Genauigkeiten in Reaktionszeittestungen [1]

Wolfgang Palm

In diesem Papier wird gezeigt, wie aus den Ergebnissen von Reaktionzeittestungen, den Mittelwerten der Reaktionszeiten, ihren Schwankungen und den Reaktionsfehlern eine Funktion namens *Genauigkeit* konstruiert werden kann. Sie kann zur Prüfung der Glaubwürdigkeit von solchen diesen Testergebnisse verwendet werden – also zur sogenannten Beschwerdevalidierung (Palm 2020a,b).

1. Normierte und nicht normierte Größen

Reaktionszeitestungen dienen beispielsweise zur Prüfung von Aufmerksamkeitsverläufen: Einfache Zeichen erscheinen auf einem Bildschirm, schnellstmoglich ist die eine oder die andere Taste zu drücken. Aufmerksamkeitstestung geschieht am besten programmiert. Die durch das Programm zu messenden oder zu zählenden Größen sind: Der zeitliche Abstand zwischen Reiz und Reaktion, die Zahl der falschen Reaktionen, die Zahl der ausgelassenen Reaktionen und die Zahl der voreiligen Reaktionen. Da die Reaktionszeiten schwanken, werden von den Programmen als Kennzahlen der Mittelwert oder der Median ausgegeben, sowie die Streuung der Werte um die mittlere Größe, die als Standardabweichung berechnet wird.

Die Vorgehensweise, Testergebnisse in Normwerten auszudrücken, ist gängige Routine. Dabei wird kaum je erwogen den basalen Mess- oder Zählergebnissen eine eigene Bedeutung zuzuweisen, obwohl diese der elementare Ausdruck der individuellen Performance sind. Doch genau das kann in gutachterlichen Fragestellungen zweckmäßig sein, sobald zu erörten ist, ob die vorliegende Performance einer Proband*in ihre bestmogliche ist, also ihrer Fähigkeit entspricht. Hierbei ist hilfreich, dass diese Ergebnisse auf direkten Messungen und Zählungen beruhen, die zu ihrer Ausführung keiner psychologischen Theorie bedürfen, weil sie nur Zeitdauer und Anzahl enthalten.

Indes weisen verschiedene Reaktionszeitaufgaben auch unterschiedliche Schwierigkeitsgrade auf, welche die Reaktionszeit eines Proband*in beeinflussen. Schwierigere Aufgaben werden eine längere Reaktionszeit erfordern, insbesondere, um fehlerfrei ausgeführt zu werden, ein Vorgang, der allerdings von Proband*in zu Proband*in unterschiedlich ausfallen wird. Daher ist der Grad der Schwierigkeit nicht unabhangig von der Stichprobe zu bestimmen und geht folglich implizit in die Verteilung der Reaktionszeiten und deren Normierung ein. Anhand dieser Normierung wird die Leistung – also die *Performance* - einer Proband*in der Bewältigung einer Schwierigkeit relativ zu anderen Leistungen plaziert. Als Basisgröße hierfür wird meist die mittlere Reaktionszeit der oft weit schwankenden Reaktionszeiten verwendet, entweder Mittelwert oder Median. Wegen dieser impliziten Abhängigkeit von der jeweiligen Aufgabenschwierigkeit sind die mittleren Werte der Reaktionszeiten einer Proband*in für verschiedene Aufgaben nur über den Umweg einer Normierung miteinander zu vergleichen. Folglich scheidet diese Größe für einen nicht normierten Vergleich von Ergebnissen aus. Selbstverständlich ist sie als Vergleichsgröße wichtig.

1 Dieses Papier ist die Umarbeitung eines früheren. Stand: Juni 2021

2. Fähigkeits -Testungen unter Begehrensvoraussetzungen

Die Ergebnisse von Reaktionszeitmessungen sind nicht bereits als solche der Ausdruck der *Fähigkeit*, des 'eigentlichen' Könnens eines Proband*in, weil diese Ergebnisse immer mitbestimmt sind von emotional-motivationalen und willentlichen Prozessen, umgangssprachlich formuliert, von Anstrengung und Mitarbeit. Sie beeinflussen die mittlere Reaktionszeit, die Reaktionszeitschwankungen und die Fehlerzahlen, gehen also mit ein in eben jene Größen, die als Ergebnisse der Testdurchführungen vorliegen. Deshalb sind die Ergebnisse – weder in Rohwerten noch in Normwerten – unbefragt als Ausdruck der Fähigkeiten eines Proband*in anzuerkennen. Doch es gibt keine Methode, mit der sich der Einfluss der Motivation aus den vorliegenden Testergebnissen eindeutig herausrechnen ließe. Dieses Problem ist besonders brisant in Testuntersuchungen, die mit Proband*in mit einem Renten- oder Entschädigungsbegehren durchgeführt werden. Solche Untersuchungen sollen klären, inwieweit Fähigkeitsminderungen vorliegen, die die Ausübung einer Berufs- oder Erwerbstätigkeit einschränken, oder – wie es juristisch heißt – sich "kausal" auf ein Ereignis (z.B. Unfall oder Gewalttat) zurückführen lassen.

Eher selten gibt es hierbei neurologische Befunde über eine Schädigung des Gehirns und somit keine Hinweise auf eine Funktionsstörung und damit verbunden, welche Fähigkeitseinschränkungen zu erwarten sind. Zudem unterliegt eine 'Breitband'-Testauswahl zur Überprüfung von Fähigkeiten gewissen situativen Beschränkungen: Wegen der geforderten Anstrengung sind einem Proband*in in der Regel an einem Tag nicht mehr als vier bis höchstens fünf Stunden Untersuchungszeit zuzumuten, in der zusätzlich zur Aufmerksamkeit auch Gedächtnis, Intelligenz und exekutive Funktionen zu prüfen sind. Ein zweiter Termin lässt sich aus anderen Gründen (Entfernung, Kosten) nur selten realisieren. Dadurch werden der Testauswahl enge Grenzen gesetzt.

Bekannte Tests zur Beschwerdevalidierung (BSV) haben bei der Aufklärung von Verfälschungstendenzen während Reaktionzeitmessungen einen beschränkten Nutzen. Einmal treten die BSV-Tests meist als Gedächtnistests auf, zum anderen muss zur Logik ihrer Interpretation klar gesagt werden: Die Auffälligkeit in *jedem einzelnen* BSV-Test *beweist* nicht, dass ein Proband in allen anderen Tests schlecht mitgearbeitet hat. Jedoch kann man auffälligen BSV-Tests ein Warnsignal entnehmen, um die Mitarbeit eines Proband*in bei anderen negativ auffälligen Testergebnissen genauer unter die Lupe zu nehmen.

Hierfür soll im Folgenden eine einfach zu handhabende mathematische Funktion konstruiert werden, die sich graphisch darstellen lässt, und die es gestattet, den Einfluss von Anstrengung bzw. Mitarbeit auf die Testergebnisse besser einzuschätzen. In diese Formel werden die nicht normierten, auf direkter Messung beruhenden Größen *Reaktionszeit*, *Reaktionszeitschwankungen* und *Fehlerzahlen* eingehen.

Zunächst erfolgt eine Darlegung von Definitionen und der daraus entwickelten Formeln, danach werden Hinweise zur Interpretation gegeben. Den Abschluss bilden einige rechtfertigende Nachgedanken.

3. Formeln für die Genauigkeiten

Nach den Handbüchern zu den einschlägigen Testverfahren dienen die Fehlerzahlen (falsche Reaktionen und/oder Auslassungen) oft als Maß für die *Sorgfalt* (So) im Ausführen der Anforderungen. Sei fz die Zahl der Fehler und Iz die Zahl der kritischen Items (auf die zu reagieren ist), so lässt sich die *Sorgfalt* quantitativ als eine Größe definieren, die proportional zu dem Quotienten (Iz - fz)/(fz+1) ist [2].

Analog lässt sich durch die Verwendung der Standardabweichung der Reaktionszeiten eine Größe *Konstanz* (Ko) definieren, die die Gleichmäßigkeit des Reagierens beschreibt: Sei m das Mittel (Mittelwert oder Median) und s die Standardabweichung, so sei Ko proportional zum Quotienten (m – s)/(s) [3].

Aus Gründen der Anschaulichkeit - die gesuchte Funktion soll ihre Interpretation anhand von Diagrammen erhalten - sollen sich die Werte der Funktion im Intervall]0 ; 1] bewegen. Die Werte sollen also die obere Grenze +1 nicht überschreiten und nach unten möglichst wenig unter 0 absinken. Die gesuchte Funktion erhält den Namen *Genauigkeit* (Gn).

Verwendet man Ko und So wie oben definiert, so werden die Quotienten mit kleinerem fz bzw. kleinerem s zwar größer, doch es entstehen durch Multiplikation oder durch Addition Werte größer als 1. Die Erfahrung zeigt zudem, dass s meist in der Größenordnung von 30% bis 60% von m liegt, der %-Satz von fz jedoch geringer ausfällt. Dadurch wird nach obiger Definition von So der Einfluss der Fehler in Gn zu gering gewichtet.

Um zu einer brauchbaren Formel im Wertebereich]0 ; 1] zu kommen, müssen die Formeln für So und Ko 'umgestülpt' werden zu So: = (Iz + fz) / (Iz – fz) und Ko: = m / (m -s). So enthält zudem eine höhere Gewichtung der Fehleranteile. Soweit – so einfach; doch wie ergibt sich ein Zusammenhang zwischen den beiden Größen So und Ko, die beide auf von einander unabhängigen Messvorgängen beruhen, auf der Messung von Zeiteinheiten [ms] und der Zählung von Fehlern?

Die Antwort lautet: Der Zusammenhang wird durch eine Formel gesetzt, die als solche durch direkte Messungen von Zeit und Anzahl leider nicht bestätigt werden kann. Sofern So und Ko voneinander unabhängig sind (geringe Korrelation), böte es sich an, die beiden als Komponenten eines Vektors zu betrachten, dessen skalare Größe aus der Summe der quadrierten Komponenten hervorgeht. Doch dieses Verfahren führt zu keinen brauchbaren Ergebnissen innerhalb des Intervalls]0 ; 1]. Besser sind einfache Mittelungsverfahren, beispielsweise das arithmetische, das geometrische und das harmonische Mittel, und zwar mit den folgenden Formeln [4]:

Formel (A): Gn: = 2 - Wurzel(Ko * So) = 2 – Wurzel[m / (m -s) * (Iz + fz) / (Iz – fz)]

---

2 Die 1 im Nenner verhindert, dass der Bruch für fz = 0 undefiniert wird.

3 Falls die Standardabweichung = Null ist, verliert die Rede vom Mittelwert ihren Sinn. Das kommt praktisch nie vor. Korrekterweise müsste an die genannten Größen der Index i angehängt werden (i = 1 ... bis ... n; n ist die Zahl der verwendeten Tests oder Testabschnitte). Der Übersichtlichkeit halber wird er jedoch weggelassen.

4 Damit in den folgenden Formeln kein Nenner mit dem Wert 0 entsteht, gilt für für diese Formel: fz < Iz; s ist positiv; und s < m. Falls die Daten dem nicht entsprechen, kann man sich fragen, ob eine sinnvolle Messung vorliegt. Für s > m wird der Ausdruck unter der Wurzel negativ und kann für die Darstellung mittels Diagramm durch einen beliebigen Wert kleiner Null ersetzt werden.

Gn kann negativ werden und nähert sich nach oben der ganzen Zahl 1 an. Die Werte aus dieser Formel existieren nur punktweise und lassen sich grafisch darstellen. Neben den absoluten Höhen sind der Verlauf bedeutsam.

Verwendet man statt des geometrischen das harmonische Mittel, so fallen die Differenzen in den Ergebnispunkten meist marginal aus:

Formel (B): Gn: = 2 – 2/(1/Ko + 1/So) = 2-2/[(m – s) / m + (Iz – fz) / (Iz + fz)]

Größer fallen die Unterschiede aus, wenn man statt der Formeln (A) oder (B) eine Formel für das arithmetische Mittel verwendet, doch deren Werte tendieren zu rasch gegen Null und fallen darunter. Zur Berechnung mittels Tabellenkalkulation empfiehlt sich die Formel (A). Die eher geringen Unterschiede zwischen den Formeln (A) und (B) sind für die Interpretation bedeutungslos. Detailgenaueres Rechnen führt also nicht notwendig zu genauerer Erkenntnis. Dochdrei Eigenschaften der Genauigkeitsfunktion sind diskussionswürdig:

(1) Infolge der Definition von So und Ko bzw. So und Ko hängt die Höhe der Genauigkeiten nicht direkt von der Höhe der mittleren Reaktionszeiten ab, weil es für einen bestimmten Quotienten 'theoretisch' eine unendliche Zahl von möglichen Eingangsgrößen für Zähler und Nenner gibt. Gn hängt allerdings von der Relation s/m ab, was beispielsweise dazu führen kann, dass eine langsame Reaktionszeit mit hoher Genauigkeit einher geht. Das mag auf den ersten Blick irritieren. Indes sind sowohl Mittelwert m als auch Standardabweichung s *nur Rechergebnisse;* diese Größen *existieren nicht* in einem Sinne wie die Messdaten. Die die Aufgaben ausführende Proband*in weiß nichts von ihrem Mittelwert und ihren Schwankungen: beide spielen in ihren Reaktionen keinen Rolle! Hingegen ist es für eine Proband*in anstrengend langsam *und* konstant zu reagieren, weil sie dazu kein vorgegebenes Maß hat, an das sie sich halten könnte. Eine menschliche Reaktion ist eben keine maschinelle Reaktion. In ihr ist keine Konstanz vorprogrammiert, Konstanz braucht Anstrengung! Zeigen die Auswertungen, das jemand im Mittel langsam und mit geringen Schwankungen reagiert hat, wodurch s/m klein ausfällt, so verweist dies auf eine gute Anstrengung einer Proband*in.

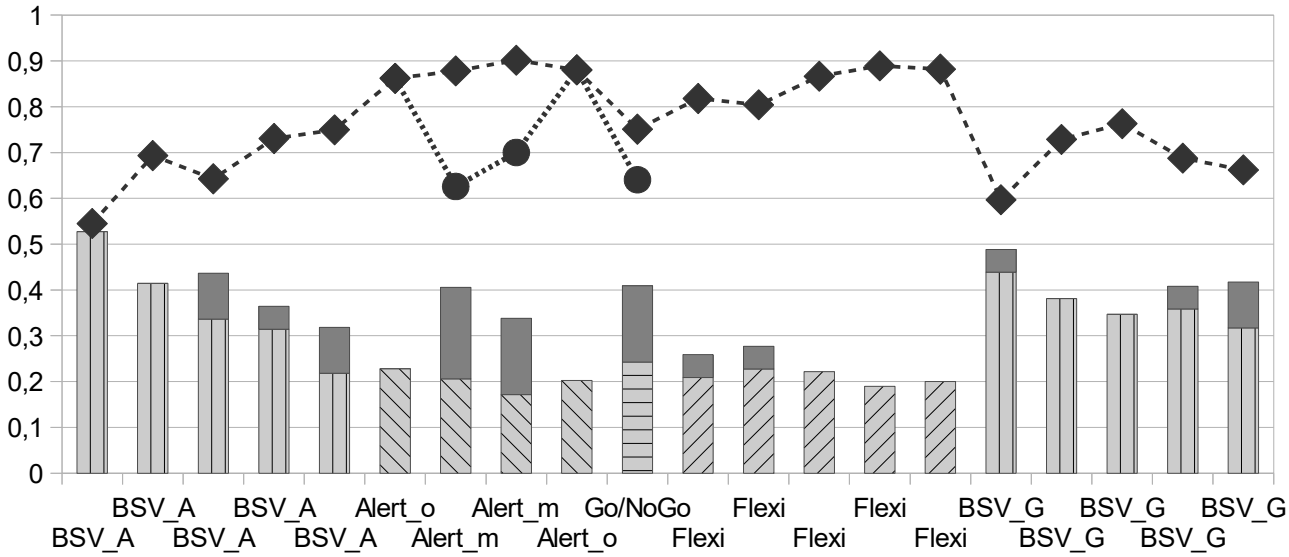
(2) Die Ergebnisse aus den Formeln (A) und (B) existieren zwar rechnerisch, weil die Formeln freie Konstruktionen sind, die weder Messergebnisse direkt abbilden, noch deren Vorhersage gestatten. Falls dies ein Problem ist, hat es bei dem hier gewählten Ansatz vermutlich keine Lösung, doch etwas daran ist indes 'hebbar'. Man kann die Differenz zwischen Formel (A) und Formel (B) grafisch fast verschwinden lassen, indem der Maximalwert nach Formel (A) auf den Maximalwert nach Formel (B) (oder umgekehrt) 'angehoben' wird und der Differenzwert zu all jenen Ergebnissen addiert wird. Das mag unbefriedigend sein, funktioniert jedoch hinreichend gut, weil die noch bestehenden, minimalen Unterschiede in der Interpretation der Genauigkeiten keine Rolle spielen.

(3) Eine weitere Schwierigkeit steckt in den Formeln für Ko und Ko. Während die Größen für Iz und fz direkt durch Abzählen gewonnen werden und dadurch eine Größe wie fz/Iz auf direkten Zähl--handlungen beruht, stehen in Zähler und Nenner des Quotienten s/m statistische Größen. Man kann deshalb einwenden, dass durch die Formeln (A) und (B) "Kraut und Rüben" verrechnet werden. Diese Schwierigkeit ließe sich möglichwerweise beheben, indem man Mediane und

Qantile verwendet. Doch dazu müssten alle einzelnen Reaktionszeiten bekannt sein, sowie die Zeitpunkte bzw. Stellen, an denen sich Fehler, Auslassungen oder voreilige Reaktionen ereignet haben. Solche detailreichen Auskünfte liefern aber nur wenige Testprogramme. In der Praxis darf man froh sein, wenn es definierbare Testabschnitte gibt, für die Mittelwerte, Standardabweichungen und Fehlerzahlen bekannt sind oder sich gut abschätzen lassen.

4. Darstellung mittels Diagrammen

Ein Test ist kein Test – so lautet eine Erfahrungsregel für die Praxis des Testens. Zur eingehenden Analyse der Mitarbeit eignet sich der Verlauf von Genauigkeiten über einige computerisierte Reaktionszeittestungen hinweg, die von einer Proband*in aufeinander folgend ausgeführt worden sind. Deren Datenplot sieht beispielsweise wie folgt aus:



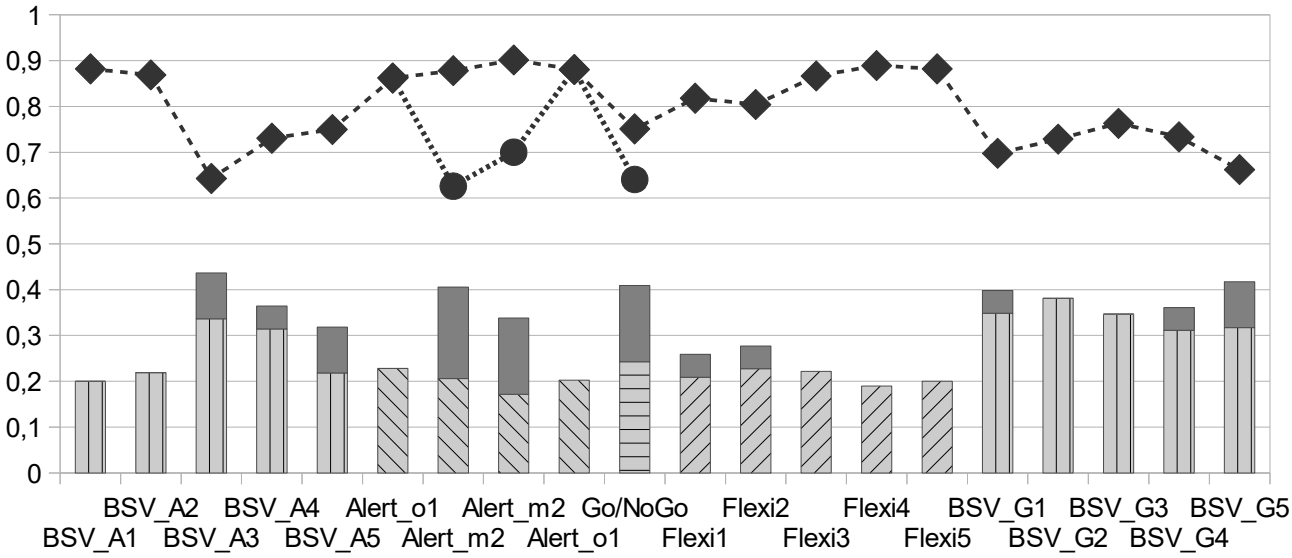
- Abb.1: Rhomben und Kreis-Symbole repäsentieren die Genauigkeiten Gn, deren obere Schranke = 1 ist. Gestreifte Balkenanteile stellen die relativen Schwankungen s/m dar; graue Balkenanteile stehen für die relativen Fehlerzahlen fz/Iz. BSV kennzeichnet zwei Tests aus der Bremer Symptom-Validierung, die Teil der TBFN ist (Heubrock & Petermann, 2013). *Alertness*, *Go/NoGo* und *Flexibilität* gehören zur Testbatterie zur Aufmerksamkeitsprüfung (TAP) (Fimm & Zimmermann, 2019). Die Nummerierung 1 ... 5 verweist auf Testabschnitte mit Iz = 20 kritischen Items. Die doppelten Symbole über *Alertness* und Go/NoGo ergeben sich aus unterschiedlichen Fehlerarten; sie werden im nächsten Abschnitt interpretiert.

Für die BSV-Tests existiert jede einzelne Reaktionszeit und jede Fehlerstelle im Testablauf, sodass jeder Genauigkeitspunkt genau zu berechnen ist. Die Größe der Abschnitte mit Iz = 20 ermöglicht den Vergleich mit den Ergebnissen für die Tests aus der TAP. Denn für die Abschnitte der *Alertness* sowie für Go/NoGo beträgt Iz = 20. Die Flexibilität wird anhand der grafischen Ergebnisausgabe nachträglich unterteilt, die Eingangswerte für die fünf s/m – Ergebnisse und für die entsprechenden Gn-Koeffizienten müssen leider geschätzt werden; die fz/Iz enthalten hingegen genaue Werte.

Die Abbildung hat Spiegelsymmetrie: je höher die Balken, desto niedriger die Genauigkeiten und

umgekehrt. Auf den ersten Blick ist zu erkennen, dass es konstante Reaktionsweise nicht gibt [5]. Erfahrungsgemäß sind Gn < 0,6 bereits stark auffällig. Denn für fz=0 und Gn=0,59 wird s/m ungefähr 0,69, d.h. die Standardabweichung beträgt bereits 69% des Mittelwerts, was beachtliche Reaktionschwankungen kennzeichnet. Kommen Fehler noch hinzu, fällt Gn rapide unter Null. Gn = 1 wird nie erreicht, weil s immer größer Null sein muss. Gn-Werte > 0,8 sind indes recht gut. Formel (A) und auch Formel (B) führen also – wie erwünscht – zu Ergebnissen in einem abbildungstechnisch günstigen Bereich.

Der Vergleich der BSV-Daten mit jenen der TAP-Testungen ist etwas heikel, weil er eine *Ausreißer-Korrektur* benötigt. Sofern sie greift, wirkt sie sich in den BSV-Testergebnissen zu Gunsten der Proband*in aus [6], wie der nächste Abbildung zu entnehmen ist.



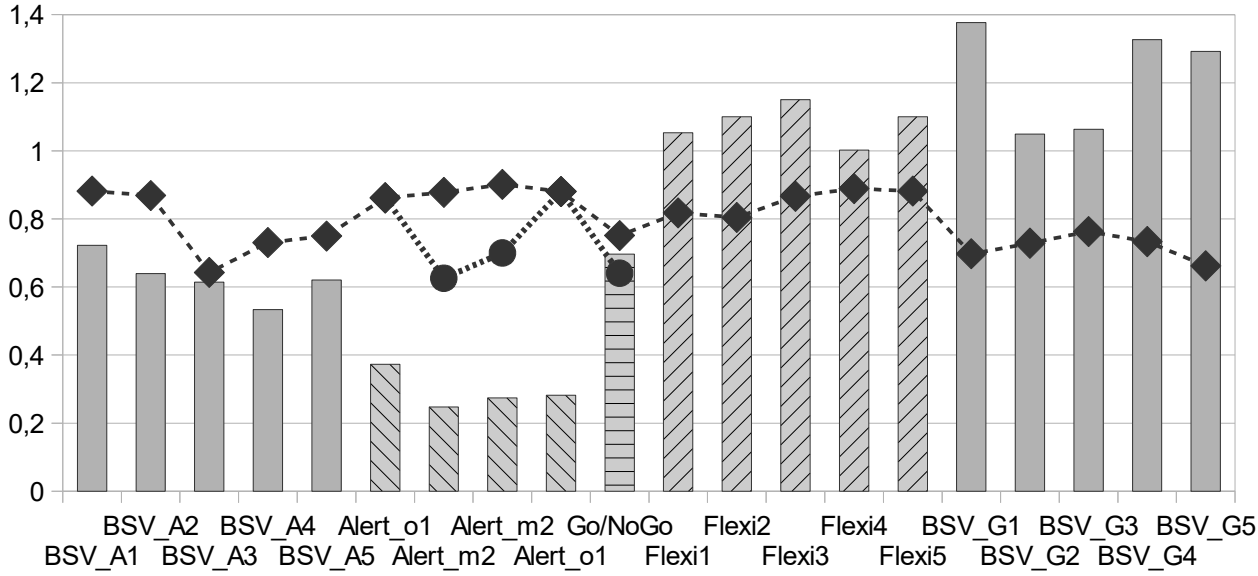
- Abb.2: Bezeichnungen wie in Abb.1, BSV-Datensätze mit zusätzlicher Ausreißer-Korrektur.

In den obigen Datensätzen erstreckt sich die Ausreißer-Korrektur vor allem auf die beiden ersten Abschnitte von BSV_A und BSV_G. Das ist insofern akzeptabel, als lediglich die ersten ein bis drei Reaktionen im ersten Abschnitt erfasst werden, die teilweise der Einübung dienen. Würden die Korrekturen in die späteren Abschnitte eingreifen, könnte ein problematisches Reagieren unkenntlich werden.

Aufschlussreich ist oftmals auch ein Vergleich der Genauigkeiten (Gn) mit den mittleren Reaktionszeiten (Rz) der Testabschnitte.

5 Lediglich selten und näherungsweise gibt es sie bei recht zwanghaften Proband*innen oder solchen, die in ihrer Berufstätigkeit sehr genau arbeiten müssen. Allerdings fällt deren Tempo in Normwerten (Median oder Mittel-wert) eher unterdurchschnittlich aus.

6 Auf https://www.psytest.de ist zu erfahren, dass der Normierung aller TAP-Testergebnisse eine *Ausreißer-Korrek-tur* voraus gegangen ist. Werte ab |2,35*s + m| gelten in jedem Datensatz als Ausreißer. Zweckensprechend sollte also mit jedem BSV-Datensatz einer Proband*in, der mit den TAP-Ergebnissen verglichen werden soll, zuvor eine Ausreißer-Korrektur durchgeführt werden. Danach müssen Mittelwert und Standardabweichung neu berechnet werden. Da sich die Korrektur stärker auf s als auf m auswirkt, wird der Quotient s/m kleiner und folglich der Wert für Gn besser.



- Abb.3: Genauigkeiten (Rhomben und Kreis-Symbole) und Mittelwerte (senkrechte Balken) der Reaktionszeiten bei Iz = 20. Die Ausreißerkorrektur (→ Abb.2) betrifft die Mittelwerte der Reaktionszeiten (Rz) und die Genauigkeiten (Gn) der BSV-Tests. Senkrechte Achse: Rz [Sekunden]; Gn [ohne Maßeinheit]

Für die drei Tests, die mit zwei Tasten bedient werden (BSV_A, *Flexibilität* und BSV_G), sind längere Reakionszeiten zu erwarten als für *Alertness* und Go/NoGo, die mit einer Taste zu bedienen sind.

4. Interpretation

Liegen alle Gn im Bereich [0,8 ; 1], so ist erfahrungsgemäß die Mitarbeit bzw. Anstrengung einer Proband*in in der Reaktionszeittestung nicht zu bemängeln. Einzelne Abweichung nach unten sollten gesondert betrachtet werden [7].

Zunächst kann mittels Formel (A) eine Vergleich der globalen Genauigkeiten Gg für die drei Tests mit Iz = 100 und Zwei-Tasten-Bedienung durchführen. Für BSV_A ist Gg = 0,766, für BSV_G ist Gg = 0,745 und für Flexibilität beträgt Gg = 0,816. Offenbar kann Probandin im schwersten Test der Serie, der *Flexibilität*, besser mitarbeiten als in den beiden einfachen Tests zur Beschwerdevalidierung! In Abb.3 fällt auf, dass die Rz für die leichte BSV_G länger ausfallen als für die erheblich schwerere *Flexibilität:* Die mittlere Reaktionszeit (Rz) in der *Flexibilität* beträgt 1063 ms; sie ist geringer als die 1216 ms für den letzten Test BSV_G. Erwartungsgemäß liegt die mittlere Rz für BSV_A mit 623 ms deutlich darunter und ist länger als die beiden mittleren Rz für *Alertness*.

7 Eine Datenanalyse fördert nur jene Information zu Tage, die bereits in den Daten enthalten ist. Logische Schlüsse sind nicht erkenntniserweiternd. Eine Interpretation fügt jedoch hinzu: Nicht nur Wissen aus dem fachlichen Hintergrund und der Erfahrung des Interpretierenden, sondern Bedeutung und menschlich-einfühlende Einschätzung, die zu einer Gesamtbewertung führt. Aus einer Interpretation ist das Personal-Menschliche nicht zu entfernen, ja es darf daraus nicht entfernt werden, um die Interpretation einer Künstlichen Intelligenz zu überlassen. Menschen dürfen nur durch Menschen eingeschätzt werden, nicht aber durch Maschinen (Precht, 2020)! Deswegen kann es auch keine allgemeine, in jedem Fall zutreffende Regel für das Interpretieren geben.

Die Kreis-Symbole kennzeichen problematische Reaktionen. In den Testabschnitten *Alert_m1* und *Alert_m2* reagiert die Proband*in insgesamt neunmal auf das Vorsignal und in Go/NoGo dreimal auf das falsche Signal. Das kann für die impulsive Entladung einer emotionalen Spannung sprechen, deren willentlich Kontrolle misslingt. Die in Go/NoGo sehr langsame mittlere Rz, die in Normwerten sehr niedrig ausfällt, bewahrt sie möglicherweise davor noch mehr falsche Reaktionen zu machen. Ist diese schlechte Performance nun ein Zeichen für Anstrengungsvermeidung oder für geschwächte Fähigkeiten [8]?

Die fünf Reaktionszeittestungen sind Teil einer längeren Testserie, davor werden drei Papier-Stift-Tests durchgeführt, danach eine Gedächtnisbatterie und eine Intelligenztestung. An deren Ende wird, nach drei bis vier Stunden Dauer, die kognitive Ermüdung geprüft. Nur selten kommt es zu einer entsprechenden Feststellung, im vorliegenden Fall jedenfalls nicht. Hinweise auf eine Hirnschädigung gibt es nicht. Insgesamt dauern die Reaktionszeitmessungen rund 45 Minuten, davon entfällt der weitaus größte Zeitanteil auf Zwischenpausen, Unterbrechungen und Probedurchgänge. Folglich ist für den zeitlichen Verlauf in Abb1 oder Abb.2 nicht vom einem negativen Einfluss einer kognitiver Ermüdung auf die Performance auszugehen. Mehr spricht für den Einfluss motivational-emotionaler Prozesse.

Für den Testablauf der BSV_A vom ersten bis zum letzten Tastendruck benötigt die Proband*in rund 3 Minuten 45 Sekunden! Die reine Reaktionszeit beträgt 1 Minute und 6 Sekunden. Aus Abb.2 geht hervor, dass die relativen Schwankungen im dritten Abschnitt nach oben springen, sodann Fehler hinzu kommen und folglich die Genauigkeiten absinken, wohingegen in Abb.3 die Rz nur gerigfügig schneller wird. Hier zeichnet sich der Einfluss emotionaler Impulse ab: die Proband*in will es schnell hinter sich bringen. In der *Alertness ohne Vorsignal* sind entsprechend der Testkonstruktion keine Fehler zu machen, Auslassungen unterlaufen ihr nicht. In der *Alertness mit Vorsignal* gibt es sodann die problematischen neun impulsiven Reaktionen auf das Vorsignal, die zum Absinken der Gn führen. Möglicherweise versucht sie in Go/NoGo die Kontrolle über ihre Impulse zu gewinnen, indem sie insgesamt langsamer reagiert, wodurch die Gn aber nicht besser wird. Insgesamt ist daher in der BSV_A, der *Alertness* und in Go/Nogo von einer emotional beeinflussten Performance auszugehen, die unterhalb des Fähigkeitsniveaus der Proband*in liegt.

Das ändert sich jedoch in der *Flexibilität*. Nach den anfänglichen zwei Fehlern steigt die Genauigkeit in einen guten Bereich. Die mittlere Rz für den ganzen Test liegt in Normwerten im unteren Normalbereich. An diesem schwersten Test der Serie gibt es nichts zu bemängeln! Dieser Test erfordert – umgangssprachlich ausgedrückt – eine in sich gesammelte Ausrichtung auf

8 Man kann an dieser Stelle nicht auf eine Diagnose zurück greifen, um zu behaupten, was jemand kann oder nicht kann. In Repliken auf Gutachten mit unerwünschtem Ausgang, steht in etwa: „Wie Herr Prof. X festgestellt hat, leidet … an der Krankheit mit Diagnose Y, die mit Einschränkungen … einher geht." Abgesehen davon, dass dies ein 'Schluss' vom Begriff einer Sache auf deren Existenz bzw. von einem Sollen auf ein Sein ist, beruht eine psychische Diagnose noch nicht einmal auf Messungen, wie sie von Medizinern zur Diagnostik von Krankheiten durchgeführt werden. Meistens wird sie auf Grund verbaler Angaben von Patienten gestellt! Eine psychische Diagnose beweist daher überhaupt nicht, was jemand kann oder nicht kann (Schneider et al., 2016). Psychologische Tests hingegen prüfen Fähigkeiten, jedoch nur unter der Voraussetzung, dass eine Proband*in optimal mitarbeitet (Linden et al., 2015) – und genau das gilt es an dieser Stelle zu überprüfen. Wäre dies einfach, wären auch die Überlegungen zur Genauigkeit überflüssig. Im Grunde stehen zur Prüfung der Mitarbeit bei Testungen (Fragebögen sind keine Tests!) nur zwei Wege zur Verfügung: die Datenanalyse und ihre Interpretation sowie die Verhaltensbeobachtung.

Wahrnehmung und motorische Ausführung. Das ist anstrengend, weil jede impulsive Reaktion zu Fehlern führen würde. Hier hat die Proband*in ihre emotionlanen Impulse also unter exekutiver Kontrolle. Wieso also davor nicht? Herausforderung führt folglich zu besserer Performance! [9] Wie zur Bestätigung dieses Satzes sinkt die Performance in der abschließenden BSV_G wiederum ab. Sind es in der Flexibilität zwei Fehler, so werden es in der BSV_G fünf Fehler. In Abb.2 nimmt das Verhältnis von s/m in allen fünf Abschnitten wieder zu. Da ist er wieder, der impulsiv-emotionale Einfluss auf die Reaktionen. Fast möchte man sagen: Die Proband*in hat schlichtweg keine Lust auf die banalen Aufgaben der BSV_G. Fazit: Abgesehen von der *Flexibilität* ist die Performance in den übrigen Tests eine unterhalb des Fähigkeitsniveaus der Proband*in.

5. Nachgedanken

Direkte Messhandlungen, die auf dem jeweiligen historisch-technischen Stand des Alltagshandelns ohne Rückgriff auf fachspezifische Theorien durchgeführt werden können, sind in der Psychologie eher selten; in der Experimentalphysik bilden sie im MKSA-System eine unentbehrliche Grundlage aller Messvorgänge (Vogel, 2012). Die Genauigkeitsfunktion beruht jedoch auf direkten Messhandlungen: Jeder 'normale' Erwachsene kann die Zeit messen und Fehler abzählen. In der psychologischen Testpraxis sind jedoch abgeleitete Messhandlungen an der Tagesordnung, sie beruhen auf mathematisch formulierten Vorannahmen über Zusammenhänge, die davon unabhängig nicht messbar sind. Ein bekanntes Beispiel ist ein Intelligenztest. Darin werden Wortschatzübungen mit Logikaufgaben durch Abzählen der Anzahl gelöster Aufgaben gleich gesetzt und eine gewisse Anzahl solcher Gleichsetzungen zu einem IQ-Koeffizienten addiert. (Die Lösung der Aufgaben setzt allerdings seitens der Testperson ein Wissen und Können voraus, das über bloßes Reagieren in Zeitmessungen und und das Zählen von Fehlern hinaus geht.) Die Rechtfertigung für die Gleichsetzung der unterschiedlichen Aufgaben mittels ganzer Zahlen erfolgt erst ex post durch die Berechnung von bestimmten Korrelationskoeffizienten (Itemschwierigkeit, Reliabilität, Validität), die an großen Stichproben gewonnen werden. So führt die Theorie zur Messhandlung und die Messhandlung bestätigt ihrerseits die Theorie - innerhalb eines korrelativen Gebildes des Wissens, das wie ein Fischernetz im Meer einer in weiterhin unbekannten 'Wirklichkeit' dahintreibt.

Für die Formeln der Genauigkeit besteht die einfachste und trivialerweise vermutlich 'immer wahre' Annahme darin, den Zusammenhang in der Testperson (der Proband*in) anzusiedeln, weil es ein- und dieselbe Person ist, die sowohl die Reaktionsschwankungen als auch die Fehlerzahlen produziert. Die für deren Zusammenfügen zu einer Einheit nötige 'intuitive Idee' könnte lauten, dass der Mensch ein schlecht funktionierender Computer sei oder ein solcher, auf dem ein fehlerhaftes Programm läuft. Denn ein gut funktionierender Computer reagiert auf einen eingehenden Reiz fehlerlos mit konstanter Verzögerung, wodurch s = 0 und fz = 0 werden und somit die höchste *Genauigkeit* erzielt wird. Das mag abschreckend klingen, doch im Grunde folgt die computerisierte Aufmerksamkeitstestung dieser Idee schon längst: Als wichtigste Größe in den einfachen Tests wie *Alertness* oder Go/NoGo, aber auch in der *Flexibilität,* gilt die mittlere Reaktionszeit (oder der Median), mit der Folge, dass man nicht so recht weiß, was die

9 Diese vielleicht merkwürdig klingende Feststellung beruht auf jahrelanger Testerfahrung mit Proband*innen.

Standardabweichungen bedeuten. Oder sind sie schlicht nur Meßfehler oder Reaktionsfehler? Falls ja, definiert die beschreibende Statistik das Seiende, das, was existiert. Indes sollte das Seiende den Grund für den Zweck statistischer Operationen bilden.

Der *Alertnesstest* konstruiert eine hochgradig reduzierte Weise menschlichen Reagierens innerhalb vorgegebener Zeitintervalle. Weniger geschieht nur noch in der *Vigilanztestung* – und entfernter vom Alltagsleben geht es kaum noch. Sturm (2005) reiht sie unter den Aspekt der Intensitätsaktivierung ein. Entsprechend einem Denken entlang des Konditionierungsparadigmas wird die Alertness ohne Warnreiz als ein "Maß an Aufmerksamkeitsaktivierung" definiert, als eine "intrinsische Fähigkeit", die "ausschließlich probandenbestimmt" sei. Deren Messung diene zudem der "tonischen Aufmerksamkeitsaktivierung", die vom physiologischen Zustand des Organismus bestimmt sei.

Abgesehen davon, dass eine Proband*in der Praxis des Testens unter dem extrinsischen Druck steht reagieren zu *müssen*, ist unklar, was mit den kryptisch anmutenden Formulierungen gemeint ist. Weder aus den neuronalen Verknüpfungen, die von bestimmten Hirnregionen ausgehen und mit ihnen rückkoppeln, noch aus der reduziertesten Reaktionszeitmessung lässt sich so etwas wie eine "intrinsische Fähigkeit" herleiten. Gehirnteilen können keine Fähigkeiten zugesprochen werden. Denn Fähigkeiten sind etwas, das wir Menschen einem anderen Menschen als lebendigen Wesen zuschreiben und nicht einem seiner Organe oder gar Teilen davon (Hacker & Bennet, 2010; Fuchs 2013). Um eine Fähigkeit zu messen, muss sich diese in der Welt äußern, eine Probandin muss also zumindest auf eine Taste tippen. Damit definiert der Messvorgang [10], was eine "Fähigkeit" ist. Eine dabei durchgeführte fMRT-Untersuchung zeigt, welche Gehrinprozesse bei der durch den Messvorgang konstruierten "Intensität der Aufmerksamkeitsaktivierung" aktiviert werden, nicht aber so etwas wie den 'Sitz einer Fähigkeit' oder gar den 'Ort im Kopf', wo sie erzeugt wird.

Macht man den Mittelwert der Reaktionszeiten zum Maß für eine Fähigkeit, so wird zu einem Fehler, was davon abweicht. Doch das ist nicht mehr als ein semantischer Trick. Tatsächlich gemessen werden nur die einzelnen Reaktionszeiten an bestimmten Zeitstellen, an denen sie im Programm 'erwartet' werden. Und, falls zwei Tasten verwendet werden, von welcher Taste sie auszugehen haben. Entstehen sie dort nicht, entsteht ein Fehler. Das Programm wird also die Zeitspanne zwischen einem 'Startzeitpunkt' und einem Eintreffen des Reaktionssignals entweder

10 Bei der Messung der Reaktionszeit handelt es sich tatsächlich um eine Messung. Nicht weil es die in der psychologischen Methodenlehre verbreitete Einteilung der Skalenniveaus, die ein Erbstück des *logischen Empirismus* ist, ebenfalls so benennt, sondern weil im Rechner eine Uhr die Dauer der Reaktion *misst*. Im Messvorgang wird ein physischer Vorgang mit einem anderen physischen Vorgang verglichen, für den es ein physikalische Erklärung gibt. Der Zweck der Messung und damit die Skalierung und die Geometrie werden freilich seitens der Menschen, die eine Messung durch führen wollen, vorgegeben. Deshalb kann eine Messung gelingen oder scheitern. Verwunderlich ist eher, dass sie gelingt. Denn die Interaktion zwischen Messinstrument und zu messendem Gegenstand beruht dabei auf (bestenfalls physikalisch gut erklärten) Naturprozessen, die nicht nur im Messprozess ablaufen, sondern darüber hinaus allgemein. In diesem Sinne wird in der Psychologie eher selten gemessen, es sei denn, man misst z.B. Puls, Hautleitfähigkeit oder EEG- Potenziale – doch die Erklärung solcher Messungen ist wiederum physikalisch. Weder mittels Tests noch mittels Fragebogen, noch in Experimenten wird im obigen Sinne gemessen. Denn den Menschen, die geprüft werden oder an Experimenten teilnehmen, werden Fragen vorgelegt oder Aufgaben, die sie durchzuführen haben, wobei richtigen oder falschen Antworten oder richtigen, teilweise richtigen oder falschen Aufgabenlösungen anschließend Zahlen zugeordnet werden. Dass eine solche Zuordnung gelingt, ist freilich nicht verwunderlich.

an der falschen Zeitstelle messen oder eine Zeitspanne lang auf eine Reaktion 'warten', um dann '0' zu codieren. Das Programm misst also gar keine Fehler! Es misst entweder die Zeitspanne bis zum Eintreffen eines Signals oder die Zeitspanne seines 'Ausbleibens'. Es misst auf jeden Fall eine Zeitspanne und keine Fehler! Es misst auch keine Mittelwerte und keine Standardabweichungen, es berechnet sie. *Fehler* haben nur Bedeutung in der menschlichen Sprache! Die Zeit hat keine Fehler! Fehler sind also die Codierung für nicht ausgegebene Zeitspannen an der *falschen*, d.h. nicht an der erwarteten Zeitstelle. Formel (A) und Formel (B) addieren daher unbekannte Zeitspannen als *Fehler* zu bekannten Zeitspannen. Flapsig gesagt, werden also in den Formeln nicht "Kraut und Rüben" addiert, sondern "Rüben" und einfache Schätzgrößen für unbekannte "Rüben". Zugegeben, das ist nicht elegant, funktioniert aber für die Praxis der Begutachtung hinlänglich gut.

6. Nachtrag (Mai 2021)

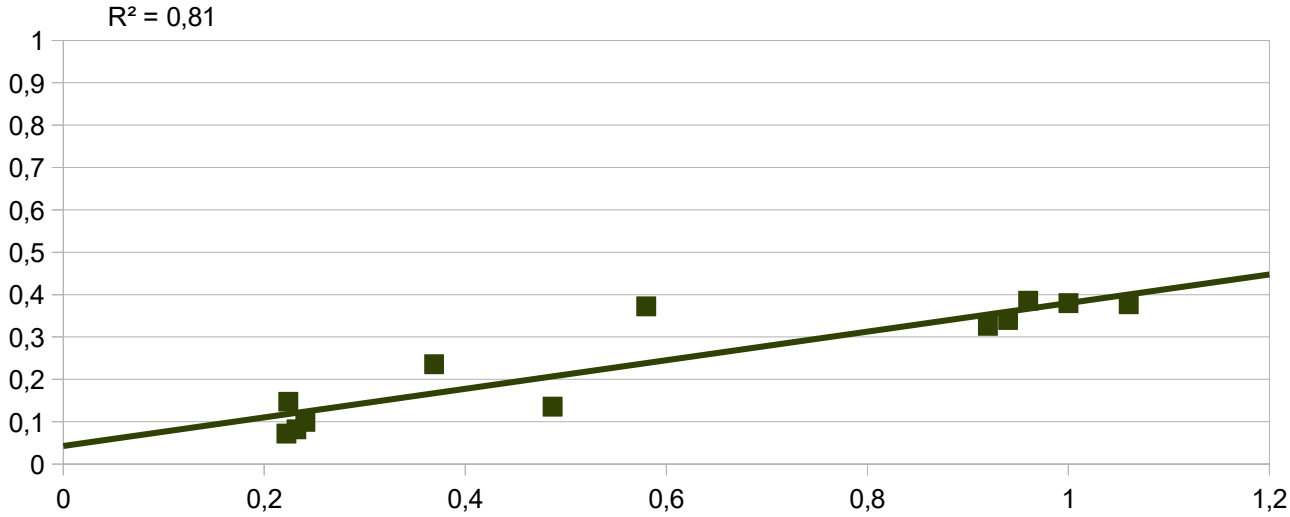
Die verwendeten, bereits einfachen Formeln ließen sich vermutlich verständlicher machen, hätte man statt Fehlerangaben die gemessenen Zeiten für Auslassungen und falsche Reaktionen zur Hand. Dann bliebe als einzig 'echter Fehler' jener bei der Benutzung zweier Reaktionstasten, ein 'Symmetriefehler'. Physikalisch gesehen gibt im zeitlichen Verlauf keine 'echten Fehler' – nicht reagieren ist dasselbe wie unendlich langsam reagieren. 'Echte Fehler' entstehen in der Raumsymmetrie durch die Verwechslung von links und rechts.

Ein praktisch weitaus wichtigerer Gesichtspunkt ist der des intraindivudellen Vergleichs. Der Verlauf der Genauigkeiten gestattet einen Vergleich der genauen Arbeitens einer Proband*in, während einer Folge von Aufgaben und über eine bestimmte Zeitspanne hinweg – unter Berücksichtigung zweier Eigenschaften der Testaufgaben:

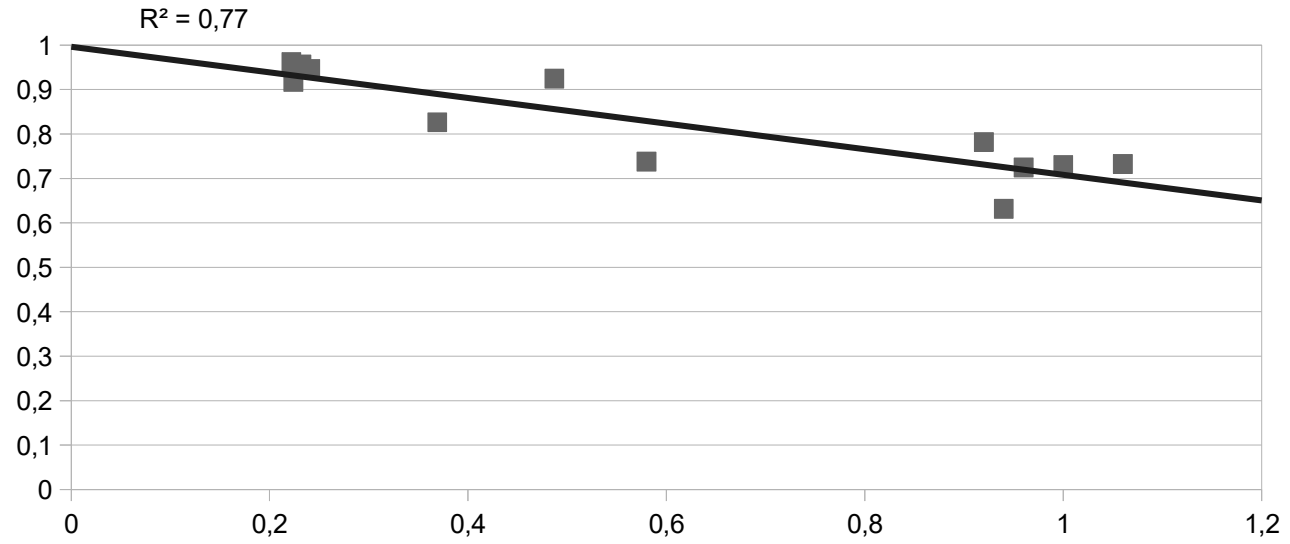
(a) Die Reaktion hat unmittelbar auf das Aufscheinen bestimmter Zeichen auf dem Bildschirm zu erfolgen; es dürfen keine, den 'kognitiven Aufwand' erhöhenden Wahrnehmungs- und Suchvorgänge nötig sein. Sind solche erfordert, wächst das Verhältnis s /m oft sprunghaft an.

(b) Die *Alertness* ohne Vorsignal ist der simpelste Reaktionszeittest. Alle andere Tests erfordern mehr 'kognitiven Aufwand'. Dieser schlägt sich in den Rohdaten in breiteren Schwankungen der Einzelergebnisse nieder, insbesondere sofern zwei Reaktionstasten zu bedienen sind. Es lässt sich empirisch, d.h. an vielen einzelnen Verläufen beispielhaft zeigen, dass unter diesen Umständen der Quotient s /m tendenziell zunimmt und folglich der Quotient Genauigkeit / mittlere Rz abnehmen muss (Abbildungen auf der nächsten Seite). Der Grund hierfür ist 'menschlicher Art': Es ist schwieriger langsamere Reaktionen konstant zu halten als schnellere.

Die normierten Testergebnisse, z.B. der weithin verwendeten Testbatterie TAP, aber auch anderer Programme, beruhen auf diversen Eichstichproben, die untereinander keinen Zusammenhang haben. In der Praxis werden die Tests aber von einer Proband*in zeitlich nacheinander ausgeführt. Folglich haben deren Ergebnisse einen inneren Zusammenhang. Aber unterschiedliche Ergebnisse einer Proband*in in Normwerten können leider auch auf der Eigenart der Stichprobe und der für die Normierung durchgeführten Ausreißer-Korrekturen beruhen. Da in die Genauigkeiten nur die nicht normierten Größen eingehen (m, s/m, Fz, Iz), kann man sie auch als *Reaktionseffektivitäten* betrachten und bei Zweifeln an den normierten Ergebnissen als Korrektiv verwenden.



- Abb. 4: Waagrechte Achse: mittlere Reaktionszeiten (m) in Sekunden; senkrechte Achse relative Schwankungen s / m. Verwendete Tests einer Proband*in: *Alertness*, Go/NoGo, *Flexibilität*, *Geteilte Aufmerksamkeit*. Die *Flexibilität* wird mit zwei Tasten bedient, was die längsten Reaktionszeiten erzeugt.



- Abb. 5: Waagrechte Achse: mittlere Reaktionszeiten (m) in Sekunden; senkrechte Achse Genauigkeiten (Max = 1). Verwendete Tests einer Proband*in: *Alertness*, Go/NoGo, *Flexibilität*, *Geteilte Aufmerksamkeit*. Die *Flexibilität* wird mit zwei Tasten bedient, was die längsten Reaktionszeiten erzeugt.

Autor und Copyright: Dr. Wolfgang Palm
Dipl.-Psych., Dipl.Phys., Psychotherapeut
Sachverständiger der Psychotherapeutenkammer BaWü
www.psy-gutachten.de
Stand des Papiers: Juni 2021

Literatur:

Bennet R M, Hacker P M S (2010). Die philosophischen Grundlagen der Neurowissenschaften. Wissenschaftliche Buchgesellschaft. Darmstadt.

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

Fuchs T (2013). Das Gehirn – ein Beziehungsorgan. 4., aktualisierte und erw. Auflage. Verlag Kohlhammer

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

Linden M, Baron S, Muschalla B, Ostholt-Corsten M (2015). Fähigkeitsbeeinträchtigungen bei psychischen Erkrankungen. Verlag Hans Huber

Palm W (2020a). Anmerkungen zur Beschwerdenvalidierung, https://www.wopalm.com/material/Beschwerdenvalidierung.pdf

Palm W (2020b). Eine kleine Evaluation der Beschwerdevalidierung. https://www.wopalm.com/material/Evaluation.pdf

Precht R D (2020). Künstliche Intelligenz und der Sinn des Lebens. Verlag Goldmann

Schneider W, Dohrenbusch R, Freyberger H.J, Henningson P, Irle H, Köllner V, Widder B (2016). Begutachtung bei psychischen und psychosomatischen Erkrankungen. Verlag Hogrefe

Sturm W (2005). Aufmerksamkeitsstörungen. Verlag Hogrefe

Vogel H (2012). Gerthsen Physik, 20. Auflage, Verlag Springer