

Reimer Kornmann

Voraussetzungen für die Prüfung von Fördereffekten ohne Kontrollgruppenvergleich

Beitrag zur Frühjahrstagung der Arbeitsgruppe Empirische Sonderpädagogische Forschung (AESF) in
Würzburg, 6. / 7. Mai 2005

Zusammenfassung

Um Effekte von Fördermaßnahmen ohne die Untersuchung einer Kontrollgruppe abzuschätzen zu können, muss das Erfolgskriterium mittels eines psychometrischen Tests erfasst werden. Dieser wird bei jeder Versuchsperson einmal vor Beginn und einmal nach Beendigung der Förderung angewendet. Im Einzelfall kann eine Förderung dann als erfolgreich gelten, wenn die Differenz zwischen der ersten und der zweiten Messung einen kritischen Wert übersteigt. Dieser ergibt sich aus dem Standardmessfehler des Tests unter Berücksichtigung des Regressionseffekts und des Einflusses der Testversiertheit. Um die so zu bildende kritische Differenz bestimmen zu können, müssen zwei Merkmale des Tests bekannt sein: zum einen die Höhe der Retest-Reliabilität, deren Bestimmung sich auf den gleichen Zeitraum zwischen Erst- und Zweitestung beziehen sollte wie die Dauer der Förderung, zum anderen die Differenz zwischen den Mittelwerten der Erst- und Zweitestung bei der Bestimmung der Reliabilität.. Ein genereller Fördereffekt kann dann behauptet werden, wenn die Zahl der erfolgreich geförderten Personen signifikant höher ist als die Zahl der nicht erfolgreich geförderten. Das Vorgehen und die dazu erforderlichen Voraussetzungen werden unter forschungspraktischen, methodischen und erkenntnistheoretischen Gesichtspunkten diskutiert.

Eine Möglichkeit, Fördereffekte ohne Kontrollgruppenvergleich zu prüfen, ist recht bekannt: Man untersucht die gleiche Personenstichprobe mehrfach über einen längeren Zeitraum hinweg. Die Intervention setzt erst ein, nachdem bereits mehrere Messungen zur Erhebung der "baseline" vorliegen. Die Messungen werden in gleichen zeitlichen Intervallen wie vor Beginn der Intervention und auch nach Beendigung derselben fortgesetzt. Per Zeitreihenanalysen lässt sich dann prüfen, ob Veränderungen der Messwerte in zeitlichem Zusammenhang mit der Intervention stehen. Hierfür haben aus unserem Kreise Lauth & Fellner (1998) ein schönes Beispiel in der Festschrift für Fritz Masendorf gegeben (S. 119).

An diese Möglichkeit hatte ich allerdings nicht gedacht, als ich meinen Beitrag anmeldete. Ich hätte präziser formulieren müssen, um dabei hervorzuheben, dass die Untersuchungen auf nur zwei Messungen, eine vor Beginn, die andere nach Beendigung der Förderung, beschränkt seien.

Den Anstoß, über diese Möglichkeit nachzudenken, verdanke ich einigen meiner Studierenden. Von ihnen gefragt, ob sie eine von ihnen durchgeführte Förderung mehrerer Kinder mit dem gleichen Programm zum Inhalt ihrer Wissenschaftlichen Hausarbeit machen könnten, verweise ich auf die bekannten Eckpunkte des Kontrollgruppen-Design: eine Messung vor, die andere nach der Förderung bei jeweils einer Versuchs- und Kontrollgruppe. Mein Hinweis auf das Erfordernis einer Kontrollgruppe provoziert nicht selten einigen Widerstand seitens der Studierenden. Die von ihnen vorgebrachten Gegenargumente leuchten durchaus ein. Sie betreffen

- den zeitlichen und organisatorischen Mehraufwand
- die Tatsache, dass Erfolg versprechende Fördermaßnahmen einer mehr oder weniger systematisch ausgewählten Gruppe gezielt vorenthalten werden.

Ein weiteres methodisches Argument kann hinzugefügt werden:

Nicht immer ist auszuschließen, dass bestimmte Effekte der Förderung auch die Kontrollgruppe erreichen, etwa wenn sich die Lehrkräfte oder die Kinder der Kontrollgruppe kundig machen, was denn mit der Versuchsgruppe geschieht. Solche Effekte erschweren es, die Nullhypothese zurückzuweisen, obwohl sie eigentlich zu verwerfen wäre, sie begünstigen also den Fehler 2. Art oder den Beta-Fehler und würden ein eigentlich wirksames Programm fälschlicherweise als unwirksam ausweisen. Dies ist sicherlich eine in vieler Hinsicht unerwünschte Konsequenz.

Mit diesem Problemkomplex wurde ich vor einigen Monaten wieder einmal konfrontiert, als eine Studierende vorschlug, eine kleine Gruppe graphomotorisch auffälliger Kinder im Vorschulalter gezielt zu fördern und sich meinem Ansinnen, eine Kontrollgruppe zu rekrutieren und zu untersuchen, argumentativ widersetze.¹ Im Verlauf unserer Diskussion griff ich auf einen Gedanken zurück, den ich in diesem Kreise bisweilen zwar angedeutet, aber nie ausführlich dargestellt habe. Ich vermute, dass meine Überlegungen unter bestimmten Bedingungen eine bessere Alternative zu der gängigen experimentell orientierten Forschungspraxis bieten könnten. Bei der Darstellung soll auch noch deutlich werden, dass das vorgeschlagene Vorgehen über die Lösung der oben angedeuteten Probleme hinaus einen weiteren forschungsmethodischen Vorteil bietet.

Der Grundgedanke entstammt der Psychometrischen Einzelfalldiagnostik (Huber, 1973):

Jedem gemessenen oder beobachteten Wert X_0 entspricht ein sogenannter "wahrer" Wert X_t .

$$X_0 \sim X_t.$$

¹ Für die gelungene Überzeugungsarbeit danke ich hier stellvertretend für weitere Studierende Frau Jennifer Soboll, deren Daten auch hier zum Teil verwendet wurden.

Dieser "wahre" Wert liegt mit einer bestimmten Wahrscheinlichkeit p innerhalb eines Konfidenzintervalls (KI). Dieses ergibt sich aus der Größe des Standardmessfehlers (S_e) des verwendeten Tests:

$$s_e = s_x \cdot z_{\alpha} \sqrt{1 - r_{tt}}$$

und wird symmetrisch um den gemessenen Wert X_0 gelegt:

$$KI = X_0 - s_e < X_t < X_0 + s_e$$

Bei der Psychometrischen Einzelfalldiagnostik wird u. a. nach der Signifikanz von intraindividuellen Veränderungen gefragt. Solche intraindividuellen Veränderungen werden angenommen, wenn bei einer wiederholten Messung der ermittelte Testwert X_{02} außerhalb des Konfidenzintervalls liegt, das den ersten Testwert X_{01} umgibt:

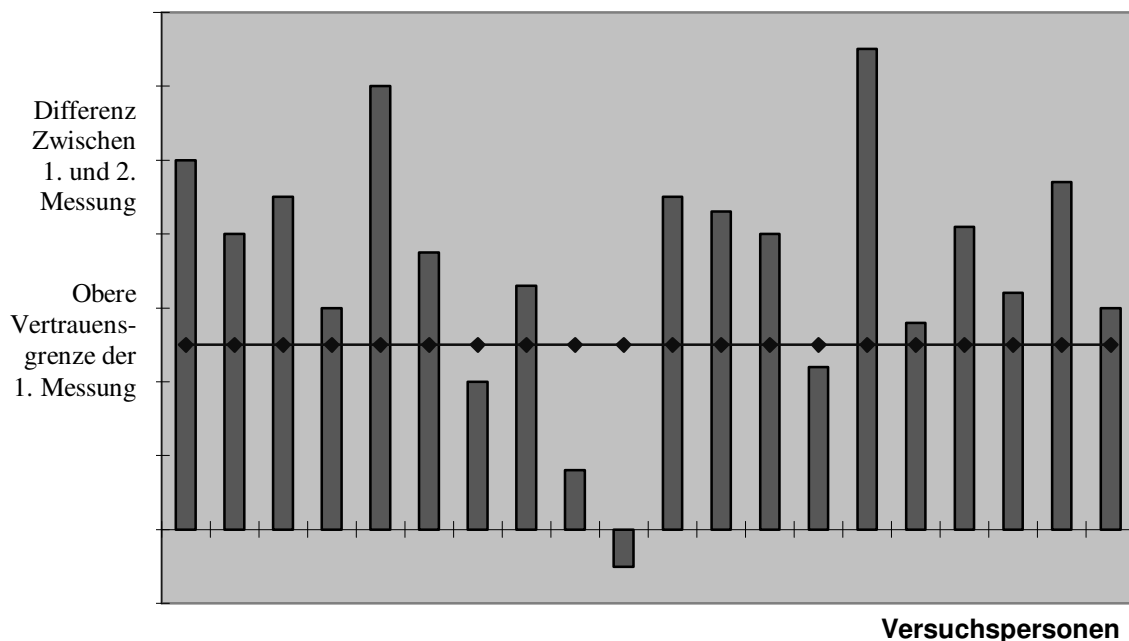
$$X_{02} > X_{01} + s_e \text{ bzw. } X_{02} < X_{01} - s_e$$

Diese Formel mag für den Hausgebrauch genügen. Sie ist aber im Hinblick auf eine bessere Realitätsangemessenheit noch etwas zu modifizieren. Auf diese Modifikationen gehe ich später ein. Festzuhalten ist aber schon eine erste wichtige Voraussetzung, die in der Praxis auch oft erfüllt ist: Die Reliabilität des Messinstruments, mit dem der Messwert X_0 ermittelt wird, muss bekannt sein. Dabei sollte es sich um die Retest-Reliabilität handeln, wobei das zeitliche Intervall zwischen Erst- und Zweitmessung etwa der Periode entsprechen sollte, welche auch zwischen der ersten und zweiten Messung im Einzelfall liegt.

Dies vorausgesetzt, lassen sich die Überlegungen schnell und einfach zu Ende führen:

Wie bei einem Kontrollgruppen-Design werden von jeder geförderten Person zwei Messwerte erhoben: einer vor Beginn der Förderung (X_{01}), der andere nach Beendigung der Förderung (X_{02}). Dann wird für jede einzelne Person die Differenz zwischen diesen beiden Messwerten gebildet. Ist diese Differenz größer als der Standardmessfehler, dann zählt die entsprechende Person zu den

erfolgreich Geförderten, ist sie geringer, dann zählt sie zu den nicht erfolgreich Geförderten. In einem letzten Schritt wird mittels eines einfachen statistischen Tests für Nominaldaten geprüft, ob sich die Anzahl der erfolgreich Geförderten von der der nicht erfolgreich Geförderten signifikant unterscheidet. Ein fiktives Beispiel zeigt nachfolgende Abbildung: Alle Versuchspersonen, deren Messwertdifferenzen oberhalb des kritischen Wertes liegen, gelten als erfolgreich gefördert, alle, deren Messwertdifferenzen darunter liegen, werden als nicht erfolgreich Geförderte klassifiziert. In unserem Beispiel ist das Verhältnis zwischen erfolgreich Geförderten zu nicht erfolgreich Geförderten 16 : 4.



Diese einfache Methode hat gegenüber dem statistischen Vergleich der Mittelwerte von Versuchs- und Kontrollgruppe einen weiteren Vorteil: Sie liefert exakte Anhaltspunkte, welche Personen in welchem Ausmaß von der Förderung profitiert haben und welche nicht. Diese zusätzliche Information ermöglicht also weitere Differenzierungen bei der dann folgenden praktischen Arbeit ebenso wie bei der theoretischen Einordnung der Ergebnisse.

Soweit also die Grundgedanken.

Ich hatte bereits angedeutet, dass das hier grob skizzierte Vorgehen noch einiger Verfeinerungen bedarf. Diese sollen nachfolgend dargestellt und an einem gerade gewonnenen praktischen Beispiel erläutert werden. Dabei handelt es sich um die eingangs schon erwähnte Förderung von vier graphomotorisch auffälligen Kindern im Vorschulalter mittels eines eigens zusammengestellten Förderungsprogramms mit Anregungen vor allem von Oussoren-Voors (2004) sowie Pauli & Kisch (2003). In einem Zeitraum von vier Wochen wurden neun Sitzungen mit den Kindern durchgeführt. Die geringe Zahl dieser Kinder spielt keine Rolle, um das weitere Vorgehen zu erläutern.

Vor Beginn der Förderung und nach Beendigung der Förderung wurden die Kinder mit der "Graphomotorischen Testbatterie (GMT)" von Rudolf (1986) untersucht.

Ich greife die Werte von Patrick heraus. Vor der Förderung erzielte er 32 Punkte und nach der Förderung 42 Punkte, er verbesserte sich also augenscheinlich um 10 Punkte.

1. Modifikation: Berücksichtigung einer möglichen Testversiertheit

Bekanntlich fallen Messergebnisse im Leistungsbereich bei einer Messwiederholung - unabhängig von der Reliabilität des Messinstruments und der Höhe des Messwerts bei der ersten Untersuchung - im Durchschnitt etwas besser aus. Dies wird mit der Tatsache erklärt, dass die untersuchten Personen mit der Situation und den Anforderungen der Untersuchung bei der zweiten Testung besser vertraut sind und deshalb einen höheren Anteil ihres Leistungspotenzials abrufen können (Schmidt, 1971). Klauer (1993) fordert daher, diesen Effekt bei der Bildung der Differenzen zwischen der ersten und der zweiten Testung zu berücksichtigen. Dazu genügt es sicher, wenn die einfache Differenz der Arithmetischen Mittel (AM), die bei der Bestimmung der Retest-Reliabilität ohnehin anfällt, zu dem Messwert der ersten Untersuchung addiert wird. Das Konfidenzintervall wird also nicht um den ursprünglich

gemessenen Wert X_{01} , sondern um den um Wert $X_{01} + (AM_2 - AM_1)$, der um die Differenz der Mittelwerte erhöht ist, gelegt.

Üblicherweise werden die Arithmetischen Mittelwerte der 2. Messung, die bei der Bestimmung der Retest-Reliabilität von Tests anfallen, in den Testhandbüchern nicht mitgeteilt. Man kann aber versuchen, sie bei den Testautoren oder Verlagen zu erfragen. Dank des freundlichen Entgegenkommens von Verlag und Testautor gelang dies im vorliegenden Falle problemlos.² Allerdings lagen nur die über alle Altersstufen hinweg gewonnenen Arithmetischen Mittelwerte und Standardabweichungen der 1. und 2. Testung vor mit $AM_1 = 77.41$ und $AM_2 = 79.39$. Die Differenz beträgt also 1.98 Punkte, Die um diesen Wert bereinigte Verbesserung von Patrick beträgt nunmehr

$$10 - 1.98 \text{ also } 8.02 \text{ Punkte,}$$

d.h. sein zweiter Messwert ist mit $X_{0\text{korr}} = 40.02$ statt mit $X_2 = 42$ anzusetzen.

Zur Berechnung des Konfidenzintervalls müssen die Standardabweichung und die Test-Retest-Reliabilität bekannt sein. Zur Standardabweichung lagen im Testmanual mehrere Angaben aus verschiedenen Teiluntersuchungen vor, die zwischen $s = 20,3$ und $s = 28,6$ variierten. Für die Altersgruppe von Patrick wurde $s = 25,8$ angegeben. Für die Retest-Reliabilität fand sich Testhandbuch der erstaunlich hohe Wert von $r_{tt} = 0.98$, der sicherlich nicht repräsentativ für ansonsten anfallende Stichproben ist. Trotzdem wurde er mangels verfügbarer Alternativen verwendet.

$$s_e = 25.8 \cdot 1.96 \sqrt{1 - 0.98} = \underline{\underline{7.15}}$$

Die von Patrick erzielte Leistungsverbesserung von 8.02 Punkten ist also größer als der Standardmessfehler, und sein bei der 2. Messung erzielter und bereinigter Wert liegt außerhalb des ermittelten Konfidenzintervalls.

$$KI_{PA} = 22.85 < 32 < 39.15 < 40.2$$

² Dem Beltz-Verlag und Herrn H. Rudolf sei auch an dieser Stelle herzlich gedankt!

2. Modifikation: Berücksichtigung des Regressionseffekts³

Bei Messwiederholungen besteht eine Regression zur Mitte, d. h. die Messwerte verschieben sich mit einer leichten Tendenz zur Mitte, die aber umso größer ist, je stärker die Messwerte der ersten Messung vom Arithmetischen Mittel differieren. Man geht also davon aus, dass der Messfehler um so größer ist, je extremer der jeweils gemessene Wert ausfällt (Lord & Novick, 1968). Zur Korrektur dieser absehbaren Tendenz wird der wahre Wert X_t regressionsanalytisch geschätzt:

$$X_t = X_0 \cdot r_u + (1 - r_u) \cdot AM$$

Diese regressionsanalytische Schätzung hat übrigens den gleichen Effekt wie die Möglichkeit, das Konfidenzintervall nicht symmetrisch, sondern in zwei ungleichen Abständen um den gemessenen Wert zu legen (vgl. Nonnally, 1967). Dabei wäre dann der zum Mittelwert hin ausgerichtete Teil des Intervalls der größere. Wird auf diese regressionsanalytische Schätzung des wahren Werts verzichtet, unterstellt man stillschweigend die Gültigkeit der sogenannten Äquivalenz-Hypothese.

Bei Patrick wurde die regressionsanalytische Schätzung für beide Messwerte vorgenommen.

$$X_{1\text{korr Pa}} = 32 \cdot 0.98 + (1 - 0.98) \cdot 77,41 = \underline{\underline{32.9}}$$

$$X_{2\text{korr Pa}} = 40.02 \cdot 0.98 + (1 - 0.98) \cdot 79,39 = \underline{\underline{40.8}}$$

Die regressionsanalytisch geschätzten wahren Werte betragen nun $X_{t1} = 32.9$ für die 1. Messung und $X_{t2} = 40.8$ für die 2. Messung. Die Differenz ist dabei mit 7.9 Punkten gegenüber der Schätzung unter der Äquivalenzhypothese etwas geringer geworden. Allerdings bezieht sich bei der regressionsanalytischen

³ Für wichtige Impulse zu den Überlegungen dieses und des nächsten Abschnitts sowie der abschließenden Diskussion danke ich meinem Kollegen Gerhard Eberle.

Schätzung der wahren Werte auch der Standardmessfehler auf dieselben. Er wird berechnet nach der Formel

$$s_{ereg} = s \cdot z_{\alpha} \sqrt{r_{tt}(1 - r_{tt})}$$

Für Patrick ergibt dies:

$$s_{ereg} = 25.8 \cdot 1.96 \sqrt{0.98 \cdot 0.02} = \underline{\underline{7.08}}$$

Das Konfidenzintervall lautet nun

$$KI_{reg\ Pa} = 25.82 < 32.9 < 39.98 < 40.8$$

Der geschätzte wahre Wert der 2. Messung $X_{t2} = 40.8$ liegt außerhalb dieser Grenzen.

3. Modifikation: Die Korrektur nach der tau-Normierung (sofern man normieren will)

Bei der üblichen sogenannten x-Normierung geht in die Normierungsgleichung die Standardabweichung ein, die an einer gezogenen Stichprobe gewonnen wurde. Genauer wäre es aber, die Standardabweichung der geschätzten wahren Werte zu berücksichtigen. Diese beziehen sich auf die gesamte Population und ergeben sich als Produkt der ermittelten Standardabweichung s_x mit der Quadratwurzel des empirisch ermittelten Reliabilitätskoeffizienten r_{tt} :

$$s_t = s_x \cdot \sqrt{r_{tt}}$$

Folglich errechnen sich die unter der tau-Normierung zu transformierenden Werte von Patrick wie folgt:

$$z_{1t\ Pa} = \frac{32.9 - 77.41}{25.8 \cdot \sqrt{0.98}} = \underline{\underline{-1.74}}$$

$$z_{2t\ Pa} = \frac{40.8 - 79.39}{25.8 \cdot \sqrt{0.98}} = \underline{\underline{-1.51}}$$

Die Differenz zwischen $z_1 = -1.74$ und $z_2 = -1.51$ beträgt $z = 0.23$. Das nach z umgerechnete Konfidenzintervall beträgt 0.28 . und hat für $z = -1.74$ die obere Grenze von $z = -1.46$.

$$KI_{Pa} = -1.97 < -1.74 < -1.46 > -1.51$$

Der von Patrick erzielte Wert von $z = -1.50$ fällt also noch knapp in diesen kritischen Bereich, so dass er unter der strengen Voraussetzung der tau-Normierung bei Berücksichtigung der Regressionshypothese und des Effekts der Testversiertheit jetzt knapp das Erfolgskriterium verfehlt.

Die hier geschilderte Prozedur lässt sich nun für jede Person, die einem bestimmten treatment unterzogen wurde, einzeln durchführen. Sie ist weit weniger aufwändig als dies aufgrund meiner recht breiten Schilderung erscheinen mag und lässt sich schnell in eine Routine überführen. Für jede Person ist dann nur noch zu bestimmen, ob sie das gesetzte Erfolgskriterium erreicht hat oder nicht. Ist die Anzahl von untersuchten Personen hinlänglich groß, lässt sich statistisch prüfen, ob die Anzahl der erfolgreich Geförderten signifikant größer ist als die der nicht erfolgreich Geförderten. Dies wäre sinnvoll, wenn die Qualität des Förderungsprogramm evaluiert werden soll.

Ich komme zum Schluss.

Aus meiner Darstellung sollten bereits die Bedingungen hervorgegangen sein, unter denen die Prüfung von Fördereffekten ohne Kontrollgruppenvergleich möglich ist. Ich fasse sie noch einmal zusammen:

1. Zur Ermittlung des Fördererfolgs ist ein Messinstrument, am besten ein psychometrisch konstruierter Test, zu verwenden, dessen Retest-Reliabilität bekannt sein muss. Auf dieser Basis ist das Konfidenz-Intervall zu berechnen. Der zeitliche Abstand zur Ermittlung dieses Kennwerts sollte in etwa auch der Dauer der Förderung entsprechen.

2. Zur Kontrolle eines möglichen Effekts der Testversiertheit müssen die Arithmetischen Mittelwerte der ersten und der zweiten Untersuchung, die zur Ermittlung der Retest-Reliabilität durchgeführt wurden, vorliegen.
3. Mögliche Regressionseffekte sollten bei der Bestimmung der individuellen Ausgangslagen und der Erfolgskriterien berücksichtigt werden. Dies ist umso wichtiger, je extremer die individuelle Ausbildung dieser Merkmale ist.

Hinzuweisen ist auch noch auf die Tatsache, dass die tau-Normierung strengere Kriterien liefert als die übliche x -Normierung. -

Soweit der nahezu unverändert gelassene Text meines Vortrags in Würzburg. Ich gehe nun auf einzelne Punkte ein, die sich aus der daran anschließenden recht lebhaften Diskussion ergeben haben.

1. Für das Beibehalten einer Kontrollgruppe wurde das Argument angeführt, dass dadurch der experimentelle Charakter der Untersuchung gewahrt bleibe. Man könne auf diese Weise den Einfluss verschiedener Faktoren mehr oder weniger genau kontrollieren bzw. ausschalten.

Grundsätzlich ist zu betonen, dass sich meine Vorschläge auf dem Feld des quasiexperimentellen Vorgehens bewegen und ihre Aussagekraft nicht den hohen Ansprüchen rein experimentellen Vorgehens genügt. Dazu wäre es beispielsweise erforderlich, zunächst die gesamte Untersuchungsstichprobe nach dem Zufallsprinzip auszuwählen, um dann die Versuchs- und Kontrollgruppe aufgrund einer weiteren Randomisierung zu bilden (vgl. dazu Bortz, 1984, S. 400ff).

Solche Voraussetzungen lassen sich im Bereich der sonderpädagogisch relevanten Forschung wohl kaum realisieren (siehe dazu auch die forschungskritischen Anmerkungen von Gadenne, 1976, S. 76ff). Immerhin kann aber mit der hier vorgeschlagenen Vorgehensweise die Frage geprüft

werden, ob signifikante Veränderungen in der beabsichtigten Richtung eingetreten sind. Sofern sich diese Frage positiv beantworten lässt, ist weiterhin der Schluss berechtigt, dass die gewählten Interventionen diese gewünschten Veränderungen nicht verhindert haben. Eine solche Aussage darf dann als Hypothese verallgemeinernd beibehalten werden, bis sie widerlegt oder modifiziert worden ist.

2. Es wurde vermerkt, dass sehr wohl Bezug auf eine Vergleichsgruppe genommen wurde, und zwar auf die Personenstichprobe, die zur Bestimmung der Retest-Reliabilität untersucht worden sei.

Dies ist selbstverständlich korrekt. Mein Vorschlag zielt also darauf ab, vorhandenes Datenmaterial im Sinne der drei eingangs aufgeführten Argumente (Arbeitersparnis, Gleichbehandlung der Versuchspersonen, Vermeidung des Beta-Fehlers) zu nutzen. Um den quasiexperimentellen Charakter der Untersuchungen zu wahren, ist im übrigen ist zu fordern, dass die Stichprobe, die zur Bestimmung der Retest-Reliabilität untersucht wurde, repräsentativ ist für die Angehörigen aller in Betracht kommenden Zielgruppen des Tests. So sollte man von einem guten psychometrischen Test erwarten, dass er diese Voraussetzung erfüllt. Zumindest fordern dies die "Standards für pädagogisches und psychologisches Testen" (Häcker, Leutner & Amelang, 1998). Will man mehrere unabhängige Variablen berücksichtigen, kann man diese durchaus in jedem Einzelfall induzieren, entsprechende Gruppenbildungen vornehmen und die entsprechenden statistischen Auswertungen mit solchen Verfahren vornehmen, die sich für den Vergleich mehrerer unabhängiger Verteilungen von Alternativdaten eignen.

3. In diesem Zusammenhang wurde eingewendet, dass nur in sehr seltenen Fällen Datenmaterial mit den erforderlichen Angaben (Retest-Reliabilitäts-Koeffizienten mit passenden Zeitintervallen und Parameter der zweiten Testung) verfügbar seien.

Auch dies lässt sich nicht bestreiten. Gerade deswegen und im Sinne der dargelegten und begründeten Argumente für einen Verzicht auf Untersuchungen von Kontrollgruppen ist zu fordern, dass alle veröffentlichten psychometrischen Tests die erforderlichen Angaben enthalten. Deswegen sollte man auch nicht vor der Forderung zurückschrecken, mehrere Retest-Reliabilitäts-Koeffizienten für unterschiedliche Zeitintervalle und Personenstichproben zu fordern. Genau dieses sehen ja auch die "Standards" (Häcker, Leutner & Amelang, 1998) vor, an denen sich Testautoren und -verlage eigentlich orientieren sollten. Der hierfür aufgebrauchte Arbeitsaufwand bei der Testentwicklung entlastet jedenfalls die forschenden Testanwender und führt zu einer erheblichen Qualitätsverbesserung sowohl der Instrumente als auch der damit möglichen Forschungsarbeit. Das in diesem Zusammenhang eingebrachte Argument, dass Retest-Reliabilitätskoeffizienten nicht allein sinnvoll zur Einschätzung der Testgüte seien und man deswegen auf Testhalbierungs-Koeffizienten bzw. Angaben zur internen Konsistenz zurückgreifen sollte, ist selbstverständlich richtig. Für den hier vorgestellten Anwendungsbereich von Tests ist aber der Aspekt der Messwiederholung und ihrer Zuverlässigkeit entscheidend.

4. Gegen das Argument, dass der Verzicht auf eine Kontrollgruppe sowohl zu einer Reduzierung des Arbeitsaufwands und der damit zusammenhängenden organisatorischen Probleme führe als auch zu einer größeren Gerechtigkeit beitrage, wurde angeführt, dass Arbeitskapazitäten stets begrenzt seien und auch stets nur ein verschwindend kleiner Teil derjenigen Personen untersucht werde, für die die Intervention in Betracht komme.

Dieser Einwand trifft unter der Voraussetzung zu, dass mit dem Forschungsergebnis selbst keine weiteren praktischen Zielsetzungen mehr verfolgt werden und Forschungsarbeit primär theoretischen Überlegungen folgt und daher nach streng experimentellem Muster anzulegen sei. Die Zielsetzungen meines Vorschlags gehen jedoch - wie bereits unter Punkt 1) vermerkt - nicht in diese Richtung.

Literatur

Bortz, J. (1984).

Lehrbuch der empirischen Forschung. Heidelberg: Springer.

Gadenne, V. (1976).

Die Gültigkeit psychologischer Untersuchungen. Stuttgart: Kohlhammer.

Häcker, H., Leutner, D. & Amelang, M. (1998).

Standards für pädagogisches und psychologisches Testen. Supplementum 1/1998 der Diagnostica und der Zeitschrift für Differentielle und Diagnostische Psychologie. Göttingen: Hogrefe und Bern: Huber.

Huber, P. J. (1973).

Psychometrische Einzelfalldiagnostik. Weinheim: Beltz.

Klauer, K. J. (1993).

Learning potential testing: the effect of retesting. In J. H. M. Hamers, A. J. J. M. Ruijsenaars & K. Sijtsma (Eds.), Learning potential assessment. Theoretical, methodological and practical issues (pp. 135-152). Amsterdam: Swets & Zeitlinger.

Lauth, G. W. & Fellner, C. (1998).

Evaluation eines multimodalen Therapieprogramms bei Aufmerksamkeitsdefizit-/ Hyperaktivitätsstörung über eine differenzierte Einzelfallforschung. In M. Greisbach, U. Kullik & E. Souvignier (Hrsg.), Von der Lernbehindertenpädagogik zur Praxis schulischer Lernförderung (S. 109-124). Lengerich: Pabst.

Lord, F. M. & Novick, M. R. (1968).

Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Weseley.

Nonnally, J. C. (1967).

Psychometric Theory. New York: McGraw-Hill.

Oussoren-Vors, R. (2004).

Schreibtanzen I. Von abstrakten Bewegungen zu konkreten Linien – für 3-8jährige Kinder. Dortmund: Borgmann.

Pauli, S. & Kisch, A. (2003).

Geschickte Hände. Feinmotorische Übungen für Kinder in spielerischer Form. Dortmund: Borgmann.

Rudolf, H. (1986).

Graphomotorische Testbatterie (GMT). Weinheim: Beltz.

Schmidt, L. R. (1971).

Testing-the-Limits im Leistungsverhalten. Möglichkeiten und Grenzen. In E. Duhm (Hrsg.), Praxis der Klinischen Psychologie, Bd. II (S. 9-29). Göttingen: Hogrefe.