Analyse der Interrater-Reliabilität beim Einsatz der Gefährdungsbeurteilung Psychischer Belastungen

Nadine SEIFERLING, Karlheinz SONNTAG, Miriam KRUG

Ruprecht-Karls-Universität Heidelberg, Arbeits- und Organisationspsychologie
Hauptstr. 47-51, 69117 Heidelberg

Kurzfassung: Die Gefährdungsbeurteilung psychischer Belastungen (GPB) ist ein objektives und konsensorientiertes Beobachtungsverfahren zur Analyse psychischer Belastung am Arbeitspatz, bei dem psychische Belastungsfaktoren von einem geschulten Team eingeschätzt werden. Seit ihrer Entwicklung wurde die GPB in verschiedenen Unternehmen unterschiedlicher Größe und Branchen eingesetzt.
Aufgrund der vorliegenden Daten wurden Interrater-Reliabilitäten für N=89 beurteilte Arbeitsplätze in vier Unternehmen analysiert. Die Ergebnisse weisen auf eine ausreichende bis gute Übereinstimmung zwischen den Beobachterurteilen hin. Erwartungsgemäß zeigt sich außerdem, dass das Ausmaß der Übereinstimmung mit der Erfahrung der Analyseteam-mitglieder wächst. Ferner werden Einflussfaktoren sowie Implikationen für die Anwendung der GPB in der Unternehmenspraxis diskutiert.

Schlüsselwörter: Gefährdungsbeurteilung psychischer Belastungen am Arbeitsplatz, GPB, objektive Analyseverfahren, Gütekriterien

1. Theoretischer Hintergrund

In einer sich wandelnden Arbeitswelt, die durch technologische Neuerungen, Dynamisierung und Informatisierung geprägt ist, verändern sich auch Belastungsmuster von Führungskräften und Mitarbeitern. Tätigkeiten werden zunehmend komplexer, kommunikativer und sind mit mehr Eigenverantwortung verbunden. In den letzten Jahren haben dabei insbesondere psychische Belastungen an Bedeutung gewonnen. Zahlreiche Studien belegen die Relevanz psychischer Belastungen für gesundheitliche Gefährdungen am Arbeitsplatz (vgl. Hasselhorn & Portuné, 2010; Sonntag, Turgut & Feldmann, 2016).

Um negativen Folgen psychischer Belastung am Arbeitsplatz vorzubeugen, ist es sinnvoll und notwendig, neben Arbeitsplatzgestaltung, -umgebung, sowie Arbeitsmitteln und Qualifikation der Mitarbeiter auch psychische Belastungsfaktoren in die Gefährdungsbeurteilung mit einzubeziehen. Diese Notwendigkeit findet seit 2013 auch in der Erweiterung von §5 des Arbeitsschutzgesetzes eine Entsprechung: „psychische Belastungen bei der Arbeit“ sind nun explizit in der Liste der potenziellen Gefährdungsfaktoren, die in die Beurteilung der Arbeitsbedingungen mit einzubeziehen sind, genannt.

Dabei lässt das Arbeitsschutzrecht bei der Wahl der Analysemethode zur Ermittlung psychischer Belastungen einen großen Spielraum, so dass verschiedene Verfahren mit unterschiedlichen Analysezugängen zur Gefährdungsbeurteilung psychischer Belastungen eingesetzt werden können. Zu den am häufigsten genutzten Verfahren zählt neben (standardisierten) Mitarbeiterbefragungen und moderierten Analyseworkshops vor allem die Methode der *Beobachtung* bzw. des

Beobachtungsinterviews (Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, 2014). Diese Verfahren bieten den Vorteil, dass sie eine objektive Beurteilung des Arbeitsplatzes ermöglichen.

Die Gefährdungsbeurteilung psychischer Belastungen (GPB) ist ein *objektives* und *konsensorientiertes Beobachtungsverfahren* zur Analyse psychischer Belastung am Arbeitspatz (Sonntag, Michel & Seiferling, 2014; Sonntag et al. 2016). Die Beurteilung psychischer Belastung erfolgt dabei durch ein eigens geschultes Analyseteam, das aus Vertretern verschiedener Fachbereiche (z.B. Arbeitsmedizin, -sicherheit und Betriebsrat) besteht. Im Rahmen einer Begehung schätzt jedes Analyseteammitglied die vorherrschende Belastung anhand standardisierter Fragen ein. In der anschließenden Konsensfindung werden die Einzelurteile unter Einbezug der verschiedenen Sichtweisen diskutiert und das Team einigt sich auf ein gemeinsam getragenes Urteil.

Gewisse Unterschiede in der Einschätzung sind daher bei der Durchführung durchaus erwünscht, um ein facettenreiches und umfassendes Bild des Arbeitsplatzes zu erhalten. Im Hinblick auf die Güte des Instruments und aufgrund ihrer Funktion als Diskussionsgrundlage für die Konsensphase ist es jedoch von großer Bedeutung, dass die eingesetzten standardisierten Fragen ein grundlegendes Maß an Objektivität und Zuverlässigkeit gewährleisten.

2. Methode

Da Messinstrumente nie „fehlerfrei“ sind, ist es insbesondere bei Beobachtungsverfahren, bei denen Einschätzungen durch menschliche Beurteiler vorgenommen werden, wichtig die Interrater-Reliabilität (Übereinstimmung der Beurteilungen) zu bestimmen. Diese gibt Aufschluss darüber, wie ähnlich (objektiv) und wie zuverlässig (reliabel) die Einschätzungen verschiedener Beurteiler (Rater) sind (Wirtz, 2004).

Ein grundlegendes Maß an Übereinstimmung zwischen verschiedenen Beobachtern stellt sicher, dass die Einzelurteile der Rater eine gute Schätzung der wahren Ausprägung der Anforderungsdimensionen der GPB und somit auch eine gute Diskussionsgrundlage für die Konsensfindung darstellen. Nachfolgend werden die Analysemethode sowie die Stichprobe dargestellt.

2.1 Interrater-Reliabilität

Zur Ermittlung der Interrater-Reliabilität wurden Intraklassenkorrelationen (ICC; engl. intraclass correlations) durchgeführt. Diese Methode stellt ein gängiges Verfahren für intervallskalierte Daten von mehr als zwei Ratern dar. Die Interrater-Reliabilität kann Werte zwischen 0 und 1 annehmen, wobei höhere Werte eine höhere Ähnlichkeit zwischen den Urteilen und höhere Reliabilität anzeigen. Wird beispielsweise ein Wert von .63 beobachtet, so sind 70% der beobachteten Varianz auf die „wahre Varianz“ und 30% auf die Messfehlervarianz bzw. auf die Unterschiede zwischen den Rater-Urteilen zurückzuführen (Hallgren, 2012). Eine reliable Messung und hohe ICC-Werte resultieren dabei nur, wenn die Varianz der Ratings zwischen den Beobachtern gering ist, und gleichzeitig die Varianz zwischen den Items verhältnismäßig groß ist. Hallgren (2012) zufolge werden Werte größer als .75 als hervorragende, Werte zwischen .60 und .74 als gute und von .40 bis .59 als ausreichende Interrater-Reliabilitäten klassifiziert (≤.40 unbefriedigend)

Für die Bestimmung der ICC existieren verschiedene Formen der Intraklassenkorrelationen. Aufgrund der Eignung für wechselnde Gruppenzusammensetzungen (unterschiedliche Zusammensetzung des Analyseteams) und des Fokus auf Ähnlichkeit (im Gegensatz zur Gleichheit der Ratings) kam in der vorliegenden Studie die unjustierte Form der ICC (ICC_{unjust}) zum Einsatz.

Bei der Analyse der Interrater-Reliabilität interessierte neben der Reliabilität der Einzelurteile eines durchschnittlichen Analyseteammitglieds auch die Veränderung über die Zeit, da aufgrund der Komplexität des Verfahrens und der steigenden Expertise mit einer Verbesserung der Rater zu rechnen ist. Um diese Annahme zu überprüfen wurden die Interrater-Übereinstimmungen jeweils zu Beginn und zum Ende des Pilotprojektes ermittelt.

2.2 Stichprobe

Die Grundlage für die Analysen bildeten Daten von Ratern aus vier verschiedenen Unternehmen unterschiedlicher Branchen. Alle Daten wurden im Rahmen von Pilotprojekten zur Beurteilung psychischer Belastungen am Arbeitsplatz, die in Kooperation mit der Arbeits- und Organisationspsychologie der Universität Heidelberg durchgeführt wurden, erhoben.

Insgesamt wurden 89 Arbeitsplätze in den verschiedenen Unternehmen betrachtet, wobei die Analyseteams jeweils aus drei bis acht Mitgliedern bestanden. Durchschnittlich waren bei jeder Begehung 5,46 Analyseteammitglieder vertreten.

3. Ergebnisse

Die Analysen der Reliabilität der Einzelurteile eines durchschnittlichen Analyseteammitglieds ergaben bei 46% der Arbeitsplätze eine gute (35%) bis hervorragende (11%) Interraterreliabilität ($ICC_{unjust} \geq .60$).

Weitere 45% der Arbeitsplatz-Einschätzungen waren ausreichend übereinstimmend und reliabel (ICC_{unjust} .40-.59). Bei 9% der Arbeitsplätze fand sich hingegen eine unbefriedigende Übereinstimmungsrate ($ICC_{unjust} \leq .40$). Abbildung 1 zeigt eine Übersicht über die Verteilung der Interrater-Reliabilitäten.

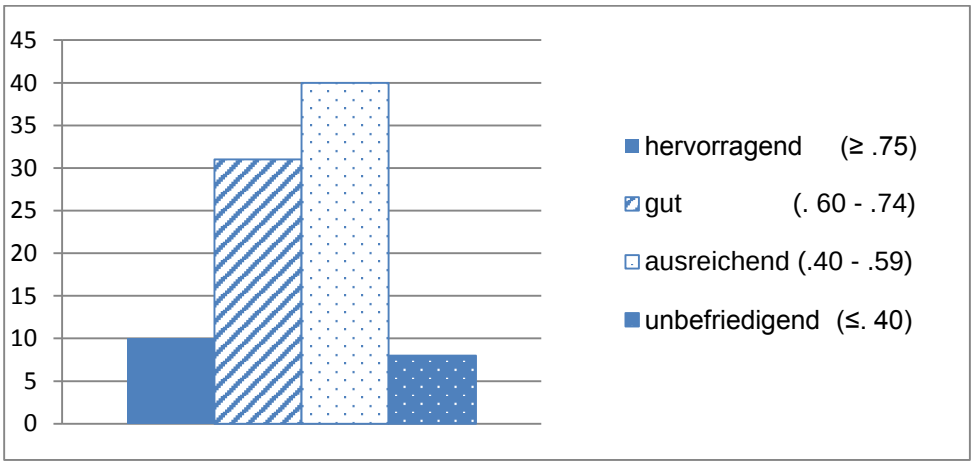

***Abbildung 1**: Häufigkeit der ICCs an 89 Arbeitsplätzen in fünf Unternehmen.*

Weiterhin wurde die Veränderung der Interrater-Reliabilitäten über die Zeit hinweg betrachtet. Aufgrund von Übungseffekten und steigender Expertise wurde erwartet, dass die ICCs über den Verlauf des Pilotprojektes hinweg zunehmen.

In Tabelle 1 sind die ICCs der jeweils ersten und letzten Begehung in den vier Unternehmen aufgeführt. In allen Unternehmen zeigte sich eine Steigerung der Interrater-Übereinstimmung im Vergleich zur ersten Begehung. Anhand der Konfidenzintervalle lässt sich feststellen, dass dieser Unterschied jedoch nur für Unternehmen D signifikant war.

Tabelle 1: *Ermittelte Interrater-Reliabilitäten (ICC_{unjust}) zu Beginn (erste Begehung) und Ende des Pilotprojektes (letzte Begehung). * p< .05; ** p<.001*

	ICC_{unjust}		
	Erste Begehung	Letzte Begehung	*Differenz*
Unternehmen A	.540**	.674**	*.134*
Unternehmen B	.284**	.452**	*.168*
Unternehmen C	.383**	.531**	*.148*
Unternehmen D	.414**	.837**	*.423**

4. Diskussion

Im Rahmen dieser Untersuchung wurde die Interrater-Reliabilität als wichtiges Gütekriterium für die Bestimmung der Objektivität und Reliabilität der Anforderungsdimensionen der GPB ermittelt. In die Analysen gingen Rater-Daten von insgesamt 89 Arbeitsplatzbegehungen in vier Unternehmen unterschiedlicher Branchen mit ein.

Die Ergebnisse weisen auf eine gute Übereinstimmung zwischen den Beobachterurteilen hin. In über 90% der Fälle waren die Interrater-Übereinstimmungen mindestens ausreichend (≥ .40). Außerdem zeigte sich im Verlauf der Pilotprojekte auch eine – wenn auch meist nicht signifikante – Steigerung der ICCs. Bei der letzten Arbeitsplatzbegehung der Pilotprojekte lagen die Interrater-Reliabilitäten aller Unternehmen über .40.

Eine deutliche Steigerung wurde in Unternehmen D erreicht. Dies ist wahrscheinlich in der geringen Fluktuation der Analyseteammitglieder begründet, da im Rahmen des Pilotprojektes vergleichsweise wenige Personen in der Anwendung des Verfahrens geschult wurden und so die Gruppenzusammensetzung wenig variierte, so dass sich das Team besonders „gut eingespielt“ gewesen sein dürfte.

Die ausreichenden bis guten Ergebnisse der ICCs zeigen, dass die Analyseteammitglieder im Rahmen der Begehungen etwas Ähnliches zuverlässig einschätzen und bestätigen damit die Güte des Analyseinstrumentes. Andererseits implizieren sie - insbesondere im Hinblick auf die Veränderungen über die Zeit - aber auch die Wichtigkeit der Schulung und Begleitung der Analyseteams in der Pilotphase um mit dem Verfahren vertraut zu werden. Dabei ist es von besonderer Bedeutung die Analyseteammitglieder in der Schulung auf mögliche Beobachtungsfehler wie z.B. Strenge- und Mildeeffekte hinzuweisen, sowie gute unternehmensspezifische Beispiele für die einzelnen Items zu entwickeln, um so ein gemeinsames, einheitliches Verständnis der Fragen zu entwickeln und zu stärken.

5. Literatur

Arbeitsschutzgesetz. *§5 Beurteilung der Arbeitsbedingungen.* Abgerufen am 05.12.2015. http://www.gesetze-im-internet.de/arbschg/__5.html.

Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (Hrsg.) (2014). *Gefährdungsbeurteilung psychischer Belastung - Erfahrungen und Empfehlungen.* Berlin: Erich-Schmidt-Verlag.

Hallgren K.A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8(1*), 23-34.

Hasselhorn, H. M. & Portuné, R. (2010). Stress, Arbeitsgestaltung und Gesundheit. In B. Badura, U. Walter & T. Helhlmann (Hrsg.), *Betriebliche Gesundheitspolitik: Der Weg zur gesunden Organisation* (S. 361-376). Berlin: Springer.

Sonntag, Kh., Michel, A. & Seiferling, N. (2014). Gefährdungsbeurteilung psychischer Belastung (GPB). In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (17. Aufl., S. 635). Bern: Verlag Hans Huber.

Sonntag, Kh., Turgut, S. & Feldmann, E. (2016). Arbeitsbedingte Belastungen erkennen, Stress reduzieren, Wohlbefinden ermöglichen. Ressourcenorientierte Gesundheitsförderung. In Kh. Sonntag (Hrsg), *Personalentwicklung in Organisationen* (4. Auflage). Göttingen: Hogrefe.

Wirtz, M. (2004). Bestimmung der Güte von Beurteilereinschätzungen mittels der Intraklassenkorrelation und Verbesserung von Beurteilereinschätzungen. *Rehabilitation, 43(6)*, 384-389.