*Thomas Günther*[1,2] *& Bruno Fimm*[3]

# Wie messe ich im Einzelfall Leistungsverbesserungen?

**Drei Methoden, um Fortschritte zuverlässig zu beurteilen**

**ZUSAMMENFASSUNG.** Nachmessungen nach erfolgter therapeutischer Intervention werden häufig durchgeführt, um den Fortschritt des Patienten zu überprüfen. Eine höhere Punktzahl bei einer Folgemessung kann das Ergebnis eines tatsächlichen Fortschritts sein. Es kann aber auch sein, dass der Patient ein besseres Ergebnis erzielt, weil er das Instrument zum zweiten Mal sieht und bearbeitet oder sich in diesem Moment besser konzentrieren kann. Aber wie kann man sicher wissen, ob der Patient relevante Fortschritte erzielt hat? Dieser Artikel beschreibt drei Methoden, die Therapeuten in der täglichen Praxis anwenden können, um den Fortschritt eines Patienten zuverlässig zu beurteilen und zu dokumentieren. Mit dem Vertrauensintervall kann mithilfe eines Zuverlässigkeitswertes aus dem Testhandbuch eine Aussage getroffen werden, in welchem Bereich der tatsächliche Punktwert liegt. Die kritische Differenz gibt an, wie viel Veränderung erfolgen muss, um von einer klinisch relevanten Verbesserung sprechen zu können. Der McNemar-Test kann anstelle der kritischen Differenz verwendet werden, wenn der verwendete Test keine Zuverlässigkeitswerte aufweist. Alle drei Methoden sind so beschrieben, dass sie einfach, schnell und mit frei verfügbaren Programmen durchgeführt werden können.

## Einleitung

Dieser Artikel soll eine Hilfestellung geben, um zwei Messungen in einem therapeutischen Setting miteinander zu vergleichen. Dies ist beispielsweise der Fall, wenn ein Therapeut in einer Nachmessung untersuchen möchte, ob sein Patient tatsächlich Fortschritte erzielt hat. Ein höherer Wert in der Nachmessung kann auf tatsächliche Fortschritte des Patienten zurückzuführen sein. Er könnte jedoch möglicherweise auch daraus resultieren, dass der Patient das Instrument zum zweiten Mal gesehen oder zum zweiten Testzeitpunkt besser geschlafen hat.

Dies ist vergleichbar mit den täglichen Schwankungen des Körpergewichts. An einem Tag kann das Gewicht, basierend auf dem natürlichen Input und Output des Körpers, bis zu 2,5 Kilogramm schwanken. Bei solchen Schwankungen stellt sich die Frage, wann eine Gewichtszunahme oder Gewichtsabnahme eine „echte" und „relevante" Veränderung ist. In diesem Beispiel (Körpergewicht) liegt die Ungenauigkeit darin, dass sich das zu Messende in einem bestimmten Rahmen verändert und nicht stabil ist. Zudem kann das verwendete Messgerät aber auch ungenau sein. Eine Badezimmerwaage ist gut geeignet, um Körpergewicht kilogrammgenau zu messen. Die gleiche Waage ist jedoch zum Kochen nicht brauchbar, wenn grammgenau gemessen werden muss. In diesem Falle ist das verwendete Instrument nicht ausreichend exakt.

Bei Instrumenten, die psychische Merkmale messen, ist diese Art der Ungenauigkeit normal (z.B. Messung von Wortschatz oder Sprachverständnis). Wie beim Körpergewicht sind psychische Merkmale zudem nicht stabil und können in Abhängigkeit von Situation, Untersucher oder Tagesform schwanken. Diese Ungenauigkeiten werden in der Testtheorie „Messfehler" genannt (*Moosbrugger & Kelava* 2012).

In einer Nachmessung soll beispielsweise der Erfolg einer Therapie überprüft werden. Beim Vergleich von Vor- und Nachmessung ist es nun entscheidend, die Größe des Messfehlers des verwendeten Testinstruments zu berücksichtigen, um zu beurteilen, ob es sich um eine signifikante Verbesserung handelt. Signifikant bedeutet, dass der Patient sich substanziell verbessert – mehr, als aufgrund von Zufall (normalen Schwankungen) und/oder Messungenauigkeiten zu erwarten wäre. Die Testentwickler kennen das Problem dieser Schwankungen und Ungenauigkeiten und versuchen die Messungenauigkeiten so klein wie möglich zu halten. Daher wird in den meisten Testverfahren ein Wert für Zuverlässigkeit (Reliabilität) angegeben.

Im folgenden Abschnitt werden zunächst unterschiedliche Zuverlässigkeitswerte beschrieben, sodass bei den späteren Verfahren der richtige Wert ausgewählt werden kann. Anschließend wird aufgezeigt, wie in kurzer Zeit und mit frei verfügbaren Programmen aufgrund der Zuverlässigkeitswerte ein Vertrauensintervall berechnet werden kann (z.B. „mit 90%iger Sicherheit liegt der Wert des Patienten in einem Bereich von X bis Y").

Im folgenden Abschnitt wird beschrieben, wie man für einen Test eine „kritische Differenz" berechnen kann. Dies ist ein Wert, um den sich ein Patient mindestens verändern muss, damit von einer signifikanten Verbesserung (oder Verschlechterung) gesprochen werden kann. Anschließend wird noch eine Methode beschrieben (McNemar-Test), mit der man relevante Veränderungen bei Tests, die sich aus dichotom bewerteten Items zusammensetzen, untersuchen kann, falls das Testverfahren keine Zuverlässigkeitswerte hat. Abschließend erfolgt eine kurze Anleitung, wie die Ergebnisse solcher Verfahren in einem Bericht zusammengefasst werden können.

## Zuverlässigkeit

Die Zuverlässigkeit bzw. Reliabilität gibt an, inwieweit eine Messung frei von Messfehlern ist. Ein Rohwert aus einem Test (z.B. Anzahl richtiger Antworten in einem Sprachtest) setzt sich nach der klassischen Testtheorie aus zwei Teilen zusammen: dem tatsächlichen (wahren) Wert des Patienten und einem Messfehler. Wäre ein Test frei von Messfehlern, wäre der

1 Zuyd University, Faculty of Health, Heerlen, The Netherlands
2 Lehr- und Forschungsgebiet für klinische Neuropsychologie des Kindes- und Jugendalters; Uniklinik RWTH Aachen
3 Klinik für Neurologie, Uniklinik RWTH Aachen

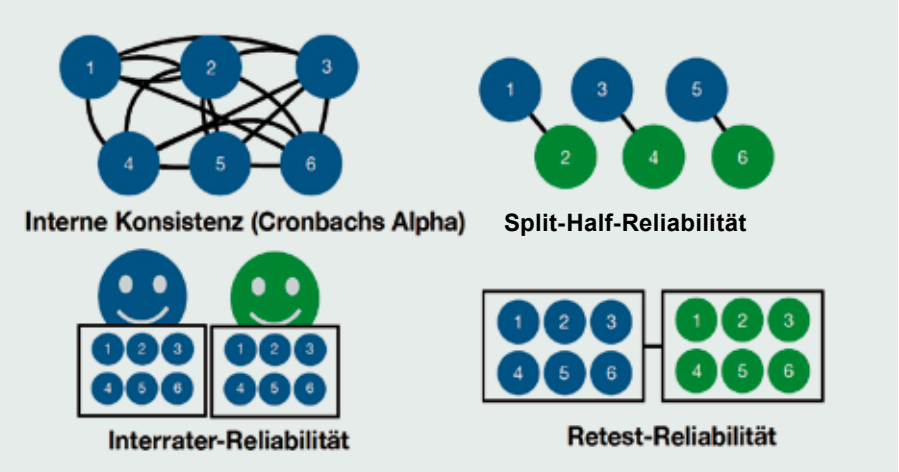
gemessene Testwert identisch mit dem wahren Wert des Patienten. Die Zuverlässigkeit eines Tests gibt daher an, inwieweit Messungen von zufälligen Faktoren beeinflusst werden. In Testhandbüchern wird die Zuverlässigkeit mit einer Zahl zwischen 0 (nicht zuverlässig) und 1 (perfekte Zuverlässigkeit, keine Messfehler) ausgedrückt. Hierbei gilt die Faustregel, dass Reliabilitäten kleiner als 0,6 unzureichend sind (*Kersting* 2006).

Die Zuverlässigkeit kann auf verschiedene Arten berechnet werden. In Testhandbüchern wird meist mindestens eine der folgenden vier Methoden verwendet, um die Zuverlässigkeit anzuzeigen (siehe auch Abb. 1):

- Die *Interne Konsistenz (Cronbachs Alpha)* gibt an, inwieweit verschiedene Items in einem Test, die dasselbe Merkmal messen sollen, dies auch tatsächlich tun. Die interne Konsistenz beschreibt demnach den Zusammenhang zwischen den einzelnen Items eines Tests und somit dessen Homogenität.
- Die *Split-Half-Reliabilität* basiert auf dem Zusammenhang von zwei Testhälften (als zwei parallele Tests konzipiert). Zum Beispiel kann die Aufteilung in zwei Hälften erfolgen, indem die geraden Items in der einen Hälfte und die ungeraden Items in der anderen Hälfte platziert werden.
- Bei der *Retest-Reliabilität* werden die Messungen zweier Zeitpunkte bei Verwendung des gleichen Testverfahrens miteinander korreliert. Eine hohe Retest-Reliabilität zeigt dabei an, dass (1) in der Normierungsstichprobe keine Leistungsunterschiede zwischen den beiden Zeitpunkten bestehen oder aber (2), dass systematische Veränderungen, die die Mehrzahl der Probanden aufweisen (z.B. steigt die Leistung für die meisten Probanden um einen konstanten Betrag) vorliegen. Die Retest-Reliabilität hängt sowohl vom Zeitintervall zwischen den beiden Messungen, vom gemessenen Merkmal (dies kann recht stabil oder aber auch situativ schwankend sein) als auch von der Güte des Testverfahrens selbst ab.
- Die *Interrater-Reliabilität* ist der Grad der Übereinstimmung zwischen verschiedenen Beurteilern. Mehrere Prüfer verwenden denselben Test bzw. dasselbe Ratingverfahren, um dieselbe Situation oder Person zu bewerten. Wenn dies zu den gleichen Ergebnissen führt, wird davon ausgegangen, dass die persönlichen Merkmale des Prüfers keinen Einfluss auf die Durchführung des Instruments haben.

Weitere Einzelheiten zur Bedeutung der Zuverlässigkeit bei psychometrischen Tests werden in *Bühner* (2011), *Lane* et al. (2015) oder *Price* (2017) beschrieben.

**Abb. 1: Schematische Übersicht der beschriebenen Zuverlässigkeitswerte**



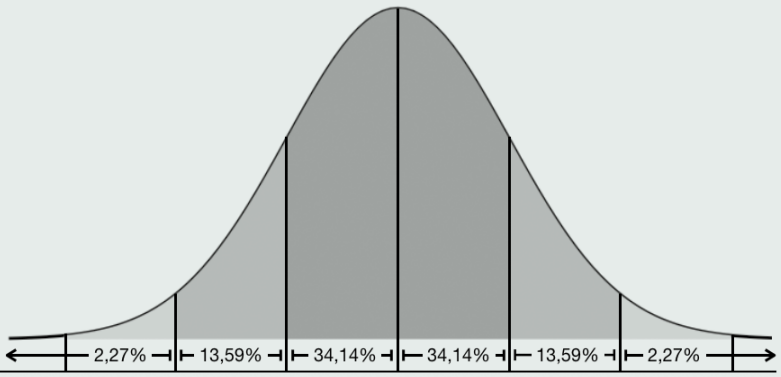

## Vertrauensintervall (VI)

Mit einem Testinstrument erhält man zunächst einen Rohwert, beispielsweise die Anzahl der richtigen Antworten. Dieser Rohwert wird bei einem Test mithilfe von Normwerttabellen in einen Normwert umgewandelt. Durch den Normwert wird das Ergebnis einer Person mit dem einer Normgruppe vergleichbar (z.B. vergleichbares Alter, Bildungsstand, Geschlecht). Es gibt verschiedene Normwerte, die alle eine relative Position auf einer Normalverteilung angeben. Die am häufigsten verwendeten sind Prozentränge sowie Standardnormen wie z.B. T-, z- und IQ-Werte (Abb. 2).

Für die Veränderungsmessung sind lediglich die Standardnormen geeignet, da man nur mit diesen (aufgrund deren Intervallskalenqualität) Differenzen berechnen kann. Prozentränge sind hierfür nicht geeignet, da sie nur Ordinalskalen-Niveau aufweisen (d.h. nur Aussagen wie „Prozentrang X ist kleiner/größer als Prozentrang Y" erlauben). Sofern das Manual eines Testverfahrens nur Prozentränge aufweist, können diese auch in eine Standardnorm umgewandelt werden (siehe hierzu: www.psychometrica.de/normwertrechner.html).

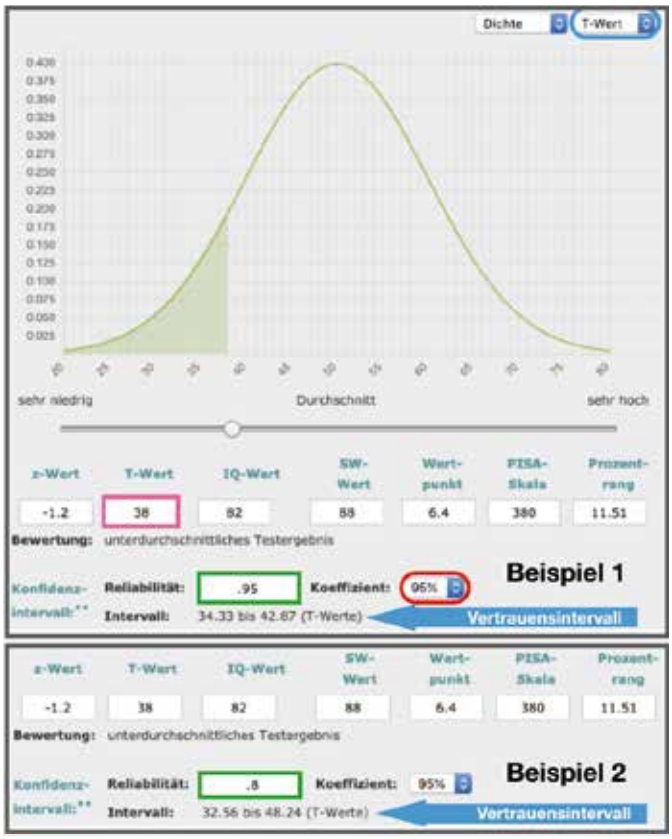
Wenn die Zuverlässigkeit eines Instruments unter 1 liegt, weist ein Testinstrument einen gewissen Grad an Ungenauigkeit auf. Aus diesem Grund ist es besser, neben der Interpretation der Standardnorm (z.B. T-Wert) auch das sog. Vertrauensintervall (VI) zu berücksichtigen. Das VI gibt mit einer gewissen Sicherheit an, in welchem Bereich die tatsächliche Punktzahl (der wahre Wert) unter Berücksichtigung des Messfehlers liegt. Das VI gibt somit auch

**Abb. 2: Übersicht von verschiedenen Standardwerten und ihrer Position auf der Normalverteilung**




| Standardabweichung | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| Prozentränge | 0,13 | 2,27 | 15,86 | 50 | 84,13 | 97,72 | 99,86 |
| z-Werte | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| T-Werte | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
| IQ | 55 | 70 | 85 | 100 | 115 | 130 | 145 |

**Abb. 3: Einfluss des Zuverlässigkeitswertes (Reliabilität) auf das VI. Beispiel 1 (oben) hat eine Reliabilität von 0,95 und Beispiel 2 (unten) von 0,8. Das VI ist unten deutlich größer (beide Abbildungen basieren auf Screenshots von www.psychometrica.de/normwertrechner.html)**




einen ersten Hinweis darauf, wie groß die Änderung in der Nachmessung sein muss, um von einer signifikanten Verbesserung sprechen zu können.

Die Berechnungen in den folgenden Beispielen wurden mit dem kostenlosen Programm „Normwertrechner" durchgeführt (www.psychometrica.de/normwertrechner.html). Das Programm kann online in einem Browser verwendet oder auf einem Rechner installiert werden (Windows/Mac). Um ein Vertrauensintervall zu berechnen, müssen die folgenden Schritte durchgeführt werden (Abb. 3):

1. Zunächst muss der Standard-Normwert bestimmt werden, für den das VI berechnet werden soll. Diesen erhält man, wenn man mit dem Testhandbuch den Rohwert in einen Normwert umwandelt. Meist sind dies Prozentränge oder T-Werte. Im Beispiel werden T-Werte verwendet (blaue Markierung). Diese Auswahl bestimmt, in welcher Einheit der VI angezeigt wird (blauer Pfeil).
2. Der Standardwert muss dann in das Programm eingetragen werden, in diesem Beispiel T-Wert = 38 (rosa Markierung). Es können auch in den anderen Feldern die entsprechenden Werte eingetragen werden. Sobald ein Wert eingetragen ist, wandelt das Programm die eingegebenen Standardwerte (z. B. T = 38) automatisch in die anderen Standardwerte um (hier Prozentrang 11,51 % und z-Wert -1,2).
3. Dann muss der Zuverlässigkeitswert (Reliabilität) aus dem Handbuch des verwendeten Tests eingetragen werden. Als Beispiel wird hier ein Wert von 0,95 verwendet (grüne Markierung). Hier sollte der Wert für Cronbachs-Alpha, Split-Half-Reliabilität oder Interrater-Reliabilität verwendet werden.
4. Als nächstes muss der Sicherheitsgrad (Konfidenzkoeffizient) angegeben werden, hier 95% (rote Markierung). Üblich sind in der klinischen Praxis 95% oder 90% Koeffizienten.
5. Das VI kann nun in dem Programm abgelesen werden (blauer Pfeil). Die tatsächliche (wahre) Punktzahl des Patienten liegt in dem Beispiel mit 95%iger Sicherheit zwischen 34,33 und 42,87. Es ist daher möglich, dass die tatsächliche Punktzahl des Patienten noch im Durchschnitt liegt (T > 40).

Einige Testverfahren geben bereits Vertrauensintervalle an. Dann braucht dieser Schritt nicht durchgeführt zu werden (siehe Testhandbuch). Das VI gibt einen ersten Hinweis darauf, wie groß die Veränderung in der Nachmessung sein muss, um klinisch relevant zu sein. Im obigen Beispiel würde man von einer sinnvollen Verbesserung ausgehen, wenn der VI der Vormessung (34,3-42,8) nicht mit dem VI der Nachmessung überlappt. Bei der oben angegebenen Reliabilität von 0,95 hätte sich der Patient erst ab einem T-Wert von 46 (VI = 43,4-50,0) signifikant verbessert.

In Abbildung 3 (unten, Beispiel 2) ist zudem deutlich zu sehen, wie sich eine abnehmende Zuverlässigkeit auswirkt. Die Zuverlässigkeit beträgt im zweiten Beispiel 0,8 (grüne Markierung). Eine Verringerung der Zuverlässigkeit erhöht den VI. Die tatsächliche Punktzahl des Patienten liegt mit diesem geringeren Zuverlässigkeitswert zwischen 32,5 und 48,2 (blauer Pfeil). In diesem Fall ist die Bewertung des Tests deutlich ungenauer und die Ergebnisse müssen mit größerer Vorsicht interpretiert werden.

## Kritische Differenz (KD)

Mit der kritischen Differenz (KD) kann für einen Test berechnet werden, wie viel Veränderung stattfinden muss, um von einem klinisch relevanten Unterschied sprechen zu können (z.B. *Schmidt-Atzert* & *Amelang* 2012). Manchmal sind die KDen im Testhandbuch bereits aufgeführt. Falls nicht, können sie schnell selber berechnet werden. Wenn die Zuverlässigkeitswerte des Tests für verschiedene Altersgruppen (bzw. je nach Geschlecht oder Bildung) unterschiedlich sind (siehe Handbuch), dann muss die KD für jede Altersgruppe bzw. Normgruppe berechnet werden. Der Vorteil dieser Methode ist, dass sie für jeden Test nur einmal durchgeführt werden muss. Wenn man einmal für ein Testverfahren die KDen berechnet hat, können diese in Zukunft für alle Patienten verwendet werden.

Zur Berechnung der KD bei Messwiederholung mit dem gleichen Testverfahren wird eine Formel verwendet (Abb. 4). Die KD kann manuell berechnet werden. Das hier vorgestellte Beispiel wird mit einer Excel-Tabelle ausgeführt, die unter folgendem Link heruntergeladen werden kann: http://download.tguenthert.de/Kritische_Differenz_D.xlsx.

**Abb. 4: Formel für die kritische Differenz (KD)**


Zur Berechnung der KD müssen drei Werte in die Formel eingegeben werden:

1. Der z-Wert ist das Maß der Sicherheit für die KD. Dies ist vergleichbar mit dem zuvor beschriebenen Sicherheitsgrad beim Vertrauensintervall. In der Regel wird ein Sicherheitsgrad von 95 % verwendet, wobei dann ein z-Wert von 1,96 eingegeben werden muss.
2. Ferner muss die Standardabweichung (SD) von dem Wert eingetragen werden, für den die KD berechnet werden soll. Die KD kann für den Rohwert berechnet werden, beispielsweise für die Anzahl der richtigen Antworten eines Tests. Dann muss die Standardabweichung des entsprechenden Wertes der dazugehörigen Normgruppe im Handbuch nachgeschlagen und in die Formel eingetragen werden. Wenn man mit Rohwerten rechnet, kann dies jedoch zu Problemen führen. Wenn der Patient beispielsweise aufgrund des Abstandes zwischen Vor- und Nachmessung in eine andere Normgrup-

pe rutscht, wird im Rohwert auch eine Verbesserung aufgrund des höheren Alters erwartet (sofern es sich um Kinder oder Jugendliche handelt). Dies wird bei der Berechnung der KD nicht berücksichtigt. Ein weiteres Problem ist, dass Mittelwerte und Standardabweichungen der verschiedenen Normgruppen in vielen Testhandbüchern nicht aufgeführt sind. Deshalb ist es besser, die KD mit Standardnormen zu berechnen. Dabei spielt der Wechsel in eine andere Normgruppe keine Rolle und die SDen für die Standardwerte sind immer gleich: Für z-Werte ist die SD = 1, für T-Werte ist die SD = 10 und für IQ-Werte ist die SD = 15 (siehe auch Abb. 2).

Die KD kann nicht für Prozentränge berechnet werden, da diese nur ordinal- und nicht intervallskaliert sind, somit keine Berechnung von Differenzen erlauben! Wenn der verwendete Test nur Prozentränge ausgibt, müssen diese zunächst in eine Standardnorm konvertiert werden (z.B. z-Wert oder T-Wert). Dies geht schnell mit dem im vorherigen Kapitel beschriebenen Normwertrechner. Im aktuellen Beispiel soll eine KD für die T-Werte eines Tests erstellt werden und daher wird in die Formel für die SD der Wert 10 eingetragen.

3. Der Zuverlässigkeitswert (Reliabilität) sollte im Handbuch des verwendeten Tests zu finden sein. Im Beispiel beträgt die Zuverlässigkeit 0,95. Da es um den Vergleich zweier Testmomente geht, ist die Retest-Reliabilität theoretisch am besten geeignet. Problematisch ist hingegen, dass sie häufig nicht untersucht wurde, dass die Stichproben für die Untersuchung meist sehr klein sind (< 100) oder dass das untersuchte Intervall deutlich anders ist als der Abstand zwischen den beiden Testungen beim eigenen Patienten. Daher wird hier überwiegend die interne Konsistenz (Cronbachs Alpha) verwendet.

**Abb. 5: Beispiel für eine Excel-Tabelle (http://download.tguenthert.de/Kritische_Differenz_D.xlsx), die eine kritische Differenz (KD, orangefarbene Kästchen) mit einer Standardabweichung von 10 (T-Wert) und einem Zuverlässigkeitswert von 0,95 berechnet (oben). Das Beispiel unten verdeutlicht die Steigerung der KD auf 12,4, wenn die Zuverlässigkeit des Tests geringer ist (hier 0,8).**

| Messwiederholung mit dem gleichen Testverfahren: | |
|---|---|
| z-Wert (1%; zweiseitige Testung): | 2,576 |
| z-Wert (5%; zweiseitige Testung): | 1,96 |
| Standardabweichung (S) der Norm- bzw. Rohwerte: | 10 |
| Reliabilität des Testverfahrens (Rel): | 0,95 |
| Kritische Differenz (1% Signifikanzniveau): | 8,15 |
| Kritische Differenz (5% Signifikanzniveau): | 6,20 |

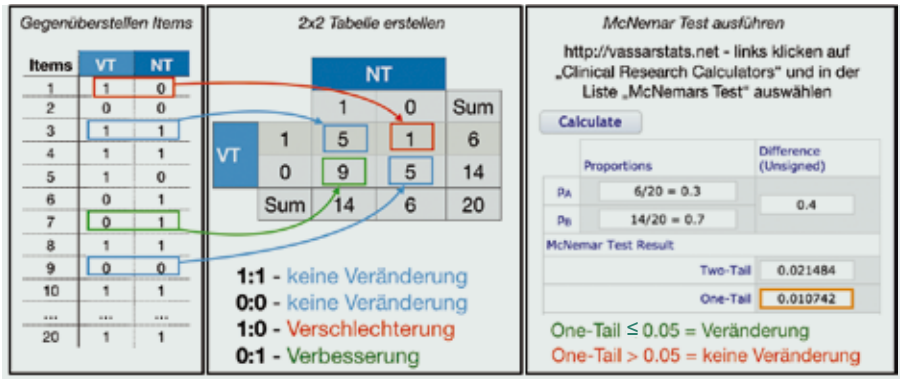
| Messwiederholung mit dem gleichen Testverfahren: | |
|---|---|
| z-Wert (1%; zweiseitige Testung): | 2,576 |
| z-Wert (5%; zweiseitige Testung): | 1,96 |
| Standardabweichung (S) der Norm- bzw. Rohwerte: | 10 |
| Reliabilität des Testverfahrens (Rel): | 0,8 |
| Kritische Differenz (1% Signifikanzniveau): | 16,29 |
| Kritische Differenz (5% Signifikanzniveau): | 12,40 |

Wenn die oben genannten Werte in die Formel eingegeben werden, erhält man eine KD von 6,2 ($KV = 1{,}96 * 10 * \sqrt{2 * (1 - 0{,}95)}$). Dies bedeutet, dass der T-Wert eines Patienten in der Nachmessung mindestens 6,2 T-Werte höher liegen muss, bevor eine klinisch relevante Verbesserung (oder Verschlechterung) angenommen werden kann. Mit der gleichen Formel und einer niedrigeren Zuverlässigkeit von 0,8 steigt die KD auf 12,4 (siehe auch Abb. 5 zum Vergleich). Jetzt muss der Patient in der Nachmessung mindestens eine Veränderung von 12,4 T-Werten aufweisen (d.h. doppelt so viel), bevor von einer relevanten Veränderung gesprochen werden kann.

## McNemar-Test

Der McNemar-Test kann verwendet werden, wenn für das verwendete Testverfahren die Zuverlässigkeit (Reliabilität) nicht überprüft wurde und der Test aus dichotomen Items besteht (weitere Einzelheiten zur Einzelfallstatistik siehe *Morley* 2017 oder *Bortz & Lienert* 2003). Bei einem dichotomen Item findet eine richtig/falsch- oder ja/nein-Bewertung statt. Ein typisches Beispiel ist ein Wortschatztest, bei dem

**Abb. 6: Durchführung des McNemar-Tests**




ein Kind 20 Gegenstände benennen muss. Bei einem korrekten Item erhält das Kind einen Punkt (1), bei einem falschen Item keinen Punkt (0). Anschließend wird ein Summenwert berechnet, z.B. 16 der 20 Items waren korrekt. Mit dem McNemar-Test kann nun überprüft werden, ob eine Verbesserung in einer Nachmessung eher zufällig oder tatsächlich signifikant ist. Eine Überprüfung der Veränderung mithilfe des McNemar-Tests kann mit den folgenden Schritten durchgeführt werden:

**Schritt 1: Gegenüberstellen der Items aus der Vor- und Nachmessung**

Abbildung 6 zeigt, wie dies bei einem Test von 20 Elementen aussehen kann. Die erste Spalte zeigt die Nummer (oder Namen) des Items (hier 1 bis 20). In der zweiten Spalte (VT) wird angegeben, ob der Patient das Item in der Vormessung korrekt (1) oder falsch (0) gemacht hat. Dasselbe wird für die Nachmessung gemacht (dritte Spalte NT).

**Schritt 2: Erstellen einer 2x2-Tabelle**

Bei der Gegenüberstellung von Vor- und Nachtestung sind pro Item vier verschiedene Kombinationen möglich:

- 1:1 – VT und NT richtig (keine Änderung)
- 0:0 – VT und NT falsch (keine Änderung)
- 1:0 – VT richtig und NT falsch (Verschlechterung)
- 0:1 – Fehler in VT und gut in NT (Verbesserung)

Aus diesen vier Optionen kann eine 2x2-Tabelle erstellt werden, um die Vor- und Nachmessung miteinander zu vergleichen (siehe Tabelle in der Mitte von Abb. 6). Diese Tabelle gibt bereits einen ersten Überblick darüber, ob sich der Patient verbessert (höherer 0:1-Wert) oder verschlechtert hat (höherer 1:0-Wert).

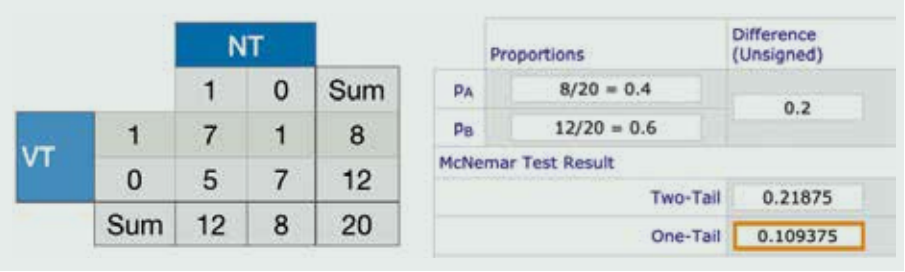
**Schritt 3: McNemar-Test ausführen**

Im McNemar-Test wird nun überprüft, ob die 0:1-Zelle (Verbesserung) ausreichend Items enthält, um von einer klinisch relevanten Verbesserung sprechen zu können. Ist die Anzahl in der 1:0-Zelle (Verschlechterung) deutlich größer als in der 0:1-Zelle, wird eine klinisch relevante Verschlechterung überprüft. Für die Berechnung werden somit lediglich die Zellen der 2x2-Tabelle verwendet, die eine Veränderung der Itemantworten anzeigen (Verbesserung, Verschlechterung).

Der McNemar-Test kann mit vielen Statistikprogrammen durchgeführt werden. Er kann jedoch auch wie folgt online erfolgen:

- Auf der Seite http://vassarstats.net auf „Clinical Research Calculators" klicken und dann in der Liste „McNemars Test" auswählen.
- Auf der dann erscheinenden Seite erscheint eine 2x2-Tabelle, in der die Zahlen aus der selbst erstellten 2x2-Tabelle eingetragen werden können.
- Anschließend auf die Schaltfläche „Calculate" klicken und es erscheint das Ergebnis (siehe Abb. 6 rechts).

**Abb. 7: Beispiel für einen McNemar-Test, in dem der Unterschied zwischen Vor- (VT) und Nachtestung (NT) nicht signifikant (p = 0,109375) und damit nicht klinisch relevant ist**




Nur die Zahl nach „One-Tail" ist für das Ergebnis wichtig. In diesem Beispiel beträgt die Wahrscheinlichkeit 1,07 % (siehe Abb. 6 rechts, 0,010742), dass kein Unterschied zwischen Vor- und Nachmessung besteht. Die Chance, dass es sich nur um einen zufallsbedingten Unterschied handelt, ist also sehr gering. Die Wahrscheinlichkeit ist kleiner als 0,05 oder 5 %. Daher wird angenommen, dass der festgestellte Unterschied ein signifikanter und somit klinisch relevanter Unterschied ist.

In einem anderen Beispiel (siehe 2x2-Tabelle in Abb. 7) waren 7 Items in VT und NT richtig (1:1) und 7 Items waren in beiden Messungen falsch (0:0). Es gab nur 1 Item, das in der VT richtig und in der NT falsch war (1:0; Verschlechterung), und 5 Elemente, die in der NT gut und in der VT falsch waren (0:1; Verbesserung). Wenn mit dieser Tabelle der McNemar-Test durchgeführt wird, ergibt sich eine Chance von 10,9 % (siehe Ergebnis bei One-Tail: 0,109375), dass die festgestellte Verbesserung zufällig ist. Dieser Wert ist deutlich größer als 5% (oder 0,05). Die Chance ist daher ziemlich groß, dass der festgestellte Unterschied auf Zufall beruht. Demnach kann in diesem Beispiel nicht nachgewiesen werden, dass sich der Patient in der Nachmessung verbessert hat.

## Berichten der Ergebnisse

Mit den beschriebenen Methoden kann überprüft werden, ob sich ein Patient verbessert hat (oder nicht). Die Verfahren können allerdings nur verwendet werden, wenn ein geeignetes Testverfahren zu Verfügung steht. In einem Bericht an Externe (z.B. Überweiser) ist es wichtig, diese Art von quantitativen Veränderungen von den qualitativen Beschreibungen zu unterscheiden. Die quantitativen Ergebnisse sind eine wichtige Ergänzung zu den qualitativen Bewertungen. Sie helfen dabei, die Wirksamkeit (oder Unwirksamkeit) einer Therapie zu belegen. Demnach sollte diese auch in der Berichterstattung aufgeführt werden. Tabelle 1 enthält Beispiele, wie die Ergebnisse in einen Bericht formuliert werden können. Es ist wichtig, dass die verwendete Methode für den Empfänger transparent ist und die Schlussfolgerung mit Ergebnissen bzw. Daten begründet ist.

**Tab. 1: Zusammenfassung der vorgestellten Verfahren, Links zu den verwendeten Internetseiten und Formulierungsvorschläge für die Berichterstattung**

| METHODE | BEISPIELFORMULIERUNG |
|---|---|
| **Vertrauensintervall (VI) bzw. Konfidenzintervall** gibt mit einer gewissen Sicherheit an, in welchem Intervall der tatsächliche Wert des Patienten liegt.<br>↳ **www.psychometrica.de/normwertrechner.html** | • „Mit einem Rohwert von 21 hat Herr Günther ein noch durchschnittliches Ergebnis erzielt (T-Wert 42; Vertrauensintervall 38-44)."<br>• „Das Vetrauensintervall in der Nachmessung (T-Wert 51, Vertrauensintervall 48-54) überlappt nicht mit dem in der Vortestung (T-Wert 42; Vertrauensintervall 38-44). Damit hat sich Herr Günther klinisch relevant verbessert." |
| **Kritische Differenz (KD)** gibt an, um wie viel Standardwerte (oder Rohwert-Punkte) sich ein Patient verändern muss, um von einem klinisch relevanten Unterschied sprechen zu können.<br>↳ **http://download.tguenthert.de/Kritische_Differenz_D.xlsx** | • „Mit einer Veränderung von 9 T-Wert-Punkten hat sich Herr Günther in der Nachmessung signifikant verbessert (Kritische Differenz = 6)."<br>• „In der Nachmessung war der Wert von Herrn Günther 9-T-Wert Punkte höher. Die Veränderung liegt allerdings unter der kritischen Differenz (KD=12). Eine klinisch relevante Verbesserung konnte in der Nachmessung daher nicht nachgewiesen werden." |
| **McNemar-Test** kann verwendet werden, wenn kein Zuverlässigkeitswert (Reliabilität) verfügbar ist und der Test aus dichotomen Items besteht.<br>↳ **http://vassarstats.net** | • „Im Test X mit 20 Items hat sich Herr Fimm bei 9 Items verbessert, bei einem Item verschlechtert und bei 10 Items waren die Ergebnisse in Vor- und Nachmessung gleich. Damit hat sich Herr Fimm in der Nachmessung signifikant verbessert (McNemar; p= 0,01; einseitig)."<br>• „Im Test X mit 20 Items hat sich Herr Fimm bei 5 Items verbessert, bei einem Item verschlechtert und bei 14 Items waren die Ergebnisse in Vor- und Nachmessung gleich. Eine klinisch relevante Verbesserung konnte damit nicht nachgewiesen werden (McNemar; p= 0,11; einseitig)." |

## : LITERATUR


Bortz, J. & Lienert, G.A. (2003). *Kurzgefasste Statistik für die klinische Forschung – Leitfaden für die verteilungsfreie Analyse kleiner Stichproben* (Vol. 2). Berlin: Springer

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion.* München: Pearson

Kersting, M. (2006). Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn. *Psychologische Rundschau* 57 (4), 243–253

Lane, S., Raymond, M.R. & Haladyna, T.M. (2015). *Handbook of Test Development.* New York: Routledge

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion.* Berlin: Springer

Morley, S. (2017). *Single Case Methods in Clinical Psychology.* New York: Routledge

Price, L.R. (2017). *Psychometric Methods.* New York: The Guilford Press

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik.* Berlin: Springer



**Prof. Dr. Thomas Günther** ist Logopäde und Psychologe. Er ist Professor im Lehr- und Forschungsgebiet für klinische Neuropsychologie des Kindes- und Jugendalters am Universitätsklinikum der RWTH Aachen. Zudem arbeitet er an der Faculty of Health der Zuyd University in den Niederlanden.



**PD Dr. Bruno Fimm** ist Psychologe und arbeitet an der Klinik für Neurologie an der Uniklinik RWTH Aachen.



**DOI** 10.2443/skv-s-2020-53020200102



**KONTAKT**
**Prof. Dr. Thomas Günther**
Universitätsklinikum der RWTH Aachen
Klinik für Psychiatrie, Psychosomatik und Psychotherapie des Kindes- und Jugendalters – Lehr- und Forschungsgebiet für klinische Neuropsychologie
Neuenhofer Weg 22
52074 Aachen
tguenther@ukaachen.de