Veränderungsmessung und Interventionsevaluation

Bodo Krause (Berlin)

1. Einleitung

Gegenstand dieses Beitrags sind Zugänge zur Evaluation von Rehabilitationsmaßnahmen für auffällige Kraftfahrer, die durch gezielte Interventionen (therapeutische Maßnahmen, Schulungsprogramme, Trainingsprogramme, Kurse) zu einem normgerechten Verkehrsverhalten (Wiederherstellung der Kraftfahreignung) geführt werden sollen.
Wesentlich ist in diesem Kontext, dass die gesetzlichen Rahmenbedingungen zunehmend entwickelt wurden und mit FeV § 70 entscheidend die Qualitätssicherung solcher Interventionen und ihre Evaluation festgeschrieben sind.

2. Arten und Ziele der Evaluation

In den letzten Jahren hat sich ein Einvernehmen darüber herausgebildet, Evaluieren „mit der Frage nach der Zielerreichung und der Wirkung unseres Tuns“ (DEZA, 2000) zu verbinden und „vor allem zur Wirkungsbeobachtung mit dem Zweck der Qualitätsentwicklung und –sicherung“ einzusetzen. „Überall da, wo das Wirksam-Werden von Leistungen in hohem Maße angewiesen ist auf das aktive Mitwirken konkreter Personen, liegt der originäre Beitrag der Evaluation zu Bewertung, Verbesserung und Steuerung" (Beywl & Taut, 2000).

Differenzierend kann (nach DEZA, 2000) Evaluation darauf gerichtet sein,

- die Zusammenarbeit insgesamt zu prüfen,
- die Wirkungen, Zielsetzungen oder Effizienz zu prüfen,
- spezifische Fragen, die mit dem Umfeld zusammenhängen, zu prüfen,
- Schlussfolgerungen und Lehren für eine nächste Projektphase zu begründen.

In allen Fragen wird zwischen einer **internen und einer externen Form der Evaluation** unterschieden, wobei die externe Form „von außen“ prüft, ob die angewandten Interventionsmethoden (Kurse, therapeutischen Interventionen) angemessen und tauglich sind (Qualitätssicherung). Externe Evaluatoren sind gleichzeitig ein Bindeglied zwischen den Auftraggebern, Programm- oder Projektautoren sowie Beteiligten und eröffnet einen Meinungsaustausch zur weiteren Qualifizierung der Maßnahmen. Eine Ergänzung durch eine interne (Selbst-) Evaluation ist häufig sinnvoll.

Der Leitfaden für die Planung von Projekt- und Programmevaluation (1997) präzisiert vier elementare Aspekte der Interventionsevaluation:

- **Relevanz,** d.h. ob das Projekt für die Zielgruppe das Richtige (Bedeutsame) tut. Dies beinhaltet auch die Frage nach der Akzeptanz der Maßnahme.
- **Verlauf,** d.h. ob das Projekt das tut, was es beabsichtigt.
- **Wirksamkeit,** d.h. ob die Aktivitäten geeignet sind, das Ziel zu erreichen. Dies kennzeichnet die Effektivität der Maßnahme hinsichtlich der Zielfunktion.
- **Effizienz,** d.h. ob es wirksam ist (Ressourcen effizient einsetzt; Kosten-Nutzen-Aspekt bzgl. der Zielerreichung).

Diesen Aspekten kommt unterschiedliche Bedeutung zu, je nachdem ob der Fokus der Evaluation auf das Ergebnis (Output, **summative Evaluation**) oder auf den Umsetzungsprozess (**formative Evaluation**) gesetzt ist. In einer früheren Arbeit (Metzler, Krause, 1997) unterscheiden wir dabei vier **Phasen der Evaluation:**

Im Sinne einer formativen Evaluation

- **die Phase 1 als Phase der Erkundung und Erprobung (Planungs- und Entwicklungsphase),** in der vor allem Bedarf, Interventionstechnik und Manual abgeklärt werden,
- **die Phase 2 (Pilotphase),** in der der Erfolg der Aktivitäten evaluiert wird

Im Sinne der summativen Evaluation:

- **die Phase 3 (Testphase)** einer kontrollierten klinischen Studie, und
- **die Phase 4 (Konsolidierungsphase) der Praxiskontrolle** nach Etablierung des Projekts, hinsichtlich der Zielerreichung, sowie deren Bedingungen und Auswirkungen.

3. Gegenstände der Veränderungsmessung und Interventionsevaluation

3.1. Unterscheidung nach Zielgruppen

Interventionen sind immer auf ausgewählte Zielgruppen gerichtet, da es nicht das universelle Interventionsprogramm für auffällige Kraftfahrer gibt und geben kann. Entscheidend sind dabei Differenzierungen der Zielpopulation, wie sie in der Literatur diskutiert sind, z.B.

- Differenzierung von Meyer-Gramcko und Sohn (1995) in Alkoholtäter und Punktetäter, bei denen unterschiedliche Ursachen vorliegen, die auch unterschiedliche Programme erfordern.

- Differenzierung von Spoerer und Ruby (1996) mit der Unterscheidung in vier Teilpopulationen, die unterschiedlich zu behandeln sind:

- Auffällige junge Fahranfänger (mit und ohne Alkoholdelikten)
- Alkoholauffällige Kraftfahrer (Erst- und Wiederholungstäter)
- Drogenauffällige Fahrer
- Mehrfachtäter

Wichtig in diesem Zusammenhang ist auch die Feststellung, dass diese Differenzierung in Teilpopulationen nicht abgeschlossen ist. So berichten Meyer-Gramcko und Sohn (1995) über das erstmalige Auftreten von Punktetäterinnen und begründen einen geschlechtsspezifischen Unterschied in den Ursachen der Auffälligkeit im Fahrverhalten und begründen damit auch eine geschlechtsspezifische Differenzierung in den Rehabilitationsmaßnahmen.

3.2. Ein Beispiel: Die Evaluationsstudie ALKOEVA (Winkler, Jacobshagen und Nickel, 1986, 1988)

Wir stellen hier als prototypisches Beispiel die Studie ALKOEVA vor, in der drei unterschiedliche Programme hinsichtlich ihrer Wirksamkeit bei der Rehabilitation wiederholt alkoholauffälliger Kraftfahrer beurteilt wurde. Es handelte sich dabei um die Programme IFT, I.R.A.K. und LEER, die mit folgenden Zielstellungen evaluiert wurden:

Ziel 1: Wird in den Kursen eine Erweiterung des Wissens hinsichtlich der Thematik „Trinken und Fahren“ erreicht?
Ziel 2: Ändern sich verkehrsspezifische Einstellungen und Haltungen zum Umgang mit Alkohol beim Führen von Kraftfahrzeugen?
Ziel 3: Führt die Kursteilnahme zu relevanten Verhaltensänderungen und hat dies einen Rückgang der Rückfallquote nach erneuter Wiedererteilung der Fahrerlaubnis zur Folge?
(Ziele zitiert nach Spoerer und Ruby, 1996).

Nimmt man beide Studien zusammen (1986 und 1988) dann kann der empirische Forschungsansatz durch folgenden Versuchsplan gekennzeichnet werden, der 3 Versuchsgruppen und eine Kontrollgruppe enthält:

Untersuchungszeitpunkt **Versuchsgruppe**	**t_0** **(Begutachtung)**	**t_1** **(nach 3 Jahren)**	**t_2** **(nach 5 Jahren)**
VG 1 (IFT)	Eignungsmängel; Kursempfehlung	Wissen Einstellung Rückfallquote	Wissen Einstellung Rückfallquote
VG 2 (I.R.A.K.)	Eignungsmängel; Kursempfehlung	Wissen Einstellung Rückfallquote	Wissen Einstellung Rückfallquote
VG 3 (LEER)	Eignungsmängel; Kursempfehlung	Wissen Einstellung Rückfallquote	Wissen Einstellung Rückfallquote
KG	Gleiches Deliktbild, aber keine Eignungsmängel	Wissen Einstellung Rückfallquote	Wissen Einstellung Rückfallquote

Wesentliche Ergebnisse waren:
- der Nachweis vermehrten Wissens über Alkohol und Fahren
- der Nachweis veränderter verkehrsrelevanter Einstellungen
- der Nachweis, dass in allen Versuchsgruppen die Rückfallquote geringer war, als in der Kontrollgruppe „geeigneter" Kraftfahrer.
- der Wirkungsnachweis der Kurse auch noch nach 5 Jahren (Nachwirkungseffekt).

Als eine der ersten umfassenderen Evaluationsstudien belegt ALKOEVA die Aussagekraft dieser empirischen Forschungsmethodik und erlaubt es auch, Anforderungen an einen Standard für Evaluationen in der verkehrspsychologischen Rehabilitation zu begründen. Bevor wir dazu kommen, wollen wir als erstes die Problematik der Veränderungsmessung als Kernstück einer Interventionsevaluation kennzeichnen.

4. Veränderungsmessung als Kernstück der Evaluation

4.1. Veränderungen bezeichnen **quantitative oder qualitative Änderungen** im Ausprägungsgrad einer oder mehrerer betrachteten Dimensionen (z.B. Veränderungen in der Reaktionsgeschwindigkeit, Veränderungen im Lebensstil, Veränderungen im Zuwendungsinteresse, Veränderungen in der Einstellung zum Trinken und Fahren ...)

4.2. Veränderungen beziehen sich also immer auf **Änderungen des individuellen Ausprägungsgrads** einer (hier psychischen) Dimension (hier des Erlebens und Verhaltens als Kraftfahrer). Veränderungen kennzeichnen also die intraindividuelle Variabilität.

4.3. Veränderungen werden über Kennwerte (Veränderungswerte) beschrieben, diese sind **abhängig vom Skalenniveau der Messskala**:

- bei metrischen Daten werden Differenzwerte und Variationsmaße verwendet,
- bei ordinalen Skalen werden neben Lageparametern (Median, Quartile) und Variationsmaßen vor allem Anordnungsänderungen und Rangplatzdifferenzen verwendet,
- bei nominalen Klassen werden Veränderungen der Klassenzugehörigkeit verwendet.

4.4. Veränderungsmessung bezeichnet nun den theoretischen Hintergrund zur Erfassung und Analyse solcher Kenngrößen von Veränderungen.

4.5. Zielstellungen für die Veränderungsmessung

a) Personen verändern sich im Zeitablauf in ihrem Erleben und Verhalten. Deshalb ist die **Erfassung dieser entwicklungsabhängigen Veränderungen** eine Zielstellung der Veränderungsmessung und damit ein Zugang zur differentiellen Entwicklungspsychologie und -diagnostik. Dies ist das Ziel einer (nicht bewusst beeinflussten) **natürlichen Veränderungsanalyse**.

b) Es gehört auch zum Gegenstandsbereich der Psychologie, menschliches Erleben und Verhalten zu beeinflussen. Dies erfolgt allgemein durch Interventionen (z.B. medikamentös, psychotherapeutisch, durch Lern- und Trainingsmethoden). Dabei entsteht die Frage nach der **Wirksamkeit dieser Interventionen**, d.h. das Ziel ist eine **Interventionseffektanalyse**

c) Beide Zielstellungen sind nicht nur für Einzelpersonen relevant sondern auch für Gruppen (Mannschaften, Teams, Berufsgruppen, ...). Wir generalisieren daher, dass sich **Veränderungsmessung auf die Erfassung und Bewertung von Änderungen in den Ausprägungsgraden relevanter Dimensionen (D_1, ..., D_r) an Untersuchungseinheiten (UE)** bezieht. Untersuchungseinheiten können dabei sowohl einzelne Personen als auch Personengruppen (z.B. Fahrgemeinschaften) sein. Hinsichtlich der Verallgemeinerbarkeit von Befunden werden die UE als Bestandteile entsprechender Populationen (Grundgesamtheiten) verstanden, dann wird aus Stichproben auf die Population geschlossen.

4.6. Theoretische Kennzeichnung der Veränderungsmessung

a) Veränderungsmessung bezieht sich also auf **wiederholte Beobachtungen von Merkmalsausprägungen** an einzelnen Untersuchungseinheiten oder einer Stichprobe. Solche Beobachtungen werden theoretisch durch Zufallsgrößen beschrieben und dienen der Modellierung der Messsituation.

b) Betrachten wir nur eine Dimension, z.B. die Körpertemperatur von Herrn X, dann ergibt sich über wiederholte Messung die Veränderung dieser Temperaturen, die z.B. als “Fieberkurve” darstellbar ist. Es entsteht eine Folge von Beobachtungswerten, die Realisierungen der **Zeitreihe Y_1, ...,Y_T** sind.

c) Die **Anzahl der Messzeitpunkte** bestimmt Typ und Aussageformen der Veränderungsmessung:

- bei zwei Messzeitpunkten spricht man von einer **Prä- und Postmessung**.
- bei mehr als zwei Messzeitpunkten besteht z.B. die Möglichkeit, zusätzlich zur Interventionswirkung auch noch die Nachwirkungen über spätere Messzeitpunkte zu beurteilen. Es entsteht der Ansatz einer **Paneluntersuchung**.
- ab 50 Messzeitpunkten spricht man dann von einer **Zeitreihenanalyse**. Sie erfasst nicht nur Veränderungen (als Trend der Beobachtungswerte) sondern gestattet es auch, Periodizitäten in den wiederholten Beobachtungswerten zu identifizieren.

4.7. Veränderungsmessung als indirekte Messmethode

a) Geht man davon aus, dass die meisten psychischen Dimensionen einer direkten Beobachtung nicht zugänglich sind, folgt, dass auch für die Veränderungsmessung die wiederholten Messungen als indirekte Messungen zu kennzeichnen sind. Damit entsteht die Situation, dass **jede einzelne Messsituation als indirekte Messsituation** zu kennzeichnen ist und sich die Veränderung selbst aus dem **Vergleich dieser indirekten Messungen** ergeben muss.

b) am einfachsten zeigt sich dies für die (metrische) Prä- und Posttestanalyse, die durch folgendes Strukturdiagramm 0000kennzeichenbar ist:

Modellstruktur für die Analyse von Prä- und Post-Messungen

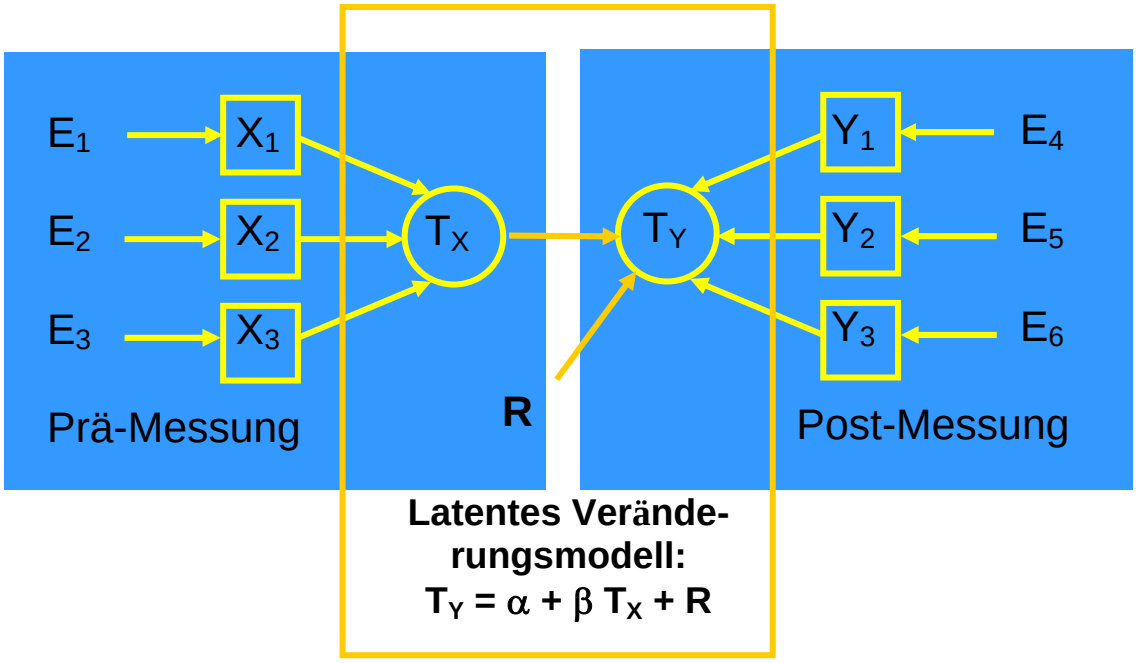

Abb. 1: Lineares Veränderungsmodell mit Fehlern in den Variablen

Prä- und Postmessung sind durch Messmodelle gekennzeichnet. Die Veränderungswirkung ergibt sich durch ein lineares Regressionsmodell hinsichtlich der Veränderungen in den latenten Eigenschaftsausprägungen prä und post. Über dieses Veränderungsmodell kann die Veränderungswirkung abgeschätzt und beurteilt werden. Dabei bleibt, wie bei Regressionen üblich, ein Rest an nichtaufklärbarer Prä- und Postbeziehung übrig, der im Sinne des Bestimmtheitsmaßes die Güte des Regressionsmodells kennzeichnet.

4.8. Komponentenanalyse der Veränderung

Ziel der Komponentenanalyse ist die Präzisierung potentielle Ursachen für die Entstehung von Veränderungen. Global werden dabei drei Komponenten unterschieden:

a) Veränderungen entstehen dadurch, dass auf die Ausgangsbedingungen der UE die natürliche Umwelt und die Messmethode einwirken und damit zu Veränderungen führen. Diese Wirkung, die immer besteht, bezeichnen wir als **Remissionswirkung** (z.B. Alterung, Progression, Regressionseffekte, Abhängigkeit vom Ausgangswert...).

b) Veränderungen entstehen auch durch die gezielten Interventionen. Diese Wirkung bezeichnen wir als **Treatmenteffekt**.

c) Veränderungen entstehen auch durch die Wirkungen der zufälligen **Fehler** in der Prä- und Postmessung, die voneinander nicht unabhängig sind, sondern verbunden in den Fehler des Veränderungskennwerts eingehen.

d) Allgemein ist also ein beobachteter Veränderungswert Δ von allen drei Komponenten abhängig:

Veränderungswert Δ = f (Treatment, Remission, Fehler).

e) Offensichtlich muss ein Veränderungsmessmodell Antwort auf folgende Fragen geben:
- wie groß ist die **Fehlerwirkung**?

- wie groß ist die **Remissionswirkung**?
- was bleibt als **"bereinigte" Treatmentwirkung** ausweisbar?

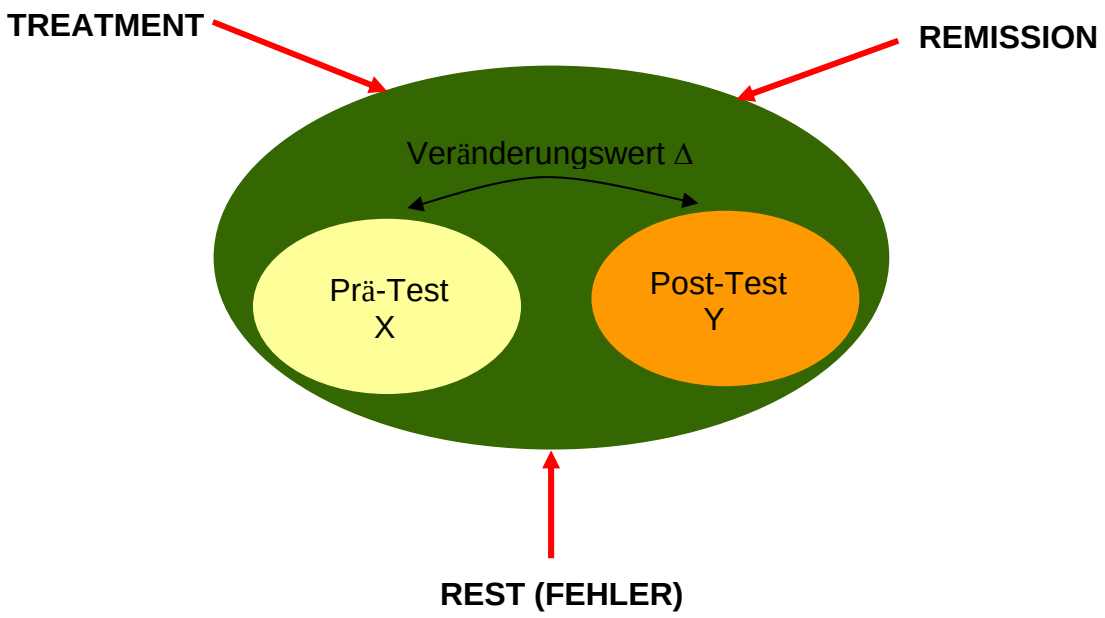

Abb. 2: Komponente der Herausbildung von Veränderungen

4.9. Das Untersuchungsdesign

Für die empirische Analyse von Veränderungen sind drei grundsätzliche Zugänge, die durch unterschiedliche Versuchspläne kennzeichenbar sind, typisch:

a) Für eine **Einzelfalluntersuchung** sind spezielle Versuchspläne erforderlich, die es gestatten, Veränderungseffekte wiederholt (möglichst) unabhängig zu beobachten und aus diesen Replikationen zu beurteilen (Ausblendpläne, Umkehrpläne, Periodenversuchspläne)
b) Für die **populationsbezogene Untersuchung** besteht das Hauptproblem nach der Fehlerabschätzung in der Trennung von Treatment- und Remissionseffekten (z.B. mittels der Kovarianzanalyse).
c) Ein alternativer Zugang ist ein **Versuchs- Kontrollgruppendesign**, bei dem die KG der Intervention unter sonst gleichen Untersuchungsbedingungen nicht unterzogen wird. Aus dem Vergleich VG-KG, genauer der Differenz der Veränderungswerte, lässt sich dann (bei metrischen Daten) die Treatmentwirkung abschätzen.

4.10. Probleme und Ansätze der Veränderungsmessung

a) Grundprobleme und Zugänge der Veränderungsmessung

- Veränderungswerte

Veränderungswerte können immer nur hinsichtlich des Datenniveaus des Beobachtungsmerkmals bestimmt werden:
- bei metrischen Daten Differenzwerte oder Streuungswerte
- bei ordinalen Daten Medianwerte, Quartile, Quartilabstände, Anordnungsänderungen
- bei nominalen Daten Häufigkeiten des Auftretens von Kategorien und Kategoriewechsel

- Statistische Hypothesentestung

Veränderungshypothesen allgemein beziehen sich dann immer auf die obigen Veränderungswerte, wobei es zwei Zugänge gibt:
- die elementarstatistische Prüfung der Veränderungswerte

- die Modellierung der Veränderungsmesssituation mit statistischer Beurteilung sowohl der Veränderungswerte als auch der Modellgüte. Hierbei sind neben der Beurteilung beobachteter Veränderungen sowohl Modellfit (Passfähigkeit des Messmodells) als auch Personenfit (Passfähigkeit der Personen) zu beurteilen.

b) Grundprobleme und Zugänge der Veränderungsmessung bei metrischen und ordinalen Daten)

(i) Die Fehler-in-den-Variablenmodelle
Fehler-in-den-Variablenmodelle kennzeichnen eine Modellklasse für metrische Daten, bei der davon ausgegangen wird, dass alle beobachteten Variablen fehlerbehaftet sind und diese Fehler auch korrelieren könnten. Unter diesen Voraussetzung erfolgte eine Modellierung der Veränderungsmesssituation wie eingangs bereits beschrieben. Vergleiche dazu Abbildung 1.

- Statistische Hypothesentestung
Im Rahmen dieses Modells können dann die wahren Ausprägungen der psychischen Eigenschaft geschätzt und deren Veränderung beurteilt werden. Den einfachsten Zugang bildet ein lineares Strukturgleichungsmodell (LISREL), das für normalverteilte Beobachtungsdaten angewendet werden kann.

(ii) Das kriterienorientierte Modell von Lander (1990, 1997), das sowohl für metrische wie ordinale Daten spezifiziert ist und einen kriteriumsorientierten Zugang zur Veränderungsmessung formuliert. Der Modellansatz umfasst folgende Komponenten:
Die Grundannahmen:
- Ein **metrisches, normalverteiltes Merkmal Y** wird vor und nach einer Intervention beobachtet.
- Der Veränderungswert $D = Y_{post} - Y_{prä}$ beinhaltet drei Wirkanteile:
  - den **Treatment- Effekt (T)** als Wirkung der Intervention,
  - den **Remissionseffekt (R)** als Wirkung weiterer Einflussgrößen und den **Fehler**, der hier kein unabhängiger Anteil ist ! (Thorndike, 1924, Thompson, 1924)
- Ziel ist die **Erschließung des Interventionseffekts** aus dem Veränderungswert.

Das lineare Lander- Modell geht von folgenden Annahmen aus:
- Der Modellansatz
Zugang ist ein Versuchsgruppen- / Kontrollgruppenvergleich bei dem die Differenzwerte

$D = Y_{post} - Y_{prä}$ folgenden additiven Modellbeziehungen genügen sollen:

für Versuchsgruppe VG: D_{VG} (Prä- Postdifferenz) = $D_T + D_R + D_E$
für Kontrollgruppe KG: D_{KG} (Prä- Postdifferenz) = $D_R + D_E$

und es gilt ein **linearer Zusammenhang** zwischen den Prä- und Postwerten, der sich in folgenden Gleichungen spezifiziert:

$Y_r = (1-b_r) * Y_0 + b_r * X'_j$ für die Remissionswirkung
$Y_v = (1-b_v) * Y_0 + b_v * X'_j$ für die Veränderung insgesamt
$Y_n = (1-b_n) * Y_0 + b_n * X'_j$ für die Nachwirkung

Dabei bezeichnet Y_0 die vorgegebene Zielgröße, X'_j sind die jeweiligen Prä-Test-Werte.

- Grundaussagen aus dem linearen Veränderungsmodell
Grundsätzliche Aussagen werden über die Prüfung von Hypothesen über die Differenzkomponenten bzw. diese Regressionskoeffizienten möglich. Die wichtigsten sind:

- H_0: D(Veränderung) = 0 bzw. $b_v = 1$ prüft Prä- Post- Homogenität

- H_0:	D(Remission) = 0	bzw. $b_r = 1$	prüft Remissionseffekt
- H_0:	D(Treatment) = 0	bzw. $b_v = b_r$	prüft den Treatmenteffekt
- H_0:	D(Treatment) = D(Remission)		prüft die Treat.-Rem.Homogenität
- H_0:	D(Ziel) = 0	bzw. $b_v = 0$	prüft Zielerreichungseffekt
- H_0:	D(Nachwirkung) = 0	bzw. $b_n = b_v$	prüft Nachwirkungseffekt

Darstellung der PPA-Ergebnisse in einem D_{ov} -D_z-Koordinatensystem

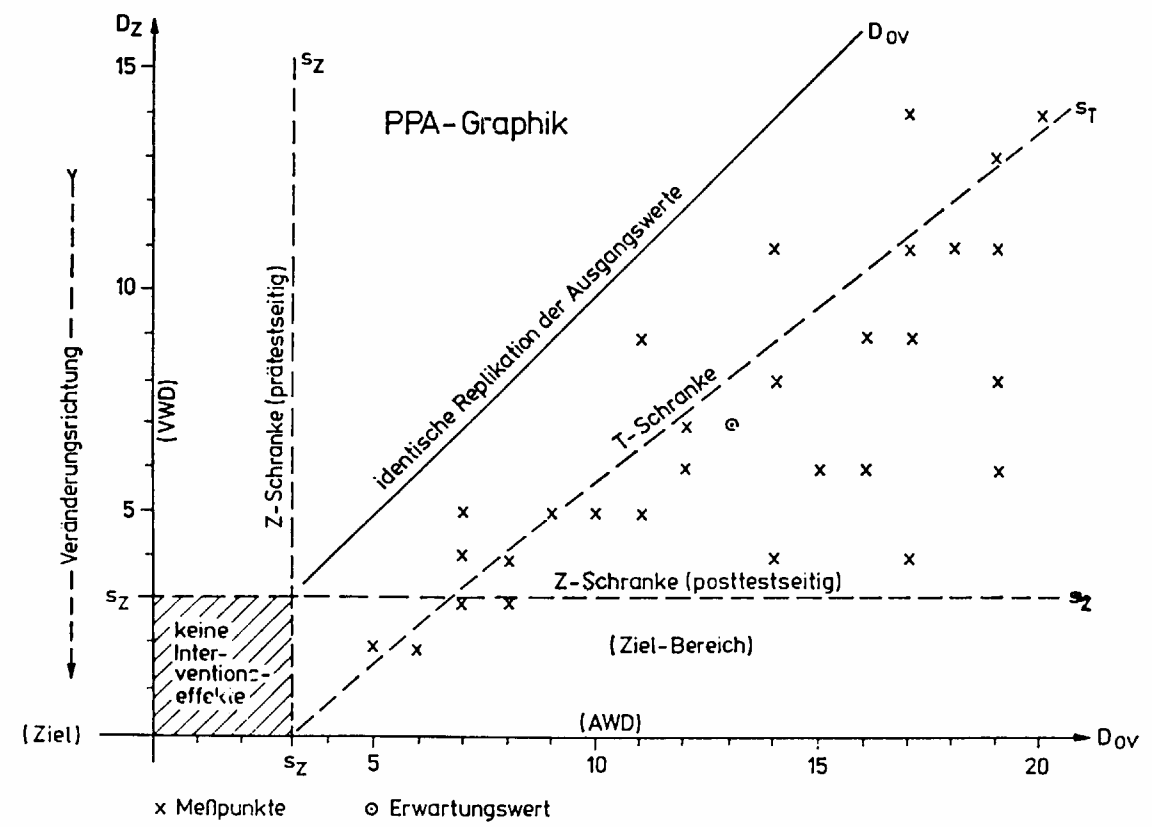

Abb. 3: Dargestellt sind in dem Bezugssystem Ausgangswertdifferenz (AWD) und Verlaufswertdifferenz (VWD) die Gerade der identischen Replikation (keine Veränderungswirkung), die Gerade des signifikanten Treatmenteffekts (T-Schranke) und das Kriterium (Z-Schranke). Die Veränderungswirkungen des Probanden (x) werden damit in 4 Klassen eingeteilt, je nachdem ob T und Z-Schranke erreicht wurden. Diese Häufigkeiten ermöglichen es, die Verfahrenseffizienz zu kennzeichnen.

c) Grundprobleme und Zugänge der Veränderungsmessung bei qualitativen Daten-Item- Antwort – Modelle

Item- Antwort- Modelle sind allgemein dadurch ausgezeichnet, dass sie durch eine Item-Antwort- Funktion die Wahrscheinlichkeit des Lösens /Positiv Antwortens als Modell kennzeichnen. Mit der Wahl der Itemantwortfunktion sind die Modellparameter und damit die Grundannahmen bestimmt. Zwei Ansätze, die als Modell die logistische Funktion verwenden, sollen dies demonstrieren:

(i) Das Veränderungsmodell von Zwindermann (1991)
Entscheidend für das logistische Testmodell ist die Annahme, dass der Ausprägungsgrad der Eigenschaft auf der latenten Dimension durch einen Eigenschaftsparameter (Fähigkeitsparameter) F_{jt} zum Zeitpunkt t gekennzeichnet wird und der Zusammenhang zwischen Antwortverhalten und Fähigkeitsausprägung durch eine Item- Antwort- Funktion (icc) folgender Form beschrieben wird:

prä: $$P(X_{ij1}=1 \mid F_{j1}) = p_{ij} = \frac{\exp(F_{j1})}{(1+\exp(F_{j1}))} = \frac{e^{F_{j1}}}{1+e^{F_{j1}}}$$

Dies ist das **Modell für die Prämessung**, wobei die Wahrscheinlichkeiten gleichzeitig den Messfehler modellieren.
Für die Darstellung der Postmessung gehen wir nun davon aus, dass sich die Eigenschaftsausprägung F_{j2} zum Zeitpunkt 2 (post) um einen bestimmten Betrag v_j verändert hat, der bei Versuchs- und Kontrollgruppe verschieden sein kann. Man erhält dann die Itemfunktion

post: $$P(X_{ij2}=1 \mid F_{j2}) = g_{ij} = \frac{\exp(F_{j1}+v_j)}{(1+\exp(F_{j1}+v_j))}$$ (**Modell der Postmessung**!!)

mit der Aufspaltung $\mathbf{v_j = q_j * T + R,}$ bei der T den Veränderungseffekt durch das Treatment und R einen zeitlich bedingten Remissionseffekt beschreiben. Dies ist das **Modell für die eigentliche Veränderungswirkung**. Durch die Gewichtsfaktoren q_j kann zwischen Versuchs- und Kontrollgruppe unterschieden werden:

$q_j = 1$ kennzeichnet die Versuchsgruppe durch die Wirkung von Treatmenteffekt T und Remissionseffekt R,

$q_j = 0$ kennzeichnet die Kontrollgruppe ohne die Wirkung eines Treatmenteffekts, also den reinen Remissionseffekt.

Eine Erweiterung des Ansatzes auf Mehrpunkterhebungen ist möglich und wird vom Autor dargestellt.
Insgesamt ist dies ein Spezialfall des linearen logistischen Testmodells (LLTM), als multivariate Erweiterung des gewöhnlichen Raschmodells, wie dies Fischer (1995) darstellt. Es ist ausgelegt für die Analyse von Veränderungen in a priori bekannten Subgruppen (hier Versuchs- und Kontrollgruppe) und verwendet allgemein die Item-Antwort-Funktion

$$P(X_{vi} = 1) = \frac{\exp\left[b_{tgi}\,(\theta_v - \beta_i - \delta_{gt})\right]}{1 + \exp\left[b_{tgi}\,(\theta_v - \beta_i - \delta_{gt})\right]}$$

mit: θ_v Fähigkeit der Person v,

β_i Schwierigkeit des Items i,

δ_{gt} Veränderung der Itemschwierigkeit für Subpopulation g zum Zeitpunkt t.

Für dieses Modell ist die Conditional Maximum Likelihood Schätzung (CML) möglich, da der Summenscore r_v einer Person eine erschöpfende Statistik ist.
Allen diesen Ansätzen ist der Grundgedanke gemeinsam, dass sich eine Veränderung in der Fähigkeit/Eigenschaft darin zeigen müsste, dass dadurch auch der Schwierigkeitsparameter sich dementsprechend verändert.

(ii) Das lineare Veränderungsmodell von Fischer
Im Weiteren betrachten wir das modifizierte lineare Raschmodell (Stefan Klein, 1999), das in unserem Kontext besonders geeignet ist und durch die Eigenschaften ausgezeichnet ist, dass es die lineare Beziehung θ_v - β_i zwischen Fähigkeits- und Schwierigkeitsparameter vermeidet. Den Unterschied verdeutlicht folgende Überlegung:
Das klassische Raschmodell, und damit auch das LLTM, erklärt die Wahrscheinlichkeit des Lösens eines Items in Abhängigkeit von der Fähigkeit F einer Person und der Schwierigkeit S eines Items:

P (Lösen eines Items) = f (F, S) (Item- Antwort- Funktion).

Diese Funktion ist die logistische Funktion f = exp (F-S) / [1 + exp(F-S)]. Die Schätzungen der Fähigkeits- und Schwierigkeitsparameter erfolgen stichprobenunabhängig und spezifisch objektiv. Die Modellgültigkeit ist beurteilbar (Möglichkeit des Modelltests).
Diese Unterscheidung, bei der die Schwierigkeit S eines Items auf eine Population P bezogen ist, wird nun aufgegeben. Wir betrachten jetzt mit Fischer (1995a,b) jede einzelne Person bzgl. jedes einzelnen Items. Jede dieser Kombinationen heißt virtuelle Person. Deren Lösungsverhalten wird durch einen gemeinsamen (aber spezifischen) **Parameter F_{ij}** gekennzeichnet, der die „Fähigkeit“ dieser virtuellen Person beschreibt.

Wir betrachten nun diese virtuellen Personen zu den einzelnen Untersuchungszeitpunkten, die dann die **virtuellen Items** bilden. Es ergeben sich folgende vier V-Items:

- V-Item 1: Versuchs- und Kontrollgruppe im Prätest
- V-Item 2: Kontrollgruppe im Posttest (d.h. ohne Intervention)
- V-Item 3: Versuchsgruppe im Posttest (d.h. mit Intervention)
- V-Item 4: Versuchsgruppe bei der Nachmessung

Für das Raschmodell, d.h. wenn diese so konstruierten V-Items eine Raschskala konstituieren, zeigen sich Veränderungen dann in den Schwierigkeitsparametern dieser vier V-Items:

- **V-Item 1**: Der Schwierigkeitsparameter, also **S_1, wird 0 gesetzt** (Normierung).

- **V-Item 2**: Der S_2- Parameter kennzeichnet die Wirkung der Veränderung der F_{ij} von Prä nach Post bei der Kontrollgruppe:

  $S_2 - S_1 = S_2$ = Remissionswirkung

- **V-Item 3**: Der S_3- Parameter kennzeichnet die Wirkung der Veränderung der F_{ij} von Prä nach Post bei der Versuchsgruppe, enthält also Remissions- **und** Treatmenteffekte. Durch Differenzbildung erhalten wir:

  $S_3 - S_2$ = Treatmentwirkung

- **V-Item 4**: Der S_4- Parameter kennzeichnet die Gesamtveränderung der F_{ij} bis zur Nachmessung, beinhaltet also Remissions-, Treatment- und Nachwirkungseffekte. Durch Differenzbildung resultiert jetzt:
  $S_4 - (S_3 - S_2) - S_2 = S_4 - S_3$ = Nachwirkung

Im Ergebnis des LLRA-Modells sind damit vergleichbare Aussagen wie mit dem Lander- Modell begründbar und statistisch prüfbar. (Einen Modellvergleich haben Klein & Krause, 1999 dargestellt.)
Generalisiert ist dies das linear logistisches Modell mit “relaxed assumptions” (LLRA) mit der Itemantwortfunktion wobei θ_{vi} die oben beschriebenen Parameter F_{ij} der virtuellen Personen und δ_{gt} die Schwierigkeitsparameter der virtuellen Items gt sind.

$$P\,(X_{vit} = x_{vit}) = \frac{\exp\,[x_{vit}\,(\theta_{vi} - \delta_{gt})]}{1 + \exp\,[x_{vit}\,(\theta_{vi} - \delta_{gt})]}$$

(iii) Modellgüte und Personenfit

Für alle indirekten Messmethoden ist die Frage nach der **Modellgültigkeit** von entscheidender Bedeutung für die Aussagekraft der Messwerte. Im Rahmen der IRT- Modelle, insbesondere bei Verwendung der logistischen Itemantwortfunktion, ergibt sich der Zugang zur Prü-

fung der Modellgüte über die Eigenschaft der **Populationsunabhängigkeit** und der **spezifischen Objektivität**. Danach sind die Itemparameter unabhängig von den Personenparametern. Somit müssen sich auch für Extremgruppen die gleichen Itemparameter ergeben. Dies ist die Grundlage des **Modelltests nach Anderson**.
Umgekehrt kann man, da im Modell ja Item- und Personenparameter gleichwertige Komponenten des Verhaltens sind, auch die **Frage nach dem Personenfit** stellen: Genauer ist dies die Frage danach, ob wirklich alle Personen ihr Verhalten auf vergleichbare Weise und damit modellkonform erzeugen. Dies ist eine Frage, die in der neueren Forschung unter den Begriffen **Fehlspezifikation** und **"differential item functioning" (DIF)** untersucht wird und mit dem Bezug auf die einzelne Person direkte Zugänge zur Differentialdiagnostik eröffnet.
Dies soll nun im Sinne einer abschließenden Problemsicht noch einmal präzisiert werden:

1. Die dargestellten Modelle ermöglichen die **Beurteilung von gruppenspezifischen Veränderungen**. Personenspezifische Aussagen sind im LLRA nicht möglich.

2. Geht man davon aus, dass sich Veränderungen nicht homogen einstellen, also der Effekt eines kognitiven Trainings von Kindern z.B. von der sozialen Herkunft abhängt, dann kann man versuchen, veränderungshomogene Subgruppen zu identifizieren und damit die unterschiedlichen Veränderungswirkungen ausweisen. Dieses Phänomen wird als **Fehlspezifikation des Veränderungsmechanismus** bezeichnet. Dies gilt insbesondere auch für die anfangs unterschieden Zielgruppen verkehrspsychologischer Rehabilitation.

Verallgemeinert gilt:
Multivariate IRT-Modelle nehmen an, dass die Schwierigkeiten eines Items durch das Zusammenwirken psychischer Eigenschaften/Fähigkeiten erklärt werden kann. **Fehlspezifikation** bedeutet dann, dass ein in der Realität vorhandener Trait 1 fälschlicherweise einem Item i zugeordnet wird. Bei normalen Raschmodellen spricht man von **„differential item functioning" (DIF)** und kennzeichnet, dass das Antwortverhalten auf ein Item unterschiedlich verursacht sein kann. Solche DIF-Phänomene können in unterschiedlicher Form, genauer in der Personen- und in der Itemebene auftreten:

Personenebene Itemebene	alle Personen	Subgruppen
alle Items	keine Fehlspezifikation	Fehlspezifikation auf Itemebene
Subgruppen von Items	Fehlspezifikation auf Personenebene	Fehlspezifikation auf Personen- und Itemebene

a) als **Fehlspezifikation allein auf der Itemebene**. Dies bedeutet, dass eine Subgruppe von Items eine andere Veränderungswirkung bei den Personen ausweist, als der Rest der Items.
b) als **Fehlspezifikation allein auf der Personenebene**, der kennzeichnet, dass die Veränderung für Subgruppen (im Extremfall für jede Person) unterschiedlich ist.
c) als **Fehlspezifikation auf Personen- und Itemebene**. Dies bedeutet, dass eine Subgruppe der Items in einer Subpopulation der Personen eine andere Veränderungswirkung ausweist, als der Rest der Items in der Population.

Für die Fehlspezifikation auf Personenebene begründen sich zwei Zielstellungen:
a) das Ziel, eine Menge von Personen zu identifizieren, auf die das geschätzte Itemantwortmodell (hier das LLTM) nicht passt. Hierzu werden Methoden zur Kennzeichnung des **Personenfit** entwickelt, und
b) das Ziel zu bestimmen, warum eine Menge aberranter Personen nicht in das geschätzte Modell passen. Zusätzlich werden Methoden entwickelt, die es gestatten, aufgrund manifester Variablen zu entscheiden, ob Personen ein aberrantes Antwortmuster aufweisen oder nicht.

Die Bedeutung der Fragestellung nach dem Personenfit soll kurz für zwei Anwendungsbereiche angedeutet werden:

Für die Konstruktion von Messmethoden werden beim Vorliegen eines ungenügenden Personenfit folgende Aussagen möglich:

- Kennzeichnung schlecht angepasster Personen, die damit aus der Kalibrierungsstichprobe entfernt werden können,
- Berücksichtigung einer Klasse unskalierbarer Personen
- Bestimmung derjenigen Antwortmuster in der Stichprobe, die besonders häufig zu einem schlechten Personenfit führen und Entdeckung von Gemeinsamkeiten in diesen Antwortmustern.

Für den Fall der Veränderungsmessung über zwei Messzeitpunkte werden folgende Hypothesen prüfbar:

- Hat sich bei einer Person j überhaupt eine Veränderung, die einen praktischen bedeutsamen Wert übersteigt, ereignet?
- Liegt die Veränderung unterhalb eines praktisch bedeutsamen Wertes?
- Liegt die Veränderung innerhalb eines bestimmten Bereiches (Zielbereich)?
- Erreicht die Veränderung einen bestimmten Bereich (Zielbereich) nicht?
- Gibt es eine „Erinnerung an den ersten Messzeitpunkt“?

Eine ausführliche Diskussion dieser Probleme mit konstruktiven Lösungsansätzen finden wir bei Ponocny (2002) und Klein (2002)

5. Evaluation als empirische Forschungsmethode

Wissenschaftlich fundierte Evaluation von Interventionsmaßnahmen geht wenigstens von zwei Voraussetzungen aus:

a) Einer **präzisen Fassung der Zielfunktion** für eine wohl definierte Population, die durch eine Merkmalscharakteristik beschrieben ist, und

b) Der begründeten **Hypothese**, dass die Interventionsmaßnahmen einen Einfluss auf das Erleben und Verhalten der Probanden der Population hinsichtlich der Zielerreichung haben.

Beide Voraussetzungen sind wesentliche Merkmale, die eine Evaluation von einem „monitoring“ unterscheiden. Das klassische Beispiel für den Unterschied beider Formen ist ein früher Ansatz aus der Schulevaluation, bei dem der Unterschied zwischen dem „monitoring“ von Schulräten und dem Ergebnis einer Evaluationsstudie im Land Baden-Württemberg dazu führte, dass das (staatliche) Evaluationsinstitut mit der Begründung aufgelöst wurde, „... dass trotz einer Anwendung der empirisch-statistischen Verfahren sich unauflösliche Widersprüche zu den Beobachtungen der Oberschulräte ergeben haben.“ (Teschner, 1979, zitiert nach Wittmann, 1985).

Wesentlich für die wissenschaftlich fundierte Evaluation verbleibt, dass beide Voraussetzungen unter methodischem Gesichtspunkt zwei Konsequenzen haben:

a) Die Kausalhypothese gestattet es, **experimentelle Versuchspläne für Evaluationsstudien** zu begründen und dabei eine willkürliche Manipulation der unabhängigen Variablen zu ermöglichen. Häufiger Anwendungsfall ist dabei ein Versuchs- und Kontrollgruppen-Design, der dem Max-Kon-Min- Prinzip eines guten Experiments (Kerlinger, 1973) genügt. Dies verweist zusätzlich auf die Anwendung von Kontrolltechniken, um konfundierende Einflussgrößen in ihrer Wirksamkeit auszuschließen oder zumindest zu kontrollieren. Gerade bei Programmevaluationen ist dies eine der schwierigsten Aufgaben.

b) Die Zielfunktion gestattet es, die **abhängige Variable zu spezifizieren** und damit den Einfluss der variierten unabhängigen Variablen hinsichtlich dieses Zielkriteriums zu beurteilen.

Beide Konsequenzen verdeutlichen, dass die Aussagekraft einer Evaluationsstudie wesentlich von der **Präzision der Zielfunktion**, der **qualitativen Begründung der Kausalhypothese** und dem **gewählten Versuchsplan** bestimmt sind. Was dies für die Evaluation verkehrspsychologischer Rehabilitationsmaßnahmen bedeuten könnte, soll nachfolgend gekennzeichnet werden:

6. Gedanken zu einem Standard für die Evaluation verkehrspsychologischer Maßnahmen.

Einen inhaltlich begründeten Standard zur Evaluation verkehrspsychologischer Rehabilitation leiten wir aus unserem allgemeinen Vorschlag (Metzler, Krause, 1997) ab. Dieser Vorschlag orientiert sich an dem Standard „Good Clinical Practice“ und begründet acht methodische Grundforderungen an eine kontrollierte klinische Studie, die unseren Anforderungen genügen. Sie spezifizieren

a) bzgl. des **Versuchsplans** die Anforderungen, dass

- dieser ein Versuchsgruppen- Kontrollgruppen- Design sein soll, der die Wirkung ggf. alternativer Behandlungsmethoden mit einer Kontrollgruppe beinhalten soll. Genau dies erfüllt die o.g. Studie ALKOEVA, die verschiedene Behandlungsprogramme mit einer unbehandelten Kontrollgruppe vergleicht. Alternativ ist auch ein kriteriumsorientierter Zugang denkbar, bei dem das Zielkriterium vorgegeben ist und dann die Interventionseffekte bzgl. der Zielerreichung bewertet werden.

- die Auswahlkriterien der Probanden (Ein- und Ausschlussregeln) angegeben und nachvollziehbar sein sollen.

- die Studie prospektiv angelegt sein muss und auch keine respektive Kontrollgruppe zulässig ist.

- die Zuteilung zu den Versuchs- und Kontrollgruppen zufällig erfolgen muss.

- den Probanden alternative Rehabilitationsmaßnahmen nicht bekannt sein dürfen. Die Erfolgsbeurteilung erfolgt dann durch neutrale Dritte.

b) bzgl. der **Zielfunktion** ist wichtig, dass

- ein **Zielkriterium** zur Beurteilung des Erfolgs gewählt wird, dass auch für den Probanden Relevanz besitzt. Dies bedeutet, das Zielkriterium von sogenannten Surrogatkriterien zu unterscheiden, die keine echte Bedeutung im Sinne der Zielfunktion für den Patienten haben. Kehren wir auch hier zu den drei Zielen der Studie ALKOEVA zurück:

Ziel 1: Wird in den Kursen eine Erweiterung des Wissens hinsichtlich der Thematik „Trinken und Fahren“ erreicht?

Ziel 2: Ändern sich verkehrsspezifische Einstellungen und Haltungen zum Umgang mit Alkohol beim Führen von Kraftfahrzeugen?

Ziel 3: Führt die Kursteilnahme zu relevanten Verhaltensänderungen und hat dies einen Rückgang der Rückfallquote nach erneuter Wiedererteilung der Fahrerlaubnis zur Folge?

Es wird deutlich, dass diese drei Ziele sehr wohl unterschiedliche Bedeutung besitzen. Für eine summative Evaluation ist das Ziel 3 entscheidend, während die Ziele 1 und 2 wohl eher als Surrogatkriterien zu betrachten sind (Dies wäre im Fall einer formativen Evaluation natürlich anders.).

- Der Leitfaden für die Planung von Projekt- und Programmevaluation (1997) präzisiert die Zielfunktion einer Evaluation durch den Begriff „SMART objectives“ (gescheite Ziele) mit der Differenzierung:
  - **S**pecific (spezifisch und konkret)
  - **M**easurable (messbar)
  - **A**ppropriate (angemessen, adäquat)
  - **R**ealistic (realistisch)
  - **T**imely (absehbar, in der gesetzten Zeit erreichbar)

  und vervollständigt so unsere Diskussion.

c) bzgl. weiterer Eigenschaften sind folgende Aspekte zu berücksichtigen:

- die **Kontrolle/ Beurteilung der Complience** sowohl auf der Seite des Probanden als auch auf Seite des Therapeutens. Wesentliche Konsequenzen bei Nichteinhaltung der Complience kennzeichnet folgende Graphik (Tab. 1 nach Metzler, Krause, 1997)

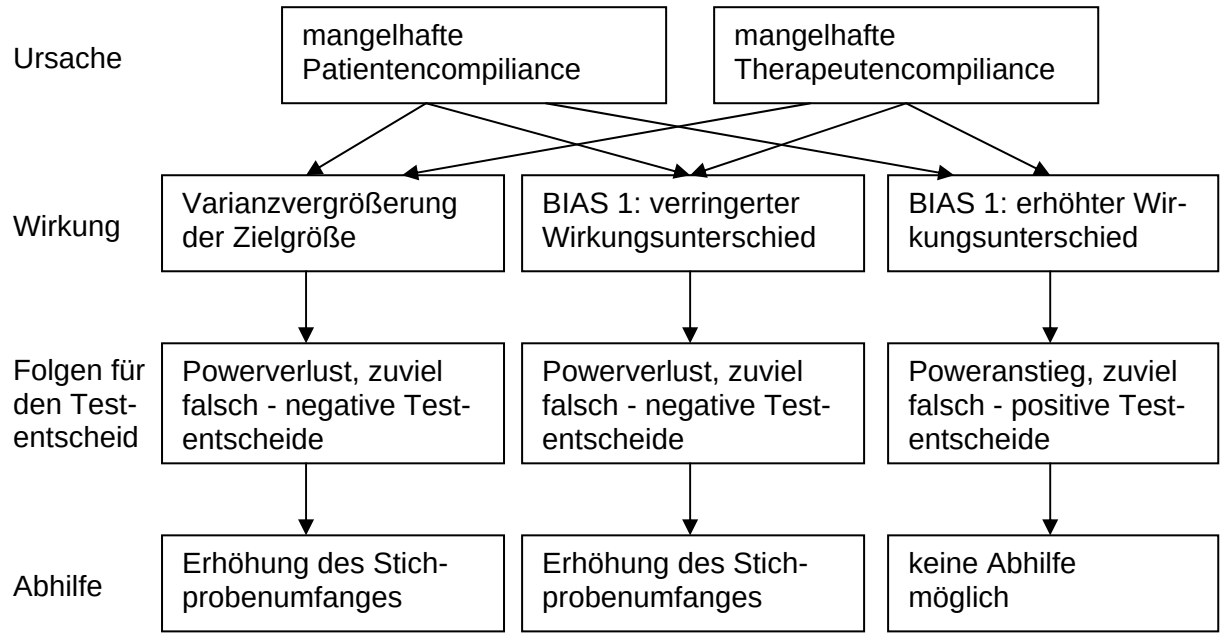

Abb. 4: Auswirkungen mangelnder Compliance von Patienten und Therapeuten auf Fehler in den beobachteten Daten.

- damit wird direkt das Problem des **Stichprobenumfangs** für die Evaluationsstudie angesprochen. Entscheidend ist hier die Wahl eines Stichprobenumfangs n_{opt}, der es gestattet, eine begründeten bedeutsamen Mindestunterschied mit großer Wahrscheinlichkeit auch auszuweisen (Power-Analyse). Dies ist zu verbinden mit einer **konfirmatorische Datenanalyse**!!.

- Die Auswirkungen unterschiedlicher Eigenschaften des Datenmaterials auf den Stichprobenumfang verdeutlicht die Abb. (Tab. 2 nach Metzler, Krause, 1997):

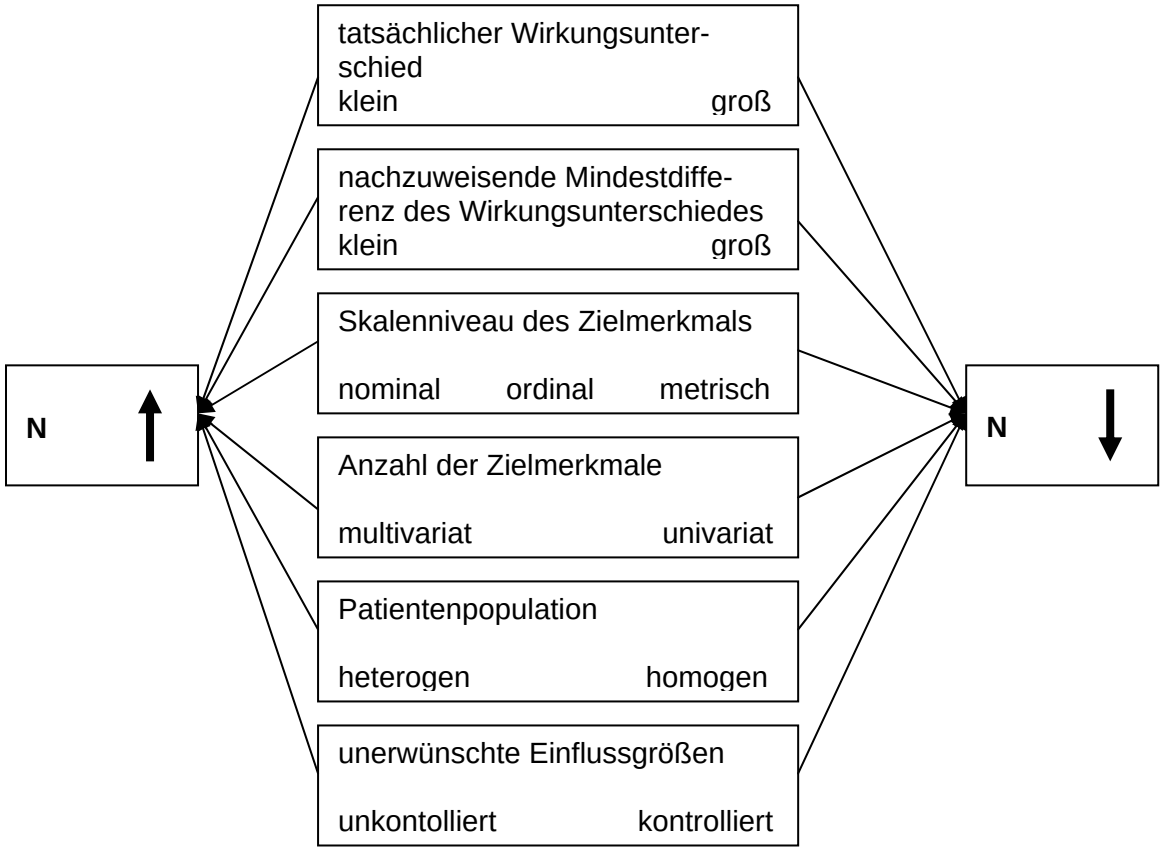

Abb. 5: Einflussgrößen auf die Größe des Stichprobenumfangs

d) Die Standards des „Joint Committee on Standards for Educational Evaluation“ (JC- Standards, Sanders, 1999) formulieren weitergreifende Standards für wissenschaftliche Evaluationen, die sich wie folgt differenzieren lassen (Beywl und Taut, 2000);
- 7 Nützlichkeitsstandards, die sichern, dass die Zielfunktion der Evaluationsnutzer erreicht wird.
- 3 Durchführbarkeitsstandards, die sichern sollen, dass eine Evaluation realistisch, gut begründet und kostenbewusst durchgeführt wird,
- 8 Korrektheitsstandards, die sich auf die rechtlichen und ethischen Bedingungen einer Evaluationsstudie beziehen.
- 12 Genauigkeitsstandards, die sichern sollen, dass fachlich fundierte Informationen über die Güte und/ oder Verwendbarkeit des evaluierten Projekts/ Programms entstehen.

7. Fazit und Konsequenzen

Als Anforderungen an einen Standard für die Evaluation von verkehrspsychologischen Rehabilitationsmaßnahmen ergeben sich folgende Konsequenzen:

a) Die Anforderung des Gesetzgebers zur Evaluation dient dem Ziel der Qualitätssicherung. Damit handelt es sich um eine summative Evaluation.

b) Im Rahmen dieser summativen Evaluation sind wesentlich das Ziel und die Zielpopulation zu kennzeichnen. Dabei erscheint eine Zielfunktion wie für die Nachschulungskurse „Mainz 77“ (Birnbaum u.a., 2002) formuliert ist, als nicht ausreichend:
„ Die Wahrscheinlichkeit einer Wiederauffälligkeit durch Fahren unter Alkoholeinfluss ist bei erstmals alkoholauffälligen Kraftfahrern, die nach einer Vorselektion an einem Nachschulungskurs „Mainz 77“ teilnehmen, trotz der Verkürzung der Sperrfrist nicht größer, als diejenige von vergleichbaren Personen, die an einem

solchen Kurs nicht teilnehmen.“ Im Sinne einer Qualitätsentwicklung wäre hier an eine Quotenvorgabe von z.B. 30% zu denken.

c) Diese Quotenvorgabe (Ziel der Rehabilitation) gestattet es auch, bei wiederholter Evaluation die Qualitätssicherung zu beurteilen bzw. deren Veränderung festzustellen. Positive Veränderungen sollten zur Anhebung der Zielquote führen, negative Veränderungen zu einer Ursachenanalyse. So könnte das erstmalige Auftreten von Punktetäterinnen (s.o.) eine solche Ursache sein, die ein modifiziertes Programm erfordert, um die Zielquote einzuhalten.

d) Der wissenschaftliche Zugang erfordert auch für die summative Evaluation einen Versuchgruppen-/ Kontrollgruppendesign, der Treatment- und Zielerreichungseffekt auszuweisen gestattet. Alternativ hierzu ist auch ein kriteriumsorientierter Zugang zur summativen Evaluation möglich.

e) Entscheidend für die Aussagekraft einer Evaluationsstudie bleibt die Kontrolle und Einhaltung der Rahmenbedingungen der Interventionsmaßnahme.

f) Die Diskrepanz zwischen Theorie und Praxis: Der Methodiker weist Wege zur Erkenntnis auf, diese sind in der Praxis nur selten in idealer Form umsetzbar. Deshalb sollten die Anmerkungen zu einem Standard als Empfehlungen verstanden sein. Abweichungen sollten dann begründbar sein und bei der Befundbewertung beachtet werden.

8. Literatur

Beywl, W. und Taut, S. (2000). Standards: Aktuelle Strategie zur Qualitätsentwicklung in der Evaluation. Vierteljahreshefte zur Wirtschaftsforschung, 69. Jahrgang, Heft 3, 358-370.

Birnbaum, D., Biehl, B., Sage, E. und Scheffel, B. (2002). Evaluation des Nachschulungskurses „Mainz 77“. NVZ, Heft 4.

DEZA (Direktion für Entwicklung und Zusammenarbeit) (2000). Externe Evaluation. DEZA, strategisches Controlling, 3003 Bern, http://www.deza.admin.ch

Fischer, G.H. (1995a). "The Linear Logistic Test Model". In: Fischer, G.H./Molenaar,I.W. "Rasch Models. Foundations, Recent Developments and Applications." 1995. S. 131-156. Berlin: Springer.

Fischer, G.H. (1995b). "Linear Logistic Models for Change". In: Fischer, G.H./Molenaar,I.W. "Rasch Models. Foundations, Recent Developments and Applications." 1995. S. 157-180. Berlin : Springer.

Fischer, G.H. (1995c). "Some neglected Problems in IRT". Psychometrika, Vol. 60, 1995,S. 459-487.

Glück, J.; Spiel, C. (1997). Zur Analyse individueller Veränderungsunterschiede mit Item-Response-Modellen - Eine Antwort auf Stelzl (1977). MPR-Online, 2, 51 -54.

Jacobshagen, W. (1997). Nachschulungskurse für alkoholauffällige Fahranfänger (NAFA). Köln: Verlag TÜV Rheinland.

Kerlinger, F.N. (1973). Foundations of behavional research (2nd. ed.). New York: Holb, Reichert und Winston.

Klein, S. und Krause, B. (1999). Validierung von Modellen der qualitativen Prä- und Posttest-analyse. In: B. Krause, P. Metzler. Empirische Evaluationsmethoden. Bd. 3. S. 7-26. Berlin: ZeE

Klein, S. (1999).Vergleich zweier unterschiedlicher Schätzmethoden bei Latent-Trait-Modellen zur Veränderungsmessung. In: B. Krause, P. Metzler. Empirische Evaluationsmethoden. Bd. 3. S. 27-46. Berlin: ZeE

Klein, S. (2002). Neue Methoden zur Entdeckung von Fehlspezifikationen bei Latent-Trait-Modellen der Veränderungsmessung. Dissertation an der Humboldt-Universität zu Berlin.

Lander, H.J. (1990). Die Abschätzung von Interventionseffekten mittels einer linearen Prä-Posttest-Analyse.. Z. Psychol., 198(2), 247 - 264.

Lander, H.J. (1997). Die statistische Beurteilung von Interventionseffekten mittels einer kategorialen Prä-Posttest-Analyse. In: B. Krause, P. Metzler. Empirische Evaluationsmethoden. Bd. 2. S. 17-28. Berlin: ZeE

Leitfaden für die Planung von Projekt- und Programmevaluation (1997). Fachbereich Evaluation, Bundesamt für Gesundheit, CH-3003 Bern, http://www.BAG.admin.ch

Metzler, P. und Krause, B. (1997). Methodischer Standard bei Studien zur Therapieevaluation. Methods of Psychological Research Online, Vol. 2, No. 1. http://www.papst-publishers.de/mpr/

Meyer-Gramcko, F. und Sohn, J.M. (1995). Qualitätsmanagement bei verkehrspsychologischen Rehabilitationsmaßnahmen. In: Risser, R. (Hrsg.). 35. BDP-Kongress für Verkehrspsychologie. Bonn: Deutscher Psychologen-Verlag

Meyer-Gramcko, F. und Sohn, J.M. (1995). 10 Jahre VerkehrspsychologischePraxis. Jahresbericht 1995. www.vpp.de/JB%201995.htm

Ponocny, I. (2002). On the applicability of some IRT models for repeated measurement designs. MPR-Online. 7. 21 - 40

Rost, J. (2002). Mixed and latent Markov models as item response models. MPR-Online, 7, 53 - 72

Sanders, J.R. (Hrsg.) (1999). Handbuch der Evaluationsstandards. Opladen: Leske + Budrich.

Spoerer, E. und Ruby, M.M. (1996). Zurück ans Steuer. Therapie und Praxis der Rehabilitation auffälliger Fahrer. Braunschweig: Rot-Gelb-Grün.

Thompson, G.H. (1924). A formula to correct for the effect of errors of measurement on the correlation of initial values with gains. J. exp. Psychol. 7, 321-324.

Thorndike, E.L. (1924). The influence of the chance imperfections of measures upon the relation of initial score to gain or loss. J. exp. Psychol. 7, 225-232.

Winkler, W., Jacobshagen, W. und Nickel, W.R. (1986). Die Wirksamkeit von Kursen für alkoholauffällige Kraftfahrer. Schlussbericht zum Forschungsprojekt 7714/4, 7714/7, 7714/10 der Bundesanstalt für Straßenwesen. Hannover: Technischer Überwachungs-Verein Hannover e.V.

Winkler, W., Jacobshagen, W. und Nickel, W.R. (1988). Die Wirksamkeit von Kursen für wiederholt alkoholauffällige Kraftfahrer. Bergisch Gladbach: Bundesanstalt für Straßenwesen, Unfall- und Sicherheitsforschung, Heft 64.

Wittmann, W.W. (1985). Evaluationsforschung. Aufgaben, Probleme und Anwendungen. Berlin: Springer-Verlag.

Wottawa, H.; Thierau, H. (1990). Lehrbuch Evaluation. Toronto: Huber.

Zwindermann, A.H. (1991). "A Generalized Rasch Model for Manifest Predictors." Psychometrika, Vol. 56. 1991. S. 589-600.