Somnologie 2012
DOI 10.1007/s11818-012-0558-9
Eingegangen: 15. Dezember 2011
Angenommen: 5. April 2012
© Springer-Verlag 2012

S. Schnieder[1] · J. Krajewski[1] · T. Esch[2] · B. Baluch[3] · B. Wilhelm[4]

[1] Experimentelle Wirtschaftspsychologie, Universität Wuppertal

[2] College at Old Westbury, Neuroscience Research Institute, State University of New York, Coburg University of Applied Sciences Division of Integrative Health Promotion, Coburg

[3] Department of Psychology, School of Health and Social Sciences, Middlesex University, London

[4] STZ eyetrial am Department für Augenheilkunde, Universität Tübingen

Nur valide oder auch akkurat?

Provisorische Bestimmung der Messgenauigkeit des pupillographischen Schläfrigkeitstests mithilfe von Selbst- und Fremdratings

Schläfrigkeit ist verantwortlich für etwa 10–20% aller tödlichen Unfälle im Straßenverkehr und besitzt somit eine hohe gesellschaftliche und volkswirtschaftliche Relevanz. Diese Public-Health-Perspektive ist jedoch nicht mit Fragestellungen der Fahrtauglichkeit zu verwechseln, also der Bestimmung des individuellen Risikos, unter hoher Schläfrigkeit tatsächlich auch zu verunglücken. Belastbare Quantifizierungen dieser individuellen Risikoperspektive konnten derzeitig noch nicht ermittelt werden. Erste Hinweise auf die zu erwartenden Größenordnungen des Risikos, im Zuge einer einzelnen Risikoexposition (z. B. eine 1 km lange Autobahnfahrt unter schwerer Schläfrigkeit) tödlich zu verunglücken, liefert die Grundrate, im deutschen Straßenverkehr überhaupt tödlich zu verunglücken: pro 200 Mio. gefahrener Kilometer ein tödlicher Unfall. Multipliziert man somit dieses extrem unwahrscheinliche Ereignis mit der relativen Risikosteigerung unter schwerer Schläfrigkeit von etwa 185 [12], lässt sich die voraussichtliche Größenordnung der Wahrscheinlichkeit eines Schläfrigkeitsunfalls abschätzen: nach durchschnittlich etwa 1 Mio. gefahrenen Kilometern ein tödlicher Unfall unter schwerer Schläfrigkeit. Erst die jährlichen 1000 Mrd. gefahrenen Kilometer in Deutschland summieren die kleinen individuellen Wahrscheinlichkeiten zu einer großen gesellschaftlich relevanten Zahl von Schläfrigkeitsunfällen [7, 9, 20]. Der Entwicklung messgenauer und verfälschungsresistenter Messinstrumente zur Erfassung von sicherheitskritischen Schläfrigkeitszuständen kommt daher aus gesellschaftlicher (Kosten-)Perspektive eine hohe Relevanz zu [3].

Ein vielversprechendes Instrument zur vorrangigen Ermittlung der momentanen Schläfrigkeit und somit beeinträchtigten Fahreignung stellt der Pupillographische Schläfrigkeitstest (PST) dar [43]. Die Erfassung der für die Diagnose von Schlafstörungen interessanten gemittelten Tagesschläfrigkeit [34] ist zwar denkbar, soll in diesem Beitrag jedoch nicht weiter verfolgt werden. Das physiologische Grundprinzip des PST basiert darauf, dass die Pupillenweite durch das vegetative Nervensystem gesteuert wird und in Dunkelheit einer ausschließlich sympathischen Regulierung unterliegt. Im Wachen und bei hoher zentralnervöser Aktivierung bleibt die Pupille unter Ausschluss von Lichteinfluss über lange Zeit stabil. Bei erhöhter Schläfrigkeit hingegen treten bereits nach wenigen Minuten deutliche Schwankungen der Pupillenweite auf, die durch eine Instabilität der sympathischen Hemmung der parasympathischen Neurone im Erdinger-Westphal-Kern verursacht werden. Die Hemmung geht vom noradrenergen Kerngebeit A1-A5 im Hirnstamm aus, Fasern verlaufen von dort auch zum Locus Coeruleus (LC; [38, 39, 41, 45]), dessen neuronale Feuerrate im engen Zusammenhang zu Schwankungen der Pupillendurchmessers steht [1]. Quantifiziert wird diese Instabilität über Kenngrößen, die die Instabilität der Pupillenweite und das Ausmaß niederfrequenter Oszillationen quantifizieren [19, 38].

Häufigkeitsbezogene Normwerte des PST [18, 39, 40] existieren ebenso wie Aussagen zur Reliabilität (Test-Restest-Reliabilität: 0,64; [18]). Die Validität des PST wurde durch eine Reihe von Validierungskriterien bestätigt:

- Medikation mit (De-)Aktivierungsinduktion [10, 11, 30],
- mit Schläfrigkeit assoziierte, klinische Störungsbilder [43],
- zirkadiane Verläufe [28, 39, 42],
- physiologische Parameter (z. B. Elektroenzephalographie, EEG; [24]),
- Multipler Schlaflatenz-Test (MSLT; [5, 15, 22, 23, 26, 33, 35, 39]) und
- subjektive Schläfrigkeitsskalen [19, 28, 38, 39, 40].

Trotz des Nachweises der generellen Validität steht die Bestimmung der für die Einzelfalldiagnostik zentralen Messgenauigkeit (Akkuratheit) des PST noch aus. Die Akkuratheit beantwortet die wesentlich bedeutsamere Frage: Wie groß ist die Unsicherheit der Messung, also die zu erwartende Abweichung eines Messwerts von einem als wahren Wert zu interpretierenden Goldstandard. Diese kann nur aus dem Vergleich von PST-Kennwerten mit

eindeutigen Referenzwerten der Schläfrigkeit („Ground-Truth-Werten") erfolgen, nicht aber über den Abgleich mit nichtschläfrigkeitsskalierten und an sich unscharfen Validierungskriterien.

Das in medizintechnischen Kontexten übliche Konzept der Ground-Truth-Referenzwerte stellt anders als das Konzept der Validierungskriterien schärfere Anforderungen an Referenzwerte. So werden als Validierungskriterien auch Maßzahlen anerkannt, die keine eindeutige Skalierung der Schläfrigkeit besitzen, wie z. B. spektrale EEG-Amplitudendichten oder Reaktionszeiten des Psychomotorischen Vigilanztests. Überdeckt werden die Unzulänglichkeiten von nicht adäquat skalierten Validierungskriterien über die Nutzung von Zusammenhangsmaßen in der Ergebnisdarstellung. Durch diesen Kunstgriff verliert sich das eigentliche Ziel der Studien, die Messgenauigkeit eines Verfahrens (d. h. um wieviel Prozent liegt ein berechneter Schläfrigkeitsmesswert von dem tatsächlichen wahren Schläfrigkeitswert entfernt) zu quantifizieren, schnell aus den Augen. Um diese Frage zu beantworten, sind somit reine Validierungskriterien unzureichend und Ground-Truth-Referenzwerte unabdingbar. Die hier skizzierte Problematik der bisher für den PST zum Einsatz gebrachten Validierungskriterien wird im Folgenden näher erläutert, um den Entwicklungsbedarf von Alternativkriterien im Sinne von Ground-Truth-Werten für den PST zu begründen.

Im Validierungsbereich der Medikation mit Substanzen zur Senkung oder Steigerung des zentralnervösen Aktivierungsniveaus können keine feinaufgelösten numerische Zuordnungen von Medikamentendosen zu Schläfrigkeitszuständen getroffen werden. Als abgesichert gilt lediglich der bei steigender Dosis (beginnend bei einem individuellen Ausgangswert) monoton steigende bzw. sinkende Schläfrigkeitsgrad. Diese Validierungsansätze können daher, ähnlich den mit Schläfrigkeit assoziierten klinischen Störungsbilderbildern (vgl. obstruktives Schlafapnoe-Syndrom, Narkolepsie, [43]), eben nur ordinale Informationen zum tatsächlichen Schläfrigkeitszustand liefern. Vergleichbare Nachteile weisen auch Validierungsansätze im zirkadianen Schlafentzugsmodell auf. Weder die Tageszeit noch die Dauer der Schlafdeprivation erlauben eine hinreichend präzise Ableitung der aktuellen individuellen Schläfrigkeit.

Analog bieten physiologische Validierungsansätze (z. B. EOG- oder EEG-basiert) bislang ebenfalls – u. a. aufgrund ihrer großen interindividuellen Variabilität – keinen ausreichend abgesicherten wahren Referenzwerte (Ground-Truth-Wert) der Schläfrigkeit. Trotz der dokumentierten Zusammenhänge von EEG- (z. B. Theta-Aktivität) mit (wiederum selbst unscharfen) Schläfrigkeitsvalidierungskriterien lässt sich konstatieren, dass EEG-Indikatoren (bzw. die multivariate und mustererkennungsgestützte Fusion vieler Einzelindikatoren, [21, 32]) zwar ausgesprochen vielversprechend, aber zum gegenwärtigen Zeitpunkt noch nicht als Referenzwerte für Schläfrigkeit geeignet sind. Begründet ist dieses darin, dass keine aus hinreichend großen Normstichproben abgeleitete Schläfrigkeitsskalierung der schläfrigkeitssensitiven EEG-Parameter vorliegt (z. B. eine erfasste spektrale Theta-Amplitudendichte von 15 dB entspricht einem Schläfrigkeitskonfidenzintervall von 7,1–7,8).

Eine ähnliche Situation findet sich für die Nutzung von MSLT-Kennwerten. Obgleich der MSLT in der Narkolepsiediagnostik als valides und in der Schlafmedizin aufgrund seiner Augenscheinvalidität als anerkanntes Verfahren gilt, wird die Validität des Verfahrens aufgrund folgender Schwierigkeiten kritisiert [16, 33, 44]: einerseits fehlende Normwerte, die populationsbezogene Häufigkeiten von MLST-Messwerten dokumentieren sowie andererseits voraussichtlich zwischen Probanden und auch innerhalb eines Probanden variierende Zuordnungsregeln von Einschlafzeiten zu Schläfrigkeitswerten. Auch wenn der MSLT teilweise als das beste unter den verfügbaren Verfahren („Goldstandard") eingeschätzt wird, erscheint es derzeitig aufgrund der genannten Argumente und insbesondere der fehlenden Zuordnungsregeln von Einschlafzeiten zu Schläfrigkeitswerten nicht als Ground-Truth-Wert einsetzbar. Zudem stößt der MSLT in Forschungsdesigns, die an der Erfassung schneller Fluktuationen von Schläfrigkeit interessiert sind, schnell an seine Grenzen, da hier seine Durchführungsdauer und Inkompatibilität mit anderen Tätigkeiten zu einer niedrigen zeitlichen Auflösung der Messungen führen.

Selbsteinschätzungsskalen hingegen besitzen unabhängig von ihrer Messgenauigkeit den entscheidenden Vorteil einer ihnen bereits immanenten Schläfrigkeitsskalierung. Die in allen sonstigen Verfahren nötige Übersetzung der Verfahrenskennwerte in Schläfrigkeitswerte kann bei Selbsteinschätzungsskalen wie der Karolinska Sleepiness Scale (KSS; [29]) entfallen. Entscheidend für die Messgenauigkeit ist hier einerseits ein hohes Maß an Sorgfalt, mit der die Introspektion durchgeführt wird und andererseits eine Anwendungssituation, die eine geringe Verfälschungsmotivation induziert [8]. Daher kann zunächst nur im Fall von reinen experimentellen Forschungssituationen von einer sinnvollen Anwendung der Selbstreporte ausgegangen werden. Zusätzlich zu den genannten Problemen können Schlafstörungen und Schlaferkrankungen zu einer Gefährdung der KSS-Validität führen und somit den Einsatz in schlafmedizinisch-diagnostischen Kontexten erschweren. Neben diesen Rahmenbedingungen reduziert insbesondere die fragwürdige Reliabilität der 1-Item-Skalen die Beurteilung dieser Verfahrensklasse als Ground-Truth-Wert.

Anders als die Selbstbeobachtungsskalen können Fremdbeobachtungsskalen über multiple Ratings diese Reliabilitätsprobleme überwinden [6, 25, 36, 37]. Ein wesentlicher Teil der schläfrigkeitsbedingten Veränderungen des autonomen Nervensystems (z. B. längere Lidschlüsse, verlangsamte Atmung) sowie der kognitiven, motivationalen und emotionalen Beeinträchtigungen (z. B. Selbststimulationsverhalten wie das Reiben des Gesichts, emotionale Verflachung über ausdrucksarme Mimik) wird in diesen Verfahren zur Abschätzung des Schläfrigkeitsgrads genutzt. Aus diesem Grund wurde das über eine gute Intra- und Interraterübereinstimmung verfügende, gut validierte Observer Rating of Drowsiness Protocol (ORD) angewendet ([36]; s. Anhang: **Tab. 3**).

Erhöht man die Anzahl der unabhängigen Messungen mit unsystematischem Fehler und gleichem wahren Wert z. B. durch mehr Fremdbeobachter oder einem zusätzlichen Selbstrating, erhöht sich die

Reliabilität einer Messung. Die Kombination von jeweils einzeln validierten, als Messwiederholungen zu interpretierenden Instrumenten, wie dem Selbst- und Fremdrating, ist aus dieser Perspektive naheliegend. Zusätzlich erschließt die Aufnahme von validierten Selbstreporten zu validierten Fremdeinschätzungen die Möglichkeit, auch nichtbeobachtbare, phänomenale Erlebenskomponenten zu integrieren. Um die Fremd- mit den Selbstbeobachtungsratings unter Beibehaltung einer aussagekräftigen Skalierung fusionieren zu können, ist eine deckungsgleiche Skalierung erforderlich [12]. Zu diesem Zweck und um eine Kompatibilität mit der (innerhalb der an Zustandsmonitoring und weniger an gemittelten Tagesschläfrigkeit interessierten Schläfrigkeitsforschung) populären KSS-Schläfrigkeitsskala zu gewährleisten, wird die ORD mit ihrem Wertebereich von 0–100 linear auf den KSS Wertebereich von 1–9 transformiert.

Ausgehend von den benannten Einschränkungen der bisher vorgestellten Validierungskriterien, v. a. dem Fehlen einer direkten Zuordnung von Messkennwerten (z. B. MSLT-, EOG-, EEG-Kennwerte) zu skalierten Schläfrigkeitswerten, sollen in der vorliegenden Studie in erster Annäherung durch die Fusion von Selbst- und Fremdratings Quasi-Ground-Truth-Werte konstruiert werden (vgl. z. B. [12]). Mithilfe dieser Quasi-Ground-Truth-Werte erfolgt die Bestimmung der Validität und auch Messgenauigkeit des PST. Erwartet werden analog zu den bisherigen Validierungsergebnissen positive Zusammenhänge zwischen Schläfrigkeitsratings und Pupillenstabilität-Standardkennwerten des PST.

Methode

Probanden

Die Teilnehmer dieser Studie setzten sich aus 30 gesunden Studenten (18 Frauen, 12 Männer; Altersdurchschnitt: 26,2 Jahre, Standardabweichung, SD: 2,2 Jahre) der Bergischen Universität Wuppertal zusammen. Die Versuchspersonen wurden zunächst über den Hintergrund und das Untersuchungsprozedere mündlich aufgeklärt und gaben eine schriftliche Ein-

Zusammenfassung · Abstract

Somnologie 2012 · [jvn]:[afp]–[alp] DOI 10.1007/s11818-012-0558-9
© Springer-Verlag 2012

S. Schnieder · J. Krajewski · T. Esch · B. Baluch · B. Wilhelm

Nur valide oder auch akkurat?. Provisorische Bestimmung der Messgenauigkeit des pupillographischen Schläfrigkeitstests mithilfe von Selbst- und Fremdratings

Zusammenfassung

Fragestellung. Ziel der vorliegenden Studie ist es, neben der Validität auch provisorisch die Messgenauigkeit des pupillographischen Schläfrigkeitstests (PST) sowohl über die Anwendung von Selbst- und Beobachterratings als auch über die aus ihnen fusionierten Werte zu bestimmen.
Methode. Mit 30 gesunden Frauen und Männern wurden zu jeweils 4 Messzeitpunkten in einer partiellen Schlafdeprivationsstudie (20.00–04.00 Uhr) insgesamt 113 PST-Messungen durchgeführt. Unmittelbar zuvor wurden ein Selbstreport (Karolinska Sleepiness Scale, KSS) und fünf videobasierte Fremdratings (Observer Rating of Drowsiness, ORD) erfasst.
Ergebnisse. Es ließen sich moderate Übereinstimmungen zwischen PST-Parametern (Pupillen-Unruhe-Index, Amplitudenspektrum) und den Validierungskriterien KSS und ORD nachweisen. Die mittels der Fusion aus KSS-Selbstreporten und Fremdbeobachtungen fusionierten gewonnenen Referenzwerte zeigten für die PST Parameter eine Korrelation von r=0,54; einen mittleren Fehler von 1,58 KSS-Punkte und einen prozentualen Fehler von 35%.
Schlussfolgerung. Die Ergebnisse stützen die Annahme der moderaten Validität und auch provisorisch die der Akkuratheit des PST. Zusätzlich legen sie nahe, dass die hier vorgeschlagene Fusionierung von Fremd- mit Selbstratings (Multiples-Rating-Ansatz) möglicherweise eine pragmatisch-effiziente Zwischenlösung zur Schätzung von abgesicherten Referenzwerten der Schläfrigkeit im Sinne eines Quasi-Ground-Truth-Kriteriums darstellen. Diese Eignung gilt voraussichtlich insbesondere für Studiendesigns, die auf die Erfassung von zeitlich feinaufgelösten Schläfrigkeitsverläufen fokussieren. Ferner deuten die Ergebnisse darauf hin, die Grenzen der bisherigen PST-Schläfrigkeitkategorien kritisch überprüfen zu lassen.

Schlüsselw örter
Pupillographischer Schläfrigkeitstest · Validität · Quasi-Ground-Truth · Schläfrigkeit · Multiples-Rating-Ansatz

Just valid or even accurate?. Determine the measurement accuracy of the pupillographic sleepiness test by applying self and observer ratings

Abstract

Objective. The purpose of the present study was to provide validation and accuracy data for the pupillographic sleepiness test (PST), on the one hand, by applying self and observer ratings and, on the other hand, by fusioned self and observer ratings as a sleepiness reference value.
Methods. A total of 30 healthy women and men participated in a partial sleep deprivation study (20.00–04.00 h) and PST measurements were conducted every 2 h for a total of 113 PST measurements. Karolinska Sleepiness Scale (KSS)-based self-reports and five video-based observer ratings of drowsiness (ORD) were measured immediately before the PST in order to provide reliable reference sleepiness values.
Results. PST parameters (pupil unrest index, power of frequency) correlated significantly with the sleepiness validation criteria the KSS and ORD used in this study. Fusioned reference values obtained from one self-report and observer ratings showed a correlation of r=0.54, a mean absolute percentage error of 1.58 KSS points, and an error of 35%.
Conclusion. Our results indicate the moderate validity of the PST. Furthermore, the proposed sleepiness reference value might serve as a feasible intermediate solution to estimate sleepiness in the sense of a reference ("quasi-ground truth") value. This might be true especially for within-subject designs with a focus on the time course of sleepiness. Moreover, the results might show the necessity to recalculate the thresholds of the current PST categories of sleepiness severity.

Keywords
Pupillographic Sleepiness Test · Validity of results · Quasi-Ground-Truth · Drowsiness · Multi-Rating Validation

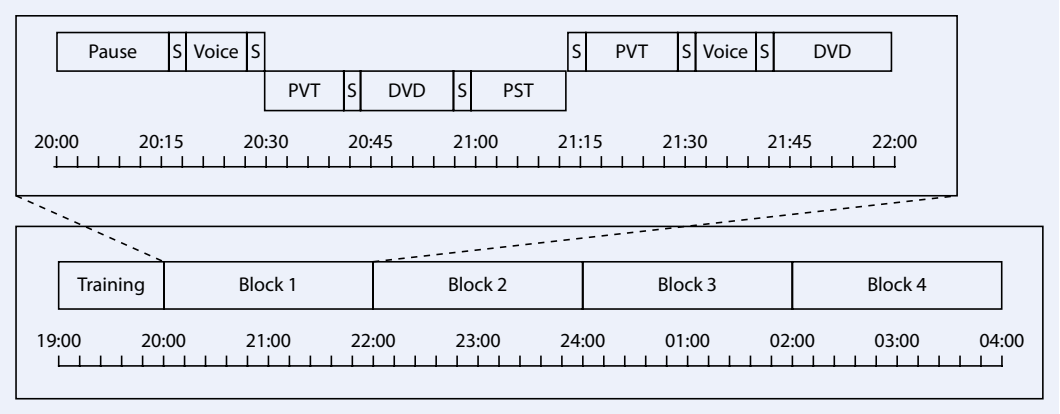

Abb. 1 ▲ Ablauf der Untersuchung. Der 2-stündige Block 1, z. B., enthält die in dieser Untersuchung genutzten Instrumente: Psychomotorischer Vigilanztest (*PVT*), Karolinska Sleepiness Scale (*S*), DVD mit Anschauen einer Autobahnfahrt (*DVD*), pupillographischer Schläfrigkeitstest (*PST*). In diesem Beitrag wird nicht weiter eingegangen auf die in der oberen Zeitleiste dargestellten Untersuchungsteile, wie z. B. Sprachaufzeichnungen (*Voice*)

verständniserklärung ab. Im Einklang mit den Verfahrensregeln zur Durchführung von psychologischen Experimenten wurde aufgrund der vollständigen Aufklärung der Probanden und der für die Probanden vorhersehbaren, nicht über die Alltagserfahrungen hinausgehenden Zumutungen kein weiteres Ethikvotum eingeholt. Als Inklusionskriterien wurde das Fehlen von jeglicher Medikation sowie eine Alkohol-, Nikotin- und Koffeinkarenz von 12 h vor Beginn der Tests angewendet. Ebenso wurden anhand eines Schlaftagebuchs Hinweise auf Insomnie, Ein- und Durchschlafstörungen und Schlafapnoe sowie über den Pittsburgh-Schlafqualitätsindex (PSQI; Cut-off des PSQI: >5; [2]) Hinweise auf eine reduzierte Schlafqualität als Ausschlusskriterien angewendet, um die möglicherweise verschlechterte Genauigkeit von Selbsteinschätzungen zu umgehen. Aufgrund der vorgenannten Kriterien wurden keine Teilnehmer ausgeschlossen.

Experimentelle Bedingungen

Einem partiellen Schlafdeprivationsparadigma (Untersuchungszeitraum 20.00–04.00 h; zum Ende der Untersuchung sind die Probanden seit 22 h wach) und der Forschungslinie des Schläfrigkeits-Zustandsmonitorings folgend, wurde in der vorliegenden Studie ein Messwiederholungsdesign verwendet. Pro Versuchsnacht durchliefen jeweils 2 Probanden das Untersuchungsprotokoll. PST-Daten wurden über die Infrarot-Video-Pupillographie alle 2 h über 4 Messzeitpunkte hinweg über eine Dauer von jeweils 11 min erhoben (21:00, 23:00, 01:00, 3:00). Insgesamt ergeben sich aufgrund vereinzelter technischer Messprobleme 113 isolierte PST-Anwendungen.

Zehn Minuten vor der PST-Messung wurden die Teilnehmer beim Anschauen eines 10-minütigen DVD-Videos (monotone Autobahnfahrt mit niedriger Verkehrsdichte) mit einer Videokamera aufgenommen. Die 10 je einminütigen Aufnahme-Epochen der Teilnehmer wurden im Nachhinein durch 5 Rater anhand des ORD (eines an Wierwille und Muttray [25] orientierten Vorgehens) beurteilt. Um den Originalverfahren-Skalenbereich von 0–100 in Deckung mit der etablierten KSS zu bringen, wurden im Nachhinein folgende auf einer linearen Transformation (KSS = (ORD x 8/100) + 1) basierende Zuordnungen der intervallskalierten KSS auf die Ratingskalen getroffen: KSS 1 = 0–6,25, KSS 2 = 6,25–18,75, KSS 3 = 18,75–31,25, KSS 4 = 31,25–43,75, KSS 5 = 43,75–56,25, KSS 6 = 56,25–68,75, KSS 7 = 68,75–81,25, KSS 8 = 81,25–93,75, KSS 9 = 93,75–100.

Eine Rating-Session der ORD beinhaltet zunächst die Sichtung von allen vier 10-min-Aufnahmen eines Probanden in doppelter Darstellungsgeschwindigkeit, um der Verschiedenartigkeit der individuellen Schläfrigkeitsreaktionen der Probanden gerecht zu werden. Anschließend wurden bei Normalgeschwindigkeit viermal 10 einminütige Epochen einzeln beurteilt. Die 5 Rater (jeweils mit Ausbildungshintergrund in Psychologie) konnten sich die Aufnahmen beliebig häufig anschauen, in den Videos zu beliebigen Stellen springen und ihre Urteile laufend korrigieren. Für 40 min zu beurteilendes Material ergab sich so Gesamtdauer der Rating-Session von etwa 90 min. Eine Verblindung der Rater lag vor, da sie keinen Einblick in die Selbstreporte und Ergebnisse der PST-Untersuchung hatten.

Um eine größtmögliche Äquivalenz des Schläfrigkeitszustands während der PST-Messung mit seinen Referenzwertmessungen zu erhalten, ist sowohl auf eine möglichst große zeitliche Nähe als auch auf einen vergleichbaren Grad der äußeren Stimulation geachtet worden. Daher erfolgte die DVD-Situation (und die ORD Messungen) in einem reizarmen, monotonen, mit der PST-Messung identischen räumlichen Setting. Aus dem gleichen Grund wurden die KSS-Selbsteinschätzungen jeweils unmittelbar vor den pupillographischen Messungen der Teilnehmer erfasst. Vor den Videoaufzeichnungen wiederum wurde ein 10-minütiger Psychomotorischer Vigilanztest (PVT) durchgeführt (**Abb. 1**).

Die Untersuchung fand in einem ruhigen, abgedunkelten Raum der Bergischen Universität Wuppertal statt. Im Fall überlanger Lidschlüsse (>10 s) und der nach Verhaltensbeobachtung durch den Untersucher abgeleiteten Einschlafereignisse wurde entsprechend der PST-Standards ein gestuftes Vorgehen mit Weckreizen praktiziert, um vollständige Messaufzeichnungen zu gewährleisten. [38,

39]. Dieses beinhaltete bei sehr starker Einschlaftendenz das Übermitteln eines Tonsignals und im Fall des Einschlafens ein leises persönliches Ansprechen. Dieses Prozedere musste jedoch bei keinem Probanden eingeleitet werden. Der Kopf der Probanden wurde durch eine Kinn- und Stirnstütze fixiert, durch eine Brille mit IR-Filter-Gläsern sahen die Versuchspersonen nur einen kleinen schwach leuchtenden Fixationspunkt (circa 0,8 m zum Probanden projiziert). Die Teilnehmer wurden instruiert, ruhig zu sitzen und diesen Punkt zu fixieren. Zwischen den einzelnen Untersuchungsdurchläufen konnten die Studienteilnehmer miteinander reden, lesen oder leichte Mahlzeiten und Getränke zu sich nehmen.

PST-Auswerteparameter

Die Aufzeichnung des spontanen Pupillenverhaltens erfolgte mittels Infrarotvideo-Pupillograph (PST, Fa. AMTech GmbH, Dossenheim). Mittels Infrarottechnologie wurde die Pupille mit einer Videokamera gescannt und die Rohdaten automatisch artefaktbereinigt, um fehlerhafte Signale zu eliminieren. Diese Signale wurden durch lineare Interpolation korrigiert. Die primären, die Pupilleninstabilität quantifizierenden PST-Parameter sind der Pupillenunruhe-Index (PUI; Angabe in mm/min) und das Amplitudenspektrum (Frequenzbereich: 0–0,8 Hz). Als Maß für das spontane Pupillenverhalten in Dunkelheit fungierte der PUI, der als die Summe der absoluten Änderungen des Pupillendurchmessers im Dunkeln, dividiert durch die Zeit, definiert ist. Die langsamen Pupillenoszillationen im Amplitudenspektrum wurden basierend auf einer Frequenzanalyse mittels Fast-Fourier-Transformation (FFT) berechnet. Hierbei wird die Datensequenz in 8 Abschnitte mit jeweils 2048 Werten unterteilt, wobei ein Abschnitt 82,92 s entspricht. In jedem Zeitfenster wird der Mittelwert des Pupillendurchmessers berechnet, von jedem Messwert subtrahiert und die FFT durchgeführt [19].

Statistische Analyse

Die Korrelationen zwischen den PST-Daten (PUI, Amplitudenspektrum) und den selbst und fremd angegebenen Daten wurden nichtparametrisch nach Spearman berechnet und hinsichtlich ihrer Signifikanz (einseitige Testung wegen der nichtexplorativen, aufgrund zahlreicher Untersuchungen eindeutig spezifizierten Zusammenhangsrichtung von PST-Parametern und Schläfrigkeit; Signifikanzschwelle, $p < 0,05$; signifikante Ergebnisse werden mit * markiert) bewertet.

Der intendierte Forschungsansatz ist an der Bestimmung der Akkuratheit einzelner PST-Messungen interessiert. Daraus ergäbe sich eine in die Signifikanzbestimmung einfließende Samplegröße von $n = 113$. Um jedoch der weitverbreiteten Perspektive einer an der Tagesschläfrigkeitsdiagnose von einzelnen Patienten interessierten Leserschaft entgegenzukommen (die mit einer Mittelung einzelner PST-Messungen arbeitet), wird für die Berechnung des Signifikanzniveaus ein Stichprobenumfang von $N = 30$ Personen zugrunde gelegt.

Ferner werden aus dem Universum potenzieller Verrechnungsmodi von Selbst- und Fremdratings zur Bestimmung des Schläfrigkeits-Gesamtscores zwei besonders plausible, prototypische Vertreter vorgestellt und empirisch überprüft:

- die frühe Fusion (unterscheidet nicht zwischen Selbst- und Fremdratings, sodass die Selbsteinschätzungen zu je 16,7% in den Gesamtscore eingehen) und
- späte Fusion (zunächst werden die 5 Fremdratings gemittelt, anschließend erst werden sie mit dem Selbstrating, das hier 50% des Gesamtscores ausmacht, verrechnet).

Des Weiteren wurden die Rater-Übereinstimmungen über die Intraclass-Korrelation und die mittlere absolute Abweichung („mean absolute error", MAE) bestimmt, indem das von einem Rater zu einer Videoepoche abgegebene Urteil von dem Gesamtmittelwert der Urteile dieser Videoepoche abgezogen wird. Anschließend wurde auch für alle anderen Ratings dieser aktuellen Videoepoche ein Abweichungswert bestimmt und alle Abweichungswerte dieser Videoepoche gemittelt. Dieses Verfahren wurde unabhängig voneinander auf alle Videoepochen angewendet und schließlich über eine Mittelung die finale MAE bestimmt. Anders als andere Übereinstimmungsmaße (z. B. auch Kappa oder Korrelationskoeffizienten) ermöglicht die MAE eine unmittelbare, geradlinige und transparente Abschätzung der Übereinstimmung in leicht nachvollziehbaren Einheiten, wie sie auch bei der Bestimmung der PST-Messgenauigkeit wünschenswert ist (hier: mittlerer Abstand der PST-Messwerte von der die Schläfrigkeitsreferenz repräsentierenden Regressionsgeraden; **Abb. 2, 3**). Im Einklang mit der Intention, leicht zu interpretierende Kennwerte der Messgenauigkeit anzubieten, wurde aufbauend auf den MAE neben der absoluten auch die mittlere prozentuelle Abweichung („mean absolute percentage error", MAPE) der über den PST erfassten Schläfrigkeit von den jeweiligen ihnen zugeordneten Referenzwerten ermittelt (z. B. beträgt bei einer PUI = 6 und einer zugeordneten KSS = 5 der Fehler 1 KSS-Punkt, was bezogen auf den Referenzwert 1/5, also 20%, entspricht).

Ergebnisse

Gütekriterien des ORD

Vorab wurde die Reliabilität der ORD über einen Rater-Übereinstimmungs-Ansatz ermittelt (Intraclass-Korrelation, $r = 0,89^*$; MAE = 0,56). Die Retestreliabilität wurde über eine erneute Anwendung des ORD nach einer Woche bestimmt und beträgt für die ermittelten Gesamtscores $r = 0,94^*$ und MAE = 0,52. Hinweise für die Validität des ORD liefern die Zusammenhänge zur Selbstbeobachtungs-KSS ($r = 0,88^*$) und auch zur mittleren Reaktionszeit des PVT ($r = 0,48^*$).

Pupilleninstabilitätskennwerte

PUI. Wenn zur Berechnung der Signifikanzen die Anzahl der Probanden (und nicht die durchaus plausible Anzahl an 113 isolierten PST-Messungen) zugrunde gelegt wird, ergeben sich keine signifikante Zusammenhänge zwischen PUI und dem KSS-Selbstrating (**Tab. 1**). Moderate Übereinstimmungen hingegen finden sich für den PUI und die transformierten ORD-Fremdratings. Die spä-

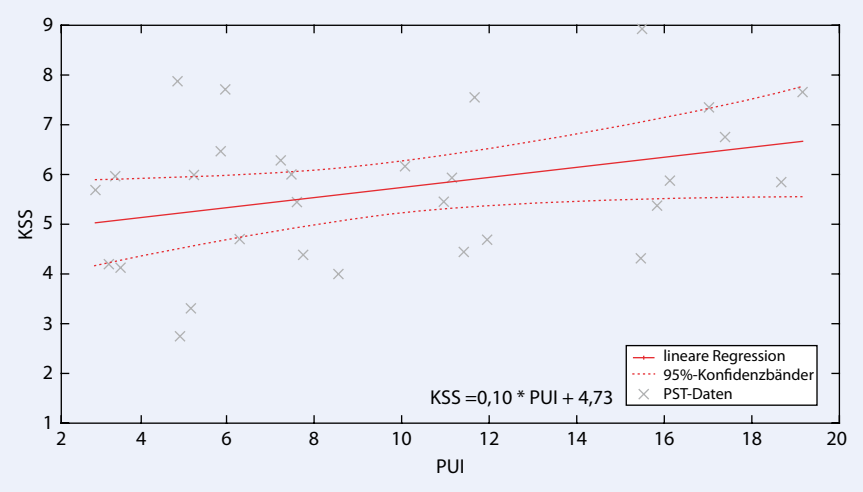

Abb. 2 ▲ Pupilleninstabilitätsbezogene PST-Parameter (PUI). Regressionsgeradengleichungen mit Konfidenzbändern für früh fusionierte KSS- und ORD-Werte; *KSS* Karolinska Sleepiness Scale; *ORD* Observer Rating of Drowsiness; *PST* pupillographischer Schläfrigkeitstest; *PUI* Pupillenunruhe-Index

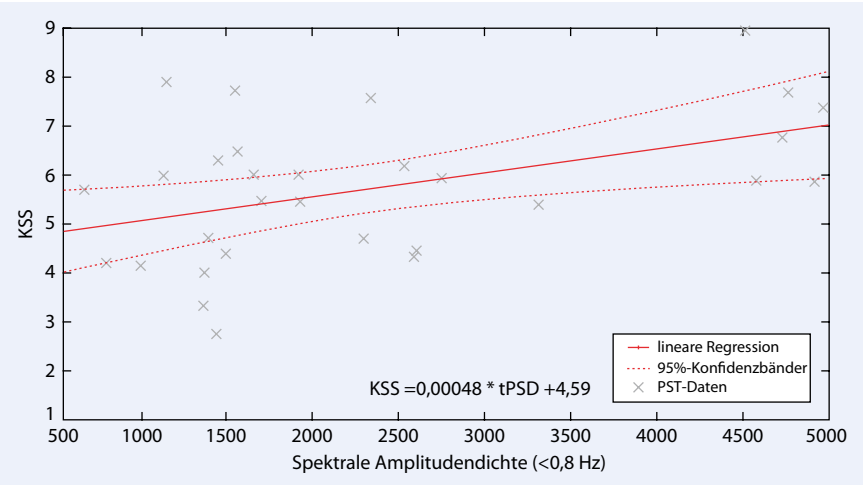

Abb. 3 ▲ Pupilleninstabilitätsbezogene PST-Parameter (tPSD). Regressionsgeradengleichungen mit Konfidenzbändern für früh fusionierte KSS- und ORD-Werte; *KSS* Karolinska Sleepiness Scale; *ORD* Observer Rating of Drowsiness; *PST* pupillographischer Schläfrigkeitstest; *tPSD* spektrale Amplitudendichte < 0,8 Hz

te Fusion aus Selbstrating (Anteil am Gesamtscore: 50%) und Fremdratings (Anteil am Gesamtscore: 50%), sowie die frühe Fusion aus Selbstrating (Anteil am Gesamtscore: 16,7%) und Fremdratings (Anteil am Gesamtscore: 83,3%) ergeben signifikante Zusammenhänge zu PUI-Scores (**Tab. 1**, **Abb. 2**).

Amplitudenspektrum < 0,8 Hz. Keine signifikante Korrelationen zeigten sich zwischen der Zunahme langsamer Oszillationen im Amplitudenspektrum und der KSS (**Tab. 1**). Wiederum ergeben sich für die ORD-Referenzwerte der späten Fusion sowie frühen Fusion aus Selbstrating und Fremdratings, auch unter Verwendung der Samplegröße von N = 30 Personen, moderate signifikante Zusammenhänge (**Tab. 1**, **Abb. 3**).

Bewertung des dreistufigen PUI-basierten PST-Kategoriensystems

Die **Tab. 2** zeigt die in Normwertuntersuchungen über perzentile Kategorisierungen ermittelte 3-Klassen-Einteilung von PUI-Werten [36] in „unauffällig" (PUI < 6,6), „kontrollbedürftig" (6,6 < PUI < 9,8) und „pathologisch" (PUI > 9,8) sowie die den PUI-Klassen jeweils zugeordneten Schläfrigkeitsreferenzwerte (vgl. auch **Abb. 2**). Folgt man den über Perzentile (bzw. Standardabweichungen) definierten Kategorien, entspricht z. B. die PUI-Kategorie „pathologisch" dem Schläfrigkeitsreferenzwert KSS > 6,16. Alternativ zu dem auf Verteilungsparametern basierenden Kategoriensystem würde ein auf den fusionierten KSS-Ratings basierendes PUI-Schläfrigkeitsmesssystem – entsprechend dem gefundenen signifikanten linearen Regressionszusammenhang – KSS = 0,10 * PUI + 4,73 (t = 4,28, p < 0,001) z. B. folgende Zuordnungen treffen: PUI = 2,7→ KSS = 0,10 * 2,7 + 4,73 = 5; PUI = 12,7→ KSS = 0,10 * 12,7 + 4,73 = 6; PUI = 22,7→ KSS = 0,10 * 22,7 + 4,73 = 7.

Diskussion

Das Ziel der vorliegenden Studie war es, einerseits weitere bislang wenig oder überhaupt nicht genutzte Validierungsansätze des PST, wie das KSS-Selbstrating und das ORD-Fremdrating (**Tab. 3**), anzuwenden. Anderseits sollten die allgemeinen Grenzen von üblichen Validierungsansätzen dokumentiert und aus ihnen der Bedarf an optimierten Referenzwertkriterien i. S. von Ground-Truth-Schläfrigkeitsreferenzwerten herausgearbeitet werden. Ferner sollte die für einen sinnvollen Einsatz von „objektiven" Messinstrumenten unbedingt erforderliche Entwicklung von Zuordnungsregeln der objektiven Messwerte zu Schläfrigkeitsskalierungen herausgearbeitet werden (s. Regressionsmodelle in **Abb. 2, 3**). Zur Diskussion gestellt wurde in diesem Zusammenhang ein aus Selbst- und Fremdbeobachtungsverfahren zusammengesetzter Multiples-Rating-Ansatz. Plausibilität erlangt die Fusionierung von Selbst- und Fremdreportmesswerten v. a. über die zu erwartende Stabilisierung der als Messwiederholung aufzufassenden Einzelmessungen. Diese Annahme der Steigerung der Reliabilität und somit auch Validität ließ sich auch anhand von empirischen Indizien, d. h. stärkeren Zusammenhängen zu PST-Kennwerten, erhärten.

So fanden sich hinsichtlich der beiden Pupillenstabilitätsparameter des PST (PUI und Amplitudenspektrum) sowohl für die frühe als auch für die späte Fusion der KSS und ORD ähnliche moderate interindivi-

Tab. 1 Validitäts- und Akkuratheitsmaße für die Standard-PST-Kennwerte PUI und tPSD

	r	MAE	MAPE (%)
PUI			
Selbstreport-KSS	0,25	1,69	41,8
Fremdreport-ORD	0,59	1,55	34,6
Späte Fusion: KSS und ORD	0,44	1,61	36,7
Frühe Fusion: KSS und ORD	0,54	1,58	35,0
tPSD			
Selbstreport-KSS	0,27	1,67	41,7
Fremdreport-ORD	0,55	1,54	34,9
Späte Fusion: KSS und ORD	0,43	1,58	36,8
Frühe Fusion: KSS und ORD	0,51	1,60	35,2

KSS Karolinska Sleepiness Scale; **MAE** „mean absolute error"; **MAPE** „mean absolute percentage error"; **ORD** Observer Rating of Drowsiness; **PST** pupillographischer Schläfrigkeitstest; **PUI** Pupillenunruhe-Index; **tPSD** spektrale Amplitudendichte < 0,8 Hz.

Tab. 2 Perzentilbasierte Kategorisierung des PUI und Zuordnung zu KSS-Werten

	PUI-Kategorien		
Referenzwerte	< 6,6 („unauffällig") (N = 11, n = 44)	6,6–9,8 („kontrollbedürftig") (N = 5, n = 28)	> 9,8 („pathologisch") (N = 14, n = 41)
Selbstreport-KSS	5,38 (1,70)	5,25 (0,75)	5,71 (1,44)
Fremdreport-ORD	5,34 (1,68)	5,21 (1,05)	6,26 (1,35)
Späte Fusion: KSS und ORD	5,36 (1,69)	5,23 (0,89)	5,98 (1,35)
Frühe Fusion: KSS und ORD	5,34 (1,68)	5,22 (0,99)	6,16 (1,34)

KSS Karolinska Sleepiness Scale; **ORD** Observer Rating of Drowsiness; **PUI** Pupillenunruhe-Index. In Klammern aufgeführt sind die jeweils zugeordneten Standardabweichungen. Die genutzten ORD-Werte werden im Sinne einer besseren Vergleichbarkeit und Fusionierbarkeit linear auf die KSS-Skalierung transformiert; **N** Anzahl der Probanden, **n** Anzahl isolierter PST-Messungen.

duelle Korrelationen zu Schläfrigkeitsreferenzwerten von bis zu r = 0,54 (MAE = 1,58 KSS-Punkte, MAPE = 35,0%; **Tab. 1**), womit selbstreportbezogene Validierungsergebnisse früherer PST-Messungen [17, 19, 28, 39, 42] i. Allg. gut repliziert werden konnten. Die hier gefundene Größenordnung der interindividuellen Zusammenhänge von PUI und Selbstreporten ist in der aus vergangenen Studien bekannten Höhe von r = 0,3. Die hier favorisierte konservative Interpretation der Samplegröße von N = 30 statt n = 113 lässt diese Korrelation nicht die Signifikanzschwelle überschreiten. Reduzierend auf die Stärke der hier ermittelten interindividuellen Korrelationen wirkt ferner die eher normalverteilte, realistische Rohwerteverteilung der Schläfrigkeitswerte, die entgegen der in den meisten Vergleichsstudien vorzufindenden extremeren Rohwerteverteilungen die Korrelationshöhe reduziert.

Zusätzlich zu den bisher formulierten Zusammenhangsgrößen wurde in dieser Studie unter Anwendung eines Quasi-Ground-Truth-Kriteriums und der verwendeten absoluten Fehlermaße eine plastische (weil in Schläfrigkeitseinheiten angegebene) Größenordnung der zu erwartenden Unsicherheit dargestellt. Über die hier vorgeschlagene, im Bereich der Mess- und Medizintechnik übliche Quantifizierung des Messfehlers kann deutlich einfacher beurteilt werden, ob zukünftig zu entwickelnde Messverfahren tatsächlich den Anforderungen ihres Einsatzkontexts gerecht werden können. Inwieweit die hier festgestellten mittleren Fehler von 1,58 KSS-Punkten oder die prozentualen Fehler von 34,9% (**Tab. 1**) den Anforderungen diverser Einsatzfelder entsprechen, kann somit in Zukunft diskutiert und auch bezweifelt werden.

Worauf die PST-Messfehler wiederum zurückzuführen sind, bleibt eine andere, noch zu klärende Frage. Eine mögliche Erklärung liefert die eher niedrige, aus einem 3-Monats-Retestintervall ermittelte PST-Retestreliabilität [18]. Kritisch anzumerken ist hierzu jedoch, dass das verwendete Retestintervall zu einer fehlenden Äquivalenz der wahren Werte der Test- und Retestmesszeitpunkte führt und somit ein wenig aussagekräftiges Maß zur Bewertung der Messgüte eines u. U. minütlich fluktuierenden Konstrukts darstellt. Weitere mögliche Messfehlerquellen des PST stellen die trotz identischer Schläfrigkeit individuell unterschiedlichen PUI-Baseline-Niveauverschiebungen und individuellen PUI-Reaktionsprofile dar. So kann sich die Verlaufsform der Steigerung der Pupillenunruhe bei wachsender Schläfrigkeit in diversen von Proband zu Proband unterschiedlichen nichtlinearen Verläufen niederschlagen. Einen vielversprechenden Ansatz zum generellen Umgang mit interindividuellen Unterschieden in psychophysiologischen Messinstrumenten liefert die mustererkennungsbasierte Detektion von unterschiedlichen Pupillen-Respondergruppen und eine auf ihr aufbauende multivariate Fusionierung vieler eher schwacher Einzelkorrrelate zu einem Gesamtmesswert [14, 31]. Ferner kann eine während der PST-Messung unbemerkte Selbststimulation die PST-Ergebnisse verzerrt haben.

Begründet werden kann die Abweichung zwischen den in dieser Studie genutzten Validierungskriterien und dem PST neben

- PST-internen Messfehlerquellen auch mit
- Veränderungen des wahren Schläfrigkeitswerts zwischen den PST- und den Referenzmessungen sowie mit
- Messfehlern der Referenzverfahren selbst.

Die unmittelbare zeitlichen Nähe sowie die nahezu identische monotone Messsituation von Fremd-, Selbstreport und PST deuten auf Veränderungen der äußeren Stimulationssituation, die die wahren Schläfrigkeitswerte im Vergleichszeitraum nicht wesentlich beeinflusst haben sollten. Eine weitere sich daran anschließende, für die praktische Verkehrssicherheitsbeurteilung zentrale, jedoch noch ungeklärte Frage bezieht sich darauf, welche Informationen eine PST-Messung bezüglich des vorangegangenen, aktuellen und zukünftigen Schläfrigkeitszustands liefert. Denkbar wäre beispielsweise, dass der PST aufgrund des Erholungseffekts seiner 10-minütigen Messdauer eine Abschätzung der Prä-PST-Schläfrigkeit, nicht jedoch der Post-PST-Schläfrigkeit erlaubt.

Bezüglich der Messfehler der eingesetzten Referenzverfahren selbst lässt sich feststellen, dass die genutzten Selbst- und Beobachtungsratings innerhalb des hier umgesetzten Forschungskontexts zwar die üblichen Validitätsanforderungen erfüllen, was jedoch über tatsächliche Messfehlerfreiheit wenig aussagt und somit im Zweifel auch eine geringe Akkuratheit der Referenzverfahren nicht ausschließt. Ferner ist neben den Messfehlern der Referenzverfahren auch die Änderung und somit Nichtäquivalenz der wahren Schläfrigkeitswerte von PST und Referenzwertmessungen eine zu berücksichtigende Fehlerquelle. Verursacht werden kann diese Nichtäquivalenz v. a. über unterschiedliche Aktivierungs- und Deaktivierungsbedingungen zwischen den Messungen sowie über die Messsituation selbst. Insbesondere die stufenweise Entspannung der Probanden während der DVD-Sitzung, die leichte Mobilisierung der Probanden im Übergang von der DVD- in die PST-Situation könnte bereits zu dieser verzerrenden, die Schläfrigkeit verringernden Stimulation geführt haben. In Zukunft könnte z. B. die Analyse nur der letzten 5 min der DVD-Messung zur Bestimmung der Referenzmessung genutzt werden. Ferner können Abweichungen der wahren Werte Selbstreport, Fremdreport und PST auch über unterschiedliche Anstrengung und Selbststimulation während der jeweiligen Messungen hervorgerufen werden.

Mögliche Einschränkungen der Generalisierbarkeit der Ergebnisse liegen in der fehlenden seniorenbezogenen Datengrundlage. So besteht ein Zusammenhang der Variable „Lebensalter“ mit erhöhter Schläfrigkeit und darüber vermittelt auch mit einer erhöhten Pupillenunruhe. Unklar bleibt jedoch, ob für die PUI-Werte bei Senioren identische Zuordnungsregeln zu Schläfrigkeitsintensitäten vorliegen (s. Regressionsmodelle in **Abb. 2, 3**) oder ob es separater altersangepasster Zuordnungsregeln bedarf. Weitere Einschränkungen der Generalisierbarkeit der Ergebnisse liegen in den Inklusionskriterien der Studie, die z. B. über ihre starken Einschränkungen (z. B. Alkohol-, Nikotin- und Koffeinkarenz) für einen großen Teil der intendierten Zielpopulation keine Aussagen treffen lassen. Neben diesen so produzierten Problemen der externen Validität ist zudem auch der Nutzenaspekt dieser Inklusionskriterien außerhalb einer schlafmedizinisch-orientierten Erfassung von allgemeiner Tagesschläfrigkeit fragwürdig. Ist die Zielsetzung einer Analyse die Erfassung der unverfälschten Tagesschläfrigkeit, können diese den aktuellen Schläfrigkeitszustand beeinflussenden Inklusionskriterien sinnvoll zum Einsatz kommen. Bei einer an punktuellen Schläfrigkeitszuständen und ihrer assoziierten Pupillenaktivität interessierten Forschungslogik sind die Einschränkungen nur dann zu legitimieren, wenn der Verdacht besteht, dass ein eigenständiger, nicht über Schläfrigkeitsveränderungen vermittelter Einfluss auf die Pupillengrößenveränderung besteht. Bislang sind nach unserem Ermessen diese Verdachtsmomente für keines der eingesetzten Inklusionskriterien der Alkohol-, Nikotin- und Koffeinkarenz belegt.

Ein weiteres Ziel dieser Studie war es, die verwendeten PST-Kategorien „unauffällig“, „kontrollbedürftig“ und „pathologisch“ einer kritischen Bewertung zu unterziehen. Legt man die Ergebnisse dieser nichtnormativen Studie zugrunde, können insbesondere „unauffällige“ eher schlecht von „kontrollbedürftigen“ Zuständen unterschieden werden. Zusätzlich werden moderate Schläfrigkeitszustände, wie aus der Regressionsgleichung in **Abb. 2** abzulesen ist, bereits ab einem KSS-Wert von 5,71 (0,10 * 9,8 + 4,73 = 5,71; also auf der KSS „etwas schläfrig“) im PST als kritisch, d. h. pathologisch, bewertet. Sollten sich diese pilothaft an einer kleineren Stichprobe ermittelten Ergebnisse auch in normativen Studien wiederholen, wären diese Grenzen sicherlich neu zu überdenken (**Abb. 2**).

Eine wesentliche Limitation der vorliegenden Studie ist der relativ kleine Messwerte- und Stichprobenumfang, der nur begrenzt geeignet ist, um über unterschiedliche Populationen und Anwendungskontexte hinweg belastbare Aussagen zu treffen. Folge dieses kleinen Datenpools ist eine verringerte statistische Power der Ergebnisse, d. h. eine vergrößerte Gefahr, ein eigentlich signifikantes Ergebnis zu Unrecht abzulehnen. Diese Power würde durch eine denkbare Alpha-Adjustierung weiter abnehmen – auch wenn sie im Fall berechneter Korrelationen zu keiner veränderten Bewertungen der Signifikanz der Ergebnisse führen würden (zur Kritik des Alpha-Adjustierungsvorgehens: [27]). Ein weiterer denkbarer Kritikpunkt ist die einseitige Hypothesentestung, für die neben der angesprochenen erhöhten statistischen Power auch das eindeutige inhaltliche Argument spricht, dass die Zusammenhangsrichtung für PST-Parameter und Schläfrigkeit aus vielen Studien bekannt ist und somit nur eine einseitige Testung sinnvoll erscheint. Weitere gewichtigere Limitationen beziehen sich auf die Verwendung des Messwiederholungsdesigns, das ohne die Variation der experimentellen Bedingungsreihenfolgen (mit einem meist linear ansteigenden Schläfrigkeitsverlauf) realisiert worden ist. Aufgrund der sich durch Messwiederholungen potenziell kumulierenden Lern-, Gewöhnungs-, und Sättigungseffekte können Konfundierungsprozesse mit schläfrigkeitsunabhängigen inneren Zuständen (wie z. B. Frustration, Langeweile, Lustlosigkeit) entstehen. Um dieser Problematik entgegenzuwirken, sollten sich zukünftige Forschungsbemühungen auf eine Verkürzung der PST-Messzeiten richten, z. B. über die Anwendung multivariater mustererkennungsbasierter Verfahren [14], in denen auch die Fusion von okulomotorischen Messgrößen wie Lid- und Blickbewegungen Berücksichtigung findet. Um die interne Validität nicht zu gefährden, könnten alternativ randomisierte Kontrollgruppendesigns ohne Messwiederholung oder Messwiederholungsdesigns mit wechselnden, auch über mehrere aufeinanderfolgende Tage verteilten Aktivierungs-und Sedierungsphasen verwendet werden. Ferner könnte die interne Validität durch die Fusion von Selbst-und Fremdratings mit weiteren KSS-skalierten psychophysiologischen Maßen verbessert werden. Dazu müssten jedoch zunächst (im Idealfall in normativen Studien) eindeutige Zuordnungsregeln, z. B. von Theta-Frequenzbandenergie im EEG zur KSS-Schläfrigkeitsskalierung, gefunden werden.

An diesem Punkt stoßen wir an eine der größten Herausforderungen der zukünftigen Schläfrigkeitsforschung, die Lösung des infiniten Regressproblems. Um dennoch Schläfrigkeitsmessungen im Fokus weiterzuentwickeln, reicht die wechselseitige Dokumentation von korrelativen Zusammenhängen unterschiedlicher nicht-

schläfrigkeitsskalierter „Validierungskriterien“ (z. B. derzeitig EEG, MSLT) nicht aus, um eines der Verfahren in den Rang eines echten Ground-Truth-Instruments zu verhelfen. Eine pragmatische Lösung des infiniten Regressproblems könnte hingegen in der Annäherung einer“quasi-ground truth“ über die Fusion multipler, messfehlerbehafteter, aber schläfrigkeitsskalierter Kriterien liegen. Wollte man jedoch Instrumente ohne vorliegende Schläfrigkeitsskalierung (z. B. Perclos und EEG-Theta-Aktivität) fusionieren, wäre das zwar nach z. B. z-Transformation mathematisch denkbar, die resultierenden Kennwerte hätten jedoch wiederum keine eindeutige Zuordnung zu Schläfrigkeitszuständen und wären daher auch nicht als „ground truth“ einsetzbar.

Fazit für die Praxis

Die hier bestimmten Selbstreportdaten replizieren die bisherigen Befunde zur Validierung des pupillographischen Schläfrigkeitstests (PST), auch wenn ihr Einsatz als Validierungskriterium mit zahlreichen Schwierigkeiten verbunden ist. Um diesen Schwierigkeiten teilweise zu begegnen, wurde das zur Validierung des PST eingesetzte Verfahrensportfolio um einen bislang vernachlässigten Validierungszugang erweitert: die zeitlich und inhaltlich fein aufgelöste und skalierte Erfassung von Schläfrigkeit mittels Fremdbeobachtungsdaten. Über die Validierungsintention hinaus sollte hier ferner die tatsächliche Messgenauigkeit und Akkuratheit der PST-Messungen über die Anwendung des (Quasi-) Ground-Truth-Konzepts bestimmt werden. Die reliabilitätssteigernde Fusionierung von Selbst- und Fremdratings (Multiples-Rating-Ansatz) diente dabei als vorläufiges Quasi-Ground-Truth-Verfahren und ergab eine in den jeweiligen experimentellen und klinischen Anwendungskontexten der PST-Messungen separat zu bewertende Messunsicherheit von 35%, die v. a. aufgrund der Unsicherheit im niedrigen PST-Messwertebereich den Bedarf einer Verbesserung der PST-Algorithmen andeutet. Abschließend sind, falls sich die Daten auch in weiteren Studien bestätigen, die in klinischen Kontexten angewendeten und über Perzentile (bzw. Standardabweichungen) definierten Pupillenunruhe-Index(PUI)-Kategorien „kontrollbedürftig“ und „pathologisch“ einer kritischen Überprüfung zu unterziehen. Somit wird über die Problematik der PST-Validierung hinaus deutlich, dass ein an perzentilen Normwerten orientiertes Vorgehen zur Bestimmung von Grenzwerten mit Vorsicht zu bewerten ist.

Korrespondenzadresse

Prof. Dr. J. Krajewski
Experimentelle Wirtschaftspsychologie, Universität Wuppertal
Gaußstr. 20, 42097 Wuppertal
krajewsk@uni-wuppertal.de

Danksagung. Wir danken Herrn Tobias Peters vom STZ *eyetrial* der Universität Tübingen für die Unterstützung und Beratung bei der Durchführung der Studie sowie für seine wertvollen Beiträge zum Manuskript.

Interessenkonflikt . Der korrespondierende Autor gibt für sich und seine Koautoren an, dass kein Interessenkonflikt besteht.

Literatur

1. Aston-Jones G, Cohen JD (2005) Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. J Comp Neurol 5:99–110
2. Buysse D J, Reynolds CF, Monk TH et al (1989) The Pittsburgh Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 28:193–213
3. Canisius S, Penzel T (2007) Vigilance monitoring – review and practical aspects. Biomed Tech 52:77–82
4. Carskadon MA, Dement WC (1982) The Multiple Sleep Latency Test: What does it measure? Sleep 5:67–72
5. Danker-Hopfe H, Kramer S, Dorn H et al (2001) Time-of-day variations in different measures of sleepiness (MSLT, pupillography and SSS) and their interrelations. Psychophysiology 38:828–835
6. Dittrich E, Brandenburg S, Thüring M (2009). Beobachtungsbasierte Erfassung von Müdigkeit im Kfz – die TUBS-Skala. In: Lichtenstein A, Stößel C, Clemens C (Hrsg) Der Mensch im Mittelpunkt technischer Systeme: 8 Berliner Werkstatt. Mensch-Maschine-Systeme. Berlin
7. Garbarino S, Nobili L, Beelke M et al (2001) The contributing role of sleepiness in highway vehicle accidents. Sleep 24:203–206
8. Horne JA, Burley CV (2010) We know when we are sleepy: subjective versus objective measurements of moderate sleepiness in healthy adults. Biol Psychol 83:266–268
9. Horne JA, Reyner LA (1995) Sleep related vehicle accidents. Br Med J 310:565–567
10. Hou RH, Langley RW, Szabadi E, Bradshaw CM (2007) Comparison of diphenhydramine and modafinilon arousal and autonomic functions in healthy volunteers. J Psychopharmacol 21:567–578
11. Huron C, Giersch A, Danion JM (2002) Lorazepam sedation, and conscious recollection: a dose-response study with healthy volunteers. Int Clin Psychopharmacol 17(1):19–26
12. Ingre M, Åkerstedt T, Peters B et al (2006) Subjective sleepiness and accident risk avoiding the ecological fallacy. J Sleep Res 15:142–148
13. Krajewski J, Batliner A, Golz M (2009) Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. Behav Res Methods 41(3):795–804
14. Krajewski J, Schnupp T, Schnieder S et al (2010) Pattern recognition methods – A novel analysis for pupillographic sleepiness test. Proceedings Measuring Behaviour 7:459–462
15. Lichstein KL, Wilson NM, Noe SL et al (1994) Daytime sleepiness in insomnia: behavioral, biological and subjective indices. Sleep 17(8):693–702
16. Littner MR, Kushida C, Wise M et al (2005) Practice parameters for clinical use of the multiple sleep latency test and the maintenance of wakefulness test. Sleep 28:113–121
17. Lowenstein O, Feinberg R, Loewenfeld IE (1963) Pupillary movements during acute and chronic fatigue. Invest Ophthalmol 2:138–157
18. Lüdtke H, Körner A, Wilhelm B, Wilhelm H (2000) Reproduzierbarkeit des pupillographischen Schläfrigkeitstests bei gesunden Männern. Somnologie 4:170–172
19. Lüdtke H, Wilhelm B, Adler M et al (1998) Mathematical procedures in data recording and processing of pupillary fatigue waves. Vision Res 38(19):2889–2896
20. MacLean AW, Davies DR, Thiele K (2003) The hazards and prevention of driving while sleepy. Sleep Med Rev 7:507–521
21. Marzano C, Fratello F, Moroni F et al (2007) Slow eye movements and subjective estimates of sleepiness predict EEG power changes during sleep deprivation. Sleep 30(5):610–616
22. Mc.Laren JW, Hauri PJ, Lin SC, Harris CD (2002) Pupillometry in clinically sleepy patients. Sleep Med 3(4):347–352
23. Merritt SL, Lloyd SR, Meyer FT, Kogan J (2003) Pupillary unrest in sleepy versus non-sleepy healthy subjects. Abstract at the 25th International Pupil Colloquium, Crete
24. Merritt S L, Schneyders HC, Patel M et al (1999) Pupil staging and EEG-defined sleepiness. Sleep 22(1):110–111
25. Muttray A, Hagenmeyer L, Unold B et al (2007) Videoanalyse der Schläfrigkeit von Fahrern. Z Arbeitswissenschaft 61(4):245–254
26. Newman J, Broughton R (1991) Pupillometric assessment of excessive daytime sleepiness in narcolepsy-cataplexy. Sleep 14(2):121–129
27. Perneger TV (1998) What's wrong with Bonferroni adjustments. Br Med J 316:1236–1238
28. Regen FM (2009) Assoziation zwischen Pupillen-Unruhe-Index und Korrelaten des zentralnervösen Aktivierungsniveaus im Wach-EEG. Dissertation, Medizinische Fakultät, Charité-Universitätsmedizin Berlin
29. Reyner LA, Horne JA (1998) Falling asleep at the wheel: Are drivers aware of prior sleepiness? Int J Legal Med 111:120–123
30. Samuels ER, Hou RH, Langley RW et al (2006) Comparison of pramipexole and modafinil on arousal, autonomic, and endocrine functions in healthy volunteers. J Psychopharmacol 20(6):756–770
31. Schnupp T, Sommer D, Krajewski J, Golz M (2010). Data analysis of the pupillographic sleepiness test utilizing computational intelligence. Proc Vigilanz 1:25–30
32. Sommer D, Golz M, Schupp T et al (2009) A measure of strong driver fatigue. Proceeding International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design 4:9–15

33. Weeß HG, Sauter C, Geisler P et al (2000) Vigilanz, Einschlafneigung, Daueraufmerksamkeit, Müdigkeit, Schläfrigkeit–Diagnostische Instrumentarien zur Messung müdigkeits- und schläfrigkeitsbezogener Prozesse und deren Gütekriterien. Somnologie 4:20–38
34. Weeß HG (2004) Diagnostik der Tagesschläfrigkeit. In: Virchow M (Hrsg) Handbuch Schlafmedizin. Dustry, Oberhaching/München
35. Wichniak AJ, Geisler P, Tracik F et al (1999) Comparison of methods for objective quantifying daytime sleepiness and its relations to quality of previous night's sleep. Sleep Res Online 2(1):591
36. Wiegand DM, McClafferty J, McDonald SE, Hanowski R J (2009) Development and evaluation of a naturalistic Observer Rating of Drowsiness protocol. National Surface Transportation Safety Center for Excellence, Blacksburg/VA
37. Wierwille WW, Ellsworth LA (1994) Evaluation of driver drowsiness by trained raters. Accid Anal Prev 26(5):571–581
38. Wilhelm B (2007) Über die Spontanoszillation der Pupille und ihre Beziehung zum zentralnervösen Aktivierungsniveau. Steinbeis Edition, Stuttgart
39. Wilhelm B, Giedke H, Lüdtke H et al (2001) Daytime variations in central nervous system activation measured by a pupillographic sleepiness test. J Sleep Res 10:1–7
40. Wilhelm B, Koerner A, Heldmaier K et al (2001) Normwerte des pupillographischen Schläfrigkeitstests für Frauen und Männer zwischen 20 und 60 Jahren. Somnologie 5:115–120
41. Wilhelm B, Widmann A, Wilhelm D et al (2009) Objective and quantitative analysis of daytime sleepiness in physicians after night duties. Int J Psychophysiol 72:307–313
42. Wilhelm B, Wilhelm H, Lüdtke H et al (1998) Pupillographic assessment of sleepiness in sleep-deprived healthy subjects. Sleep 21(3):258–265
43. Wilhelm H, Lüdtke H, Wilhelm B (1998) Pupillographic sleepiness testing in hypersomniacs and normals: Graefes Arch Clin Exp Ophthalmol 236(10):725–729
44. Wise MS (2006) Objective measures of sleepiness and wakefulness: application to the real world? J Clin Neurophysiol 23(1):39–49
45. Yoss RE, Moyer NJ, Hollenhorst RW (1970) Pupil size and spontaneous pupillary waves associated with alertness, drowsiness, and sleep. Neurology 20:545–554

Anhang

Tab. 3 Observer Rating of Drowsiness (ORD): Verhaltens- und Manierismen-Checkliste

	Keine	Wenig	Moderat	Extrem
Augen/Augenbrauen:				
Reiben/Kratzen				
Leerer Blick/Starren				
Gesenkte Augenlider				
Exzessives, häufiges Blinzeln				
Langsames Zufallen der Augenlider				
Unfokussiertes Rollen der Augen				
Glasiger Blick				
Deutliches Anheben der Augenbrauen				
Deutliches Senken der Augenbrauen				
Körper:				
Sich plumpsen lassen, krummes Sitzen, schief angelehnt sein				
Seufzen				
Sich strecken				
Schwacher eingefallener Muskeltonus				
(Rastlose) Veränderungen der Körperposition				
Mund:				
Gähnen				
In die Lippe beißen, Lecken der Lippen				
Zungenbewegungen				
Gesicht:				
Reiben und halten				
Verzerren des Gesichts				
Schwacher Muskeltonus im Gesicht				
Nacken/Kopf:				
Kratzen, glätten				
Reiben/halten				
Kopf anlehnen (nach hinten, zur Seite)				
Positionsveränderungen des Kopfs				
Kopfnicken, langsames Wegnicken				