Föderation Deutscher Psychologenvereinigungen

Berufsverband Deutscher Psychologinnen und Psychologen e. V.

DGPs Deutsche Gesellschaft für Psychologie

Stellungnahme zu Berg und Schubert
Commentary to Berg and Schubert

Diagnostik- und Testkuratorium der Föderation Deutscher Psychologenvereinigungen
verabschiedet am 28. Juni 2022

Abstract

Berg und Schubert (2019) kritisieren in ihrem Beitrag die Richtlinien zur Bewertung von Testverfahren und -geräten zur Wiederherstellung der Kraftfahreignung. Diese Richtlinien würden, so die Kritik, die „theoriegeleiteter Validierung" ignorieren. In dieser Stellungnahme zeigen wir, dass die Richtlinien diesen Zugang zur Validierung im Gegensatz zu Berg und Schuberts (ebd.) Behauptung durchaus berücksichtigen. Wir spezifizieren die Anforderungen an eine „theoriegeleitete Validierung", vor allem mit Blick auf eine empirische Untermauerung. Zudem weisen wir noch einmal auf die Bedeutung des in den Richtlinien gewählten ganzheitlichen Validierungsansatzes hin und gehen allgemein auf aktuelle Qualitätsstandards in der Testvalidierung ein. Dabei wird betont, dass auch die von Berg und Schubert (ebd.) kritisierten korrelativen Zugänge zur Validierung eine starke Theorie benötigen. Zusammenfassend lässt sich sagen, dass die von Berg und Schubert (2019) vorgebrachte Kritik unbegründet ist und auf einer selektiven Darstellung der Richtlinien sowie verschiedener Aussagen anderer Autoren beruht.

Abstract

Berg and Schubert (2019) criticize in their contribution the guidelines for the evaluation of test procedures and equipment for restoring motor fitness. According to the critics, these guidelines would ignore "theory-based validation". In this commentary, we show that the guidelines take this approach to validation into account, as opposed to Berg and Schubert's (ibid.) assertion. We specify the requirements for a "theory-driven validation", especially with a view to empirical support. In addition, we once again point out the importance of the holistic validation approach chosen in the guidelines and generally address current quality standards in test validation. It is emphasized that the correlative approaches to validation criticized by Berg and Schubert (ibid.) also require a strong theory. In summary, it can be said that the criticism expressed by Berg and Schubert (2019) is unfounded and is based on a selective presentation of the guidelines and various statements by other authors.

Stellungnahme zu Berg und Schubert

In ihrem Beitrag „Zur Bedeutung der theoriegeleiteten Validierung von Testverfahren für die Fahreignungsbegutachtung" kritisieren Berg und Schubert (2019) die Auslegung des Begriffs Validität bei der Bewertung von Testverfahren nach der Richtlinie zur Bestätigung der Eignung von Testverfahren und -geräten und der Eignung der Kurse zur Wiederherstellung der

Kraftfahreignung, Verkehrsblatt-Verlautbarung (VkBl. Amtlicher Teil, Heft 6-2017, S. 277ff)[1]. Die von den Autoren fokussierten Ausführungen zur Validität wurden von einer BASt-Expertengruppe zur Einrichtung unabhängiger Stellen zur Prüfung von Verfahren und Maßnahmen formuliert. In dieser Expertengruppe wirkten auch Verteter*innen des Diagnostik-Testkuratoriums der Föderation Deutscher Psychologenverbände (DTK) mit. Das Verfahren zur Bestätigung der Eignung von Testverfahren und -geräten zur Wiederherstellung der Kraftfahreignung weist explizit große Übereinstimmungen mit dem allgemeinen Testbeurteilungssystem des Diagnostik- und Testkuratoriums (2018) auf. Ziel der vorliegenden Stellungnahme ist es, die von Berg und Schubert (ebd.) bezüglich der Validität vorgebrachten Kritiken zu prüfen. Vor allem geht es dabei um den Begriff der theoriegeleiteten Validierung und dem aktuellen Standard zur Validierung.

Kritikpunkte von Berg und Schubert

Die Autoren behaupten, dass in den kritisierten Richtlinien eine nicht mehr zeitgemäße Verwendung des Begriffs Validität genutzt würde, um Testverfahren zu beurteilen. Sie führen als vermeintliche Alternative zur Konstrukt- und Kriteriumsvalidiering, den Begriff der „theoriegeleiteten Validierung" ins Feld. Eine genaue Definition dieses Begriffes erfolgt nicht, lediglich die Feststellung, dass der Begriff so nicht direkt, sondern nur sinngemäß in verschiedenen Qualitätsstandards zur Beurteilung von Messverfahren enthalten sei. Es lässt sich aber aus den zitierten, eigenen Arbeiten der Autoren erschließen, dass es hierbei darum geht, eine starke Theorie zu haben, die genau darüber informiert, wie sich das zu messende Konstrukt bei der Bearbeitung der Testaufgaben auswirkt. Die Testkonstruktion solle dann auf Basis dieser theoretischen Annahmen erfolgen und so die Basis für eine valide Testwertinterpretation bilden. Die Autoren beziehen sich stark auf Arbeiten von Borsboom, Mellenbergh, and Van Heerden (2004) sowie auf die aktuellen APA Standards (AERA APA & NMCE, 2014).
Schließlich stellen die Autoren die Befürchtung auf, dass durch die Fokussierung auf klassische Validitätsansätze ein „theoriegeleitet validiertes" Verfahren keine Anerkennung in der anstehenden Begutachtung finden könnte. Im Folgenden werden wir zu diesen Punkten Stellung nehmen.

Der Begriff Validität in den Bewertungsrichtlinien – Fehlt die „theoriegeleitete Validierung"?

Die Richtlinien wurden vor dem Hintergrund der bestehenden Gesetzeslage erstellt. Diese sieht vor, dass zur Fahreignungsprüfung die fünf Eigenschaften Belastbarkeit, Orientierung, Konzentration, Aufmerksamkeit und Reaktionsfähigkeit begutachtet werden müssen. Um den Nachweis zu führen, dass ein Testverfahren Werte ergibt, die im Sinne dieser Konstrukte interpretiert werden können, wurde in der BASt Expertenrunde abgewogen, welche Validitätsaspekte relevant seien. Ein Anspruch, dass diese Eigenschaften Fahreignung vorhersagen sollen bzw., dass die Testergebnisse bezüglich dieses Kriteriums prädiktiv sein sollen, besteht nicht, was einen Verzicht auf Evidenz bezüglich der Kriteriumsvalidität bedeutet. Auf Seite 9 ihres Beitrags führen Berg und Schubert (2019) Gründe für die Unterschätzung theoriegeleiteter Validität innerhalb der Fahreignungsdiagnostik auf und suggerieren, die Richtlinien würden einen kriterienorientierten Validierungsansatz verfolgen. Dies ist, wie beschrieben, schlicht falsch. Neben diesem Aspekt der Validität diskutierte die Expertenrunde weitere, weit verbreitete Standards zur Validierung. Hierzu gehören u.a. Inhaltsvalidität, Konstruktvalidität und faktorielle Validität. Berg und Schubert (2019) haben Recht, dass diese Begriffe nicht mehr in allen professionellen Richtlinien so verwendet werden. Beispielsweise steht in den APA Standards Evidenz auf Basis des Testinhalts, Evidenz auf Basis des Antwortprozesses, Evidenz auf Basis der internen Struktur, Evidenz auf Basis der Beziehungen zu anderen Variablen. In

[1] Aus Gründen der Lesbarkeit ab hier nur noch als Richtlinien bezeichnet.

der DIN 33430 und in dem hier relevanten DIN Screen (Kersting, 2018) wird von Konstruktgültigkeit, Kriteriumsgültigkeit und Inhaltsgültigkeit gesprochen. Ähnliche Formulierungen finden sich auch im Qualitätsstandard der European Federation of Psychological Associations (EFPA Board of Assessment, 2013). Dabei ist die Validität eine Erkenntnis, die sich aus Theorie und empirischer Evidenz ergibt. "Validity refers to the degree to which evidence and theory support the interpretation of tests scores for proposed uses of the tests." (AERA APA & NCME Standards, 2014, S. 11). „Konstruktvalidität" ist ein Oberbegriff, von Validität sprechen wir, wenn intendierte Interpretationen der Messwerte durch theoretische und/oder empirische Belege gerechtfertigt werden können (Kane, 2013). Die „theoriegeleitete Validierung" ist eine von vielen Zugängen zur Validität, sie trägt zur Validierung bei, kann und soll aber keinesfalls ein Ersatz für andere Validierungsstrategien sein. Es lässt sich also festhalten, dass ein internationaler Konsens bezüglich verschiedener Validitätsaspekte und Validierungsstrategien besteht. Keiner der genannten Standards sieht vor, einen dieser Validitätsaspekte als hinreichend anzusehen. Vielmehr geht es darum, Evidenz aus verschiedenen Bereichen zu sammeln, die kumulativ und im Einklang mit der intendierten Testverwendung (Ziegler, 2014) die Testwertinterpretation stützt. Diese Qualitätsstandards setzen die Richtlinien direkt um, in Tabelle 3 der Richtlinien sind entsprechend verschiedene Evidenzquellen aufgelistet.
Die von Berg und Schubert (2019) hervorgehobene „theoriegeleitete Validierung", findet sich in der Tat in fast allen Qualitätsstandards[2]. Insofern spricht nichts dagegen, auf diese Validierungsstrategie hinzuweisen. Die Behauptung, dass dieser Aspekt in den aktuellen Richtlinien nicht auftauchen würde, widerspricht allerdings eindeutig den Tatsachen – es sei denn, man klammert sich allein an den Begriff / das Wort und verschließt sich dem eindeutigen Sinn der Richtlinien. In Tabelle 3 der Richtlinien wird unter dem Punkt Konstruktdefinition und Testkonstruktion explizit gefordert: „Es ist nachvollziehbar, wie die Items erzeugt wurden, um das Konstrukt zu operationalisieren (Cronbach, Meehl, 1955)". Der Verweis auf Cronbach und Meehl ist hier wichtig, da dort zu lesen ist, dass zu spezifizieren sei, wie sich Unterschiede im zu messenden Konstrukt in den Aufgaben manifestieren. Auf diese Quelle weisen auch Berg und Schubert (ebd.) hin, allerdings um die Richtlinien zu kritisieren, ignorieren dabei jedoch, dass auf eben genau diese Quelle und eben genau aus dem von ihnen angeführten Grund in den Richtlinien ebenfalls verwiesen wird.
Es lässt sich also festhalten, dass die Hauptkritik von Berg und Schubert (ebd.), das Vernachlässigen von „theoriegeleiteter Validierung" in den Richtlinien, sachlich und fachlich unbegründet ist.

Aktuelle Standards und die Bedeutung „theoriegeleiteter Validierung"

Zunächst ist anzumerken, dass der Begriff der „theoriegeleiteten Validierung", wie von Berg und Schubert (2019) treffenderweise ausgeführt wird, keine breite Akzeptanz erfahren hat. Auch hier ist aber zwischen dem Begriff und dem Inhalt, dem Sinn zu unterscheiden. Während sich der Begriff der „theoriegeleiteten Validierung" nicht etabliert hat, wurde die Idee, dass eine Theorie existieren soll, die den Antwortprozess aus Konstruktsicht erklärt, in die meisten Qualitätsstandards integriert. Berg und Schubert (ebd.) beziehen sich mit ihrer Kritik auf einen „historischen" Zustand. Borsboom, Mellenbergh und Van Heerden haben in ihrer Publikation von 2004 zu Recht die damalige Praxis bemängelt, bei der Testentwicklung und -bewertung den Antwortprozess und damit den kausalen Zusammenhang zwischen Konstrukt und Aufgabe (im Sinne von Berg und Schubert: „theoriegeleitete Validierung") zu vernachlässigen. Diese Kritik

[2] Wie von Berg und Schubert (2019) korrekt angeführt, finden sich diese Überlegungen auch in den aktuellen APA Standards (2014), nicht jedoch wie angegeben unter 1.8, sondern unter Standard 1.12 auf S. 26. Unter der von Berg und Schubert (ebd.) genannten Nummer (Standard 1.8) wurde der Topic in den „alten" Standards von 1999 abgehandelt. Die Autoren wechseln kontinuierlich die Referenz, so sprechen sie auf Seite 8 von der „deutschen Abfassung" der Standards und beziehen sich dabei auf die Publikation von Häcker, Leutner und Amelang (1998). Dabei handelt es sich um eine Übersetzung der 1985 herausgegebenen Version der Standards.

war vor 15 Jahren berechtigt, aktuelle Qualitätsstandards berücksichtigen diesen Aspekt aber ausdrücklich.

Die von Berg und Schubert (2019) weiterhin zumindest suggerierte Annahme, dass eine „theoriegeleitete Validierung" empirische Belege obsolet mache, entspricht jedoch in keinem Fall dem Kern der Arbeit von Borsboom, Mellenbergh et al. (2004). Diese betonen in der Tat die Rolle des Antwortprozesses, genauer die Wichtigkeit eines Modells, das erklärt, wie sich Unterschiede in der zu messenden Variable im Antwortverhalten wiederspiegeln und, dass dieses Modell bereits die Testkonstruktion beeinflussen solle.

S. 1068: „Fortunately, there are various recent developments in theoretically inspired modeling (Embretson, 1994; Jansen & van der Maas, 1997; Mislevy & Verhelst, 1990; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002; Wilhelm & Schulze, 2002) that show how much is gained when one starts to consider the processes involved in item response behavior and to utilize advanced test theory models that have been developed in the past century." (Borsboom, Mellenbergh et al., S. 1068).

Es wird aber eben nicht gesagt, dass es keine empirische Prüfung dieser Modelle geben muss. Die Autoren stellen die Theoriebildung sachlogisch an erster Stelle vor Statistik und Methodik, die aber keinesfalls entfallen, sondern an zweite Stelle gesetzt werden. Die von Berg und Schubert (2019) ins Feld geführten Autoren betonen ausdrücklich die Wichtigkeit einer empirischen Prüfung und nennen spezielle statistische Methoden, die sich dafür eignen:

S. 1067/68: "This does not mean that methodological and psychometric techniques are irrelevant to validation research but that the primary source for understanding how the test works must be substantive and not methodological."

Dass der Theorie gegenüber „blinder" Empirie bei der Validierung ein Primat einzuräumen ist, ist eine Aussage, die sich in vielen Qualitätsstandards zur Validität findet (Ziegler, Lämmle, 2017). Auch die Richtlinien zur Beurteilung von Testverfahren zur Fahreignung vertreten eindeutig diese etablierte Position. Hochwertige diagnostische Qualitätsstandards wie die Richtlinien begrenzen sich aber nicht auf einen Aspekt der Validität, sondern berücksichtigen zahlreiche Validitätsaspekte. Berg und Schubert (2019) behaupten nun mehr oder minder direkt, dass nur die Strategie der theoriegeleiteten Validierung zeitgemäß sei, andere Validierungsstrategien hingegen unzeitgemäß. Weiter oben haben wir bereits ausgeführt, dass die von Berg und Schubert (2019) kritisierten Validitätsaspekte und –strategien im Einklang mit den etablierten Qualitätsstandards stehen. Vor allem konvergente und diskriminate Validitätsbelege erlauben die Feststellung, ob ein Testwert im Sinne des angestrebten Konstrukts interpretiert werden kann oder eben eher im Sinne eines anderen Konstrukts (Wehner, Roemer et al., 2018). Hierzu ist, wie gefordert, eine starke Theorie notwendig. In den Richtlinien findet sich folgerichtig in Tabelle 3 bei konvergenter Validität die Forderung nach Korrelationen mit Testwerten von Tests mit *demselben Messanspruch*. Für die diskriminante Validität wird explizit auf *das nomologische Netz* verwiesen. Zudem wird ebenfalls auf korrelationsverzerrende Einflüsse hingewiesen. Für alle drei Aspekte ist eine starke Theorie notwendig, um a-priori Hypothesen bezüglich der Zusammenhänge aufstellen zu können bzw. diese korrekt interpretieren zu können. Die Kritik von Berg und Schubert (2019), dass der Richtlinienansatz sich auf empirische (zumeist korrelative) Belege fixiere und nicht theoriegeleitet und in Folge dessen nicht mehr zeitgemäß sei, ist gegenstandslos und zurückzuweisen[3].

[3] So ist auch der Hinweis von Berg und Schubwert (2019), die Richtlinien verwiesen nur auf Lehrbücher, die einen anderen als einen korrelativen Zugang zur Validität nicht erwähnen würden falsch. Bei der Quelle Ziegler and Hagemann (2015) handelt es sich zum einen um einen Artikel, zum anderen wird hier explizit ein experimenteller Zugang beschrieben.

Konsequenzen „theoriegeleiteter Validierung"

„Theoriegleitete Validierung" bedarf also neben einer starken Theorie auch empirischer Belege. Solche empirischen Belege lassen sich beispielsweise aus einer experimentellen Testvalidierung (Krumm, Hüffmeier, et al., 2017) oder aus spezifischen Modellierungsansätzen gewinnen. Diese Modellierungsansätze beruhen nicht selten auf Modellen der Probabilistischen Testtheorie und erfordern daher sehr viele Testpersonen, um aussagekräftige Ergebnisse zu erzielen. Experimentelle Prüfungen erfordern, so die verbreitete Annahme, geringere Stichprobenumfänge. Diese Annahme ist häufig unzutreffend. In den vergangenen Jahren wurden wir durch die so genannte „Replikationskrise in der Psychologie" daran erinnert, dass bei psychologischen Studien im Allgemeinen und bestimmten Experimenten im Besonderen das Risiko besteht, dass sich Ergebnisse nicht replizieren lassen (Klein et al., 2014). Dieser Umstand erfordert neben besserer theoretischer Fundierung vor allem auch eine a-priori Stichprobenplanung sowie Replikationsanstrengungen. In dem von Berg und Schubert (2019) aufgeführten Beispiel für einen empirischen Ansatz zur theoriegeleiteten Validierung (Berg, 2018) finden sich keine Angaben zu Stichprobengrößen. Ohne diese Angaben lässt sich die Qualität der Studie nicht beurteilen, dies ist ein Grund dafür, dass Qualitätsstandards eine detaillierte Beschreibung der empirischen Untersuchungen (inklusive Stichprobenbeschreibung) fordern. In der Quelle (Berg, 2018) wird ausgeführt, dass theoriegeleitete Unterschiede zwischen verschiedenen Aufgabentypen untersucht und erwartete Unterschiede gefunden wurden. Dabei beträgt der kleinste angegebene t-Wert -4.29. Auf Basis des zugehörigen p-Werts lässt sich schätzen, dass es mindestens 7 Personen pro Bedingung gewesen sein müssen, falls es sich um ein between Design handelt. Daraus lässt sich wiederum schätzen (siehe Abbildung 1), dass die Effektstärke d = -2.29 beträgt (bei einer Gesamtstichprobe von 200 wäre es immer noch d = -.61). Die größte geschätzte Effektstärke basierend auf den Angaben wäre 13.97 (bei einer Gesamtstichprobe von 200 immer noch 3.7). Derlei Effekte sind in der Psychologie eher ungewöhnlich und erfordern daher unbedingt eine Replikation. Nimmt man die geschätzte Effektstärke und erwartet eine robuste Replikation benötigt man mindestens 129 Personen (Schönbrodt & Perugini, 2013). Dies wäre dann aber nur die Prüfung eines Aufgabentyps. Nimmt man an, dass zur Lösung komplexer kognitiver Aufgaben verschiedene psychologische Prozesse verwandt werden (Carpenter, Just, et al., 1990), ist es nicht weit hergeholt, wenn man für eine „theoriegleietete Validierung" eine Vielzahl von Experimenten und somit einen sehr hohen Bedarf an Testpersonen ableitet. Andernfalls wäre es schwierig, die verschiedenen Aufgaben gemeinsamen Anteile von den spezifischen zu trennen. Dieser hohe Aufwand mag es sein, der dazu führt, dass viele Validierungsstudien den Ansatz scheuen.

Anonymität der Gutachter*innen

Schließlich beklagen Berg und Schubert, es sei nicht sinnvoll, die Gutachter*innen, welche anhand der Leitlinien die Verfahren beurteilen, anonym zu halten. In der Tat wird der Aspekt der Gutachteranonymität auch in der Wissenschaft immer wieder angeregt debattiert. Dennoch hat sich dieses sogenannte blind peer-review Verfahren als momentaner Standard in der Wissenschaft etabliert. Dies betrifft nicht nur die Begutachtung von Artikeln, sondern auch die Begutachtung von Forschungsanträgen. Dies ist jedoch sicher ein schwaches Argument, auch hier Anonymität wahren zu wollen. Dass dadurch die Unabhängigkeit der Gutacher*innen gewährleistet bleibt ist sicher unbestritten. Zudem sei angemerkt, dass gerade die von Berg und Schubert kritisierten Leitlinien dazu beitragen, das Verfahren insgesamt transparent zu machen und die Beurteilungsrichtlinien a priori klar zu machen. Daher ist die Sorge vor Intransparenz sicher unbegründet

Abschlussbemerkung

Die Autoren referenzieren die APA und AERA Standards, was völlig richtig und wichtig ist. Dort steht allerdings, dass es „die" Validität im Sinne des Ergebnisses einer einzigen Validierungsstrategie nicht gibt. Vielmehr wird unter Validieren die Ansammlung verschiedenster Evidenz verstanden. Welche Evidenz dies ist, wird in den Standards ebenfalls expliziert (S. 14):

„… each type of evidence … is not required in all settings. Rather, support is needed for each proposition that underlies a proposed test interpretation for a specified use".

Dies bedeutet nichts anderes als dass die Validitätsbelege zum intendierten Testzweck passen müssen. Warum ist es notwendig, verschiedene Evidenz zu sammeln? Berücksichtigt man die Überlappungen zwischen Konstrukten, vor allem im kognitiven Bereich, so ist anzunehmen, dass distinkte aber überlappende Konstrukte teilweise auf ähnlichen psychologischen Prozessen beruhen. Nur durch die Akkumulation von Evidenz kann es daher gelingen, gemeinsame und spezifische Konstruktanteile interpretativ zu trennen. Als Beispiel sollen ein verbaler und ein numerischer Test zum Schlussfolgernden Denken dienen. Die zugrundeliegenden psychologischen Prozesse werden sich stark ähneln. Für Aufgaben beider Tests ist ein hochgradig ähnliches Antwortmodell plausibel. Würde man nur „theoriegeleitet validieren", bestünde das Risiko, den starken Zusammenhang zwischen den beiden Tests zu verkennen. Erst durch eine konvergente und diskriminante Validierung lassen sich die gemeinsamen und spezifischen Testanteile interpretativ trennen.
Zwar schreiben auch Berg und Schubert (2019) zunächst, dass sie „theoriegeleitete Validierung" nicht zur einzig wahren Quelle ernennen wollen, kommen dann aber zum Schluss, es könne passieren, „dass für theoriegeleitet validierte Testverfahren, die wissenschaftlich hergeleitet und validiert sind und seit Jahren in der Praxis angewendet werden, die Bewertung unter dem Punkt Validierung lautet: "nicht vorhanden". … die Träger laufen Gefahr, ihre Anerkennung ab dem Jahre 2020 fachlich ungerechtfertigt aus formalen Gründen entzogen zu bekommen." (Berg, Schubert, 2019, S.9). Wie eben ausgeführt, fordern aktuelle Qualitätsstandards aus gutem Grund eine Akkumulation von Evidenz und eben nicht nur eine Art der Validitätsevidenz. Insofern wird die Schlussfolgerung der Autoren den eigenen Ansprüchen und Argumenten nicht gerecht. Abschließend lässt sich festhalten, dass die vorgebrachte Kritik zurückzuweisen ist. Der in den Richtlinien angewandte Validierungsansatz entspricht international anerkannten, aktuellen Qualitätsstandards und beinhaltet vollumfänglich den von Berg und Schubert (2019) geforderten Ansatz der theoriegeleiteten Validierung – auch wenn der Begriff / das Wort in den Richtlinien nicht genutzt wird. Die theoriegeleitete Validierung stellt einen wertvollen Zugang zur Qualitätssicherung dar, darf aber eben nicht der einzige Zugang sein. Ein Verfahren, welches nur diese Art des Validitätsbelegs bietet wird demnach völlig zurecht abgelehnt werden. Es ist zu hoffen, dass die angesetzten Richtlinien zu einer Qualitätssteigerung führen werden. Dies ist sicher im Sinne der Verbraucher, auch, wenn es nicht immer im Interesse der Testanbieter sein mag.

6. Literatur

AERA, APA, & NMCE. (2014). *Standards for Educational & Psychological Testing*. Washington, DC, USA.

Berg, M. (2018). Experimentelles Denken in der Diagnostik kognitiver Funktionen. *Sitzungsberichte der Leibniz-Sozietät der Wissenschaften zu Berlin, 135*, 165-178.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404-431.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

EFPA Board of Assessment. (2013). *EFPA REVIEW MODEL FOR THE DESCRIPTION AND EVALUATION OF PSYCHOLOGICAL AND EDUCATIONAL TESTS*. Retrieved from

Embretson, S. (1994). Applications of Cognitive Design Systems to Test Development. In C. R. Reynolds (Ed.), *Cognitive Assessment: A Multidisciplinary Perspective* (pp. 107-135). Boston, MA: Springer US.

Jansen, B. R., & van der Maas, H. L. J. (1997). Statistical tests of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321-357.

Kersting, M. (2018). Zur Information über und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens – Die DIN SCREEN Checkliste 1, Version 3. In Diagnostik und Testkuratorium (Ed.), *Personalauswahl kompetent gestalten*. Berlin: Springer.

Klein, R. A., Ratliff, K. A., Vianello, M. A. J., R. B., Bahník, S., Bernstein, M. J., …, & Nosek, B. A. (2014). Data from Investigating Variation in Replicability: A "Many Labs" Replication Project. *Journal of Open Psychology Data, 2*(1). doi:10.5334/jopd.ad

Krumm, S., Hüffmeier, J., & Lievens, F. (2017). Experimental Test Validation. *European Journal of Psychological Assessment*, 1-8. doi:10.1027/1015-5759/a000393

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195-215. doi:10.1007/bf02295283

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of research in personality, 47*(5), 609-612. doi:10.1016/j.jrp.2013.05.009

Süß, H. M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability--and a little bit more. *Intelligence, 30*(3), 261-288.

Wehner, C., Roemer, L., & Ziegler, M. (2018). Construct Validity. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1-3). Cham: Springer International Publishing.

Wilhelm, O., & Schulze, R. (2002). The relation of speeded and unspeeded reasoning with mental speed. *Intelligence, 30*(6), 537-554.

Ziegler, M. (2014). Stop and state your intentions!: Let's not forget the ABC of test construction. *European Journal of Psychological Assessment, 30*(4), 239-242. doi:10.1027/1015-5759/a000228

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items: Pitfalls and loopholes. *European Journal of Psychological Assessment, 31*(4), 231-237. doi:10.1027/1015-5759/a000309

Ziegler, M., & Lämmle, L. (2017). Validity. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1-7). Cham: Springer International Publishing.

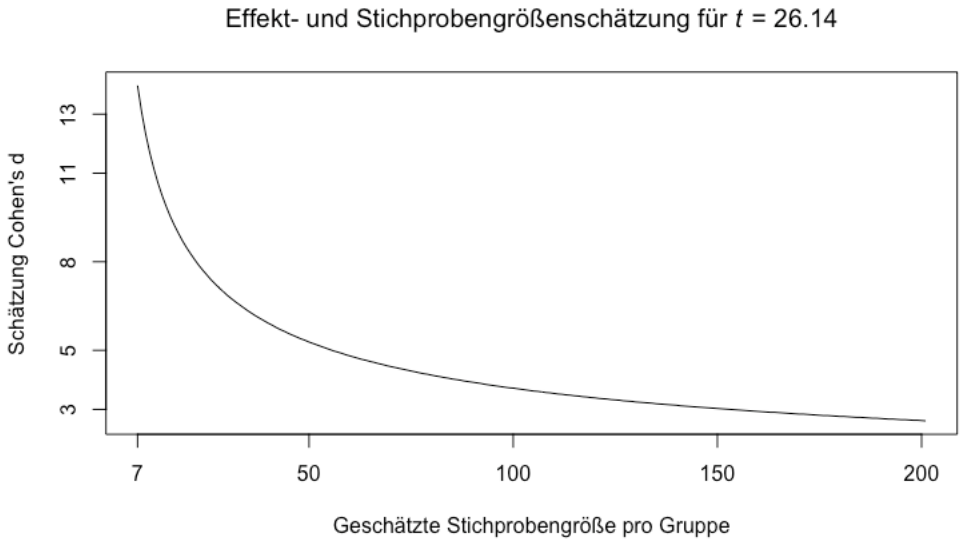

Abbildung 1. Schätzungen der Effekt- und Stichprobengröße aus Angaben von Berg (2018). Alle Schätzungen sind hier https://osf.io/p5zfk/?view_only=650a5229fd4c48b5af8531fd012b3303 dokumentiert.

Hinweis: Bitte zitieren Sie diesen Text wie folgt:

Diagnostik- und Testkuratorium (DTK) (2019). Stellungnahme zu Berg und Schubert (bezüglich der Richtlinien zur Bewertung von Testverfahren und -geräten zur Wiederherstellung der Kraftfahreignung). Zeitschrift für Verkehrssicherheit, 3/2019, 199-202. (DTK: C. Hagemeister, M. Kersting (Vorsitzender), F. Lang, N. Stenzel, K.-O. Tietze und M. Ziegler). Berlin: Föderation Deutscher Psychologenvereinigungen.