Eine kleine Evaluation der Beschwerdevalidierung

Wolfgang Palm

1. Zur Entstehung der Daten

Die nachfolgend analysierten Daten entstanden in den Jahren 2009 bis 2020 durch Testungen in Untersuchungen zur Erstellung von Gutachten. Zu mehr als 90% dienten diese Gutachten der Einschätzung des Grades der Berufsunfähigkeit der getesteten Proband*innen. In die Reihe von Tests und Fragebögen wurden zwei Tests aus der *Bremer Symptomvalidierung* (BSV) eingefügt, der auditorisch - visuelle Test A (BSV-A) und der visuelle Kurzzeitgedächtnistest A (BSV-G). Beide dienten der Beschwerdevalidierung (-> Anm. 1).

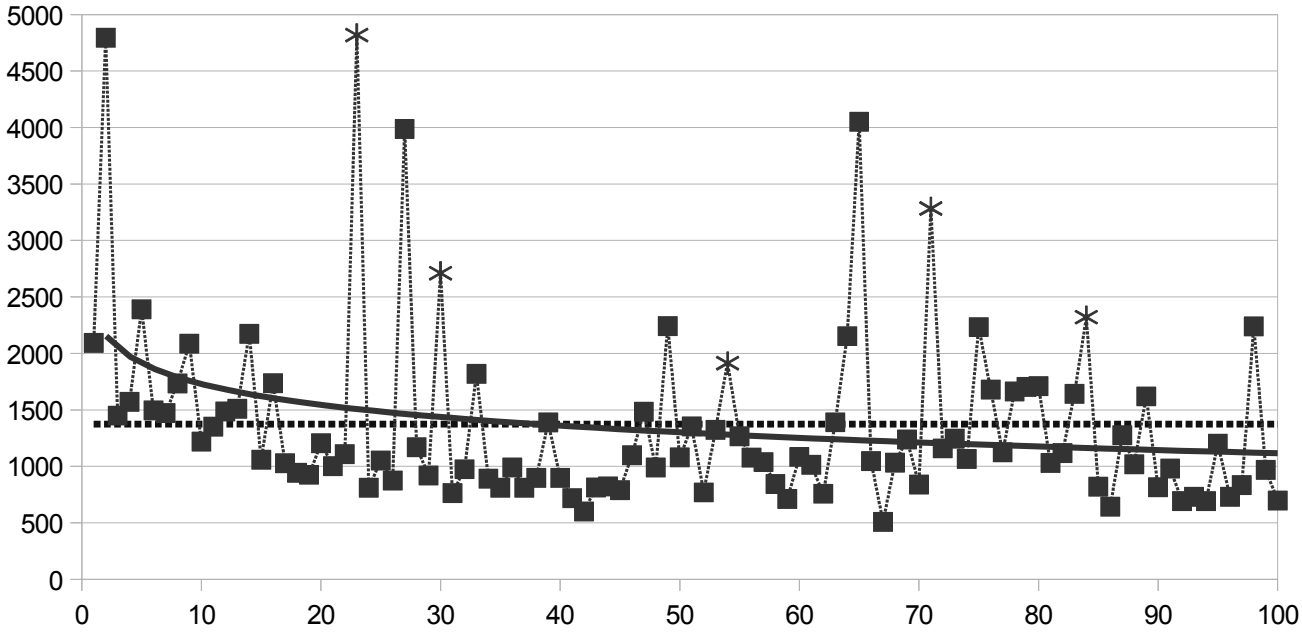
Deren Aufgaben sind reichlich simpel: Benutzt werden zwei Tasten einer PC-Tastatur, die eine für den linken, die andere für den rechten Finger. Die Aufgabe in der BSV-A lautet: die eine Taste ist zu drücken, sobald ein Punkt auf dem Bildschirm erscheint und ein dumpfer Ton zu hören ist. Die andere Taste ist zu drücken, sobald nur ein Punkt zu sehen ist. Im Grunde braucht man nur auf den Ton zu achten. Die Aufgabe in der BSV-G ist ein wenig anspruchsvoller: Zuerst erscheint - kürzer als eine Sekunde - ein Bild. Im knappen zeitlichen Abstand dazu erscheinen sodann zwei Bilder. Ist das zuerst gezeigte Bild ein Ausschnitt des rechten Bildes, dann ist die rechte Taste zu drücken; falls es ein Ausschnitt des linken Bildes ist, ist die linke Taste zu drücken. Die Aufgaben werden in beiden Tests je 100 Mal durchgeführt. Sie erfordern von einer Proband*in 'wahrzunehmen, zu entscheiden und zu reagieren' bzw. 'wahrzunehmen, zu vergleichen, zu entscheiden *und* zu reagieren'. Als Ergebnisse werden Reaktionszeiten und Fehlerzahlen aufgezeichnet. Für die BSV-A sind etwas kürzere Reaktionszeiten zu erwarten als für die BSV-G. Der motorische Anteil an der Messung beschränkt sich auf einen schnellen Tastendruck, ein Finger sollte dazu auf jeder Taste liegen.

Da die Tests BSV-A und BSV-G einigen Tests aus der TAP ähneln (-> Anm.1), wurden sie in die Folge von Aufmerksamkeitsprüfungen eingereiht, die BSV-A an den Anfang gesetzt und die BSV-G ans Ende gestellt. Dazwischen wurden vier bis fünf Tests durchgeführt, in den letzten Jahren die Tests *Alertness*, Go/NoGo, *Flexibilität*, *Geteilte Aufmerksamkeit,* von denen lediglich die *Flexibilität* zwei Tasten benötigt. Wie im weiteren gezeigt wird, hat diese zeitliche Reihung einen erheblichen Einfluss auf die Ergebnisse der Beschwerdevalidierung.

2. Der Datenpool

Der vorhandene Datenpool besteht aus 263 Datensätzen zu je 100 Daten, erzeugt von 151 Proband*innen, darunter 39% weibliche. Für alle wurden Gutachten erstellt. Der Altersdurchschnitt beträgt rund 45 Jahre, die Standardabweichung rund +/- 10 Jahre. Jeder einzelne Datensatz wurde händisch von der Textdatei in die Tabellenkalkulation übertragen und dort ausgewertet. Da der Test BSV-G von Beginn an verwendet wurde, der Test BSV-A aber erst später regelmäßig dazu kam, ist der Datenpool für den Test BSV-A um 33 Datensätze kürzer als der für die BSV-G. Datensätze, die nur für den einen Test vorhanden sind und solche mit mehr als 20 Fehlern wurden heraus genommen. Der Vorher-Nachher-Vergleich umfasst daher noch 2*112 Datensätze. Ein

Datensatz enthält 100 einzelne Daten, deren Datenplot beispielhaft wie folgt aussieht:



- Abb.1: Datensatz mit 100 Reaktionszeiten, der zeitliche Verlauf berinnt bei '0' und endet bei '100'. Das quadratische Symbol kennzeichnet die richtigen, das Stern-Symbol die falschen Reaktionen. Die fein punktierte Linie verdeutlicht lediglich das zeitliche Nacheinander. Die Maßeinheit für die senkrechte Achse ist Millisekunden [ms]. Fünf *Ausreißer* sind deutlich zu erkennen. Die punktierte Gerade liegt auf der Höhe des *Mittelwerts*. Besser an den Datensatz angepasst ist die durchgängige Linie für die *logarithmische Regression*.

Die durch die Symbole gekennzeichneten Messwerte wurden durch das Messinstrument 'programmierter Rechner' generiert. Die statistischen Kennwerte - Mittelwerte und Standard-abweichungen - die aus den Daten berechnet wurden, existieren nicht in dem Sinne wie die Messwerte; sie entstehen durch Rechenoperationen, sind also nicht das Ergebnis physischer Messvorgänge. Mediane entstehen bekanntlich durch hälftiges Abzählen des Datensatzes.

Wesentlich für menschlichen Reaktionen ist, dass sie wegen ihrer kognitiven Vorbereitung und der motorischen Ausführung ein Minimum an Zeit nicht unterschreiten können. Dieses Minimum schwankt etwa um die 200 Millisekunden. Deshalb befinden sich auch die schnellsten Zeiten eines Plots immer oberhalb der waagrechten Nulllinie. Langsamere Reaktion hätten indes beliebig Zeit zur Verfügung, würde diese nicht durch die Programme der Testserie begrenzt; die der BSV warten hingegen auf die Eingaben. Da der Mensch keine Maschine ist, seine Reaktionen oft emotional getriggert und von Vorstellungen begleitet sind, muss sich eine Proband*in bemühen, ja geradezu anstrengen, um möglichst konstant zu reagieren. Dennoch wird auch bei bester Anstrengung die zweite oder nachfolgende Reaktion nicht so ausfallen, wie die erste oder voraus gegangene; folglich wird der Plot der 100 Daten immer eine Folge schwankender Werte aufweisen. Dabei wird die zeitliche Folge des Auf- und Ab der 100 Daten, das also, was faktisch existiert, unterschiedliche Ausformungen annehmen.

Wegen der allzu menschlichen Messwertschwankungen ist die statistische Größe *Mittelwert* eine nur unzulängliche Beschreibung eines Datensatzes. Denn derselbe Mittelwert kann einmal

entstehen, weil die kurzen Zeiten weit unten im Plot aufscheinen, die oberen jedoch etwa um die 2000 ms und höher plaziert sind. Zum anderen aber auch, weil die unteren Zeiten etwas höher liegen, dafür aber der Abstand zu den oberen enger wird. Man benötigt also mindestens noch eine zweite beschreibende Größe, gewöhnlich ist dies die *Standardabweichung*, die ebenfalls nur rechnerisch existiert. Ein höherer Mittelwert (oder Median) kann mit weiteren Schwankungen einher gehen, während ein niedrigerer Mittelwert auf kleineren Schwankungen hervor gehen kann, was nach unten allerdings nur begrenzt möglich ist.

Der oben abgebildete Datensatz zeigt eine weitere Eigenart, den *Ausreißer*. Dessen Definition haftet eine gewisse Willkür an, die zu diversen mathematischen Formeln führt (-> Anm. 2). Man kann auch fragen, ob es in den Datensätzen, die mittels der genannten Programme entstehen, überhaupt Ausreißer geben kann. Denn schließlich wird jeder Datensatz durch das fortlaufende Rearieren ein- und derselben Proband*in erzeugt, sodass schwerlich zu erkennen ist, welcher Wert des Datensatzes *zufällig* sein könnte. Indes enthält die oben beschriebene Testreihe einige Tests aus der TAP, in denen programmintern bereits *Ausreißer-Korrekturen* für Mediane, Mittelwerte und Standardabweichungen durchgeführt werden. Um Vergleiche der statistischen Kennwerte zu ermöglichen, werden diese Korrekturen bei der Auswertung der BSV-Daten ebenfalls durchgeführt. Sie reduzieren die Zahl der Elemente im Datensatz auf durchschnittlich 97. Da sich große Ausreißer stärker auf die Standardabweichungen auswirken als auf die Mittelwerte und/oder die Mediane, werden auch die korrigierten Quotienten S/M merklich kleiner als die nicht korrigierten(-> Anm. 2).

3. Die Ergebnisse (-> Anm. 3)

	Mg = M(Mp)	**S(Mp)**	**M(Sp)**	**S(Sp)**	**M(Fp)**	**S(Fp)**	**M(Sp/Mp)**	**M(Gp)**
BSV-A	788,72	405,20	250,52	215,27	6,28	4,76	0,32	0,71
BSV-G	1302,15	692,67	488,69	339,78	4,15	4,43	0,38	0,65

- Tabelle 1: Die Abkürzungen bedeuten: Mg ist der Mittelwert aller Reaktionszeiten, berechnet über den Mittelwert aller 112 Mittelwerte der Datensätze. S(Mp) ist die Standardabweichung dieser Mittelwerte Mp von Mg. M(Sp) ist der Mittelwert aller Standardabweichungen der 112 Datensätze, S(Sp) die Standardabweichung von diesem Mittelwert. M(Fp) ist der Mittelwert aller Fehlerzahlen der Datensätze und S(Fp) deren Standardabweichung. Sp/Mp bezeichnet die Relation von Standardabweichung zu Mittelwert für die Reaktionszeiten pro Datensatz und M(Sp/Mp) ist deren Mittelwert. Schließlich bezeichnet M(Gp) den Mittelwert aller Genauigkeitskoffizienten pro Datensatz .

Der Vergleich der Spalten Mg und S(Mp) führt zum erwarteten Ergebnis, dass mittlere Zeiten und Standardabweichungen im etwas anspruchsvolleren Test BSV-G größer ausfallen als im Test BSV-A. Interessant daran ist lediglich, dass in beiden Tests die S(Mp) 50% bis 53% der Größe der Mittelwerte Mg haben, die S(Sp) sogar 70% bis 86%, was auf beachtliche individuelle Unterschiede hinweist. Das Verhältnis der relativen Schwankungsbreite Sp/Mp pro Datensatz beschreibt die Veränderungen zwischen erstem und zweitem Test: In rund 74% der Datensätze werden

Sp/Mp in der BSV-G größer. Zwar scheinen deren Mittelwerte M(Sp/Mp) mit 0,31 in der BSV-A und 0,36 in der BSV-G eng beeinander zu liegen, doch der t-Test wird signifikant. Gleichzeitig nehmen aber in 65,18% der verglichenen Datensätze die Fehlerzahlen ab, auch hierfür fällt der t-Test signifikant aus. Damit zusammenhängend werden in 60,71% der paarweise verglichenen Datensätze die Genauigkeitswerte in der BSV-G kleiner, diesmals jedoch ohne Signifikanz. Die Abnahme der Fehler in der BSV-G geht also mit einer Verbreiterung der relativen Schwankungen einher.

Im einzelnen ergeben sich jedoch abweichende Relationen. Nur bei 4 Proband*innen (3,6%) nehmen die Fehler Fp und die Quotienten Sp/Mp zu. Bei 24 Proband*innen (21,4%) nehmen sowohl die Fehler Fp ab, als auch die Quotienten Sp/Mp zu. Insgesamt nehmen bei 28 Proband*innen (25%) die Fehler Fp zu; 84 Proband*innen (75%) weisen größere Quotienten Sp/Mp auf. Schließlich sind unter den 73 Proband*innen (65,2%) mit geringeren Fehlern Fp nur 13 mit besseren größeren Quotienten Sp/Mp. Daraus folgt: Individuelle Vorhersagen der Reaktionen von Proband*innen von BSV-A auf BSV-G sind praktisch unmöglich.

71% der Proband*innen waren zum Zeitpunkt ihrer Untersuchungen zwischen 35 und 55 Jahren alt. Die Korrelationen zwischen Alter und Mp, Sp, Fehlern Fp und Genauigkeiten Gp erreichen höchsten r=0,257. Der Einfluss des Alters auf die Ergebnisse ist also schwach und kann vernachlässigt werden. Diese Feststellung mag für Untersuchungen, die wegen Rentenbegehren durchgeführt wurden, überraschen.

Ein wichtiger Zusammenhang fällt indes hoch aus: die Korrelation zwischen Mp und Sp, zwischen Mittelwerten und Standardabweichungen der Datensätze. Für BSV-A ergibt sich r(Mp,Sp)=0,845, für BSV-G ist r(Mp,Sp)=0,913. Höhere Zeiten gehen mit höheren Reaktionszeitschwankungen einher – und umgekehrt. Dieser Zusammenhang ist nicht trivial und bedarf einer eigenen Erklärung.

4. Diskusssion

4.1 Die Programme für BSV-A und BSV-G geben ab dem Wert von 4 Fehlern bei der Ergebnisausgabe einen Warnhinweis auf unzureichende Mitarbeit (Anstrengungsvermeidung) aus. Doch in der am Anfang der Testfolge platzierten, einfacheren BSV-A entstehen im Mittel deutlich mehr als 4 Fehler. In der am Ende der Testfolge platzierten, etwas anspruchsvolleren BSV-G im Mittel aber kaum mehr als 4 Fehler. Für die BSV-A gibt es insgesamt 36 Proband*innen (32,14%), die 0 bis 3 Fehler machten, für die BSV-G sind es hingegen 67 (59,82%), also mehr als die Hälfte. Weitere Verteilungen enthält die folgenden Tabelle:

BSV-A	Fp < 4	Fp> 4	Fp < 4	Fp > 4
Anzahl	28 (25%)	39 (34,8%)	10 (8,95%)	35 (31,25%)
BSV-G	Fp < 4	Fp < 4	Fp > 4	Fp > 4

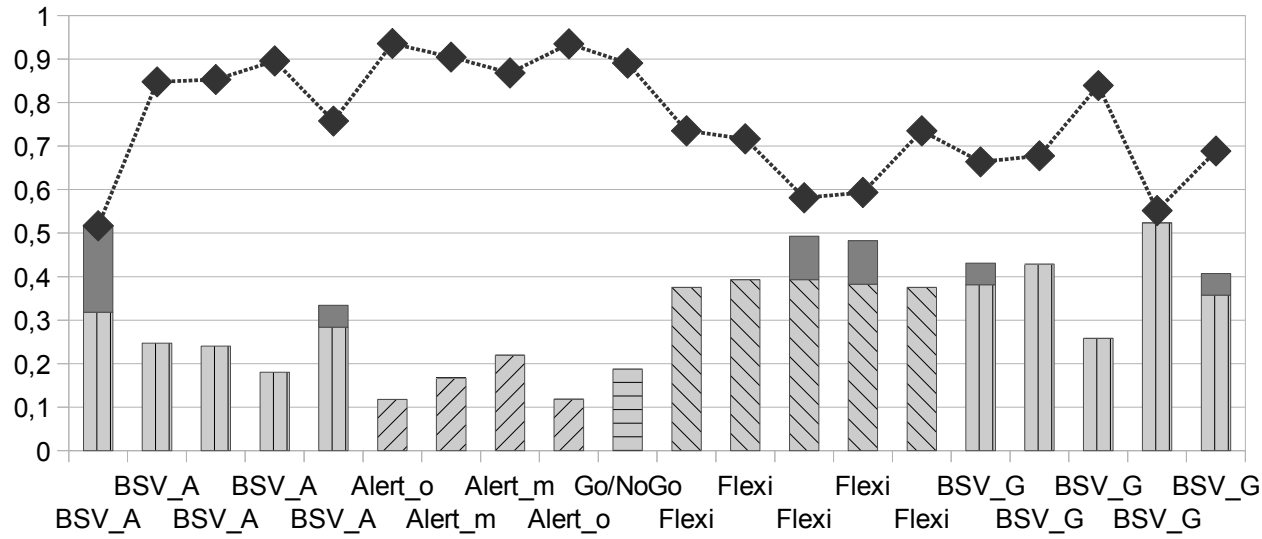
- Tabelle 2: Verhältnis der Fehlerzahlen in den beiden Tests der Beschwerdevalidierung BSV-A und BSV-G

Demnach gibt es 35 Proband*innen (31,25%) mit zwei Warnhinweisen. Knapp ein Drittel der Proband*innen aus der Stichprobe (N=112) gerät dadurch in den Verdacht mangelhafter Mitarbeit. Allerdings werden zwischen den beiden Tests vier bis fünf weitere Tests durchgeführt, die Reaktionszeiten messen und Fehler sowie Auslassungen zählen, teils mit einfacheren, teils mit schwierigeren Aufgaben. Was also führt zu der höheren Fehlerzahl in der BSV-A? Der Gebrauch einer PC-Tastatur ist mittlerweile so selbstverständlich wie das Autofahren. Plausibel erscheint die folgende Interpretation:

Anfangs sind die Reaktionen der Proband*innen von der Aufregung vor dem, was auf sie zukommt, beeinflusst. Jedoch wird danach die 100-fache Wiederholung derselben einfachen Reaktion langweilig. Einfache Aufmerksamkeitsprüfungen, zu denen man die beiden Tests auch zählen kann, sind prinzipiell langweilig. Sie erfordern eine rasche Reaktion auf eine simple Vorgabe und die Aufrechterhaltung einer Erwartungsspannung in den BSV-Tests zwischen 3 bis 6 Minuten. Kaum jemand ist für eine solche Testung positiv motiviert, die Proband*innen 'müssen' zu einer Untersuchung für ein Gutachten, weil sie sonst keine Rente bekommen. Diese Zwangsbedingung ähnelt der alltäglichen des 'Geldverdienen-Müssens'. Gegen Ende der Testfolge haben sich die meisten jedoch auf die Anforderungen eingestellt, bereits in den Tests der TAP weniger Fehler gemacht, und sie führen auch die BSV-G konzentrierter durch. Denn die meisten Proband*innen haben bereits im Alltag gelernt sich mit ihrer Tätigkeit zu arrangieren, auch wenn sie diese nicht gerne machen. Nach dieser Überlegung ist die im Vergleich rum 1,7-fache Fehlerzahl in der BSV-A gegenüber der BSV-G durch die Platzierung des Tests in der Testreihe bedingt. Erhöht man beispielsweise den Schwellenwert in der BSV-A auf 6 Fehler, dann sinkt der Wert in der letzten Spalte auf 29 Proband*innen (25,9%) ab. Rund ein Viertel alle Proband*innen produzieren zwei Warnsignale bezüglich ihrer Mitarbeit. Die Ergebnisse dieser 'kleinen Evaluation' weisen insofern darauf hin, dass Zeitpunkt eines Beschwerdevalidierungstests im Zusammenhang mit anderen Tests gut überlegt werden sollte.

4.2 Betrachtet man den Mittelwert Mp als adäquaten Ausdruck eines individuellen Reaktionstempos, so entsteht die Frage, welche Bedeutung der Standardabweichung Sp überhaupt noch zukommt. Der Mittelwert Mp ist jedoch nur eine rechnerische Größe, die in den Daten nicht oder vielleicht zufällig vorkommt; dasselbe gilt allerdings gilt für die Standardabweichung Sp. Dennoch erfasst die Sp noch etwas von der Wirklichkeit der Messdaten, nämlich dass sie mehr oder weniger hin und her schwanken (-> Abb.1). Wodurch immer diese auch Schwankungen erzeugt werden, eine Proband*in braucht eine möglichst standhafte willentliche Anstrengung, um die Schwankungsbreite und damit die Standardabweichung Sp eng zu halten. Also drückt sich ihre Anstrengung und damit ihre Mitarbeit nicht nur in den Fehlern Fp, sondern auch in der Standardabweichung Sp aus. Dieser Grundgedanke führte zur Entwicklung der Formel für die Genauigkeiten (-> Anm. 3). Ein Beispiel für deren Verlauf zeigt Abb.2:

- Abb.2: Testreihe einer Proband*in: BSV-A, Alertness, Go/NoGo, Flexibilität und BSV-G. Jeder der 20 Balken steht für 20 Reaktionen im jeweiligen Test. Gestreifte Balkenstücke kennzeichnen die relativen Schwankungen Sp/Mp, graue Balkenstücke stehen für die relativen Fehlerzahlen (Fp/20). Die Symbole markieren die Höhe der jeweilgen Werte für die Genauigkeiten, die nach oben nie den Wert 1 erreichen. Werte unter 0,6 sind bereits stark negativ auffällig. Balkenhöhe hj = Sp(j)/Mp(j) + Fp(j)/20., j = 1,2 ...20.



Der Datenverlauf in Abb. 2 zeigt beinahe Spiegelsymmetrie: je höher die Balken, desto niedriger die Genauigkeiten, und umgekehrt. In der BSV-A entstehen insgesamt 5 Fehler, in der BSV-G lediglich 2. Ist die Fehlerzahl einziges Kriterium, so wird die Proband*in nur in der BSV-A auffällig, nicht aber in der BSV-G. Doch der Verlauf der Genauigkeiten erteilt eine abweichende Auskunft: In der BSV-A ereignen sich 4 Fehler innerhalb der ersten 20 Reaktionen, was als Einübungseffekt interpretiert werden darf. In der BSV-G sind die Genauigkeiten niedriger, was auf Schwierigkeiten in der Mitarbeit verweist. Die *Flexibilität* ist für viele Proband*innen der schwiergste Test in der Testabfolge, hier könnten in der Mitte Ermüdungserscheinungen aufgetreten sein.

4.3 Da Mittelwerte und Standardabweichungen meist routinemäßig berechnet werden, gelten letztere oft als Ungenauigkeiten in der Bestimmung von Reaktionszeiten. Doch das trifft hier nicht zu, weil die durch Messhandlungen gewonnen Messdaten große Schwankungen aufweisen, bei denen die Einflüsse der Messfehler der Messapparatur vernachlässigbar klein sind. Die mittlere Relationen von Standardabweichungen zu Mittelwerten betragen 31% bzw. 36% und das nur, weil die *Ausreißer* aus den Datensätzen liminiert worden sind. In einigen Datensätzen würde sonst das Verhältnis Sp/Mp mehr als 60% ausmachen. Im einzelnen erreichten die Reaktionszeiten in den BSV Tests eine Größenordnung bis zu 7 Sekunden, ja manchmal sogar noch mehr. Doch solche Extremwerte wurden als Ausreißer eliminiert (-> Anm. 4). Nach unten besteht die bereits genannte Schranke für einzelne Reaktionen bei etwas unterhalb 200 Millisekunden. In den vorliegenden Daten ist die kürzeste Zeit für Mp 422 Millisekunden, für Sp beträgt sie rund 70 Millisekunden.

Die den Menschen möglichen Reaktionsweisen bestimmen also die Verteilung der Messdaten. Die hohen Korrelationen zwischen den Sp und den Mp in beiden Tests sind so zu interpretieren, dass größere Mittelwerte aus größeren Schwankungen hervorgehen. Das Seiende bestimmt die Berechnungen und nicht umgekehrt (-> Anm. 5).

Das Kernproblem der Beschwerdevalidierung ist jedoch, dass auch an den durch die BSV-Tests gemessenen Reaktionszeiten und Fehlerzahlen *im Einzelfall* nicht abzulesen ist, in welchem

Ausmaß daran Wollen, Motivation und kognitive Funktionen beteiligt sind. Entsprechende Unterscheidung gelingen nur begrifflich, dazu benötigt werden mathematische Operationen, Modelle der kognitiven Neurowissenschaften und vorallem eine genaue Beobachtung der Arbeitsweise von Testproband*innen, die wiederum Rückwirkungen auf deren Verhalten hat. Insgesamt ist die Fehlerzahl, wie gezeigt worden ist, ein bedingt aussagekräftiges Kriterium (-> Anm. 6).

5. Abschließende Überlegungen

Es sollte sich von selbst verstehen: auffällige Ergebnisse in Tests zur Beschwerdevalidierung sind Warnsignale, die mehr oder weniger kräftig ausfallen können; aber es sind nicht mehr als Warnsignale bezüglich der Mitarbeit einer Proband*in. Keinesfalls beweist ein solches Signal, dass alle Testergebnisse unglaubwürdig sind, nur weil die Proband*in einem Beschwerdevalidierungstest auffällig geworden ist. Häufige Praxis in Begutachtungen ist es indes, dass eine Proband*in eine Gutachter*in zu überzeugen habe, mit der Folge, dass bereits bei einem Zweifel seitens der Gutachter*in alle anderen Ergebnisse (beispielsweise die einer ausführlichen 'neuropsychologischen' Testung) als nicht interpretierbar erachtet werden.

Die vorliegende kleine Evaluation beschäftigt sich mit den Ergebnissen einer jahrelangen Verwendung von Beschwerdevalidierungstest (Fragebögen sind keine Tests), in denen Reaktionszeiten gemessen werden und in denen eine bestimmte Fehlerzahl als Kriterium gilt, jenseits dessen eine Warnung ausgegeben wird. Wie gezeigt, ist die Fehlerzahl ein ungenügendes Maß, zu berücksichtigen sind auch die Reaktionszeitschwankungen, die über die Einführungen von Genauigkeiten geschehen (Palm 2020b). Zudem sind Reihenfolgeeffekte bei der Plazierung der Tests zu beachten, die vermutlich auch bei der Verwendung von Gedächtnistests zur Beschwerdevalidierung auftreten. Wie oben gezeigt, treten im ersten und leichteren Test durchschnittlich mehr Fehler auf als im zweiten und schwereren, umgekehrt fallen aber die relativen Schwankungen im zweiten Test erheblich größer aus als im ersten. In der Folge wird der erste Test bei schlechteren Genauigkeiten öfters wegen der Fehlerzahlen auffällig als der zweite. (Beide Tests bilden Anfang und Ende in einer Serie von Aufmerksamkeitstestungen).

Zu den methodischen Aspekten führt die Frage: Was ist ein *Fehler*? In den BSV-Tests ist der Fehler eindeutig realisiert durch die Verwendung von zwei Reaktionstasten. Ein Fehler wird registriert, sobald mit der falschen Taste reagiert worden ist. (Unter den von mir oft verwendeten Tests aus der TAP ist dies auch beim Test *Flexibilität* der Fall.) Konsequenterweise darf es dann auch keine Zeitbeschränkung für die zu erwartende Reaktion geben, was ebenfalls in den BSV-Tests realisiert ist. Das hat zur Folge, dass der Untersucher bei der Ausführung dabei bleiben sollte, weil es sonst vorkommen kann, dass eine Proband*in mit einer Reaktion wartet und wartet und wartet .... Eine andere Art einen Fehler eindeutig zu machen, ist in den TAP Tests *Alertness mit Vorsignal* und Go/NoGo realisiert: die Reaktion auf das falsche Zeichen. Darüber hinaus wird problematisch, was ein Fehler sein soll. Was ist ein *Ausreißer* und was eine *Auslassung*? Beider Definition erfordert eine Zeitbeschränkung für die Reaktionen, doch die Zeit, so wie wir sie alltäglich durch die zuhandenen Uhren messen, kennt keine Fehler! Ausreißer und Auslassung können nur deshalb vom Programm gezählt werden, weil sie durch einprogrammierte Zeitbegrenzungen entstehen,

deren Wahl bestimmten Zwecken gehorcht (-> Anm 2 und 4). Ab welcher Zeitdauer ist die einzelne Reaktionszeit einer Testperson beispielsweise ein Ausreißer? Da alle Reaktionen einer Testperson von dieser einen Person in zeitlicher Reihenfolge nach Vorgabe des Programms ausgeführt werden, kann doch nicht davon ausgegangen werden, dass bestimmte Reaktionen zufällig sind. Folglich ist die Wahl der statistischen Methode zur Beseitigung von Ausreißern lediglich von den Zwecken bestimmt, die das Programm erfüllen soll, jedoch nicht von den personspezifischen Reaktionsweisen. Dieser Einwand gilt also auch für die in der vorgelegten Datenanalyse verwendete Ausreißer-Korrektur, um die kennzeichnenden Parameter mit denen der TAP-Tests bei in der Berechnung der Genauigkeiten vergleichbar zu machen (Palm 2020b).

Die Standardauswertung von Reaktionszeitmessungen in der Aufmerksamkeitstestung (z.B. In den TAP-Tests) führt zu Medianen und Mittelwerten sowie zu Standardabweichungen (SD). Dadurch werden die einzelnen Datensätze zwar zwecks Normierung vergleichbar gemacht, vielfach ist dieses Vorgehen aber für den einzelnen Datensatz eine eher mäßig gute Auswertung. Angepasstere Auswertungen scheinen jedoch nicht zu einer einfachen Vergleichbarkeit zu führen. Doch die Auswertung durch Mittelwert (Median) und SD führt zu einer Interpretationstendenz, in der Mittelwert (Median) das Maß für das Reaktionstempo ist, die SD in ihrer Interpretation offen und unklar bleibt. Dabei erreichen in den hier vorgelegten Daten die SD oft mehr als 60% des Mittelwerts! Da die Grundlage der Berechnungen, die Datensätze der Rohwerte, derart große Schwankungen aufweisen, dürfen die SD nicht als Schwankungen der Reaktionszeit interpretiert werden. Dabei dabei kommt es tendenziell zu einer Verkehrung: Nicht die Rohdaten geben Auskunft über das Seiende, sondern die statistische Prozedur bestimmt wie das Seiende auszusehen hat (-> Anm 5).

6. Anmerkungen

(1) Zur Bedeutung der Beschwerdevalidierung in Gutachten zwecks Rentenbegehren siehe Palm (2020a): Im vorliegenden Papier wird nur die Beschwerdevalidierung mittels Messung von Reaktionszeiten und Fehlern behandelt, Gedächtnistests und Fragebögen bleiben außen vor. Die Benennung des BSV-G ist irreführend, beide Tests verlangen in erster Linie die Aufmerksamkeitsausrichtung auf die Items. Wer die minimale Merkleistung im BSV-G nicht mehr erbringen kann, kommt schwerlich noch allein im Alltag zurecht und kann folglich auch nicht allein zur Untersuchung erscheinen. Die BSV ist Teil der *Testbatterie zur Forensischen Neuropsychologie*, Harcourt Test Services, 2. Auflage 2007. Das Programm erstellt für jede Proband*in und jeden Testdurchgang eine eigene, einfache Textdatei. Aus diesen Textdateien wurden die Daten für dieses Papier entnommen, und zwar die Reaktionszeiten und die Fehler für jeden einzelnen Tastendruck, der von einer Proband*in ausgeführt wurde. Erst dieses detaillierte Rohmaterial ermöglichte die hier vorgelegte 'kleine Evaluation'.
Andere Programme, wie z.B. die bekannte und viel verwendete Testbatterie zur Aufmerksamkeitsprüfung (TAP: Fimm & Zimmermann, 2019), gestatten dem Testleiter keinen einfachen Zugriff auf das Rohmaterial; sie geben nur die berechneten Kennwerte aus: die Mediane, die Mittelwerte, Standardabweichungen, Fehlerzahlen und Auslassungen. Die TAP wird mittels zweier großer Tasten bedient, nicht aber mit einer marktüblichen PC-Tastatur.
Bei fast allen Proband*innen bestand eine Depressions-Diagnose, in zwei oder drei Fällen war eine Psychose codiert und vier Personen erlitten Unfälle mit Hirnschädigungen.

(2) Hierzu z.B. Lohninger (2012). Spline-förmige, logarithmische, exponentielle oder auch lineare Regressionslinien würden die Verläufe der Reaktionszeiten passender beschreiben, allerdings müssten die Parameter der

Formeln für jeden einzelnen Datensatz spezifisch angepasst werden. Eine Berechnung der Varianz würde dadurch komplizierter werden. Der Rechenaufwand würde nicht nur sprunghaft ansteigen, man müsste auch fragen, welche Parameterwahl noch zu einem sinnvollen Vergleich der Datensätze führen könnte. Die in der TAP verwendete Korrekturformel zur Entfernung von *Ausreißern* lautet: 2,35*Sp+Mp. Werte oberhalb dieses Grenzwerts gelten als *Ausreißer*; sie werden aus dem Datensatz entfernt, für den 'gereinigten' Datensatz werden sodann die korrigierten Mp und Sp berechnet. Die Zahl der Elemente pro Datensatz nimmt dadurch ab.

(3) Für die folgenden Ausführungen gilt: Das 'p' für *Proband*in* hinter M, S, F oder G bezieht sich auf jeweils einen *Datensatz* eine Probandin mit durchschnittlich 97 Elementen. Der Übersichtlichkeit wegen werden Summenzeichen und laufende Indizes i = 1, 2 …. n weggelassen. .
Die Formel für die *Genauigkeiten* lautet: Gp = 2 – {[(Iz + Fp)/(Iz -Fp)]*[Mp/(Mp – Sp)]}^0,5 . Gp kann maximal 1 werden, erreicht diesen Wert aber nie, weil Sp nie gleich null wird. d.h. weil ein Mensch keine Maschine ist. Werte unterhalb 0,6 sind stark auffällig und verweisen auf Schwierigkeiten in der Mitarbeit. (Palm 2020b). Für die durchgeführten t-Tests gelten folgende Parameter: α =0,1, N=112, paarweiser, zweiseitiger Vergleich, H0: Differenz = 0.

(4) In den TAP-*Alertness*-Tests fällt der Quotient Sp/Mp oft kleiner aus. Auch die relativen Fehlerzahlen in *Alertness* und Go/NoGo fallen meist geringer aus, die Zahl der verlangten Reaktionen ist ebenfalls geringer. Dadurch werden die Genauigkeiten besser. Die in vielen Tests einprogrammierten Begrenzungen (z.B. ein Abstand von einer Sekunde zwischen zwei Items) lassen keine längeren Reaktionszeiten zu. Nur die *Flexibilität* gestattet längere Reaktionszeiten. Zeiten von z.B. 7 Sekunden sind in der BSV nur möglich, weil das Programm auf die Eingabe 'wartet'! Die Zeitbegrenzung erscheint zwar auf den ersten Blick als zweckmäßig, doch sie unterstellt, dass man vor der Testung einer neuen Proband*in immer schon weiß, innerhalb welcher Zeitspanne die Ergebnisse zu liegen haben. Tun sie es nicht, werden sie zu *Auslassungen*. Wird die *Ausreißer-Korrektur* auf die Rohdaten der Stichprobe eines Test angewendet, so wird der ca. 68% der Daten umfassende Normalbereich zeitlich enger. Da in der bekannten Formel für Transformation in die Standard-Normalverteilung z = (M – X)/S das M langsamer schrumpft als das S, wird der z-Wert für eine, mit dem nun 'gereinigten' Test geprüfte neue Proband*in, leichter nach oben – in der Praxis aber öfters nach unten aus dem Normalbereich herausfallen. Immer wieder habe ich mich gewundert, dass auch fit wirkende Studenten, die ich getestet habe, in der TAP-*Alertness* noch nie über die Mitte des Normalbereichs hinaus gekommen sind. Welche Genies einer sog. 'Normalpopulation' befinden sich denn da oberhalb? Von Patienten einer Klinik werden kaum Ergebnisse im oberen Normalbereich und noch höher erwartet; solche waren auch in den vielen Klinikberichtenberichten, die ich in mehr als 13 Jahren Tätigkeit als Sachverständiger gelesen habe, kaum je zu finden. Da fällt es dann auch nicht auf, dass zeitliche Begrenzungen und Ausreißer-Korrekturen zu Normwerten führen, die langsamere Reaktionszeiten in den Normwerten möglicherweise noch schlechter aussehen lassen.

(5) Diese Sachlage wiederspricht einer Idee, wonach eine Proband*in ein bestimmbares psycho-motorisches Grundtempo habe, das auf den Mittelwert oder den Median abgebildet werde. Mittels Reaktionszeittests ist es jedenfalls nicht zu *messen*. Nicht nur wegen der bereits angeführten Argumente, wonach die Messwertschwankungen keine Fehler sind und die Elimination von *Ausreißern* schwer zu begründen ist. Welche Zeitspanne sollte ein solches Tempo umfassen, wenn bespielsweise der Range der Messwerte zwischen 400 ms und 5000 ms liegt? Der auf den ersten Blick leicht zu übersehende, 'tiefere Grund' liegt allerdings darin, dass die *Zeit* gemessen wird - also eine *physikalische* Basisgröße - und kein *physisches Tempo*, was immer das sei. Die physikalische Basisgröße *Zeit* kann noch nicht einmal als Annäherung betrachtet werden, weil das, woran angenähert werden soll, unklar ist. Zur Messung der *Zeit* braucht man allerdings auch kein *neuropsychologisches Konstrukt*, alltäglich tut's die Uhr, auch die im Rechner eingebaute – ein technisches Gerät.

(6) Ein Grundproblem der Testtheorie wurde bereits in Huber (1973) und Sixtl (1985) analysiert. Es wird bis heute so 'gelöst': Man ersetzt die notwendigerweise fehlenden individuellen Parameter mittels aufwendiger Berechnungen durch die Fehlervarianzen der Eichstichproben und 'bestimmt' damit die Messgenauigkeit im Einzelfall. Bei der Festsetzung von Cutt-Off-Schwellen anhand von Sensitivität und Spezifität verfährt man insoweit analog, als man die Stichproben-Parameter als Maßstab für den Einzelfall nimmt. Wegen dieser,

über die Rechenmethoden hinweg oft schwer zu erkennenden Begründungsproblematik, sind auch zwei Warnsignale kein Beweis für *Aggravation,* sollten jedoch als ein Hinweis zur genauen Betrachtung des Datenverlaufs genutzt werden. Insbesondere der Verlauf der Genauigkeiten Gp hilft bei einer differenzierten Einschätzung, leider aber nur in Folgen von programmierten Tests, die die Reaktionszeiten (Rz) messen. Mit den herkömmlichen Papier-Stift-Versionen (z.B. ZVT, FWIT, d2) sind diese einzelnen Rz nicht zu messen und die Fehlerzahlen Fp oft nur durch genaues Mitzählen des Untersuchers zu erhalten.

Im übrigen sollte klar unterschieden werden zwischen *Mitarbeit (Anstrengung)* in Tests und *nicht authentischen Beschwerden* in Fragebögen und mündlichen Angaben (Merten et al., 2019). 'Querschlüsse' von einer Art auf die andere sind unzulässige Kurzschlüsse! Es gibt keine Zwangsläufigkeit, wonach ein Zweites geschehen muss, nur weil sich ein Erstes ereignet hat. Letztlich ist *Aggravation* eine Beurteilung im Einzelfall, quasi ein 'Indizienbeweis', der auf einer sorgfältigen Abwägung von Warnhinweisen und Beobachtungen beruhen sollte, nicht aber auf den Einfühlungen und den Mitschwingungen einer Gutachter*in (Palm 2020a). Schließlich geht es für die Proband*in um etwas, worauf sie einen Anspruch hat: Entweder per Gesetz (vorm Sozialgericht) oder auf Grund langjährig entrichteten Beiträgen zu einer Berufsunfähigkeits-Versicherung.

Zwischen Tests und Fragebogen besteht eine fundamentaler Unterschied, obwohl sie oft nach der gleichen Testtheorie (in der Praxis bis heute die klassische) konstruiert worden sind. In vielen Klinikberichten steht etwas von Tests, die durchgeführt worden seien, doch aufgeführt werden nur klinische Fragebögen. Die von Huber (1973) formulierte Ersetzung von unbekannten Individualparametern durch Gruppenparameter und 'seine technische Lösung' (Willmes & Fimm 2020) stellt m.E. keine wesentliche Einschränkung der Testinterpretation dar, sobald man davon ausgeht, dass Tests Fähigkeiten erheben und Fähigkeiten auf Zuschreibungen beruhen, die ja bereits im Alltag durchgeführt werden. Die Fähigkeiten eines Individuums können gar nicht anders festgestellt werden als durch den Vergleich mit Anderen und der daraus erfolgenden Zuschreibung. Insofern geht es gar nicht um den 'wahren Testwert' eines einzelnen Menschen, sondern um seine Plazierung innerhalb einer Gruppe bzw. einer Stichprobe. Folglich besagt das Vertrauensintervall auch nichts über den wahren Testwert für die jeweilige Fähigkeit, sondern benennt nur über die Verlässlichkeit, mit der die Stelle der Plazierung angegeben werden kann. Dadurch findet eine Verschiebung statt, die Frage ist nun: Was besagt ein Testergebnis für das Alltagsleben einer Person? Die zentrale Frage nach der Validität ist eben die prognostische. Eine weitere Frage lautet: Inwiefern besagen Testergebnisse einer sog. 'neuropsychologischen Testung' etwas über Funktionen, die dem Gehirn beigelegt werden? Schließlich sitzt ja nicht das Gehirn vor Tastatur oder der Vorlage, sondern der ganze Mensch mit Kopf, Armen und Beinen. Auch der neuropsychologische Test ist eben nur ein Test!

Autor und Copyright: Dr. Wolfgang Palm, Dipl.-Psych, Dipl.-Phys., Psychotherapeut
Sachverständiger der Psychotherapeutenkammer Baden-Württemberg
Email: praxis @wopalm.com
Stand des Papiers und der Links: November 2020

6. Literaturverweise

Fimm B, Zimmermann P (2019). Testbatterie zur Aufmerksamkeitsprüfung (TAP). https://www.psytest.de

Heubrock D, Scholl H, Petermann F (2013). Die differentielle Validität neuropsychologischer Testverfahren zum Nachweis nicht-authentischer Störungen. *Z. f. Neuropsychologie*, vol.24, 229-238

Heubrock D, Petermann F (2013). Testbatterie zur Forensichen Neuropsychologie. Pearson Tests Deutschland. https://www.pearsonclinical.de/tbfn.html

Huber H P (1973). Psychometrische Einzelfalldiagnostik. Beltz-Verlag

Lohninger H (2012): Ausreißertests. http://www.statistics4u.info/fundstat_germ/cc_outlier_tests.html

Merten T, Dohrenbusch R (2016). Psychologische Methoden der Beschwerdenvalidierung. In W Schneider, R. Dohrenbusch et al (Hrsg), Begutachtung bei psychischen und psychosomatischen Erkrankungen, 2. überarb. u. erw. Aufl., Verlag Hogrefe

Merten T, Giger P, Merckelbach H, Stevens A (2019). Handbuch zum Self-Report Symptom Inventory – deutsche Version, Verlag Hogrefe

Miller M (2017). Begutachtung im Fachgebiet Psychiatrie/Psychotherapie. In J Francke, A Gagel, D Bieresborn (Hrsg), Der Sachverständigenbeweis im Sozialrecht, 2. Aufl., Verlag Nomos

Palm W (2020a). Anmerkungen zur Beschwerdenvalidierung. https://www.wopalm.com/wp-content/uploads/2020/01/Beschwerdenvalidierung.pdf

Palm W (2020b). Genauigkeit in der Reaktionszeittestung. https://www.wopalm.com/wp-content/uploads/2020/09/Genauigkeit.pdf

Sixtl F (1985). Notwendigkeit und Möglichkeit einer neuen Methodelehre der Psychologie. *Z.f. experimentelle und angewandte Psychologie,* vol. 32(2), 320-339

Willmes K, Fimm B (2020). Einzelfalldiagnostik. Hogrefe-Verlag