nicht eindeutig feststellen. Dieser Plan ermöglicht die Feststellung der Generalisierbarkeit von Altersverläufen über verschiedene Kohorten.

		UV B: Alter (hypothesenrelevanter Faktor)			
		B ₁	B ₂	B _k
UV A : Kohorte (Kontrollfaktor)	A ₁	Vp 1, 2, ... 12	Vp 1, 2, ... 12	Vp 1, 2, ... 12	Vp 1, 2, ... 12
	A ₂	Vp 13, ... 23	Vp 13, ... 23	Vp 13, ... 23	Vp 13, ... 23

	A _j	Vp 40 50	Vp 4050	Vp 4050	Vp 40 50

Es handelt sich um einen zweifaktoriellen quasi-experimentellen Plan (durch organismische UV A und meßwiederholte UV B) mit möglichen Störungen der internen Validität (z. B. durch Testzeiteffekte); **UV A: Kohorte, Kontroll-Faktor**; UV B: Alter, hypothesenrelevanter Faktor.

Der Mangel des Längsschnittplans (Störung der internen Validität durch Konfundierung von Alter und Testzeit) ist somit nicht aufgehoben. Vielmehr wird für die eindeutige Interpretierbarkeit der Ergebnisse dieses Plans eine **Nullinteraktion von Alters- und Kohorteneffekten angenommen** (Fehlen von Testzeiteffekten). Die Hinzunahme des Kohortenfaktors dient also nicht der Kontrolle eines möglichen Testzeiteffekts (interne Validität), sondern der **Prüfung des Anwendungsbereichs der Hypothese (externe Validität)**. Die Stufen der UV A repräsentieren im engeren Sinn systematische Replikationsstudien.

2) Testzeitsequenzplan VPL2QQ

Bei einem Testzeitsequenzplan (Testzeit-Sequenz-Methode → **Querschnittsequenzplan**) werden mehrere Altersstufen zu mehreren aufeinanderfolgenden Testzeiten untersucht, d.h., Testzeit und Alter der Vpn werden systematisch variiert. Es handelt sich also um einen Plan **VPL 2 QQ** mit hypothesenrelevanter UV B „Alter“ und Kontrollfaktor UV A „Testzeit“, dessen Stufen als systematische Replikationsstudien angesehen werden können: Für den gleichen Altersbereich werden für verschiedene Testzeitpunkte Daten im Querschnitt erhoben (sukzessive Durchführung mehrerer Querschnittuntersuchungen → Sequenzen von Querschnittuntersuchungen), wobei der Vergleich der Daten der verschiedenen Testzeiten (Testzeitdifferenzen) der Kontrolle der Altersdifferenzen dient (vgl. TRAUTNER, S. 262, S. 271 f); **die verschiedenen Testzeiten bzw. Kohorten stellen also gegenseitig Kontrollgruppen dar**. Dieser Plan erlaubt Aussagen über a) die durchschnittlichen **Altersdifferenzen** für die untersuchten Testzeiten und b) die durchschnittlichen **Testzeitdifferenzen** für die untersuchten Altersstufen. Altersdifferenzen und Testzeitdifferenzen sind hier also **gegenseitig kontrolliert** (vgl. TRAUTNER, S. 262) (vgl. Vorteil von zweifaktoriellen Versuchsplänen durch Funktion der beiden Faktoren als wechselseitige Kontrollfaktoren, s.o.). Die Kohortendifferenzen lassen sich nicht eindeutig feststellen. Dieser Plan ermöglicht die Feststellung der Generalisierbarkeit von Altersunterschieden über verschiedene Testzeiten.

		UV B: Alter (hypothesenrelevanter Faktor)			
		B ₁	B ₂	B _k
UV A : Testzeit (Kontrollfaktor)	A ₁	Vp 1, 2,12	Vp 51,61	Vp 1, 2, 12	Vp 154,162
	A ₂	Vp 13, 23	Vp 62, 72	Vp 13,23	Vp 163,176

	A _j	Vp 40 50	Vp 88102	Vp 4050	Vp 191201

Es handelt sich um einen zweifaktoriellen quasi-experimentellen Plan (organismische UV A und UV B) mit möglichen Störungen der internen Validität (z. B. Kohorteneffekte); **UV A: Testzeit, Kontroll-Faktor**; UV B: Alter, hypothesenrelevanter Faktor.

Der Mangel des Querschnittplans (Störung der internen Validität durch Konfundierung von Alter und Kohorte) ist somit nicht aufgehoben. Vielmehr wird für die eindeutige Interpretierbarkeit der Ergebnisse dieses Plans eine **Nullinteraktion von Alters- und Testzeiteffekten angenommen** (Fehlen von Kohorteneffekten). Die Hinzunahme des Testzeitfaktors dient also nicht der Kontrolle eines möglichen Kohorteneffekts (interne Validität), sondern der **Prüfung des Anwendungsbereichs der Hypothese (externe Validität)**. Die Stufen der UV A repräsentieren im engeren Sinn systematische Replikationsstudien.

2. Zu BALTES' Sequenzplänen

1) Längsschnittsequenzplan VPL2QQ(W)

Der **Längsschnittsequenzplan** von BALTES ist identisch mit dem Kohortensequenzplan bei SCHAIE (vgl. TRAUTNER, S. 271).

		UV B: Alter (hypothesenrelevanter Faktor)			
		B ₁	B ₂	B _k
UV A : Kohorte (Kontrollfaktor)	A ₁	Vp 1, 2, ... 12	Vp 1, 2, ... 12	Vp 1, 2, 12	Vp 1, 2, 12
	A ₂	Vp 13, 23	Vp 13, 23	Vp 13, 23	Vp 13, 23

	A _j	Vp 40 50	Vp 4050	Vp 4050	Vp 40 50

Es handelt sich um einen zweifaktoriellen quasi-experimentellen Plan (organismische UV A und meßwiederholte UV B); **UV A: Kohorte, Kontroll-Faktor**; UV B: Alter, hypothesenrelevanter Faktor.

Der Mangel des Längsschnittplans (Störung der internen Validität durch Konfundierung von Alter und Testzeit) ist somit aufgehoben (Kohorte und Testzeit werden nicht unterschieden). Die Hinzunahme des Kohortenfaktors dient also der Kontrolle eines möglichen Testzeiteffekts und damit der **Erhöhung der internen Validität** der Ergebnisse. Durch zusätzliche systematische Replikationsstudien ist der Anwendungsbereich der Hypothese (externe Validität) zu prüfen.

2) Querschnittsequenzplan VPL2QQ

Der **Querschnittsequenzplan** von BALTES entspricht dem Testzeitsequenzplan bei SCHAIE (vgl. TRAUTNER, S. 271). → Als Kontrollfaktor statt Testzeit hier Kohorte!

		UV B: Alter (hypothesenrelevanter Faktor)			
		B ₁	B ₂	B _k
UV A : Kohorte (Kontrollfaktor)	A ₁	Vp 1, 2,12	Vp 51,61	Vp 1, 2, 12	Vp 154, ... 162
	A ₂	Vp 13, 23	Vp 62, 72	Vp 13,23	Vp 163,176

	A _j	Vp 40 50	Vp 88102	Vp 4050	Vp 191201

Es handelt sich um einen zweifaktoriellen quasi-experimentellen Plan (organismische UV A und UV B); **UV A: Kohorte, Kontroll-Faktor**; UV B: Alter, hypothesenrelevanter Faktor.

Der Mangel des Querschnittplans (Störung der internen Validität durch Konfundierung von Alter und Kohorte) ist somit aufgehoben (Kohorte und Testzeit werden nicht unterschieden). Die Hinzunahme des Kohortenfaktors dient also der Kontrolle eines möglichen Kohorteneffekts und damit der **Erhöhung der internen Validität** der Ergebnisse. Durch zusätzliche systematische Replikationsstudien ist der Anwendungsbereich der Hypothese (externe Validität) zu prüfen.

Frage 38: Wie versucht SCHAIE den Mangel des **klassischen Längsschnittplans** zu eliminieren? Welche **Konsequenzen** hat dieses Vorgehen? (1 x gefragt) bzw. Diskutieren Sie **Vor- und Nachteile** des Kohortensequenzplans von SCHAIE im **Vergleich** zum **klassischen Längsschnittplan!** (1 x gefragt)

Ausgehend von den Mängeln der konventionellen Stichprobenpläne hat SCHAIE einen Versuch unternommen, die Untersuchung der eigentlichen entwicklungspsychologischen Fragestellung mit Hilfe anderer Stichprobenpläne zu ermöglichen. Zu diesem Zweck entwickelte er ein erweitertes **dreifaktorielles Entwicklungsmodell**. Entwicklungsvorgänge sollen seiner Ansicht nach grundsätzlich in ihrer Abhängigkeit vom Alter der untersuchten Individuen, der Kohorte, der diese Individuen angehören, und dem Zeitpunkt der Untersuchung der Individuen betrachtet werden. Er faßt also **Veränderung** auf als Funktion der drei Faktoren **Alter, Kohorte** und **Testzeit**: $V = f(A, K, T)$.

Innerhalb dieses allgemeinen Entwicklungsmodells können die konventionellen Untersuchungspläne (Querschnitt-, Längsschnitt- und Zeitwandeluntersuchung) als Sonderfälle betrachtet werden. Die mit ihnen zu erzielenden Ergebnisse können nicht als reine Effekte von jeweils einer der drei Komponenten interpretiert werden, da die einzelnen Komponenteneffekte jeweils nur unabhängig von einem der beiden übrigen bestimmt werden können und mit dem jeweils anderen konfundiert sind. Für die Ergebnisse aus **klassischen Längsschnittuntersuchungen** bedeutet dies, daß sich die gefundenen Unterschiede in den Al-

tersmittelwerten aus Altersdifferenzen und Testzeitdifferenzen zusammensetzen, daß also eine **Konfundierung von Alters- und Testzeitdifferenzen** vorliegt.

Daher erweiterte SCHAIE diese drei in seinem Entwicklungsmodell enthaltenen konventionellen Stichprobenpläne zu drei Sequenzmodellen (sequentiellen Strategien der Stichprobenselektion): den Kohortensequenzplan, den Testzeitsequenzplan und den Quersequenzplan. Der Begriff der Sequenz bezieht sich darauf, daß die drei Komponenten Alter, Kohorte und Testzeit jeweils in einer bestimmten Reihenfolge untersucht werden, um sie gegenseitig zu kontrollieren.

Mit dem **Kohortensequenzplan** werden mehrere Kohorten in mehreren aufeinanderfolgenden Altersstufen untersucht (Längsschnittsequenzen). SCHAIE geht davon aus, daß er Aussagen über

- a) die durchschnittlichen Altersdifferenzen für die untersuchten Kohorten und
- b) die durchschnittlichen Kohortendifferenzen für die untersuchten Altersstufen erlaubt und daß Alters- und Kohortendifferenzen gegenseitig kontrolliert sind.

Die Testzeitdifferenzen lassen sich mit diesem Design nicht eindeutig feststellen, da hierzu verschiedene Kohorten gleichen Alters zu gleichen Testzeitpunkten untersucht werden müßten (was nicht möglich ist!). Daher ist das Fehlen von Testzeitdifferenzen bzw. das Fehlen einer Wechselwirkung von Alter und Kohorte (Nullinteraktion) die Voraussetzung für eine eindeutige Interpretation der Ergebnisse (die Testzeit wird also auf Null gesetzt).

SCHAIE empfiehlt die Verwendung des Kohortensequenzplans, wenn es darum geht, die Generalisierbarkeit von Altersverläufen über verschiedene Kohorten festzustellen (vgl. TRAUTNER, S. 262). (Überprüfung des Anwendungsbereichs der Hypothese)

Kohortensequenzplan VPL2QQ (W)

		UV B: Alter (hypothesenrelevanter Faktor)			
		B ₁	B ₂	B _k
UV A : Kohorte (Kontrollfaktor)	A ₁	Vp 1, 2, ... 12	Vp 1, 2, ... 12	Vp 1, 2, 12	Vp 1, 2, 12
	A ₂	Vp 13, 23	Vp 13, 23	Vp 13, 23	Vp 13, 23

	A _j	Vp 40 50	Vp 4050	Vp 4050	Vp 40 50

Es handelt sich um einen zweifaktoriellen quasi-experimentellen Plan (durch organismische UV A und meßwiederholte UV B) mit möglichen Störungen der internen Validität (z. B. durch Testzeiteffekte); **UV A: Kohorte, Kontroll-Faktor**; UV B: Alter, hypothesenrelevanter Faktor.

Der Mangel des Längsschnittplans (Störung der internen Validität durch Konfundierung von Alter und Testzeit) ist somit nicht aufgehoben. Vielmehr wird für die eindeutige Interpretierbarkeit der Ergebnisse dieses Plans eine **Nullinteraktion von Alters- und Kohorteneffekten angenommen** (Fehlen von Testzeiteffekten). Die Hinzunahme des Kohortenfaktors dient also nicht der Kontrolle eines möglichen Testzeiteffekts (interne Validität), sondern der **Prüfung des Anwendungsbereichs der Hypothese (externe Validität)**. Die Stufen der UV A repräsentieren im engeren Sinn systematische Replikationsstudien.

Der **Vorteil** des Kohortensequenzplans von SCHAIE gegenüber dem klassischen Längsschnittplan liegt in der Möglichkeit, Alters- und Kohortendifferenzen gemeinsam zu analysieren, sie darüber zu kontrollieren und so zumindest ansatzweise reine Altersdifferenzen zu beschreiben. Er sollte immer dann angewendet werden, wenn begründete Annahmen bestehen, daß bei der Untersuchung von Entwicklungsphänomenen Kohorteneffekte auftreten werden.

Frage 39: Veranschaulichen Sie den **Längsschnittsequenzplan nach BALTES** graphisch! Welche **Probleme** des Längsschnittplans werden dadurch **korrigiert**, und welche bleiben **bestehen?** (1 x gefragt)

Mit dem Längsschnittsequenzplan werden mehrere Kohorten in mehreren aufeinanderfolgenden Altersstufen untersucht (Längsschnittsequenzen). BALTES geht in seinem zweifaktori-

ellen Entwicklungsmodell davon aus, daß sich Alter und Testzeit für jede einzelne Kohorte meßtechnisch auf den gleichen Abschnitt des Zeitkontinuums beziehen und damit keine inhaltlich verschiedene Bedeutung haben (vgl. PETERMANN, 1978, S. 49), so daß eine der beiden Variablen zu dessen Kennzeichnung ausreicht. Da in der Entwicklungspsychologie dem Alter als Dimension der zeitlichen Einteilung ontogenetischer Veränderungen im Vergleich zur Testzeit die größere Bedeutung zukommt, entscheidet BALTES sich für die Beibehaltung der Komponente Alter. Allerdings sind auch in seinem allgemeinen Entwicklungsmodell alle drei Komponenten Alter, Kohorte und Testzeit enthalten; nur wird die Testzeit nicht als eigenständiger Faktor verwendet, sondern mit der Kohorte zusammengefaßt (K'). Er präzisiert also den Begriff „Lebensalter“, indem er Entwicklung/Veränderung (V) als eine Funktion von Alter (A) und Kohorte (K') betrachtet: $V = f(A, K')$ (vgl. PETERMANN, 1978, S. 49).

Längsschnittsequenzplan VPL2QQ(W) nach BALTES

		UV B: Alter (hypothesenrelevanter Faktor)			
		B ₁	B ₂	B _k
UV A : Kohorte (Kontrollfaktor)	A ₁	Vp 1, 2, ... 12	Vp 1, 2, ... 12	Vp 1, 2, 12	Vp 1, 2, 12
	A ₂	Vp 13, 23	Vp 13, 23	Vp 13, 23	Vp 13, 23

	A _j	Vp 40 50	Vp 4050	Vp 4050	Vp 40 50

Es handelt sich um einen zweifaktoriellen quasi-experimentellen Plan (organismische UV A und meßwiederholte UV B); **UV A: Kohorte, Kontroll-Faktor**; UV B: Alter, hypothesenrelevanter Faktor.

Durch die Sequenz mehrerer Längsschnittuntersuchungen wird das Problem der Störung der internen Validität durch Konfundierung von Alter und Testzeit des klassischen Längsschnittplans behoben (Kohorte und Testzeit werden nicht unterschieden). Die Hinzunahme des Kohortenfaktors dient also der Kontrolle eines möglichen Testzeiteffekts und damit der **Erhöhung der internen Validität** der Ergebnisse.

Allerdings ist die Vermeidung der Konfundierung von Alter und Testzeit nur unter der Voraussetzung möglich, daß der Begriff der Kohorte über seine ursprüngliche Bedeutung hinaus (Zeitintervall, in dem eine Population geboren wurde) ausgedehnt wird auf alle für eine bestimmte Kohorte bis zu einem bestimmten Testzeitpunkt aufgetretenen gemeinsamen zeitlichen (lebensgeschichtlichen) Einflüsse. In den Begriff „Kohorte“ werden also die in der historischen Zeit lokalisierbaren Testzeiteffekte mit hineingenommen. In ihrer letzten Konsequenz bedeutet die Hineinnahme der (Test-)Zeitdimension in den Kohortenbegriff, daß Angaben über Entwicklungsprozesse (Altersverläufe) immer nur bezogen auf eine bestimmte Kohorte Gültigkeit beanspruchen können (vgl. TRAUTNER, S. 273 f).

Der Anwendungsbereich der Hypothese muß durch zusätzliche systematische Replikationsstudien überprüft werden.

Frage 40: Worin unterscheiden sich SCHAIE und BALTES in ihren Vorgehensweisen zur **Abbildung** von **entwicklungsbedingten Veränderungen**? (1 x gefragt)

SCHAIE und BALTES unterscheiden sich in ihren Vorgehensweisen zur Abbildung von Entwicklungsvorgängen 1) hinsichtlich des zugrunde gelegten Entwicklungsmodells und 2) in bezug auf die Zielsetzung bei der Hinzunahme des Kontrollfaktors in die Sequenzpläne.

Ausgehend von den Mängeln der konventionellen Stichprobenpläne hat SCHAIE einen Versuch unternommen, die Untersuchung der eigentlichen entwicklungspsychologischen Fragestellung mit Hilfe anderer Stichprobenpläne zu ermöglichen. Zu diesem Zweck entwickelte er ein erweitertes **dreifaktorielles Entwicklungsmodell**. Entwicklungsvorgänge sollen seiner Ansicht nach grundsätzlich in ihrer Abhängigkeit vom Alter der untersuchten Individuen, der Kohorte, der diese Individuen angehören, und dem Zeitpunkt der Untersuchung der Individuen betrachtet werden. Er faßt also **Veränderung** auf als Funktion der drei Faktoren **Alter, Kohorte und Testzeit**: $V = f(A, K, T)$.

Innerhalb dieses allgemeinen Entwicklungsmodells können die konventionellen Untersuchungspläne (Querschnitt-, Längsschnitt- und Zeitwandeluntersuchung) als Sonderfälle betrachtet werden. Die mit ihnen zu erzielenden Ergebnisse können nicht als reine Effekte von jeweils einer der drei Komponenten interpretiert werden, da die einzelnen Komponenteneffekten jeweils nur unabhängig von einem der beiden übrigen bestimmt werden können und mit dem jeweils anderen konfundiert sind. Für die Ergebnisse aus **klassischen Längsschnittuntersuchungen** bedeutet dies, daß sich die gefundenen Unterschiede in den Altersmittelwerten aus Altersdifferenzen und Testzeitdifferenzen zusammensetzen, daß also eine **Konfundierung von Alters- und Testzeitdifferenzen** vorliegt.

Daher erweiterte SCHAIE diese drei in seinem Entwicklungsmodell enthaltenen konventionellen Stichprobenpläne zu **drei Sequenzmodellen** (sequentiellen Strategien der Stichprobenselektion): den **Kohortensequenzplan**, den **Testzeitsequenzplan** und den **Quersequenzplan**. Der Begriff der Sequenz bezieht sich darauf, daß die drei Komponenten Alter, Kohorte und Testzeit jeweils in einer bestimmten Reihenfolge untersucht werden, um sie gegenseitig zu kontrollieren.

BALTES geht in seinem **zweifaktoriellen Entwicklungsmodell**, nach dem **Veränderung** eine Funktion von **Alter** und **Kohorte** ist, davon aus, daß sich Alter und Testzeit für jede einzelne Kohorte meßtechnisch auf den gleichen Abschnitt des Zeitkontinuums beziehen und damit keine inhaltlich verschiedene Bedeutung haben (vgl. PETERMANN, 1978, S. 49), so daß eine der beiden Variablen zu dessen Kennzeichnung ausreicht. Da in der Entwicklungspsychologie dem Alter als Dimension der zeitlichen Einteilung ontogenetischer Veränderungen im Vergleich zur Testzeit die größere Bedeutung zukommt, entscheidet BALTES sich für die Beibehaltung der Komponente Alter. Allerdings sind auch in seinem allgemeinen Entwicklungsmodell alle drei Komponenten Alter, Kohorte und Testzeit enthalten; nur wird die Testzeit nicht als eigenständiger Faktor verwendet, sondern mit der Kohorte zusammengefaßt (K'). Er präzisiert also den Begriff „Lebensalter“, indem er Veränderung (V) als Funktion der zwei Faktoren Alter (A) und Kohorte (K') betrachtet: $V = f(A, K')$ (vgl. PETERMANN, 1978, S. 49).

So gelangt BALTES zu **zwei Sequenzmodellen**, dem **Längsschnittsequenzplan** (ist identisch mit dem Kohortensequenzplan bei SCHAIE) und dem **Querschnittsequenzplan** (entspricht dem Testzeitsequenzplan bei SCHAIE).

Die Sequenzpläne von SCHAIE und von BALTES kann man als quasi-experimentelle **zweifaktorielle Kontrollgruppenpläne** bezeichnen, da die Sequenzen von Längs- bzw. Querschnittuntersuchungen als Untersuchungen mit Experimental- und Kontrollgruppen aufgefaßt werden können. Der hypothesenrelevante Faktor (UV B) ist, wie bei den konventionellen Längs- bzw. Querschnittplänen auch, immer das Alter. Als zweiter Faktor (UV A) kommt ein **Kontrollfaktor** dazu, dessen Stufen als systematische Replikationsstudien angesehen werden können.

Bei SCHAIE dient die Hinzunahme des Kontrollfaktors der **Prüfung des Anwendungsbereichs der Hypothese**, also der **externen Validität**, während sie bei BALTES die **Erhöhung der internen Validität** der Ergebnisse zum Ziel hat.

Die Fragen zur **Kausalanalyse** fehlen.