Katholische Hochschule Nordrhein-Westfalen
Standort KÖLN
Fachbereich Gesundheitswesen

Bachelor-Thesis zur Erlangung des Grades „Bachelor of Science“

im Studiengang Pflegemanagement

Instrumente zur Beurteilung der Teamperformance – Übersicht und Gütekriterien

vorgelegt von:

Veronika Diamant

am: 26. Mai 2026

Erstprüfer: **Prof. Dr. Andreas Becker**

Zweitprüfer: **Severin Federhen**

Vorwort

Aus Gründen der besseren Lesbarkeit wird auf die Verwendung der Sprachformen aller Geschlechter verzichtet. Die verwendeten Personenbezeichnungen gelten im Sinne des generischen Maskulinums für alle Geschlechter.

Inhaltsverzeichnis

1. Einleitung

Die moderne Gesundheitsversorgung ist durch eine zunehmende Komplexität gekennzeichnet, die sich aus dem Zusammenspiel verschiedener Faktoren ergibt: demografische Entwicklungen, technologische Innovationen, interdisziplinäre Zusammenarbeit sowie steigende Anforderungen an Effizienz und Qualität der Patientenversorgung (St. Pierre, 2018). In Hochrisikobereichen wie Notaufnahmen, Operationssälen oder Intensivstationen sind medizinische Teams mit dynamischen, oft unvorhersehbaren Situationen konfrontiert, in denen präzise Entscheidungen unter Zeitdruck getroffen werden müssen (Gawronski et al., 2022). Die Fähigkeit, in solchen Kontexten effektiv zusammenzuarbeiten, ist dabei nicht allein von technischen Fertigkeiten abhängig, sondern in hohem Maße von sogenannten **Non-Technical Skills (NTS)** – kognitiven, sozialen und persönlichen Ressourcen, die technische Kompetenzen ergänzen (Flin et al., 2008).

Trotz des wachsenden Bewusstseins für die Bedeutung von NTS bleibt die **Patientensicherheit** eine zentrale Herausforderung im Gesundheitswesen. Studien zeigen, dass **70–80 % der medizinischen Fehler** nicht auf mangelnde fachliche Expertise, sondern auf Defizite in der Teamarbeit zurückzuführen sind (Cooper et al., 2010; St. Pierre, 2018). Typische Ursachen sind Kommunikationsfehler (z. B. unklare Anweisungen, fehlendes Closed-Loop-Feedback), mangelnde Situationswahrnehmung, unzureichende Führung oder unklare Rollenverteilungen (Gawronski et al., 2022). Diese Erkenntnisse unterstreichen die Notwendigkeit, NTS systematisch zu erfassen, zu trainieren und zu evaluieren – insbesondere in **simulierten Umgebungen**, die ein sicheres und kontrolliertes Setting für die Analyse von Teamverhalten bieten.

High-Fidelity-Simulationen (HFS) haben sich als wirksames Instrument etabliert, um klinische Szenarien realitätsnah nachzubilden und Teamprozesse zu trainieren (Dieckmann et al., 2017). Im Gegensatz zu traditionellen Ausbildungsmethoden ermöglichen HFS die Reproduktion kritischer Situationen ohne Risiko für Patienten, was sie ideal für die Schulung und Bewertung von NTS macht. Allerdings setzt eine valide und reliable Erfassung von NTS den Einsatz **standardisierter Assessment-Instrumente** voraus, die sowohl theoretisch fundiert als auch praktisch anwendbar sind.

Trotz der Verfügbarkeit zahlreicher Bewertungstools – wie dem **Anaesthetists' Non-Technical Skills (ANTS)**-System (Fletcher et al., 2003), dem **Non-Technical Skills for Surgeons (NOTSS)**-Framework (Yule et al., 2006) oder dem **Team Emergency Assessment Measure**

(TEAM) (Cooper et al., 2010) – bestehen erhebliche Unterschiede in deren **psychometrischen Eigenschaften** (Validität, Reliabilität, Praktikabilität) sowie in ihrer Eignung für verschiedene klinische Kontexte. Viele Instrumente sind entweder zu komplex für den Einsatz in der Praxis, weisen unzureichende Reliabilitätswerte auf oder sind nicht ausreichend für multiprofessionelle Teams validiert (Freytag et al., 2019). Diese Heterogenität erschwert die Auswahl geeigneter Tools und limitiert die Vergleichbarkeit von Studienergebnissen.

Vor diesem Hintergrund stellt sich die Frage, wie NTS in Healthcare-Teams während HFS **systematisch, valide und praxisorientiert** erfasst werden können. Die vorliegende Arbeit adressiert diese Lücke, indem sie bestehende Assessment-Instrumente evaluiert, ihre Stärken und Limitationen analysiert und Empfehlungen für deren Einsatz in klinischen und edukativen Settings ableitet.

2. Zielsetzung

Ziel dieser Bachelorarbeit ist die **systematische Evaluation von Instrumenten zur Erfassung von Non-Technical Skills (NTS) in Healthcare-Teams während High-Fidelity-Simulationen (HFS)**. Im Mittelpunkt steht die Frage, welche Tools sich für die **valide, reliable und praktikable Bewertung von Team-NTS** eignen und wie diese in der klinischen Praxis sowie in der Aus- und Weiterbildung eingesetzt werden können.

Konkret werden folgende **Forschungsfragen** untersucht:

a **Welche Charakteristika weisen publizierte Instrumente zur Messung von Team-NTS in HFS auf?**

- Welche Dimensionen von NTS (z. B. Situationswahrnehmung, Entscheidungsfindung, Kommunikation) werden abgedeckt?
- Wie sind die Instrumente strukturell aufgebaut (z. B. Kategorien, Elemente, Verhaltensanker)?
- Für welche Zielgruppen (z. B. Ärzte, Pflegekräfte, multiprofessionelle Teams) und Settings (z. B. Anästhesie, Chirurgie, Notfallmedizin) sind sie konzipiert?

b **Welche psychometrischen Eigenschaften (Validität, Reliabilität, Praktikabilität) weisen diese Instrumente auf?**

- Wie valide sind die Tools in Bezug auf Inhalts-, Konstrukt- und Kriteriumsvalidität?
- Wie reliabel sind die Instrumente (z. B. Interrater-Reliabilität, interne Konsistenz)?
- Wie praktikabel sind sie für den Einsatz in Simulationen und klinischen Settings (z. B. Zeitaufwand, Schulungsbedarf, Anwendbarkeit in Echtzeit)?

c **Eignen sich die Instrumente für den praktischen Einsatz in klinischen und edukativen Kontexten?**

- Welche Tools sind für **formative Assessments** (z. B. Feedback in Trainings) geeignet?

- Welche eignen sich für **summative Assessments** (z. B. Prüfungen, Zertifizierungen)?
- Wie lassen sich die Instrumente an spezifische Kontexte (z. B. Trauma-Teams, Geburtshilfe) anpassen?

Auf Basis dieser Analyse wird ein **praxisorientiertes Rahmenmodell** entwickelt, das Klinikern, Ausbildern und Forschern als Leitfaden für die Auswahl und Anwendung geeigneter NTS-Assessment-Tools dient. Die Arbeit leistet damit einen Beitrag zur **Standardisierung der NTS-Evaluation** und zur **Verbesserung der Patientensicherheit** durch gezielte Teamtrainings.

3. Methodik, Aufbau und Struktur

Um eine fundierte Analyse der Instrumente zur Erfassung non-technischer Fähigkeiten (NTS) in Healthcare-Teams während High-Fidelity-Simulationen (HFS) zu ermöglichen, stützt sich diese Arbeit auf eine **selektive, nicht-systematische Literaturrecherche**. Angesichts des begrenzten Rahmens einer Bachelorarbeit wurde bewusst eine fokussierte Herangehensweise gewählt, die eine vertiefte Auseinandersetzung mit zentralen Aspekten der Thematik erlaubt. Im Folgenden werden das Studiendesign sowie die Suchstrategie und Auswahlkriterien der verwendeten Literatur dargelegt.

a **Methodik: Studiendesign, Suchstrategie**

Die Literaturrecherche erfolgte primär in Texten, Artikeln, Studien und Büchern über die wissenschaftliche Datenbank PubMed, die Internetsuchmaschine Google Scholar, manuelle Suchanfragen sowie die Bibliothek der Katholischen Hochschule Köln.

Die Suchstrategie umfasste Schlüsselbegriffe wie *„Crew Resource Management (CRM)“, „Nicht-technische Fähigkeiten (Non-Technical Skills, NTS)“, „Simulationstraining“, „Patientensicherheit“* sowie in Kombination mit spezifischen Termini wie *„Teamarbeit“, „Leistungsanalyse“, „Medizinisches Notfallteam“, „Healthcare professionals“, „Team performance“, „observation tool“, „High Fidelity Simulation Training“, „Assessment“, „Evaluation“, „Human Factor*“, „Ressourcenmanagement“, „Ressourcennutzung“, „Aufgabenmanagement“, „menschlicher Fehler“, „nicht-technische Fähigkeit*“, „Intersektorale Zusammenarbeit“, „Gesundheitswesen“, „Führung“, „Entscheidungsfindung“, „Situationsbewusstsein“, „Kommunikation“.* Zur Eingrenzung des Suchraums wurde die Publikationssprache auf Deutsch und Englisch beschränkt.

Die initiale Auswahl der Literatur erfolgte anhand der Titel und Abstracts der identifizierten Publikationen. Ergänzend wurde eine manuelle Handrecherche durchgeführt, um weitere relevante Quellen zu erschließen. Im Rahmen eines Schneeballsystems konnten durch die Analyse der zitierten Literatur zusätzliche einschlägige Werke ermittelt werden. Als zentrale Referenzquellen dienten insbesondere die Publikationen von St. Pierre, darunter „Simulation in der Medizin. Grundlegende Konzepte – Klinische Anwendung“ (2018) sowie „Human Factors und Patientensicherheit in der Akutmedizin“ (2020), die über die Springer-Plattform zugänglich sind.

Im Rahmen der sprachlichen Qualitätssicherung wurde das KI-gestützte Tool *DeepL Write* eingesetzt (DeepL SE, o. J.). Die Anwendung beschränkte sich ausschließlich auf orthografische und stilistische Optimierungen. Sämtliche inhaltlichen Formulierungen, Argumentationsschritte und fachlichen Aussagen wurden eigenständig entwickelt, kritisch reflektiert und verantwortet. Als methodisches Design wurde eine selektive Literaturanalyse gewählt, um den aktuellen Forschungsstand strukturiert aufzuarbeiten und potenzielle Forschungslücken zu identifizieren.

b **Aufbau und Struktur**

Die vorliegende Arbeit ist in mehrere Kapitel unterteilt und widmet sich der systematischen Darstellung von Instrumenten zur Beurteilung der Teamperformance. Besonderes Augenmerk liegt dabei auf der Beschreibung zentraler Bewertungsverfahren sowie der Analyse ihrer Gütekriterien im Hinblick auf eine zuverlässige und valide Erfassung non-technischer Kompetenzen. Zu Beginn werden in Kapitel 1 die Einleitung und in Kapitel 2 die Zielsetzung der Arbeit dargelegt.

Kapitel 4 führt in die zentralen Begriffe und theoretischen Grundlagen ein, die für das Verständnis der in dieser Arbeit behandelten Beurteilungsinstrumente zur Teamperformance erforderlich sind. Nach einer Definition und Einordnung der Non-Technical Skills (NTS) werden Human Factors und Crew Resource Management (CRM) als wesentliche theoretische Rahmenkonzepte erläutert. Anschließend werden Beobachtung, Bewertung und Kompetenzentwicklung im Kontext standardisierter Verfahren sowie die Anforderungen an valide und reliabel einsetzbare Beurteilungsinstrumente beschrieben. Kapitel 5 schließt mit der Darstellung von Behavioral-Marker-Systemen, die als strukturierte, verhaltensbasierte Instrumente zur Bewertung non-technischer Kompetenzen dienen und die zentralen Ergebnisse der Arbeit zusammenführen. In Kapitel 6 erfolgt eine kritische Diskussion der Ergebnisse. Basierend auf den Ergebnissen und der Diskussion werden in Kapitel 7 Empfehlungen für die Praxis formuliert. Kapitel 8 enthält eine Zusammenfassung der Arbeit, während Kapitel 9 das Abstract umfasst. Abschließend folgen das Quellenverzeichnis, der Anhang, das Abbildungsverzeichnis, das Tabellenverzeichnis sowie das Abkürzungsverzeichnis bzw. Glossar.

4. Grundlagen

Im Folgenden werden zentrale Begriffe und Konzepte erläutert, die für das Verständnis und die Wirksamkeit von **Behavioral Marker Systems** wesentlich sind. Diese Systeme werden in der klinischen Praxis, in Simulationstrainings, in der Aus- und Weiterbildung sowie in der Forschung eingesetzt, um **non-technische Kompetenzen** strukturiert zu beobachten, zu bewerten und weiterzuentwickeln (St. Pierre, 2018, S. 172).

Nach einer Definition der **Non-Technical Skills (NTS)** (vgl. Kapitel 4.1) widmet sich Abschnitt 4.2 dem **Crew Resource Management (CRM)** und den **Human Factors**. Darauf aufbauend werden die Beobachtung, Bewertung und Kompetenzentwicklung mithilfe strukturierter Tools dargestellt (vgl. Kapitel 4.3) sowie Anforderungen an qualitativ hochwertige Beurteilungsinstrumente erläutert (vgl. Kapitel 4.4). Abschließend wird das Konzept der **Behavioral Marker Systems** vorgestellt (vgl. Kapitel 4.5).

4.1 Non-Technical Skills (NTS): Definition, Bedeutung, Dimensionen

Im wissenschaftlichen Kontext werden Non-Technical Skills (NTS) – im deutschen Sprachraum als nichttechnische Fertigkeiten bezeichnet – als eine essenzielle Komponente der beruflichen Kompetenz definiert, die maßgeblich über die Sicherheit und Effizienz in Hochrisikoumgebungen entscheidet. Berufliche Kompetenzen umfassen nach St. Pierre (2018) all jene Fähigkeiten, die Personen befähigen, in vertrauten wie auch neuen beruflichen Situationen angemessen zu handeln. Während technische Kompetenzen vergleichsweise gut messbar sind, lassen sich nichttechnische Kompetenzen – wie Führungs-, Team- oder Kommunikationsfähigkeiten – nur über beobachtbares Verhalten erfassen und stellen daher eine deutlich größere Bewertungsherausforderung dar. Für ihre systematische Beschreibung und Messung werden wissenschaftlich fundierte Verhaltensbeobachtungssysteme herangezogen (St. Pierre, 2018, S. 172). Die folgende Darstellung erläutert detailliert die Definition, die klinische sowie systemische Bedeutung und die strukturellen Dimensionen dieses Konzepts.

4.1.1 Definition der nicht-technischen Fähigkeiten

Non-Technical Skills werden fachübergreifend als kognitive, soziale und persönliche Ressourcen definiert, welche die fachlich-technischen Fertigkeiten („technical skills“) ergänzen (Cooper et al., 2010, S. 31; Gawronski et al., 2022). Während technische Fertigkeiten die konkreten prozeduralen Maßnahmen (z. B. eine Intubation oder das Anlegen einer Saugglocke) definieren, ermöglichen NTS deren zuverlässige und situationsgerechte Anwendung unter variierenden und oft belastenden Bedingungen.

Ein zentrales Merkmal der NTS ist ihre Beobachtbarkeit. Es handelt sich nicht um vage Persönlichkeitsmerkmale, sondern um konkrete Verhaltensweisen, die zur Erreichung klinischer Exzellenz oder zur Vermeidung von Fehlern beitragen (St. Pierre, 2018, S. 160). In der modernen Ausbildung werden NTS explizit als Teil der Fachkompetenz begriffen und nicht lediglich als „Soft Skills“ abgetan. Berufliche Kompetenz wird hierbei als die Fähigkeit verstanden, in vertrauten wie in neuartigen Situationen situationsangemessen reaktions- und handlungsfähig zu sein (St. Pierre, 2018, S. 160).

4.1.2 Bedeutung für die Patientensicherheit und Leistung

Die Relevanz von NTS wurde historisch durch Untersuchungen der NASA in der Luftfahrt begründet, die zeigten, dass ca. 70 % aller Unfälle nicht auf mangelndes Fachwissen, sondern auf Defizite in der Teaminteraktion, Kommunikation und Entscheidungsfindung zurückzuführen waren (Wauben et al., 2011, S. 159–160). Diese Erkenntnisse ließen sich direkt auf die Medizin übertragen:

- **Fehlerprävention** (Gawronski et al., 2022, S. 2)**:** Studien belegen, dass menschliche Faktoren (Human Factors) zu 43–70 % an unerwünschten Ereignissen in der Akutmedizin beteiligt sind. NTS fungieren hierbei als Sicherheitsnetz und Fehlerabwehrmaßnahmen (engl. „error countermeasures“) indem sie die Eskalation kleinerer Probleme verhindern (St. Pierre, 2018, S. 171).
- **Korrelation mit klinischer Leistung** (Brogaard et al., 2024, S. 3)**:** Untersuchungen bei Vakuumextraktionen zeigten eine Dosis-Wirkungs-Beziehung: Teams mit exzellenten NTS-Werten erreichten mit einer Wahrscheinlichkeit von 81 % eine hohe Leitlinienadhärenz, während Teams mit durchschnittlichen NTS dies nur in 12 % der Fälle taten.

- **Kompensation von Belastungsfaktoren** (St. Pierre, 2020, S. 303)**:** In komplexen Umgebungen wie Notaufnahmen oder Operationssälen (OPs) können NTS die negativen Auswirkungen von Zeitdruck, Unterbrechungen und hoher Arbeitslast abmildern, indem sie durch effektive Teamarbeit die kognitiven Ressourcen einzelner Teammitglieder entlasten.

4.1.3 Dimensionen der Non-Technical Skills

Die Struktur der NTS wird in der Literatur konsistent in drei Hauptbereiche unterteilt, die wiederum in spezifische Elemente gegliedert werden können (St. Pierre, 2018, S. 162).

A. Kognitive Fähigkeiten

Diese umfassen die mentalen Prozesse der Informationsverarbeitung und Planung:

I. **Situationsbewusstsein (Situation Awareness)** (Wauben et al., 2011, S. 161)**:** Das Sammeln von Informationen, das Verstehen der aktuellen Lage und die Antizipation zukünftiger Zustände

- Überwacht den Zustand von Patienten
- Überwacht Teammitglieder, um Sicherheit zu gewährleisten und Fehler zu verhindern
- Überwacht die Umgebung hinsichtlich Sicherheit und Verfügbarkeit von Ressourcen (z. B. Ausrüstung)
- Überwacht den Fortschritt in Richtung Ziel und identifiziert Veränderungen, die den Versorgungsplan verändern könnten
- Fördert die Kommunikation, um sicherzustellen, dass die Teammitglieder ein gemeinsames mentales Modell haben

II. **Entscheidungsfindung (Decision Making)** (St. Pierre, 2018, S. 162)**:** Das Erkennen von Optionen, das Abwägen von Risiken sowie die ständige Re-Evaluation des gewählten Plans

III. **Aufgabenmanagement (Task Management):** Die Planung, Prioritätensetzung, Ressourcennutzung und das Einhalten von Standards

B. Soziale und interpersonelle Fähigkeiten

Hierbei steht die Interaktion innerhalb des multidisziplinären Teams im Vordergrund:

I. **Kommunikation** (St. Pierre, 2018, S. 170)**:** Der Austausch präziser Informationen, die Nutzung von „Closed-Loop"-Kommunikation und die Bildung gemeinsamer mentaler Modelle („Shared Mental Models")

- Gibt kurze, klare, spezifische und zeitgerechte Informationen an Teammitglieder
- Sucht Informationen aus allen verfügbaren Quellen
- Verwendet Rückbestätigungen (Check-backs), um übermittelte Informationen zu verifizieren
- Nutzt SBAR, Call-outs und Übergabetechniken für eine effektive Kommunikation mit Teammitgliedern

II. **Teamarbeit und Koordination:** Gegenseitige Unterstützung, Rollenklärung und das Lösen von Konflikten

- Stellt ein Team zusammen
- Weist Rollen und Verantwortlichkeiten der Teammitglieder zu bzw. macht sie kenntlich
- Zieht Teammitglieder zur Verantwortung
- Bezieht Patienten und Angehörige als Teil des Teams ein
- Bietet aufgabenbezogene Unterstützung und Hilfe
- Gibt zeitnahes und konstruktives Feedback an Teammitglieder
- Setzt sich wirksam für die Patientensicherheit ein (z. B. mittels durchsetzungsfähiger Aussage/Assertive Statement, Two-Challenge Rule oder CUS)
- Nutzt die Two-Challenge Rule oder das DESC-Skript zur Konfliktlösung

III. **Führung (Leadership) und Followership:** Die Übernahme von Verantwortung, Aufgabenverteilung und das proaktive Einbringen als Teammitglied

- Identifiziert Teamziele und -vision
- Setzt Ressourcen effizient ein, um die Teamleistung zu maximieren
- Gleicht die Arbeitsbelastung im Team aus
- Delegiert Aufgaben bzw. Aufträge, sofern angemessen
- Führt Briefings, Huddles und Debriefings durch
- Lebt Teamverhalten vor

C. Persönliche Ressourcen

Diese Dimension betrifft die individuelle Selbststeuerung und Widerstandsfähigkeit:

I. **Stress- und Müdigkeitsmanagement:** Die Fähigkeit, die eigene Leistungsfähigkeit unter Druck aufrechtzuerhalten und Anzeichen von Überlastung zu erkennen

II. **Metakognition:** Die Fähigkeit zur Selbstreflexion über die eigenen Denkprozesse, um beispielsweise Fixierungsfehler frühzeitig zu erkennen

Non-Technical Skills bilden das strukturgebende Gerüst, das die sichere Anwendung medizinischer Expertise in komplexen, dynamischen und risikoreichen Systemen erst ermöglicht.

4.2 Crew Resource Management (CRM) / Human Factors

Das Konzept des **Crew Resource Management (CRM)** und die wissenschaftliche Disziplin der **Human Factors** bilden das theoretische Rückgrat für das Verständnis von Sicherheit in Hochrisikoumgebungen wie der Akutmedizin (Kelly et al., 2023, S. 3). Während Human Factors die Interaktion zwischen Menschen und anderen Elementen eines Systems untersucht, stellt CRM die praktische Anwendung dieser Erkenntnisse zur Optimierung der Teamleistung dar (St. Pierre, 2018, S. 418).

4.2.1 Ursprung in der Luftfahrt

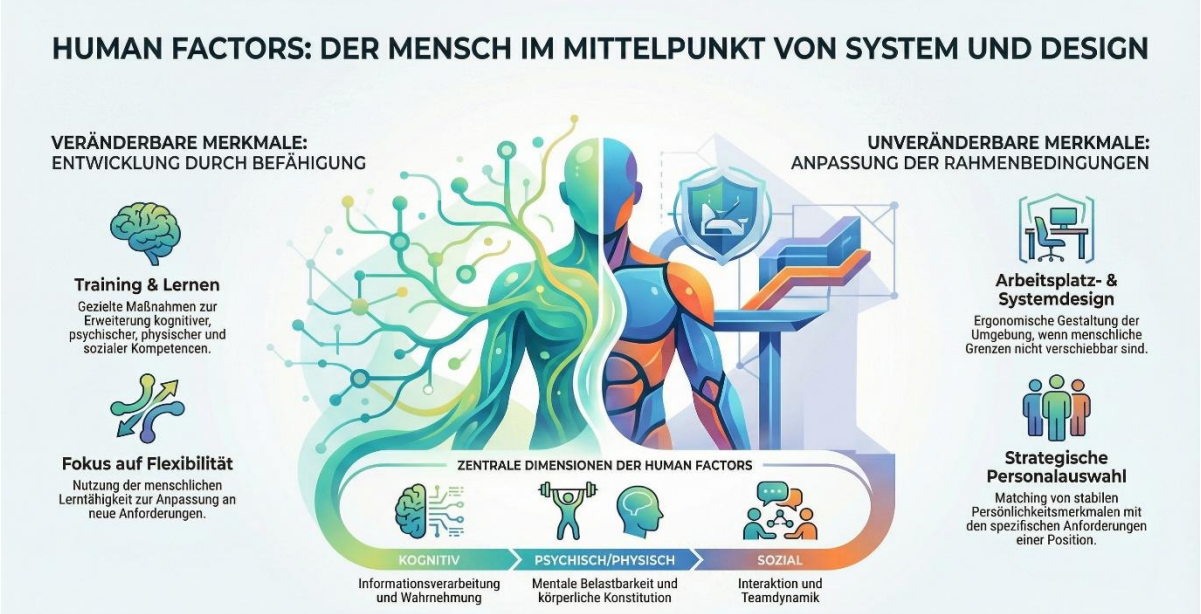
Die Ursprünge des CRM liegen in der zivilen und militärischen Luftfahrt am Ende der 1970er-Jahre. Untersuchungen der NASA zeigten damals auf, dass ca. 70 % bis 80 % aller Flugunfälle nicht auf mangelndes technisches Wissen oder Geräteversagen, sondern auf Defizite in der Kommunikation, Teaminteraktion und Entscheidungsfindung zurückzuführen waren. Als Reaktion wurde das ursprüngliche „Cockpit Resource Management“ entwickelt, welches später zum umfassenderen „Crew Resource Management“ erweitert wurde, um alle an Bord verfügbaren Ressourcen (Mensch, Technik und Information) optimal zu nutzen.

4.2.2 Übertragung auf Medizin und Pflege

Seit den 1990er-Jahren erfolgte eine erfolgreiche Adaption dieser Konzepte auf die Akutmedizin, wobei insbesondere die Anästhesiologie mit der Entwicklung des „Anesthesia Crisis Resource Management“ (ACRM) eine Pionierrolle einnahm. Dieser Transfer war motiviert durch die Erkenntnis, dass auch im Gesundheitswesen schätzungsweise 70 % der Behandlungsfehler auf menschliche Faktoren zurückzuführen sind, die oft in komplexen, multidisziplinären Settings auftreten. Krankenhäuser, insbesondere Notfallstationen und OPs, werden heute als **Hochzuverlässigkeitsorganisationen (HRO)** begriffen, in denen CRM als „Kodex der Zu-

sammenarbeit“ fungiert, um trotz hoher Dynamik und Zeitdruck Patientensicherheit zu gewährleisten (St. Pierre, 2018, S. 181).

Abbildung 1: Training vs. menschliche Eigenschaften



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. St. Pierre (2018, S. 181)

4.2.3 Zentrale Prinzipien des CRM

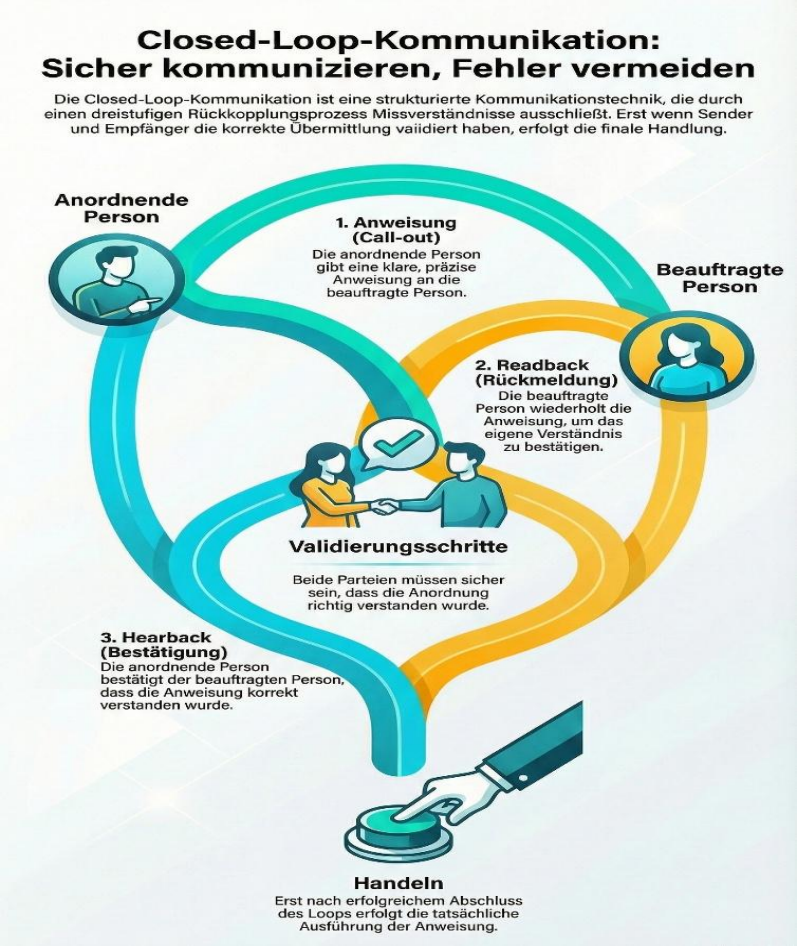
Die Prinzipien des CRM zielen darauf ab, die Leistung des Teams durch die gezielte Anwendung nichttechnischer Fertigkeiten zu steigern (vgl. Abb. 1). Zu den wesentlichen Kernbereichen gehören:

- **Kommunikation:** Hierbei stehen Techniken wie die geschlossene Kommunikation (**Closed-Loop** bzw. Read-back/Hear-back) zur Vermeidung von Missverständnissen sowie strukturierte Übergabewerkzeuge wie **SBAR** (Situation, Background, Assessment, Recommendation) im Vordergrund.

Geschlossene Kommunikationsschleife aus Anweisung, „hearback“ und „readback“. Beide Kommunikationspartner sind sich sicher, dass ihr Verständnis der Situation nicht auf Vermu-

tungen oder Erwartungen basiert, sondern vom Kommunikationspartner explizit bestätigt wurde (vgl. Abb. 2).

Abbildung 2: Closed-Loop Communication



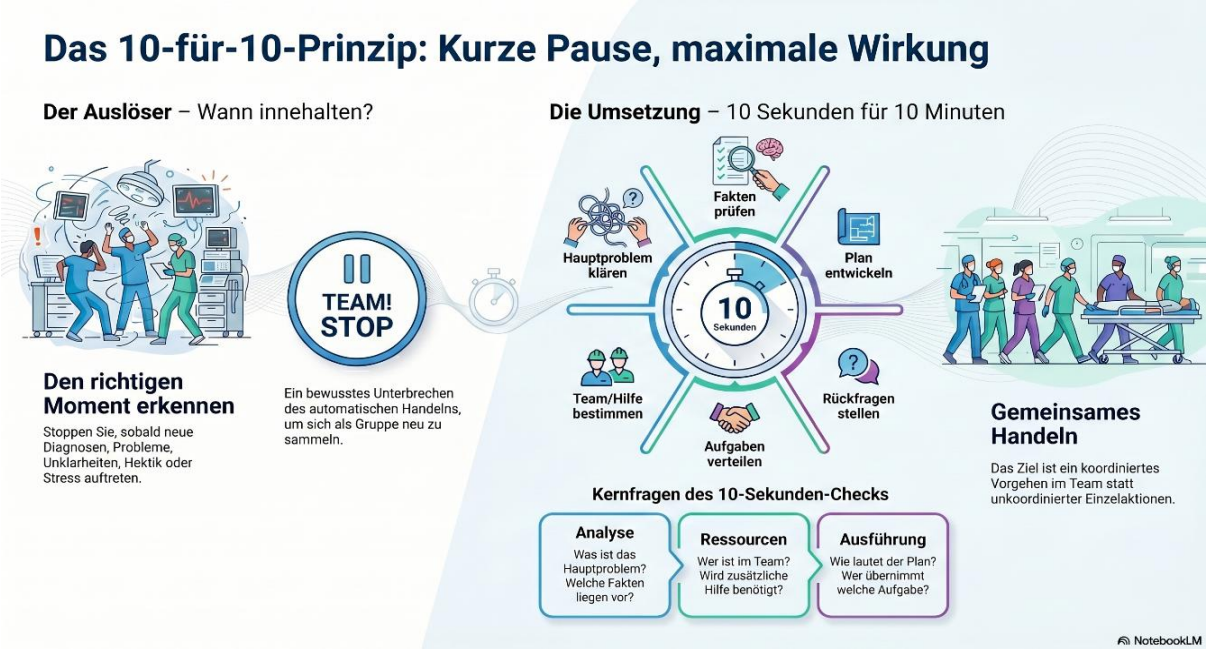
Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. St. Pierre (2020, S. 252)

- **Führung und Teamwork:** CRM fordert eine klare Rollenverteilung zwischen Führung („Teamleader“) und Geführten („Followership“) sowie die aktive Einbeziehung aller Teammitglieder. Wesentliche Werkzeuge sind hierbei Briefings zur Planung und Debriefings zur Reflexion der Teamleistung (St. Pierre, 2020, S. 253, 2020, S. 372).
- **Situationsbewusstsein:** Dies umfasst das Sammeln von Informationen, das Verstehen der aktuellen Lage und die Antizipation zukünftiger Entwicklungen. Ein bekanntes

Prinzip ist das **„10-für-10-Prinzip“** (10 Sekunden für die nächsten 10 Minuten), welches zur regelmäßigen Reevaluation der Situation aufruft (Kranz & Regener, 2023, S. 26).

- **Entscheidungsfindung:** Die Nutzung strukturierter Modelle (z. B. FOR-DEC oder DECIDE) hilft dabei, unter Stress fundierte Entscheidungen zu treffen und Fixierungsfehler zu vermeiden (St. Pierre, 2020, S. 202).

Abbildung 3: 10-Sekunden-für-10-Minuten-Prinzip



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. www.inpass.de

Ziel: Fehlerreduktion, Teamleistung und Sicherheit Das primäre Ziel des CRM ist die Reduktion vermeidbarer Fehler durch die Stärkung der interpersonellen und kognitiven Fertigkeiten (Gawronski et al., 2022, S. 2). Durch die Etablierung einer proaktiven **Sicherheitskultur** sollen Teams in die Lage versetzt werden, Fehler frühzeitig zu erkennen (Detection) und deren Auswirkungen abzuschwächen (Mitigation), bevor ein Schaden am Patienten entsteht. CRM-Trainings fördern zudem die Resilienz der Organisation, indem sie die Flexibilität und Adaptivität des Handelns in unvorhergesehenen Krisensituationen stärken(St. Pierre, 2020, S. 352).

4.2.4 CRM als theoretische Basis für Behavioral Marker Systems

Da CRM-Fertigkeiten sich in konkreten, beobachtbaren Verhaltensweisen äußern, bildet CRM die direkte theoretische Grundlage für die Entwicklung von **Behavioral Marker Systems** (BMS). BMS wie ANTS (Anästhesie) oder NOTSS (Chirurgie) nutzen die Taxonomien des CRM, um diese sonst schwer fassbaren Kompetenzen durch geschulte Rater objektiv messbar und bewertbar zu machen. Damit fungiert CRM als inhaltlicher Rahmen, während die Behavioral Marker Systems das methodische Instrumentarium zur Evaluation und gezielten Weiterentwicklung dieser überlebenswichtigen Fertigkeiten bereitstellen (St. Pierre, 2018, S. 172).

4.3 Beobachtung, Bewertung und Kompetenzentwicklung

Die systematische Erfassung von Non-Technical Skills (NTS) stellt eine Grundvoraussetzung für die Professionalisierung und Patientensicherheit in der modernen Akutmedizin dar (Gawronski et al., 2022, S. 5). Im Folgenden wird dargelegt, warum die Beobachtung dieser Fertigkeiten unerlässlich ist, weshalb subjektive Einschätzungen als unzureichend gelten und wie strukturierte Instrumente den Lernprozess in Simulation und Praxis unterstützen.

4.3.1 Notwendigkeit der Beobachtung von NTS

Berufliche Kompetenz wird wissenschaftlich als die Fähigkeit definiert, sowohl in vertrauten als auch in neuartigen Situationen situationsangemessen reaktions- und handlungsfähig zu sein (St. Pierre, 2018, S. 160). Da NTS – im Gegensatz zu manifesten Merkmalen wie dem Alter – sogenannte latente Konstrukte darstellen, sind sie nicht unmittelbar messbar. Sie müssen stattdessen aus dem beobachtbaren Handeln erschlossen und darauf aufbauend evaluiert werden (St. Pierre, 2018, S. 161). Die explizite Beobachtung ist zudem notwendig, um den Theorie-Praxis-Transfer zu sichern und sicherheitsrelevante Verhaltensweisen wie Kommunikation, Führung und Teamwork gezielt lehr- und weiterentwickelbar zu machen. Nur durch die Identifikation spezifischer Verhaltensweisen können diese als „Error Countermeasures“ (Fehlergegenmaßnahmen) im klinischen Alltag verankert werden (Fletcher et al., 2003, S. 581).

4.3.2 Insuffizienz subjektiven Feedbacks

Obwohl erfahrene Kliniker klinische Exzellenz oft intuitiv erkennen („man sieht sie, wenn man sie sieht“), fällt es ohne strukturelle Hilfsmittel schwer, präzise zu benennen, welche konkreten Fertigkeiten für diese Leistung verantwortlich sind (St. Pierre, 2018, S. 168). Ein rein subjekti-

ves Feedback, das auf dem „Bauchgefühl“ basiert, ist für eine fundierte Kompetenzentwicklung aus mehreren Gründen ungeeignet:

- **Wahrnehmungsverzerrungen:** Die menschliche Wahrnehmung unterliegt komplexen Verzerrungen. Häufige Urteilsfehler sind **der Halo-Effekt**, bei dem ein hervorstechendes Merkmal die gesamte Bewertung überstrahlt, **die Tendenz zur** Mitte bei Unsicherheit sowie **der Milde-Effekt** aufgrund von Empathie.
- **Diskrepante Wahrnehmung**: Studien zeigen, dass verschiedene Berufsgruppen (z. B. Chirurgen vs. OP-Pflege) dieselbe Situation im OP hinsichtlich Kommunikation und Teamarbeit fundamental unterschiedlich wahrnehmen und bewerten.
- **Unpräzises Vokabular**: Ohne Struktur bleibt Feedback oft „schwammig“ und konzentriert sich instinktiv auf technisch-fachliche Aspekte, während die entscheidenden interpersonellen Faktoren unberücksichtigt bleiben.

4.3.3 Notwendigkeit standardisierter Instrumente

Um diese Defizite zu kompensieren, ist der Einsatz standardisierter Instrumente (wie Behavioral Marker Systems) zwingend erforderlich. Diese Tools dienen der Objektivierung und intersubjektiven Nachprüfbarkeit von Leistungen. Durch methodische Regeln und eine vollständige Dokumentation (Transparenz) wird sichergestellt, dass Beobachtungen unabhängig vom jeweiligen Rater verlässlich sind. Zudem stellen diese Instrumente **ein einheitliches Vokabular** zur Verfügung. Diese „gemeinsame Sprache“ ist die Basis für eine strukturierte Ausbildungskommunikation und ermöglicht es, mentale Modelle innerhalb eines Teams abzugleichen. Validierte Messinstrumente erlauben es zudem, die Effektivität von Trainingsmaßnahmen objektiv zu evaluieren.

4.3.4 Beobachtung und Bewertung in verschiedenen Settings

- **Simulation**: Hier dient die Beobachtung primär dem **formativen Assessment**, also der Rückmeldung zur individuellen Entwicklung. Simulationen schaffen einen geschützten Lernraum, in dem Fehler als wertvolles Lernwerkzeug genutzt werden können, ohne Patienten zu gefährden. Das anschließende **Debriefing** gilt als „Herz und Seele“ des Trainings; hier werden mittels Techniken wie „Advocacy and Inquiry“ die den Handlungen zugrunde liegenden mentalen Modelle analysiert (St. Pierre, 2018, S. 463).

- **Ausbildung**: Der Einsatz von Checklisten und Logbüchern unterstützt den Paradigmenwechsel von „See One, Do One“ hin zu **„See One, Practice Many, Do One“**. Instrumente wie der OSCE (Objective Structured Clinical Examination) ermöglichen dabei eine objektive Bewertung klinisch-praktischer Fertigkeiten (St. Pierre, 2018, S. 108).
- **Klinische Praxis**: Im Alltag erschweren Zeitmangel und hierarchische Strukturen oft ein unmittelbares Feedback. Strukturierte Beobachtungstools können hier helfen, den Fokus auf spezifische NTS (z. B. Entscheidungsfindung) zu legen, sobald der Ausbildungsstatus der Mitarbeiter dies zulässt (St. Pierre, 2018, S. 169).

Zusammenfassend lässt sich festhalten, dass erst durch die Überführung vager Verhaltenskonzepte in messbare, standardisierte Kategorien eine systematische Kompetenzentwicklung möglich wird. Dies schafft die logische Notwendigkeit für die im nächsten Abschnitt behandelten Behavioral Marker Systems.

4.4 Anforderungen an Beurteilungsinstrumente

Die Evaluation von **Non-Technical Skills (NTS)** erfordert Instrumente, die hohen psychometrischen Standards genügen, um valide Aussagen über die Kompetenz von Individuen oder Teams treffen zu können (Bortz & Döring, 2006, S. 193). In der wissenschaftlichen Literatur werden hierfür primär die klassischen Testgütekriterien sowie anwendungsspezifische Anforderungen wie Praktikabilität und Beobachtbarkeit herangezogen (Bortz & Döring, 2006, S. 195).

a **Objektivität**

Objektivität bezeichnet das Ausmaß, in dem die Ergebnisse eines Beurteilungsinstruments **unabhängig vom Testanwender** (Beobachter oder Rater) sind. In der medizinischen Simulation und Praxis wird zwischen drei Formen differenziert:

- **Durchführungsobjektivität:** Das Ergebnis darf nicht durch das Verhalten des Untersuchungsleiters beeinflusst werden. Dies wird durch standardisierte Instruktionen und einheitliche Rahmenbedingungen (z. B. vordefinierte Simulationsszenarien) sichergestellt.

- **Auswertungsobjektivität:** Verschiedene Rater müssen bei derselben Leistung zur identischen Punktzahl gelangen. Strukturierte Bewertungsschemata und dichotome Codierungen (Ja/Nein) minimieren hierbei den individuellen Spielraum.
- **Interpretationsobjektivität:** Ein vorab festgelegter Bewertungsmaßstab (z. B. Normwerte oder klare Bestehensgrenzen) stellt sicher, dass aus denselben Daten identische Schlussfolgerungen gezogen werden.

b **Reliabilität**

Die Reliabilität (Zuverlässigkeit) kennzeichnet den Grad der **Messgenauigkeit** eines Instruments. Ein Instrument ist reliabel, wenn es unter gleichen Bedingungen konsistente Ergebnisse liefert.

- **Interrater-Reliabilität:** In der NTS-Forschung ist dies das kritischste Kriterium. Sie wird häufig mittels der **Intraklassenkorrelation (ICC)** oder des **Kappa-Koeffizienten** berechnet, um die Übereinstimmung zwischen verschiedenen Beobachtern zu quantifizieren. Werte über 0,80 gelten als gut, wobei in der Praxis oft Werte ab 0,70 als akzeptabel angesehen werden.
- **Methoden zur Steigerung:** Um eine hohe Reliabilität zu erreichen, ist ein **systematisches Ratertraining** einschließlich einer Kalibrierung an Expertenstandards unerlässlich, um Effekte wie die „Hawk/Dove“-Tendenz (systematisch zu strenger oder milder Bewertung) zu vermeiden.

c **Validität**

Die Validität ist das wichtigste Gütekriterium und gibt an, ob ein Instrument tatsächlich das misst, was es zu messen vorgibt.

- **Inhaltsvalidität:** Die Items müssen das Zielkonstrukt (z. B. Teamarbeit) in all seinen Facetten repräsentieren. Dies wird oft durch Expertenkonsens (z. B. Delphi-Verfahren) bei der Tool-Entwicklung gesichert.

- **Kriteriumsvalidität:** Hierbei wird der Zusammenhang zwischen dem NTS-Testergebnis und einem Außenkriterium (z. B. der klinischen Behandlungsqualität) untersucht. Studien belegen eine Korrelation: Teams mit hohen NTS-Werten erreichen mit 81 % Wahrscheinlichkeit eine hohe klinische Performance.
- **Konstruktvalidität:** Sie prüft, ob das Instrument theoriebasierte Hypothesen bestätigt, etwa durch den Nachweis von konvergenter Validität (Übereinstimmung verschiedener Methoden für dasselbe Konstrukt).

d **Praktikabilität**

Die Praktikabilität (Feasibility) bewertet, ob ein Instrument im klinischen Alltag oder in der Ausbildung **handhabbar** ist. Wesentliche Faktoren sind:

- **Zeitaufwand:** Ein ideales Tool zur unmittelbaren Rückmeldung sollte schnell ausfüllbar sein (z. B. in unter einer Minute).
- **Benutzerfreundlichkeit:** Das Design muss für die Anwender intuitiv und die Sprache verständlich sein. In Studien wird die Praktikabilität oft durch Surveys erhoben, wobei Tools wie TEAM oder ANTS hohe Usability-Raten aufweisen.

e **Beobachtbarkeit**

Da Kompetenzen latente Konstrukte sind, können sie nur aus dem **beobachtbaren Handeln** erschlossen werden. Ein gutes Beurteilungsinstrument muss sich daher strikt auf Verhaltensweisen beschränken, die tatsächlich sichtbar oder hörbar sind.

- Verhalten, das in einer spezifischen Situation nicht erforderlich war oder nicht gezeigt wurde, darf explizit nicht bewertet werden (Kennzeichnung als „Nicht beobachtbar“).
- Das Prinzip der Beobachtbarkeit schließt Wertungen aus, die auf bloßem „Bauchgefühl“ oder nicht belegbaren Vermutungen über die innere Einstellung basieren.

f **Klare Verhaltensanker**

Klare Verhaltensanker (Behavioral Markers) sind das Herzstück strukturierter Beobachtungssysteme. Es handelt sich um **konkrete Beispiele für positives und negatives Verhalten**, die den Skalenstufen zugeordnet sind.

- **Funktion:** Sie dienen als Referenzpunkte, die den Interpretationsspielraum für die Rater minimieren und eine „gemeinsame Sprache“ im Team etablieren.
- **Beispiel:** Für das Element „Prioritätensetzung“ wäre ein positiver Marker das „Verbalisieren der wichtigsten Punkte eines Falles“, während ein negativer Marker die „Ablenkung durch die Anleitung Auszubildender“ wäre.
- **Nutzen:** Durch diese Konkretisierung können typische Beobachtungsfehler wie der **Halo-Effekt** (ein Merkmal überstrahlt die Gesamtbewertung) oder die **Tendenz zur Mitte** wirksam reduziert werden.

Zusammenfassend lässt sich festhalten, dass ein hochwertiges Beurteilungsinstrument als Oberbegriff für Werkzeuge zur Bewertung, beispielsweise Checklisten, Ratingskalen, Fragebögen, Beobachtungsbögen oder Tests, die subjektive Wahrnehmung durch Standardisierung möglichst objektivieren, durch Verhaltensanker im Sinne von Behavioral Marker präzisieren und zugleich in der praktischen Anwendung ökonomisch ausgestaltet sein sollte.

4.5 Behavioral Marker Systems

Behavioral Marker Systems (BMS) stellen das methodische Bindeglied zwischen theoretischen Konzepten der Human Factors und der praktischen Messbarkeit menschlicher Leistung dar. Sie dienen dazu, die sonst schwer fassbaren **nichttechnischen Fertigkeiten (NTS)** in Hochrisikoumgebungen wie der Medizin objektivierbar, lehrbar und evaluierbar zu machen (St. Pierre, 2018, S. 172).

a **Definition und Zielsetzung**

Wissenschaftlich werden Behavioral Marker als **„beobachtbare, nichttechnische Verhaltensweisen“** definiert, die maßgeblich zu einer exzellenten oder unzureichenden Arbeitsleistung beitragen (Fletcher et al., 2003, S. 581). Da Konzepte wie „Teamfähigkeit“ oder „Situationsbewusstsein“ sogenannte **latente Konstrukte** sind, die nicht direkt gemessen werden können, fungieren BMS als Indikatoren, um diese Kompetenzen aus dem sichtbaren Handeln abzuleiten.

Das **primäre Ziel** dieser Systeme ist die Schaffung einer **„gemeinsamen Sprache“ (Common Language)** für multidisziplinäre Teams. BMS sollen:

- einen strukturierten Rahmen für das **Feedback und Debriefing** nach Simulationen oder Realeinsätzen bieten,
- als **„Error Countermeasures“** (Fehlergegenmaßnahmen) dienen, indem sie sicherheitskritisches Verhalten identifizierbar machen,
- die **objektive Messung** von Kompetenzentwicklungen in der Aus- und Weiterbildung ermöglichen.

b **Struktur und Aufbau**

Ein Behavioral Marker System folgt in der Regel einer **hierarchischen Struktur**, die meist in drei Ebenen unterteilt ist:

1. **Kategorien:** Übergreifende Cluster von Fertigkeiten (z. B. Teamarbeit, Entscheidungsfindung).
2. **Elemente:** Spezifische Untergruppen innerhalb einer Kategorie, die das Konstrukt präzisieren (z. B. „Prioritätensetzung“ innerhalb des Aufgabenmanagements).
3. **Verhaltensmarker (Behavioral Markers):** Konkrete Verhaltensbeispiele, die als **Ankerpunkte** für positives und negatives Handeln dienen.

Die Bewertung erfolgt meist auf einer **vierstufigen Likert-Skala** (z. B. 1 = schlecht bis 4 = gut). Eine essenzielle Komponente ist die Option **„NB“ (Nicht beobachtbar)**, die sicherstellt, dass nur Verhalten bewertet wird, das in der spezifischen Situation tatsächlich gefordert war oder gezeigt wurde.

c **Etablierte Beispiele**

Für verschiedene medizinische Fachdisziplinen wurden spezifische Systeme entwickelt, da Verhaltensmarker den jeweiligen Arbeitskontext widerspiegeln müssen:

- **ANTS (Anaesthetists’ Non-Technical Skills):** Eines der ersten medizinischen Systeme, gegliedert in Aufgabenmanagement, Teamarbeit, Situationsbewusstsein und Entscheidungsfindung.
- **NOTSS (Non-Technical Skills for Surgeons):** Speziell für Chirurgen entwickeltes System zur Bewertung von Führung, Kommunikation und kognitiven Prozessen im OP.
- **NOTECHS:** Ursprünglich in der Luftfahrt zur Bewertung von Piloten entwickelt und später für die Chirurgie adaptiert.

- **TEAM (Team Emergency Assessment Measure):** Ein Tool, das primär die Leistung des **gesamten Teams** in Notfallsituationen bewertet, statt nur die individuellen Fertigkeiten einzelner Akteure.
- **SPLINTS:** Ein System für die spezifischen nichttechnischen Fertigkeiten der Operationspflege.

d **Einsatzbereiche**

Die Einsatzgebiete von BMS haben sich weit über die Anästhesiologie hinaus ausgedehnt. Sie werden heute im **Operationssaal**, in der **Notfallmedizin**, der **Geburtshilfe** sowie in der **Pädiatrie** angewendet. Im militärischen Sanitätsdienst (z. B. **Bundeswehr**) werden BMS zudem um spezifische Marker erweitert, wie etwa die Kommunikation unter Kampfbedingungen oder den Umgang mit extrem begrenzten Ressourcen.

e **Nutzen für Praxis, Ausbildung und Forschung**

- **Nutzen für die Praxis:** BMS unterstützen die **Sicherheitskultur** in Kliniken, indem sie strukturierte Fallanalysen (z. B. Morbidity & Mortality Meetings) ermöglichen. Sie dienen als Werkzeug zur Identifikation systemischer Schwachstellen und latenter Fehler.
- **Nutzen für die Ausbildung:** BMS sind das „Herzstück“ des simulationsbasierten Lernens. Sie erlauben es Instruktoren, über das „Bauchgefühl“ hinaus präzises, **evidenzbasiertes Feedback** zu geben. Zudem helfen sie, individuelle Trainingsbedarfe im Rahmen einer Standortbestimmung zu identifizieren.
- **Nutzen für die Forschung:** In der Wissenschaft dienen BMS als validierte Messinstrumente, um die **Wirksamkeit von Trainingsinterventionen** nachzuweisen. Studien konnten so eine „Dosis-Wirkungs-Beziehung“ belegen: Teams mit exzellenten NTS-Werten (gemessen via BMS) haben eine signifikant höhere Wahrscheinlichkeit (81 % vs. 12 %), eine hohe klinische Behandlungsqualität zu erreichen.

5. Ergebnisse: Instrumente zur Evaluation von Non-Technical Skills

Nicht-technische Fähigkeiten (NTS) stellen einen wesentlichen Bestandteil professionellen Handelns in sicherheitskritischen medizinischen Kontexten dar. In Fachbereichen wie der Anästhesiologie, Chirurgie, Notfallmedizin, Geburtshilfe oder Reanimationsversorgung hängt die Qualität der Patientenversorgung nicht allein von technischem Wissen und praktischen Fertigkeiten ab, sondern in erheblichem Maß auch von Kompetenzen wie Situationsbewusstsein, Entscheidungsfindung, Kommunikation, Teamarbeit und Führung. Defizite in diesen Bereichen stehen in engem Zusammenhang mit Behandlungsfehlern, unerwünschten Ereignissen und einer eingeschränkten Patientensicherheit. Dennoch wurden nicht-technische Fähigkeiten in der medizinischen Aus- und Weiterbildung lange Zeit nicht in gleichem Maß systematisch berücksichtigt wie technische Fertigkeiten.

Vor diesem Hintergrund wurden verschiedene Instrumente entwickelt, um nicht-technische Fähigkeiten strukturiert zu beobachten, zu bewerten und im Rahmen von Feedback- und Trainingsprozessen gezielt zu fördern. Insbesondere Behavioral Marker Systems (BMS) bieten hierfür einen verhaltensorientierten und systematischen Zugang. Sie ermöglichen es, abstrakte Kompetenzbereiche in beobachtbare Verhaltensweisen zu überführen und damit für Ausbildung, Simulation und Leistungsbeurteilung nutzbar zu machen. Die verfügbaren Instrumente unterscheiden sich jedoch hinsichtlich ihres theoretischen Hintergrunds, ihrer Zielgruppe, ihres Aufbaus, ihres Detaillierungsgrads, ihrer psychometrischen Eigenschaften sowie ihrer Einsatzbereiche. Während einige Verfahren professionsspezifisch, etwa für Anästhesisten oder Chirurgen, entwickelt wurden, sind andere teambezogen oder interprofessionell angelegt.

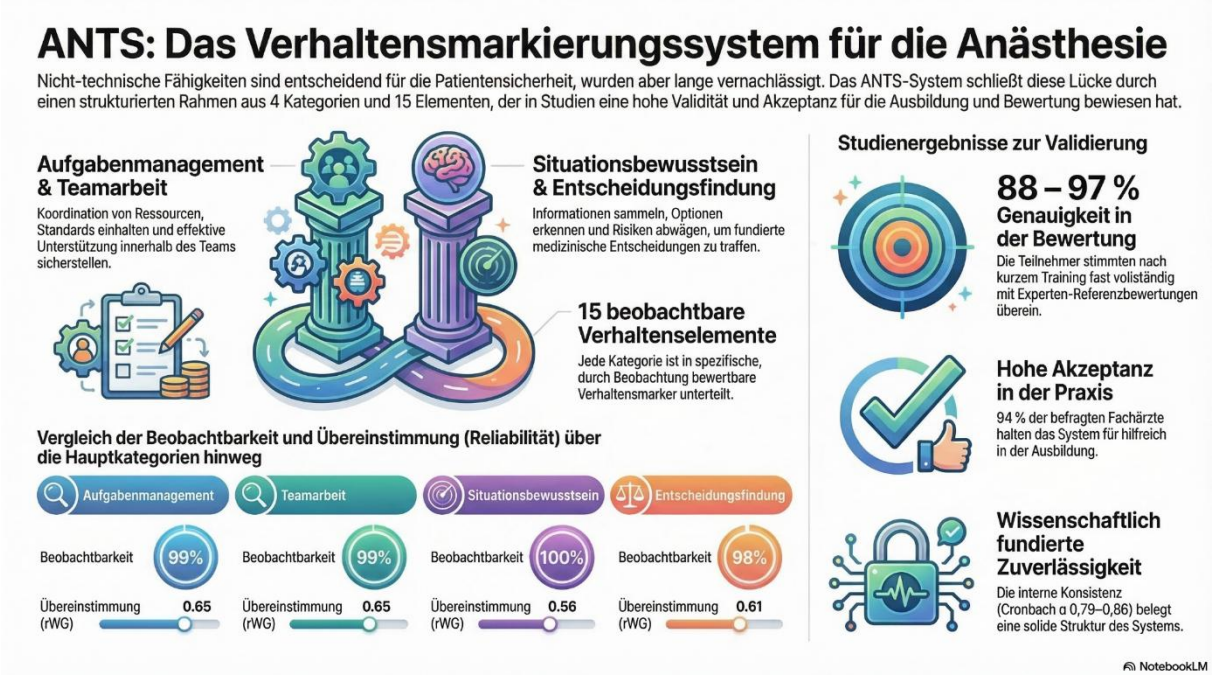
Ziel dieses Kapitels ist eine vergleichende Analyse ausgewählter Instrumente zur Erfassung nicht-technischer Fähigkeiten in unterschiedlichen medizinischen Kontexten. Im Fokus stehen dabei ihre Entwicklung, strukturellen Merkmale, zugrunde liegenden Kompetenzdimensionen, psychometrischen Eigenschaften, Anwendungsbereiche sowie Limitationen. Auf diese Weise sollen Gemeinsamkeiten und Unterschiede der Instrumente systematisch herausgearbeitet und ihre Eignung für die klinische Praxis, die medizinische Ausbildung, simulationsbasierte Trainings sowie formative und gegebenenfalls summative Beurteilungsverfahren bewertet werden. Für eine einheitliche Darstellung werden Behavioral-Marker-Systeme in ihrer hierar-

chischen Struktur aus Kategorien, Elementen und Verhaltensmarkern beschrieben (vgl. Kapitel 4.5 Behavioral Marker Systems, b. Struktur und Aufbau).

5.1 Anaesthetists' Non-Technical Skills (ANTS): Bewertung eines Verhaltensmarkierungssystems für die Anästhesie

Quelle: Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' non-technical skills (ANTS): evaluation of a behavioural marker system. Br J Anaesth. (2003) 90:580–8.

Abbildung 4: Anaesthetists' Non-Technical Skills (ANTS)



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Fletcher et al. (2003)

Das Instrument **Anaesthetists' Non-Technical Skills (ANTS)** ist ein verhaltensbasiertes Beurteilungssystem zur Erfassung nicht-technischer Fähigkeiten in der Anästhesie. Seine Entwicklung ist vor dem Hintergrund zu verstehen, dass nicht-technische Fähigkeiten zwar als wesentlich für eine sichere anästhesiologische Praxis gelten, in der traditionellen Ausbildung jedoch lange Zeit nicht explizit adressiert wurden. Im zugrunde liegenden Beitrag wird betont, dass die wachsende Einsicht in die Bedeutung dieser Fähigkeiten mit der Notwendigkeit einhergeht, sie nicht nur zu trainieren, sondern auch auf der Basis wissenschaftlich fundierter

Kompetenzrahmen und valider Messinstrumente zu erfassen. Vor diesem Hintergrund wurde ANTS als spezifisches Behavioral Marker System für die Anästhesie konzipiert und in einer experimentellen Studie auf seine grundlegenden psychometrischen Eigenschaften und seine Anwendbarkeit hin untersucht.

Die Entwicklung des ANTS-Systems erfolgte auf der Grundlage psychologischer Forschungsmethoden, mit dem Ziel, die für die anästhesiologische Praxis relevanten nicht-technischen Fähigkeiten systematisch zu identifizieren und in eine praktikable, hierarchisch strukturierte Taxonomie zu überführen. Ausgangspunkt war zunächst eine Literaturrecherche, in deren Rahmen sechs bereits existierende Verhaltensmarkierungssysteme aus Anästhesie und Notfallmedizin identifiziert wurden. Diese Systeme entsprachen jedoch nicht vollständig den Anforderungen des Projekts, insbesondere weil sie teilweise auf Teamebene und nicht auf Individualebene angelegt waren. Gleichwohl wurden ihre Inhalte und Strukturen analysiert, um wiederkehrende Fähigkeitsbereiche und thematische Schwerpunkte herauszuarbeiten. Aufbauend darauf wurden kognitive Aufgabenanalyse-Interviews mit 29 beratenden Anästhesisten durchgeführt. Diese wurden gebeten, besonders herausfordernde Fälle oder kritische Ereignisse aus ihrer Praxis zu schildern und diejenigen Fähigkeiten zu benennen, die sie für gute anästhesiologische Praxis als zentral erachteten. Die Interviewdaten wurden anschließend mit einem Grounded-Theory-Ansatz ausgewertet, um aus dem Material induktiv jene nicht-technischen Fähigkeiten abzuleiten, die als konstitutiv für sicheres und effektives Handeln in der Anästhesie gelten können.

Auf Basis dieser Analysen entwickelte das Projektteam, bestehend aus drei Psychologen und drei beratenden Anästhesisten, eine vorläufige Taxonomie. Deren hierarchische Grundstruktur orientierte sich am europäischen Luftfahrt-System NOTECHS, wurde jedoch an die spezifischen Anforderungen des anästhesiologischen Handlungsfeldes angepasst. Der Prototyp wurde in mehreren Schritten weiter verfeinert, unter anderem durch erneute Codierung eines Teils der Interviews, durch die Analyse von Zwischenfallberichten aus der Anästhesie sowie durch Beobachtungen im Operationssaal. Zu jedem Fähigkeitsbereich wurden schließlich Beispiele für gutes und schlechtes Verhalten formuliert und als konkrete Handlungsaussagen in das Instrument integriert. Zusätzlich flossen Ergebnisse zu Einstellungen britischer Anästhesisten gegenüber Teamarbeit und Sicherheit in die Überarbeitung des Systems ein. Bereits in dieser Entwicklungsphase zeigt sich ein zentrales Merkmal des Instruments: Es erfasst ausschließlich solche Kompetenzanteile, die über beobachtbares Verhalten erschlossen werden können. Fähigkeiten wie Stressmanagement oder Selbstbeherrschung wurden zwar in den Interviews als relevant benannt, jedoch bewusst aus der Taxonomie ausgeschlossen, da sie

nicht hinreichend eindeutig durch äußerlich beobachtbares Verhalten identifizierbar seien. Ebenso wurde auf eine eigenständige Kategorie „Kommunikation“ verzichtet, weil Kommunikation im ANTS-System als Querschnittsmerkmal mehrerer Fertigkeitsbereiche aufgefasst und nicht als isolierte Einzeldimension modelliert wird.

Die Struktur des ANTS-Systems ist hierarchisch aufgebaut und umfasst vier Hauptkategorien mit insgesamt 15 Elementen. Die erste Kategorie, **Task management**, beinhaltet die Elemente *Planning and preparing*, *Prioritizing*, *Providing and maintaining standards* sowie *Identifying and utilizing resources*. Diese Kategorie fokussiert damit insbesondere auf planendes, strukturierendes und ressourcenbezogenes Handeln. Die zweite Kategorie, **Team working**, umfasst die Elemente *Co-ordinating activities with team members*, *Exchanging information*, *Using authority and assertiveness*, *Assessing capabilities* sowie *Supporting others* und bildet damit zentrale interaktionelle und kooperative Aspekte der Arbeit im Team ab. Die dritte Kategorie, **Situation awareness**, setzt sich aus den Elementen *Gathering information*, *Recognizing and understanding* sowie *Anticipating* zusammen und adressiert damit die Wahrnehmung, Verarbeitung und vorausschauende Einordnung situationsrelevanter Informationen. Die vierte Kategorie, **Decision making**, besteht aus *Identifying options*, *Balancing risks and selecting options* sowie *Re-evaluating* und repräsentiert den Bereich der handlungsleitenden Entscheidungsprozesse. Ergänzt werden diese Elemente durch Verhaltensmarker, die sowohl gutes als auch problematisches Verhalten exemplarisch beschreiben. Diese Marker stellen das Kernstück des Instruments dar, da sie abstrakte Kompetenzbereiche in konkrete, beobachtbare Handlungen übersetzen und dadurch die Grundlage einer systematischen Bewertung schaffen.

Die Bewertung erfolgt auf einer vierstufigen Ratingskala mit den Abstufungen *poor*, *marginal*, *acceptable* und *good*. Zusätzlich steht die Kategorie *not observed* zur Verfügung, falls ein Verhalten in einer konkreten Situation nicht beobachtbar war oder die jeweilige Fähigkeit situativ nicht erforderlich wurde. Die Einbeziehung dieser Zusatzoption ist methodisch bedeutsam, da das Instrument damit anerkennt, dass nicht alle nicht-technischen Fähigkeiten in jeder Situation gleichermaßen sichtbar werden müssen. Die Bewertung erfolgt zunächst auf Ebene der einzelnen Elemente und anschließend auf Ebene der übergeordneten Kategorien. Damit folgt ANTS einer analytischen Bewertungslogik, die zunächst konkrete Verhaltensindikatoren erfasst und daraus eine verdichtete Kategoriebewertung ableitet.

Die experimentelle Evaluation des Instruments wurde mit 50 beratenden Anästhesisten aus 17 schottischen Krankenhäusern durchgeführt. Die Teilnehmer verfügten über eine Berufserfahrung von einem bis über 25 Jahre, im Mittel acht Jahre, und waren in unterschiedlichem

Ausmaß in Ausbildung und Beurteilung eingebunden. Für 72 % der Teilnehmer stellte die Studie den ersten Kontakt mit dem ANTS-System dar.

Vor der eigentlichen Bewertung erhielten sie ein etwa vierstündiges Training, das Hintergrundwissen zu Human Factors und nicht-technischen Fähigkeiten, eine Einführung in das ANTS-System sowie Hinweise zur verhaltensorientierten Bewertung und zu möglichen Urteilsverzerrungen umfasste. Eine Kalibrierung der Rater auf einen gemeinsamen Bewertungsstandard erfolgte bewusst nicht, da andernfalls nicht mehr ausschließlich das Instrument selbst, sondern auch der Kalibrierungsprozess mit untersucht worden wäre. Im Anschluss bewerteten die Teilnehmer acht videobasierte, simulierte Anästhesieszenarien. Die Szenarien waren im Vorfeld so konzipiert worden, dass alle 15 Elemente des Instruments auf unterschiedlichen Leistungsniveaus dargestellt wurden. Als Vergleichsmaßstab dienten Referenzbewertungen, die von drei Projektanästhesisten zunächst einzeln und anschließend konsensuell erstellt worden waren.

Hinsichtlich der Validität zeigen die Ergebnisse, dass das ANTS-System von den Teilnehmern als inhaltlich weitgehend vollständig wahrgenommen wurde. Alle Befragten bestätigten, dass das System die zentralen gezeigten nicht-technischen Verhaltensweisen adressiere. Die große Mehrheit sah weder fehlende noch überflüssige Kategorien oder Elemente. Damit spricht die Studie für eine hohe inhaltliche Abdeckung der als relevant erachteten nicht-technischen Fähigkeiten in der Anästhesie. Auch die Beobachtbarkeit der erfassten Fertigkeiten wurde insgesamt positiv bewertet. Im Mittel waren 13 der 15 Elemente in mehr als 80 % der Fälle beobachtbar, und alle vier Kategorien wurden in über 95 % der Fälle erkannt. Besonders gut beobachtbar waren die Elemente *Gathering information* und *Recognizing and understanding*, die in allen Szenarien beobachtet werden konnten, während *Assessing capabilities* mit 66 % die geringste Beobachtbarkeit aufwies. Zudem fiel es den Teilnehmern leichter, beobachtetes Verhalten den übergeordneten Kategorien als den einzelnen Elementen zuzuordnen. Dies deutet darauf hin, dass die Kategoriestruktur intuitiv zugänglich ist, während die trennscharfe Differenzierung zwischen einzelnen Elementen mehr Übung und Systemkenntnis erfordert.

Die Reliabilität des Instruments wurde aus mehreren Perspektiven untersucht. Die Interrater-Übereinstimmung wurde mittels des rwg-Koeffizienten berechnet. Auf Elementebene lagen die Werte zwischen 0,55 und 0,67, auf Kategorienebene zwischen 0,56 und 0,65. Die geringsten Übereinstimmungen zeigten sich für die Kategorie **Situation awareness** sowie insbesondere für das Element *Recognizing and understanding*, während höhere Werte für *Identifying and utilizing resources* sowie für die Kategorien **Task management** und **Team working** erreicht

wurden. Die Autoren interpretieren diese Werte als akzeptabel für ein neu entwickeltes Instrument, dessen Anwender nur ein begrenztes Training erhalten hatten. Zugleich weisen sie darauf hin, dass insbesondere kognitive Konstrukte wie Situationsbewusstsein schwieriger über Verhalten zu erfassen sind und daher geringere Raterübereinstimmungen nicht überraschen. Zudem sei davon auszugehen, dass sich die Interrater-Reliabilität mit größerer Vertrautheit mit dem Instrument und mit intensiverem Training verbessern werde.

Neben der Interrater-Übereinstimmung wurde auch die Genauigkeit der Bewertungen untersucht, also die Nähe der Raterurteile zu den Referenzbewertungen. Hier zeigte sich, dass über 88 % der Urteile innerhalb eines Skalenpunkts von der Referenzbewertung abwichen. Die mittlere absolute Abweichung lag je nach Element zwischen 0,49 und 0,84. Diese Werte sprechen nach Einschätzung der Autoren für eine insgesamt gute Genauigkeit des Instruments, auch wenn bei einzelnen Elementen Grenzunsicherheiten zwischen benachbarten Skalenstufen sichtbar wurden. Solche Abweichungen werden im Dokument vor allem auf Unsicherheiten bei der praktischen Anwendung der Skala zurückgeführt und als durch zusätzliche Schulung und Kalibrierung potenziell reduzierbar eingeschätzt.

Die interne Konsistenz des Systems wurde mit Cronbachs Alpha sowie über Korrelationen zwischen Elementen und Kategorien bestimmt. Die Alphawerte lagen zwischen 0,79 und 0,86 und deuten damit auf eine solide interne Konsistenz der Kategorien hin. Zugleich zeigte sich, dass 13 der 15 Elemente am stärksten mit ihrer jeweils vorgesehenen Kategorie korrelierten. Für zwei Elemente war dies nicht in allen Szenarien der Fall. Insgesamt wird die Struktur des Instruments im Dokument dennoch als tragfähig eingeschätzt. Die Befunde sprechen dafür, dass die Elemente innerhalb der Kategorien inhaltlich zusammengehören, ohne redundant zu sein.

Ein weiterer zentraler Aspekt der Evaluation betrifft die Benutzerfreundlichkeit und Akzeptanz des Instruments. Die Ergebnisse hierzu fallen ausgesprochen positiv aus. Sämtliche Teilnehmer gaben an, dass ANTS hilfreich sei, um Beobachtungen zu strukturieren. Die große Mehrheit hielt das Instrument zudem für geeignet, um Consultants bei der Ausbildung jüngerer Anästhesisten zu unterstützen, und beurteilte es auch als potenziell nützlich für die Beurteilung von Weiterbildungsassistenten. Ebenso wurde es von den meisten Befragten als geeignet angesehen, das Lernen im Operationssaal zu unterstützen. Hinsichtlich seiner konkreten Gestaltung wurden sowohl die Benennung der Kategorien und Elemente als auch die Beschreibungen und Verhaltensmarker überwiegend als klar und sinnvoll bewertet. Auch die vierstufige Skala wurde von der Mehrheit als ausreichend flexibel eingeschätzt, wenngleich einzelner Teilnehmer längere oder kürzere Skalen bevorzugt hätten. Insgesamt lässt sich daraus ableiten,

dass ANTS nicht nur psychometrisch grundsätzlich tragfähig, sondern auch aus Anwendersicht gut handhabbar ist.

Als Anwendungsbereiche des Instruments benennt das Dokument vor allem den Einsatz in simulatorgestützten Trainings, in Human-Factors-Kursen sowie in der regulären klinischen Ausbildung. Das System könne Beobachtungen strukturieren, Feedback präzisieren und die Wirksamkeit von Trainingsmaßnahmen messbar machen. Darüber hinaus könne es eine gemeinsame Sprache für einen Bereich der anästhesiologischen Praxis bereitstellen, der bislang vielfach nur implizit beschrieben wurde. Gerade diese sprachliche und konzeptuelle Strukturierungsleistung ist für die Aus- und Weiterbildung von besonderer Bedeutung, da sie nichttechnische Fähigkeiten als legitimen und systematisch bearbeitbaren Bestandteil professioneller Kompetenz sichtbar macht.

Gleichzeitig weist das Dokument auf mehrere Limitationen hin. Eine zentrale Einschränkung besteht darin, dass die Evaluation unter experimentellen Bedingungen und auf Basis geskripteter Videoszenarien erfolgte. Die Ergebnisse beziehen sich daher nicht auf reale klinische Settings, sondern auf eine kontrollierte Testsituation. Hinzu kommt, dass die Rater nur eine begrenzte Schulung erhielten und keine Kalibrierung stattfand. Dies dürfte sich sowohl auf die Vertrautheit mit dem Instrument als auch auf die Genauigkeit und Übereinstimmung der Bewertungen ausgewirkt haben. Besonders für kognitive Domänen wie **Situation awareness** zeigte sich, dass die Erfassung über beobachtbares Verhalten mit methodischen Schwierigkeiten verbunden ist. Darüber hinaus betonen die Autoren, dass nicht alle Elemente in jeder Situation beobachtbar sein müssen, entweder weil sie für die konkrete Situation nicht relevant sind oder weil sich die zugehörigen Verhaltensweisen nur sehr subtil äußern. Schließlich liegt nach den im Dokument dargestellten Ergebnissen noch keine umfassende Feldvalidierung vor. Insbesondere fehlen Daten zur prädiktiven Validität und zur Übereinstimmung mit anderen Leistungsmaßen im realen klinischen Einsatz.

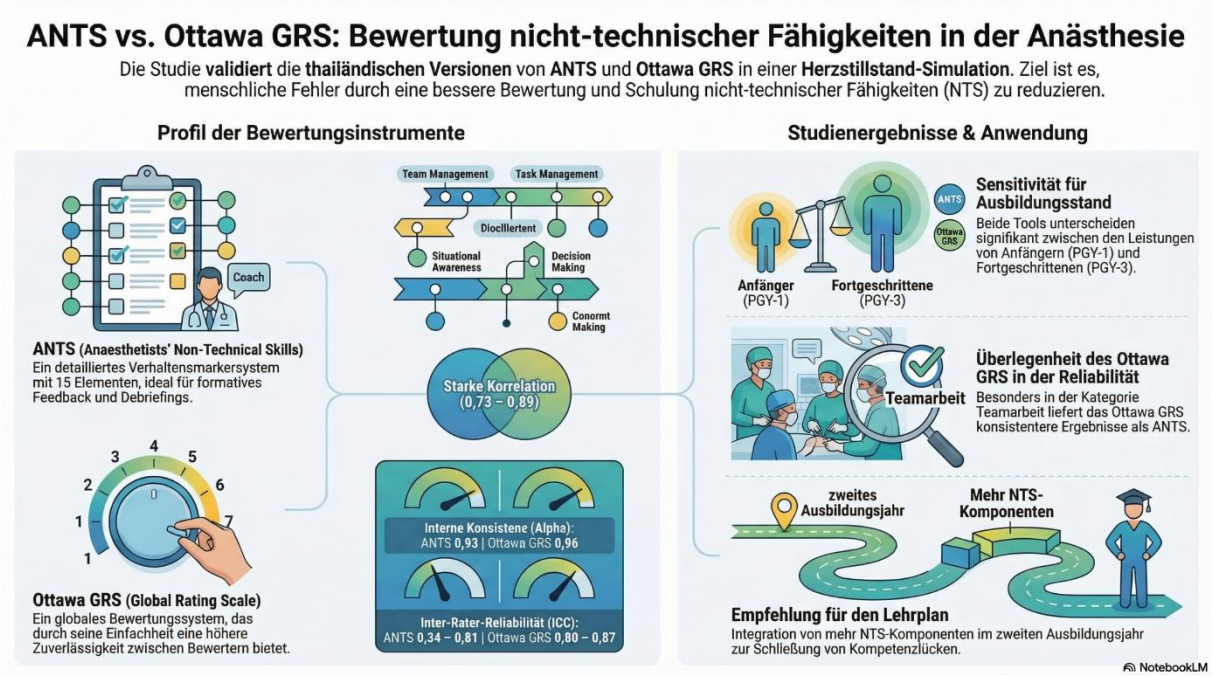
Zusammenfassend lässt sich festhalten, dass ANTS im vorliegenden Dokument als systematisch entwickeltes, fachspezifisches und verhaltensorientiertes Beurteilungsinstrument beschrieben wird, das die zentralen nicht-technischen Fähigkeiten in der Anästhesie in vier Kategorien und 15 Elemente gliedert und mithilfe konkreter Verhaltensmarker bewertbar macht. Die experimentelle Evaluation weist auf eine gute inhaltliche Vollständigkeit, eine überwiegend hohe Beobachtbarkeit, akzeptable Interrater-Übereinstimmungen, gute Genauigkeit, solide interne Konsistenz sowie eine hohe Akzeptanz und Benutzerfreundlichkeit hin. Zugleich zeigt die Studie, dass die Anwendung des Instruments eine angemessene Schulung der Beurteiler voraussetzt und dass insbesondere für seine Nutzung in realen Trainings- und Praxissituatio-

nen weitere Felduntersuchungen erforderlich sind. Damit stellt ANTS nach den im Dokument vorliegenden Befunden ein vielversprechendes Instrument dar, um nicht-technische Fähigkeiten in der Anästhesie systematisch zu beobachten, zu bewerten und in die Ausbildung zu integrieren.

5.2 Anaesthetists' Non-Technical Skills (ANTS) - Operating room, emergency und Ottawa Global Rating Scale (Ottawa GRS) - Operating room, emergency

Quelle: Jirativanont T, Raksamani K, Aroonpruksakul N, Apidechakul P, Suraseranivongse S. Validity evidence of non-technical skills assessment instruments in simulated anaesthesia crisis management. Anaesth Intensive Care. (2017) 45:469–75.

Abbildung 5: Anaesthetists' Non-Technical Skills (ANTS) - Operating room, emergency und Ottawa Global Rating Scale (Ottawa GRS) - Operating room, emergency



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Jirativanont et al. (2017)

5.2.1 ANTS (Operating room, emergency)

Im vorliegenden Dokument wird das Instrument *Anaesthetists' Non-Technical Skills* (ANTS) als ein etabliertes, spezifisch für Anästhesisten entwickeltes Verfahren zur Beurteilung nicht-technischer Kompetenzen beschrieben. Die Autoren verorten das Instrument im Bereich des simulierten anästhesiologischen Krisenmanagements und untersuchen dessen Validität im Rahmen eines standardisierten Simulationssettings. Bereits in der Einleitung wird ANTS als ein anerkanntes Instrument charakterisiert, das von Anästhesisten und Psychologen im Vereinigten Königreich entwickelt wurde und für die strukturierte Beurteilung anästhesierelevanter nicht-technischer Kompetenzen konzipiert ist. Im Unterschied zu globalen Beurteilungsskalen wird ANTS im Dokument als inhaltlich spezifisches und stärker ausdifferenziertes Instrument dargestellt, das über mehrere essenziellen Kategorien mit detaillierten Elementen verfügt.

Für die Durchführung der im Dokument beschriebenen Studie wurde das Instrument nicht neu entwickelt, sondern für den lokalen Kontext adaptiert. Nach Freigabe durch die zuständige Ethikkommission wurde die englische Version zunächst in die thailändische Sprache übersetzt. Diese Übersetzung erfolgte durch einen mit dem Training nicht-technischer Kompetenzen vertrauten Anästhesisten, der in beiden Sprachen sicher war. Anschließend wurde eine Rückübersetzung ins Englische durch einen zweiten, unabhängigen Anästhesisten vorgenommen. Die rückübersetzte Version wurde danach mit der Originalfassung unter Einbezug eines muttersprachlichen professionellen Übersetzers verglichen. Die dabei vorgenommenen Anpassungen dienten dazu, die inhaltliche Bedeutung präzise wiederzugeben. Das Dokument macht deutlich, dass mit diesem Vorgehen insbesondere sichergestellt werden sollte, dass die Adaptation die ursprüngliche Intention des Instruments wahrt und gleichzeitig sprachlich verständlich und kulturell anschlussfähig bleibt.

Die Inhaltsvalidität der adaptierten Fassung wurde anschließend mit Hilfe eines Expertengremiums überprüft. Fünf Experten aus dem Bereich nicht-technischer Kompetenzen, darunter drei Anästhesisten, ein Chirurg und ein Notfallmediziner, bewerteten die Eignung der ANTS-Kategorien und -Elemente mittels des Item Objective Congruence Index. Das Dokument berichtet, dass sämtliche Kategorien und Elemente des Instruments als relevante Bestandteile essenzieller nicht-technischer Fertigkeiten anerkannt wurden. Die ursprüngliche Struktur des Instruments blieb unverändert; lediglich einzelne Formulierungen in der thailändischen Übersetzung wurden sprachlich präzisiert. Diese Befunde sprechen nach Darstellung der Autoren für eine inhaltlich angemessene Übertragbarkeit des Instruments in den untersuchten Kontext.

Strukturell ist ANTS im Dokument durch vier übergeordnete Kategorien gekennzeichnet, die sich in insgesamt fünfzehn Elemente gliedern. Die erste Kategorie, *Task management*, um-

fasst die Elemente *Planning and preparing*, *Prioritising*, *Providing and maintaining standards* sowie *Identifying and utilising resources*. Die zweite Kategorie, *Team-working*, besteht aus *Coordinating with team*, *Exchanging information*, *Using authority and assertiveness*, *Assessing capabilities* und *Supporting others*. Der Bereich *Situation awareness* wird durch die Elemente *Gathering information*, *Recognising and understanding* sowie *Anticipating* abgebildet. Die vierte Kategorie, *Decision-making*, umfasst *Identifying options*, *Balancing risk and selecting options* und *Re-evaluation*. Diese Struktur weist darauf hin, dass das Instrument nicht lediglich globale Einschätzungen des Verhaltens zulässt, sondern verschiedene Facetten nicht-technischer Kompetenzen differenziert abbildet. Im Dokument wird ANTS daher als eine „professional, specific non-technical skills checklist“ beschrieben, deren besondere Stärke in ihrer hohen inhaltlichen Spezifität liegt.

Die psychometrischen Eigenschaften des Instruments werden im Dokument anhand eines umfassenden Validitätsrahmens diskutiert. Die Autoren orientieren sich an einem Fünf-Domänen-Modell der Validität, das Testinhalt, Response Process, interne Struktur, Beziehungen zu anderen Variablen und Konsequenzen der Testung umfasst. Im Bereich der internen Struktur wird zunächst die interne Konsistenz des Instruments mit Cronbachs Alpha berichtet. Für ANTS ergab sich ein Wert von 0,93. Dieser wird im Dokument als Hinweis auf eine hohe interne Konsistenz und eine akzeptable Eindimensionalität des Tests interpretiert. Gleichzeitig verweisen die Autoren darauf, dass Alpha-Werte oberhalb von 0,90 auch auf mögliche Redundanzen zwischen einzelnen Elementen hinweisen können. Die hohe interne Konsistenz ist damit einerseits ein Ausdruck psychometrischer Stabilität, andererseits aber auch Anlass zur kritischen Reflexion über potenzielle Überlappungen innerhalb des Instruments.

Ein zentrales Ergebnis betrifft die Interrater-Reliabilität, die mittels Intraclass Correlation Coefficients bestimmt wurde. Das Bewertungsverfahren wurde von zwei geschulten Ratern durchgeführt, die beide über Erfahrung als Anästhesisten, Simulationstrainer und Instruktoren für nicht-technische Kompetenzen verfügten. Vor der eigentlichen Datenerhebung wurden Trainingsvideos gemeinsam analysiert, bis für beide Instrumente in der Übungsphase ein ICC von über 0,8 erreicht worden war. Trotz dieses intensiven Ratertrainings zeigen sich für ANTS auf Kategorienebene differenzierte Ergebnisse. Für *Task management* wird ein ICC von 0,79 berichtet, für *Team-working* ein Wert von 0,34, für *Situation awareness* 0,81 und für *Decision-making* 0,70. Das Dokument interpretiert diese Werte dahingehend, dass *Team-working* nur eine geringe Reliabilität aufweist, während die übrigen Kategorien im moderaten bis guten Bereich liegen. Auch auf Ebene der einzelnen Elemente ergibt sich ein heterogenes Bild mit Werten zwischen 0,33 für *Exchanging information* und 0,78 für *Anticipating*. Die Autoren führen

diese Unterschiede unter anderem auf die besondere Komplexität sozialer und teambezogener Verhaltensweisen zurück, die schwerer konsistent zu beobachten und zu bewerten seien. Hinzu komme, dass die größere Detailtiefe des ANTS-Instruments dazu führe, dass nicht jede einzelne Verhaltenskomponente in jedem Szenario in gleicher Deutlichkeit beobachtbar sei.

Von besonderer Bedeutung ist die Frage, ob das Instrument in der Lage ist, unterschiedliche Kompetenzniveaus zu unterscheiden. Im Dokument wird hierzu untersucht, ob sich die ANTS-Werte zwischen Teilnehmern verschiedener Ausbildungsjahre unterscheiden.

In die Studie wurden 70 Anästhesie-Weiterbildungsassistenten einbezogen, die sich auf das erste, zweite und dritte Ausbildungsjahr verteilten. Die Ergebnisse zeigen, dass sich für alle Elemente und Kategorien signifikante Unterschiede, insbesondere zwischen dem ersten und dritten Ausbildungsjahr, nachweisen lassen. So liegen beispielsweise die Mittelwerte für *Task management* bei 2,5 im ersten und 3,4 im dritten Ausbildungsjahr, für *Situation awareness* bei 2,5 versus 3,6 und für *Decision-making* bei 2,4 versus 3,2. Das Dokument interpretiert diese Befunde als Evidenz dafür, dass ANTS zwischen unterschiedlichen Ausbildungsständen differenzieren kann und damit Beziehungen zu externen Variablen in theoretisch plausibler Weise abbildet.

Zusätzlich wird die Beziehung des ANTS-Instruments zur Ottawa Global Rating Scale untersucht. Dazu wurden die einzelnen ANTS-Elemente durch ein Expertengremium den Kategorien der Ottawa GRS zugeordnet. Auf dieser Grundlage wurden proportionale Vergleichswerte gebildet und Spearman-Korrelationen berechnet. Die Korrelationskoeffizienten fallen hoch aus und liegen bei 0,89 für *Leadership*, 0,82 für *Problem-solving*, 0,88 für *Situation awareness*, 0,85 für *Resource utilisation* und 0,73 für *Communication*. Diese Ergebnisse werden im Dokument dahingehend interpretiert, dass beide Instrumente in einem hohen Maß konvergente Informationen liefern. Für ANTS bedeutet dies, dass seine Messwerte in starkem Zusammenhang mit jenen eines weiteren etablierten Beurteilungsinstruments stehen, was die Validitätsargumentation zusätzlich stützt.

Der Anwendungsbereich von ANTS liegt im vorliegenden Dokument im simulierten anästhesiologischen Krisenmanagement, das in einem als Operationssaal konfigurierten Simulationslabor durchgeführt wurde. Das Szenario bestand aus einem unerwarteten Herzstillstand während der Ausleitung einer Anästhesie. Zentral ist dabei, dass die Autoren das Szenario so entwickelten, dass alle Aspekte der beiden Instrumente in Form erwarteter Performanzverhaltensweisen in den Ablauf eingebettet wurden. Die Simulation wurde standardisiert vorbereitet, mit Konföderierten besetzt und videografiert. In diesem Zusammenhang wird deutlich, dass ANTS

vor allem für die differenzierte Beobachtung und anschließende Rückmeldung in einem formativen Ausbildungssetting geeignet erscheint. Das Dokument hebt hervor, dass die spezifische Struktur des Instruments besonders für Debriefings von Vorteil sei, da die detaillierten Elemente konkrete Anknüpfungspunkte für Feedback und Verhaltensreflexion bieten.

Gleichzeitig benennt das Dokument mehrere Einschränkungen. Erstens wird darauf hingewiesen, dass nicht-technische Kompetenzen nicht vollständig von Wissen und technischen Fähigkeiten getrennt werden können. Gerade in einem Krisenszenario beeinflussen Fachwissen, Erfahrung und psychomotorische Kompetenz die Gesamtleistung mit. Zweitens zeigen sich deutliche Probleme in der Interrater-Reliabilität einzelner teambezogener Kategorien, insbesondere im Bereich *Team-working*. Drittens macht das Dokument deutlich, dass die höhere Detailtiefe des Instruments zwar didaktisch nützlich, praktisch jedoch anspruchsvoll ist. Nicht alle relevanten Verhaltensweisen können in einem einzelnen Video gleich gut identifiziert werden. Darüber hinaus wird eingeräumt, dass durch die Videobewertung einzelne Performanzaspekte übersehen worden sein könnten und dass Unterschiede in der Vertrautheit mit der Simulationsumgebung, insbesondere zugunsten der Teilnehmer des dritten Ausbildungsjahres, die Ergebnisse beeinflusst haben könnten. Insgesamt gelangt das Dokument dennoch zu der Einschätzung, dass ANTS über eine substanzielle Validitätsevidenz für den Einsatz im simulierten anästhesiologischen Krisenmanagement verfügt. Seine besondere Stärke liegt in der inhaltlichen Spezifität und in seiner Eignung für formatives Feedback, während seine höhere Komplexität zugleich psychometrische und praktische Herausforderungen mit sich bringt.

5.2.2 Ottawa GRS (Operating room, emergency)

Neben dem ANTS-Instrument untersucht das vorliegende Dokument die *Ottawa Global Rating Scale* (Ottawa GRS) als zweites Verfahren zur Beurteilung nicht-technischer Kompetenzen im simulierten anästhesiologischen Krisenmanagement. Die Ottawa GRS wird im Dokument als Instrument zur Erfassung von *crisis resource management skills* beschrieben, das ursprünglich für *acute care physicians* entwickelt wurde. Im Unterschied zum professionsspezifischen ANTS-Instrument ist die Ottawa GRS somit breiter angelegt und nicht exklusiv auf die Anästhesie zugeschnitten. Das Dokument betont, dass die Skala über klar definierte Bewertungsstufen für ihre jeweiligen Kategorien verfügt und bereits vor der vorliegenden Untersuchung als valides und reliables Instrument galt.

Wie beim ANTS-Instrument erfolgte für die Studie zunächst eine sprachliche und kulturelle Adaption der Ottawa GRS. Das Verfahren entsprach demjenigen der ANTS-Adaption. Die eng-

lische Originalfassung wurde in die thailändische Sprache übersetzt, rückübersetzt und anschließend mit der Originalversion abgeglichen. Auch hier war das Ziel, die inhaltliche Bedeutung der Kategorien und Bewertungsdimensionen zu erhalten und zugleich eine sprachlich verständliche sowie kulturell angemessene Fassung zu erzeugen. Die anschließende Prüfung durch ein Expertengremium ergab, dass auch die Ottawa GRS in ihrer thailändischen Version sämtliche relevanten Bestandteile essenzieller nicht-technischer Kompetenzen abbildet. Nach Angaben des Dokuments blieb die ursprüngliche Rahmenstruktur unverändert, während die sprachliche Formulierung präzisiert wurde. Die Autoren interpretieren dies als Unterstützung der Inhaltsvalidität des Instruments.

Im Unterschied zu ANTS ist die Ottawa GRS im Dokument deutlich globaler strukturiert. Ausgewiesen werden sechs Kategorien, nämlich *Overall performance*, *Leadership*, *Problem-solving*, *Situation awareness*, *Resource utilisation* und *Communication*. Anders als beim ANTS-Instrument werden im Dokument keine weiteren Unterelemente innerhalb dieser Kategorien spezifiziert. Dies verweist auf den grundlegend anderen Charakter des Instruments. Während ANTS nicht-technische Kompetenzen in einer Vielzahl einzelner Komponenten beobachtbar macht, operiert die Ottawa GRS auf einem stärker zusammenfassenden Bewertungsniveau. Das Dokument beschreibt sie entsprechend als *global rating scale* mit gut definierten Skalenstufen. Außerdem wird hervorgehoben, dass sie leichter verständlich sei und eine schnellere Beurteilung ermögliche. Diese Merkmale legen nahe, dass die Ottawa GRS weniger auf differenzierte Verhaltensanalyse als vielmehr auf eine übergreifende Gesamtbeurteilung ausgerichtet ist.

Die psychometrischen Ergebnisse der Studie fallen für die Ottawa GRS insgesamt sehr günstig aus. Die interne Konsistenz wurde ebenso wie beim ANTS-Instrument mittels Cronbachs Alpha bestimmt. Für die Ottawa GRS wird ein Wert von 0,96 berichtet. Nach Interpretation der Autoren deutet dieser Wert auf eine sehr hohe interne Konsistenz und auf eine akzeptable Eindimensionalität des Instruments hin. Zugleich wird, analog zu ANTS, darauf verwiesen, dass Alpha-Werte oberhalb von 0,90 auch auf mögliche Redundanzen hindeuten können. Dennoch wird die interne Struktur des Instruments insgesamt als sehr stark eingeschätzt.

Besonders deutlich zeigt sich die Stärke der Ottawa GRS in der Interrater-Reliabilität. Die ICC-Werte liegen bei 0,86 für *Overall performance*, 0,83 für *Leadership*, 0,84 für *Problem-solving*, 0,87 für *Situation awareness*, 0,80 für *Resource utilisation* und 0,86 für *Communication*. Das Dokument ordnet diese Werte als moderat bis gut ein und hebt hervor, dass die Ottawa GRS in allen erfassten Bereichen konsistentere Interrater-Werte aufweist als das ANTS-Instrument. Diese höhere Reliabilität wird damit erklärt, dass die Ottawa GRS weniger detaillierte Elemente

umfasst und Bewerter dadurch eher in der Lage sind, das gezeigte Verhalten stabil und einheitlich einzuschätzen. Während ANTS aufgrund seiner Detailtiefe höhere Anforderungen an die differenzierte Beobachtung stellt, scheint die globalere Anlage der Ottawa GRS die intersubjektive Übereinstimmung zu erleichtern.

Auch im Hinblick auf die Beziehungen zu anderen Variablen weist die Ottawa GRS im Dokument eine deutliche Stärke auf. Die Skala konnte signifikante Unterschiede zwischen Teilnehmern verschiedener Ausbildungsjahre nachweisen. Für alle Kategorien zeigen sich hochsignifikante Unterschiede, insbesondere zwischen dem ersten und dritten Ausbildungsjahr. Beispielsweise liegt der Mittelwert für *Overall performance* bei 3,9 im ersten und 5,7 im dritten Ausbildungsjahr, für *Leadership* bei 3,9 versus 5,9, für *Situation awareness* bei 3,7 versus 5,7 und für *Communication* bei 3,8 versus 5,6. Diese Ergebnisse deuten darauf hin, dass die Ottawa GRS in hohem Maß geeignet ist, Unterschiede in Ausbildungsstand und Erfahrung sichtbar zu machen. Damit liefert das Instrument im Dokument eine überzeugende Evidenz für seine diskriminative Validität.

Darüber hinaus wird die Ottawa GRS in Beziehung zum ANTS-Instrument gesetzt. Hierzu ordnete ein unabhängiges Expertengremium die ANTS-Elemente den Kategorien der Ottawa GRS zu. In der Analyse ergaben sich starke positive Zusammenhänge zwischen den vergleichbaren Bereichen beider Instrumente. Die hohe Korrelation mit den entsprechenden ANTS-Bereichen wird im Dokument als Hinweis darauf gewertet, dass die Ottawa GRS im Kontext der Anästhesie eine vergleichbare konzeptuelle Grundlage erfasst, obwohl sie globaler strukturiert ist. Die Autoren folgern daraus, dass beide Instrumente in gewissem Umfang austauschbar verwendet werden können. Gerade für die Ottawa GRS ist dies von Bedeutung, weil sie ursprünglich nicht spezifisch für die Anästhesie konzipiert wurde, im vorliegenden Dokument jedoch eine klare Anschlussfähigkeit an anästhesiespezifische nicht-technische Kompetenzen zeigt.

Der Einsatzbereich der Ottawa GRS ist im Dokument derselbe wie beim ANTS-Instrument, nämlich das simulierte Krisenmanagement im anästhesiologischen Operationssaalsetting. Auch hier erfolgte die Bewertung anhand videografierter Performanzen in einem standardisierten Simulationsszenario mit einem intraoperativen Herzstillstand. Das Dokument hebt hervor, dass die Ottawa GRS insbesondere durch ihre Praktikabilität überzeugt. Nach Abschluss der Bewertungen wurden die Rater ausdrücklich nach ihrer Einschätzung der Alltagstauglichkeit der Instrumente gefragt. Im Ergebnis berichten die Autoren, dass die Ottawa GRS im Vergleich zu ANTS einfacher anzuwenden sei. In der Diskussion wird dies weiter präzisiert, indem die Ottawa GRS als leichter verständlich, schneller einsetzbar und benutzerfreundlicher beschrie-

ben wird. Gerade in komplexen Assessmentsituationen wird diese Handhabbarkeit als wichtiger Vorteil hervorgehoben.

Besonders betont wird im Dokument auch die Eignung der Ottawa GRS für summative Beurteilungszwecke. Während ANTS vor allem für Debriefings und formative Rückmeldungen empfohlen wird, erscheint die Ottawa GRS aufgrund ihrer globalen Struktur und der einfach verständlichen Terminologie besser für den zusammenfassenden Vergleich von Leistungen geeignet. Zudem wird ausdrücklich darauf verwiesen, dass sie nicht auf die Anästhesie beschränkt ist, sondern auch in anderen Fachgebieten und Professionen eingesetzt werden kann. Ihre breitere Einsetzbarkeit stellt damit einen wesentlichen Unterschied zum ANTS-Instrument dar.

Gleichwohl weist das Dokument auch auf Begrenzungen der Ottawa GRS hin. Ihre globale Struktur erleichtert zwar die Anwendung, bedeutet aber zugleich eine geringere inhaltliche Spezifität. Dadurch bietet sie weniger konkrete Anhaltspunkte für detaillierte Rückmeldungen im Debriefing. Im Vergleich zu ANTS ist sie somit weniger geeignet, einzelne Facetten nichttechnischer Performanz im Detail sichtbar zu machen. Hinzu kommen die generellen Limitationen subjektiver Verhaltensbeurteilungen, die im Dokument ausführlich diskutiert werden. Dazu zählen die Abhängigkeit von einem gut konstruierten Szenario, die Notwendigkeit intensiven Ratertrainings, die Möglichkeit von Beobachtungsverlusten bei der Videoauswertung sowie der Einfluss der Vertrautheit mit der Simulationsumgebung. Auch für die Ottawa GRS gilt daher, dass ihre Aussagekraft nicht isoliert vom Assessmentkontext betrachtet werden kann.

Insgesamt kommt das Dokument zu einer positiven Bewertung der Ottawa GRS im Bereich des simulierten anästhesiologischen Krisenmanagements. Das Instrument weist eine sehr hohe interne Konsistenz, durchgängig gute Interrater-Reliabilität und eine deutliche Fähigkeit zur Unterscheidung zwischen unterschiedlichen Ausbildungsständen auf. Im Vergleich zu ANTS wird die Ottawa GRS als praktischer, schneller und benutzerfreundlicher eingeschätzt. Ihre Stärke liegt insbesondere in der globalen, übersichtlichen Bewertung und in der summativen Eignung auch über die Anästhesie hinaus. Gleichzeitig ist ihre geringere inhaltliche Spezifität zu berücksichtigen, wenn differenzierte, verhaltensnahe Rückmeldungen im Vordergrund stehen. Das Dokument gelangt daher zu dem Ergebnis, dass beide Instrumente über hinreichende Validitätsevidenz verfügen, die Ottawa GRS jedoch im untersuchten Setting die höhere Reliabilität und die größere praktische Anwendbarkeit aufweist.

5.3 Anaesthesiologists' Non-Technical Skills in Denmark (ANTSdk) 2015

Quelle: Jepsen RMHG, Spanager L, Lyk-Jensen HT, Dieckmann P, Østergaard D. Customisation of an instrument to assess anaesthesiologists' non-technical skills. Int J Med Educ. (2015) 6:17–25.

Abbildung 6: Anaesthesiologists' Non-Technical Skills in Denmark (ANTSdk) 2015

Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Jepsen et al. (2015)

Im hochgeladenen Dokument von Jepsen et al. (2015) wird die Entwicklung und Anpassung des Instruments **Anaesthesiologists' Non-Technical Skills in Denmark (ANTSdk)** beschrieben. Im Mittelpunkt der Arbeit steht die Frage, wie das ursprünglich in Schottland entwickelte Instrument **Anaesthetists' Non-Technical Skills (ANTS)** an den dänischen anästhesiologischen Operationssaalkontext angepasst werden kann. Die Studie ist damit primär als Instrumentenentwicklungs- und Anpassungsarbeit zu verstehen und weniger als psychometrische Validierungsstudie im engeren Sinn. Gleichwohl liefert sie umfangreiche Informationen zur konzeptuellen Entwicklung, zur Struktur, zu den Domänen, Elementen und Verhaltensmarkern sowie zu Anwendungsbereichen und Grenzen des Instruments.

Ausgangspunkt der Arbeit ist die Annahme, dass nicht-technische Kompetenzen eine zentrale Rolle für die Patientensicherheit spielen. Das Dokument verweist darauf, dass unzureichender Einsatz nicht-technischer Kompetenzen bei mehr als 70 % der innerklinischen unerwünschten

Ereignisse eine Rolle spielt. Nicht-technische Kompetenzen werden dabei als kognitive und soziale Fähigkeiten beschrieben, die medizinisches Wissen und technische Fertigkeiten ergänzen und damit zu sicherem und effizientem Handeln beitragen. Gerade im Operationssaal werden Kommunikationsprobleme und Defizite in Teamarbeit als bedeutsame Risikofaktoren dargestellt. Vor diesem Hintergrund argumentieren die Autoren, dass die technische Expertise von Anästhesiologen zwar notwendig, aber nicht hinreichend für eine sichere Patientenversorgung sei. Daraus ergibt sich die Notwendigkeit, geeignete Instrumente zu entwickeln beziehungsweise bestehende Instrumente so anzupassen, dass nicht-technische Kompetenzen im jeweiligen kulturellen und organisatorischen Kontext beobachtet, beschrieben und beurteilt werden können.

Die Autoren begründen die Entwicklung von ANTSdk damit, dass im dänischen Kontext zum Zeitpunkt der Studie keine ausreichend reliablen und validen Instrumente zur Verfügung standen, die Lernen und Assessment nicht-technischer Kompetenzen in der anästhesiologischen Weiterbildung unterstützen konnten. Obwohl mit ANTS bereits ein etabliertes Instrument existierte, wurde davon ausgegangen, dass eine unmittelbare Übernahme nicht möglich sei. Frühere Arbeiten zur dänischen Anpassung von Instrumenten für Chirurgen sowie Pflegeanästhesisten hatten bereits gezeigt, dass Unterschiede in Aufgaben, Verantwortlichkeiten und kulturellen Erwartungen zwischen Dänemark und Schottland bestehen. Diese Annahme wird im Dokument zusätzlich theoretisch über die **Activity Theory** gerahmt. Nach diesem Verständnis kann die Grundtätigkeit der Anästhesie als über Kontexte hinweg ähnlich betrachtet werden, während konkrete Handlungen und deren Ausführung in besonderem Maß durch nationale, institutionelle und kulturelle Bedingungen geprägt sind. Diese Perspektive unterstützt die Argumentation, dass eine Anpassung des Instruments insbesondere auf der Ebene konkreter Verhaltensausprägungen erforderlich ist.

Methodisch handelt es sich um eine explorative qualitative Studie in zwei Schritten. Im ersten Schritt wurden sechs semi-strukturierte Gruppeninterviews mit insgesamt 31 Mitgliedern des multiprofessionellen Operationsteams durchgeführt. Die Stichprobe umfasste Anästhesiologen, Pflegeanästhesisten, Chirurgen sowie instrumentierende Pflegekräfte aus verschiedenen operativen Fachbereichen. Die Interviews wurden mono-professionell durchgeführt, um zu vermeiden, dass Hierarchien oder interprofessionelle Abhängigkeiten die Äußerungen beeinflussen. Ziel dieser Interviews war es, die Wahrnehmungen der Beteiligten darüber zu erfassen, wie sich Anästhesiologen im Operationssaal verhalten sollten beziehungsweise nicht verhalten sollten. Die Teilnehmer wurden ausdrücklich gebeten, ihre Aussagen anhand von Beispielen guten und schlechten Verhaltens zu illustrieren. Im zweiten Schritt wurde der auf Grundlage dieser Interviews entwickelte Prototyp des Instruments in regionalen Weiterbildungsgre-

mien mit erfahrenen Anästhesiologen, Anästhesiologen und Weiterbildungsassistenten diskutiert. Diese Diskussionen dienten der Prüfung der Verständlichkeit, der Face Validity und der Ergänzung weiterer Behavioral Markers.

Die Auswertung der Interviewdaten erfolgte mittels **directed content analysis**, wobei die Struktur des ursprünglichen ANTS-Instruments als Ausgangspunkt diente. Die Interviewzitate wurden zunächst den vier ursprünglichen ANTS-Kategorien zugeordnet und anschließend innerhalb dieser Kategorien induktiv nach inhaltlicher Ähnlichkeit gebündelt. Diese Bündel wurden paraphrasiert und deduktiv bestehenden ANTS-Elementen zugeordnet. Zitatbündel, die in keines der vorhandenen Elemente passten, wurden separat geprüft und bildeten die Grundlage für die Entwicklung eines neuen Elements. Gleichzeitig wurden aus den Interviewzitaten konkrete Verhaltensmarker abgeleitet. Der gesamte Analyseprozess war von iterativen Abstimmungen im Forschungsteam geprägt, was im Dokument als wichtiger Bestandteil der Instrumentenentwicklung hervorgehoben wird. Damit kombinierte die Studie eine deduktive Orientierung am Ausgangsinstrument mit einer induktiven Offenheit für kontextspezifische Unterschiede.

Die Struktur des resultierenden Instruments wird im Dokument sehr klar beschrieben. Die finale Version von ANTSdk besteht aus **vier Kategorien**, **16 Elementen** und **131 Behavioral Markers**. Die vier Kategorien lauten **Situation Awareness**, **Decision Making**, **Team Working** und **Leadership**. Auffällig ist, dass die ursprüngliche ANTS-Kategorie **Task Management** in der dänischen Anpassung in **Leadership** umbenannt wurde. Diese Veränderung ist nicht nur terminologischer Natur, sondern verweist auf eine veränderte inhaltliche Akzentuierung. Die Teilnehmer der Weiterbildungsgremien betonten nach Darstellung des Dokuments ausdrücklich die Führungsrolle des Anästhesiologen im Operationssaal. In diesem Zusammenhang wurde auch das Element **Using authority and assertiveness** aus dem Bereich Team Working in die Kategorie Leadership verschoben. Dadurch wird sichtbar, dass Führung im dänischen Setting stärker als dynamisches, situationsabhängiges Steuerungsverhalten verstanden wird.

Die vier Kategorien werden im Dokument nicht nur benannt, sondern auch inhaltlich definiert. **Situation Awareness** umfasst die Aufrechterhaltung einer dynamischen Aufmerksamkeit für die Situation, wobei Informationen von Patienten, Team und Geräten einbezogen und zukünftige Entwicklungen antizipiert werden sollen. Zusätzlich gehört hierzu das Bewusstsein für die eigenen Fähigkeiten und die kontinuierliche Einschätzung des eigenen Handelns. **Decision Making** beschreibt die Beurteilung einer Situation, das Treffen einer Entscheidung, die Kommunikation und Umsetzung des Plans sowie die wiederholte Neubewertung und Anpassung der Strategie. **Team Working** meint die Unterstützung von Zusammenarbeit durch sichere

Kommunikation, die Koordination von Aufgaben unter Berücksichtigung der Kompetenzen der Teammitglieder, das Herstellen eines gemeinsamen Situationsverständnisses und die Beachtung von Faktoren, die die Leistungsfähigkeit des Teams beeinflussen. **Leadership** wird als Organisieren und Priorisieren von Ressourcen und Aktivitäten beschrieben, einschließlich der Übernahme einer führenden oder nicht-führenden Rolle je nach Situation und mit Fokus auf Sicherheit und Qualität der Arbeit. Diese Definitionen zeigen, dass die Kategorien als übergeordnete Kompetenzdimensionen zu verstehen sind, die den Rahmen für die darunterliegenden Elemente bilden.

Die Elemente des Instruments sind auf der Ebene der vier Kategorien weiter ausdifferenziert. Im Bereich **Situation Awareness** umfasst ANTSdk die Elemente **Gathering information**, **Recognising and understanding contexts**, **Anticipating and thinking ahead** sowie das neu hinzugefügte Element **Demonstrating self-awareness**. Der Bereich **Decision Making** besteht aus **Identifying options**, **Choosing, communicating and implementing decisions** und **Reassessing decisions**. **Team Working** beinhaltet die Elemente **Exchanging information**, **Assessing competencies**, **Coordinating activities** und **Supporting others**. Im Bereich **Leadership** finden sich **Planning and preparing**, **Prioritising**, **Identifying and utilising resources**, **Using authority and assertiveness** sowie **Providing and maintaining standards**. Damit bleibt die Anzahl der Kategorien im Vergleich zum Ursprungsinstrument stabil, während auf Elementebene gezielte Veränderungen vorgenommen wurden, um den dänischen Kontext angemessener abzubilden.

Eine zentrale Besonderheit von ANTSdk liegt in der großen Zahl und der differenzierten Ausgestaltung der **Behavioral Markers**. Das Dokument berichtet, dass der Prototyp zunächst 69 Verhaltensmarker umfasste, die im Zuge der Diskussion in den regionalen Weiterbildungsgremien um weitere 62 Marker ergänzt wurden, sodass die finale Fassung 131 Marker enthielt. Diese Marker stellen beobachtbare Beispiele für gutes oder schlechtes Verhalten dar und sind damit die konkrete Grundlage für Beurteilung und Feedback. Im Dokument wird hervorgehoben, dass rund 47 % der Behavioral Markers in ANTSdk von jenen des ursprünglichen ANTS-Instruments abwichen. Die Veränderungen werden vor allem mit kulturellen und organisatorischen Besonderheiten des dänischen Operationssaals erklärt. Besonders hervorgehoben werden Marker, die sich auf das Bewusstsein für eigene Kompetenzen, auf systematisches Arbeiten und auf das aktive Ansprechen drohender Fehler beziehen. Im Fließtext werden hierzu beispielhaft Marker wie „introduces himself or herself to new team members and states his or her competencies“, „appears calm“, „summarises the situation for the team when needed; for example, using ABCDE systematics“, „uses systematics in planning the task“, „says that a mistake is about to occur“ und „justifies when guidelines are not followed“ genannt.

Für **Gathering information** gilt als gutes Verhalten, sich auf die spezifische Situation zu konzentrieren, während schlechtes Verhalten darin besteht, bei der Informationssammlung keine Systematik zu verwenden. **Recognising and understanding contexts** wird positiv dadurch gekennzeichnet, relevante Veränderungen im Zustand eines Patienten dem Team mitzuteilen und angemessene Reaktionen sicherzustellen; negativ ist es, solche Veränderungen nicht anzusprechen. Bei **Anticipating and thinking ahead** wird als gutes Verhalten beschrieben, Teammitglieder darauf hinzuweisen, wenn sich eine Situation kritisch entwickeln könnte, während das Zurückweisen von Fragen nach Alternativplänen als schlechtes Verhalten gilt. Für das neue Element **Demonstrating self-awareness** stehen Marker wie „knows own limits“ positiv und „exhibits inappropriate behaviour in relation to the situation“ negativ. Im Bereich **Decision Making** ist es beispielsweise positiv, die Situation mithilfe von Systematiken wie ABCDE zusammenzufassen und die verfügbaren Handlungsoptionen zu nutzen; negativ sind das Nichtberücksichtigen von Differentialdiagnosen oder das Ausblenden des Teams bei relevanten Entscheidungen. In **Team Working** erscheinen das Vorstellen mit Namen und Kompetenz, das Reagieren auf Überforderungssignale von Teammitgliedern, die kompetenzorientierte Aufgabenverteilung und ruhiges Auftreten als günstige Marker, während zu viele Anweisungen auf einmal, das Unterlassen von Hilferufen trotz unzureichender Teamkompetenzen, Passivität in der Koordination oder verwirrtes Auftreten als ungünstig beschrieben werden. Im Bereich **Leadership** zählen systematische Planung, flexible Prioritätensetzung, ressourcenangemessenes Handeln, das Ansprechen möglicher Fehler und das begründete Abweichen von Leitlinien zu den positiven Markern, während fehlende Alternativpläne, unnötiges Verlassen des Operationssaals, Überforderung der Ressourcen, mangelndes Einfordern von Ruhe oder starres Festhalten an unpassenden Vorgaben als negative Marker gelten. Diese Marker verdeutlichen, dass ANTSdk stark verhaltensorientiert ist und konkrete Beobachtungspunkte bereitstellt.

Ein besonders bedeutsames Ergebnis der Anpassungsarbeit ist die Einführung des neuen Elements **Demonstrating self-awareness**. Das Dokument betont, dass zahlreiche Interviewaussagen darauf verwiesen, dass mangelndes Bewusstsein für die eigenen Grenzen erhebliche negative Folgen für Team und Patientensicherheit haben könne. Deshalb wurde dieses Thema nicht lediglich, als Behavioral Marker innerhalb eines bestehenden Elements belassen, sondern zu einem eigenständigen Element innerhalb von Situation Awareness ausgebaut. Diese Entscheidung hebt die Autorenschaft auch theoretisch hervor, indem sie auf mögliche kulturelle Besonderheiten in Dänemark verweist. Unter Bezug auf Hofstede wird argumentiert, dass in einer stärker egalitären und partizipativ geprägten Kultur die reflektierte Darstellung eigener Kompetenzen und Grenzen eine besonders wichtige Rolle spielen könne. Auch die

bereits erfolgte dänische Anpassung anderer Instrumente habe ähnliche Veränderungen gezeigt. In diesem Zusammenhang erscheint ANTSdk nicht als bloße Übersetzung von ANTS, sondern als inhaltlich modifizierte Fassung mit eigenständigem Profil.

Hinsichtlich der psychometrischen Eigenschaften im engeren Sinne liefert das Dokument nur begrenzte Evidenz. Es werden keine klassischen Kennwerte wie Interrater-Reliabilität, interne Konsistenz oder Konstruktvalidität in quantitativer Form berichtet. Die Studie ist vielmehr auf die qualitative Entwicklungsphase konzentriert. Gleichwohl finden sich Hinweise auf **Face Validity** und **inhaltliche Passung**. Die Teilnehmer der Weiterbildungsgremien bewerteten den Prototyp als verständlich und nutzbar im dänischen Ausbildungskontext, was im Dokument ausdrücklich als Unterstützung der Face Validity interpretiert wird. Zudem wird berichtet, dass Datensättigung nach den Interviews und Diskussionen erreicht wurde. Die inhaltliche Validität wird dadurch gestützt, dass die identifizierten Verhaltensweisen vollständig innerhalb der vier ursprünglichen ANTS-Kategorien verortet werden konnten, zugleich aber Raum für die Ergänzung eines neuen Elements bestand. Damit zeigt das Dokument vor allem Evidenz für die **inhaltliche Anpassung** und die **konzeptionelle Plausibilität** des Instruments. Eine weitergehende psychometrische Prüfung bleibt einer nachfolgenden Untersuchung vorbehalten.

Das Dokument beschreibt außerdem die Bewertungscharakteristika des Instruments. Für Kategorien und Elemente wurde eine **fünfstufige Likert-Skala** von „much below average“ bis „much above average“ eingeführt. Zusätzlich wurde ein **siebenstufiges globales Rating** von „poor“ bis „excellent“ ergänzt. Die Autoren begründen diese Entscheidung damit, dass eine fünfstufige Skala eine feinere Differenzierung beobachteten Verhaltens ermögliche und einem bei der vierstufigen Originalskala beobachteten Ceiling-Effekt entgegenwirken könne. Die globale Skala soll darüber hinaus die Einschätzung der Gesamtleistung fördern und nicht bloß eine rechnerische Mittelung der Kategorien widerspiegeln. Zusätzlich wurde auf dem Ratingformular Raum für **Freitextnotizen** vorgesehen, da numerische Werte allein nur begrenzten Informationsgehalt hätten. In der Diskussion wird betont, dass gerade die Kombination aus numerischer und narrativer Rückmeldung für Lern- und Feedbackprozesse besonders wertvoll sei. Dies deutet darauf hin, dass das Instrument von Beginn an sowohl für die strukturierte Beobachtung als auch für formative Rückmeldungen gedacht war.

Der hauptsächliche **Anwendungsbereich** von ANTSdk liegt nach dem Dokument in der **anästhesiologischen Facharztweiterbildung im Operationssaal**. Es soll dazu beitragen, eine gemeinsame Sprache für nicht-technische Kompetenzen zu etablieren, die Beurteilung entsprechender Verhaltensweisen zu strukturieren und Rückmeldungen zu präzisieren. Im Dokument wird hervorgehoben, dass ANTSdk dadurch auch eine bildungsbezogene Funktion übernimmt, weil es Begriffe und Konzepte bereitstellt, mit denen klinisches Handeln reflektiert wer-

den kann. Darüber hinaus wird argumentiert, dass das Instrument vier der sieben CanMEDS-Rollen, nämlich Collaborator, Communicator, Manager und Professional, mit konkreten beobachtbaren Verhaltensweisen unterlegt und dadurch für kompetenzorientierte Weiterbildung anschlussfähig ist. ANTSdk ist somit nicht nur ein Bewertungsinstrument, sondern zugleich ein Instrument zur Bewusstseinsbildung und Strukturierung von Lernzielen.

Gleichzeitig reflektiert das Dokument mehrere Limitationen der Studie und des entwickelten Instruments. Eine erste methodische Einschränkung ergibt sich daraus, dass die bestehenden ANTS-Kategorien als Ausgangspunkt der Analyse verwendet wurden. Die Autoren diskutieren selbst, ob dies den Erkenntnisprozess zu stark vorstrukturiert haben könnte. Sie bewerten diesen Umstand als Stärke, weil das Ausgangsinstrument auf einer soliden empirischen Basis beruht, räumen aber ein, dass alternative Strukturierungen dadurch möglicherweise weniger wahrscheinlich wurden. Eine weitere Limitation betrifft die Stichprobenziehung. Die Teilnehmer wurden als interessierte Freiwillige über lokale Leitungspersonen gewonnen, was zu einer Verzerrung zugunsten besonders engagierter oder thematisch sensibilisierter Personen geführt haben könnte. Hinzu kommt, dass die Interviews an nur einem Universitätsklinikum durchgeführt wurden. Zwar wurde versucht, durch Einbezug verschiedener Berufsgruppen und Fachrichtungen eine möglichst breite Perspektive zu sichern, dennoch ist die Übertragbarkeit auf andere Häuser, insbesondere Nicht-Universitätskliniken, begrenzt. Die Entscheidung, die Interviews mono-professionell durchzuführen, reduzierte zwar mögliche Hierarchieeffekte, könnte aber auch den interprofessionellen Reflexionsgewinn eingeschränkt haben. Schließlich weisen die Autoren darauf hin, dass die Reihenfolge der Themen im Interviewleitfaden dazu geführt haben könnte, dass über einige Kategorien weniger intensiv gesprochen wurde. Darüber hinaus fehlen in diesem Entwicklungsstadium noch quantitative psychometrische Prüfungen, sodass Aussagen über Reliabilität und weitergehende Validität noch nicht getroffen werden können.

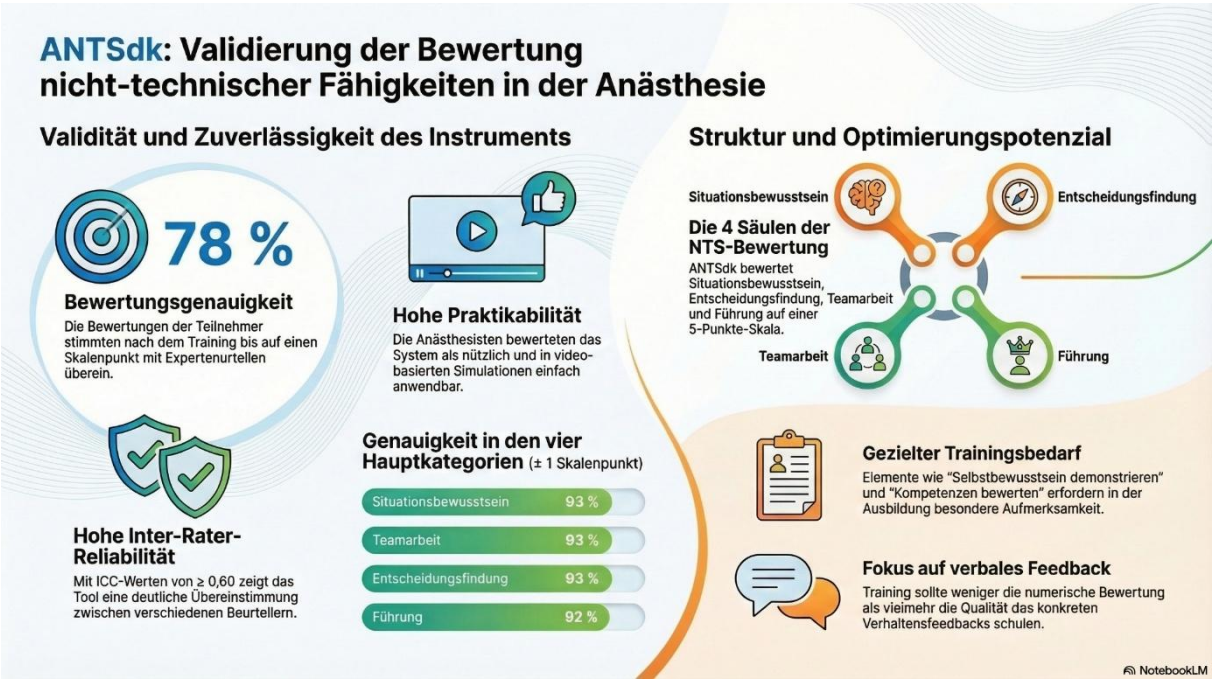
Zusammenfassend zeigt das Dokument, dass ANTSdk als kontextspezifisch angepasste dänische Version des ANTS-Instruments entwickelt wurde, um nicht-technische Kompetenzen von Anästhesiologen im Operationssaal differenziert erfassen zu können. Das Instrument umfasst vier Kategorien, 16 Elemente und 131 Behavioral Markers. Wesentliche Veränderungen gegenüber dem Original bestehen in der Umbenennung von Task Management zu Leadership, der Verschiebung des Elements Using authority and assertiveness in diesen Bereich sowie in der Aufnahme des neuen Elements Demonstrating self-awareness. Die Entwicklung des Instruments basiert auf semi-strukturierten Gruppeninterviews, direkter Inhaltsanalyse und Diskussionen mit Vertretern der anästhesiologischen Weiterbildung. Das Dokument liefert insbesondere Evidenz für die inhaltliche Passung, Verständlichkeit und strukturelle Angemessen-

heit des Instruments. Klassische psychometrische Kennwerte werden in dieser Studie noch nicht berichtet, was eine wesentliche Grenze der Arbeit darstellt. Dennoch bietet die Studie eine tragfähige Grundlage für die weitere Evaluation und Anwendung von ANTSdk in der anästhesiologischen Weiterbildung und im formativen Feedback zu nicht-technischen Kompetenzen.

5.4 Anaesthesiologists' Non-Technical Skills in Denmark (ANTSdk) 2016

Quelle: Jepsen RMHG, Dieckmann P, Spanager L, Lyk-Jensen HT, Konge L, Ringsted C, et al. Evaluating structured assessment of anaesthesiologists' non-technical skills. Acta Anaesthesiol Scand. (2016) 60:756–66.

Abbildung 7: Anaesthesiologists' Non-Technical Skills in Denmark (ANTSdk) 2016



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Jepsen et al. (2016)

Im hochgeladenen Dokument von Jepsen et al. (2016) wird das Instrument **Anaesthesiologists' Non-Technical Skills in Denmark (ANTSdk)** im Hinblick auf seine strukturellen Merkmale sowie seine Validitätsevidenz untersucht. Das Instrument ist als dänische Anpassung des ursprünglichen ANTS-Systems konzipiert und wurde entwickelt, um nicht-technische Kompetenzen von Anästhesiologen im Rahmen der Weiterbildung strukturiert zu beurteilen. Der Bei-

trag ordnet ANTSdk in den größeren Kontext kompetenzorientierter medizinischer Ausbildung ein und betont, dass nicht-technische Kompetenzen wie Situationsbewusstsein, Teamarbeit oder Führung als wesentliche Voraussetzungen sicherer anästhesiologischer Versorgung gelten. Vor diesem Hintergrund wird die Notwendigkeit hervorgehoben, über geeignete Instrumente zu verfügen, mit denen sich diese Kompetenzen systematisch beobachten, bewerten und für Feedbackprozesse nutzbar machen lassen.

Bezüglich der Entwicklung des Instruments wird im Dokument erläutert, dass ANTSdk nicht als vollständig neues Verfahren entstanden ist, sondern als **kulturelle und kontextspezifische Anpassung** des ursprünglichen Instruments **Anaesthetists' Non-Technical Skills (ANTS)**. Die Autoren verweisen darauf, dass ANTSdk auf der Grundlage multipler Interviews mit Vertretern des Operationsteams entwickelt wurde. Diese Anpassung wurde vorgenommen, weil Bewertungsinstrumente nach Darstellung des Dokuments nicht ohne Weiteres über kulturelle und organisationale Kontexte hinweg einsetzbar sind, sondern für neue Settings erneut Validitätsevidenz benötigen. ANTSdk wurde demnach entwickelt, um bestehende Verfahren der Ausbildungsbeurteilung in der anästhesiologischen Facharztweiterbildung zu ergänzen und insbesondere strukturierte Beobachtung sowie Rückmeldung zu nicht-technischen Kompetenzen zu ermöglichen. Für die theoretische Rahmung der Validierungsarbeit greifen die Autoren auf **Messick's Framework** zurück, das verschiedene Evidenzbereiche der Validität unterscheidet. In der vorliegenden Untersuchung stehen insbesondere **response process** und **internal structure** im Mittelpunkt, während content validity und erste Aspekte der internen Struktur bereits in einer früheren Arbeit erhoben worden waren.

ANTSdk besteht aus **vier Kategorien**, nämlich **Situation Awareness**, **Decision Making**, **Team Working** und **Leadership**. Diese vier Kategorien werden durch insgesamt **16 Elemente** operationalisiert. Die Kategorie **Situation Awareness** umfasst die Elemente **Gathering information**, **Recognising and understanding contexts**, **Anticipating and thinking ahead** sowie **Demonstrating self-awareness**. Der Bereich **Decision Making** setzt sich aus **Identifying options**, **Choosing, communicating and implementing decisions** und **Reassessing decisions** zusammen. Die Kategorie **Team Working** enthält die Elemente **Exchanging information**, **Assessing competencies**, **Coordinating activities** und **Supporting others**. Der Bereich **Leadership** umfasst **Planning and preparing**, **Prioritising**, **Identifying and utilising resources**, **Using authority and assertiveness** sowie **Providing and maintaining standards**. Durch diese Struktur wird deutlich, dass ANTSdk nicht allein auf globalen Eindrucksurteilen basiert, sondern verschiedene Teilaspekte nicht-technischer Performanz differenziert abbildet.

Auch die Bewertungslogik des Instruments ist im Dokument präzise dargestellt. Sowohl die Kategorien als auch die einzelnen Elemente werden mit einer **fünfstufigen Likert-Skala** eingeschätzt. Die Skala reicht von **1 = much below average** über **2 = below average**, **3 = acceptable**, **4 = above average** bis **5 = much above average**. Zusätzlich kann **N/A** vergeben werden, wenn ein Verhalten in der betreffenden Situation nicht erforderlich war. Ergänzt wird diese Struktur durch einen **global rating score (GRS)**, der auf einer **siebenstufigen Likert-Skala** von **poor** bis **excellent** vergeben wird. Darüber hinaus enthält das Formular Raum für **schriftliche Feedbacknotizen**, die dazu dienen, spezifische beobachtete Handlungen festzuhalten. In der Bildunterschrift zu Figure 1 wird hervorgehoben, dass die numerischen Ratings dazu verwendet werden können, die Entwicklung der Fähigkeiten über die Zeit hinweg zu verfolgen, während die schriftlichen Rückmeldungen konkrete Hinweise darauf geben sollen, welche Handlungen ausreichend waren oder verbessert werden sollten. Dadurch wird ANTSdk im Dokument nicht nur als Bewertungsinstrument, sondern zugleich als Instrument zur **strukturierenden Rückmeldung** und damit zur formativen Förderung verstanden.

In Bezug auf die Verhaltensmarker weist das Dokument darauf hin, dass jedes Element Beispiele für gutes und schlechtes Verhalten umfasst. Eine vollständige Liste dieser Behavioral Markers wird im Artikel jedoch nicht im Wortlaut abgedruckt. Das bedeutet, dass sich aus dem Dokument sicher ableiten lässt, dass ANTSdk ein markerbasiertes Instrument ist, die genaue Formulierung der einzelnen Marker jedoch nur indirekt erschlossen werden kann. Sichtbar und explizit genannt sind die Kategorien und Elemente, die als strukturelle Bezugspunkte der Verhaltensbewertung fungieren. Der Hinweis auf den User Guide, der im Text genannt wird, deutet darauf hin, dass die Marker dort weiter ausformuliert sind, diese Quelle liegt im hochgeladenen Material jedoch nicht vor. Für die Analyse im Rahmen des vorliegenden Dokuments bleibt daher festzuhalten, dass die konkrete Beobachtung auf behavioralen Beispielen basiert, die im Artikel selbst allerdings nicht vollständig dokumentiert werden.

Zur Prüfung der psychometrischen Eigenschaften wurde eine explorative Studie mit videobasierten Simulationsszenarien durchgeführt. Hierfür wurden **neun anästhesiologische Simulationsvideos** erstellt, die unterschiedliche Fälle und unterschiedliche Ausprägungen nichttechnischer Kompetenzen über die vier Kategorien hinweg zeigten. Die Szenarien waren so angelegt, dass die gesamte Skala ausgenutzt werden konnte. Die Szenarien wurden bewusst locker geskriptet, um ein Verhalten zu ermöglichen, das an realer klinischer Erfahrung orientiert ist. Die Aufnahmen erfolgten in einem full-scale Simulations-OP mit einem manikin-basierten Simulator. Ein externer Facharzt für Anästhesiologie sowie ein Psychologe prüften, ob die gezeigten Verhaltensweisen klinisch glaubwürdig und den Standards entsprechend waren. Zu-

sätzlich wurde durch ein Expertengremium aus Anästhesiologen, Anästhesisten und einem Psychologen ein Satz von **Referenzratings** entwickelt, indem die Videos zunächst unabhängig voneinander bewertet und anschließend bis zu einem Konsens diskutiert wurden. Diese Referenzratings dienten dazu, die Genauigkeit der Bewertungen durch die Studienteilnehmer zu prüfen.

An der Untersuchung nahmen **19 Anästhesiologen** aus neun Krankenhäusern teil, darunter überwiegend Fachärzte mit zum Teil Verantwortung für Weiterbildung. Die Teilnehmer beurteilten zunächst alle neun Videos individuell anhand des ANTSdk-Formulars. Anschließend erhielten sie eine Schulung, die aus einer Vorlesung zu Human Factors, ANTSdk und Beobachtungs- bzw. Bewertungsgrundlagen bestand und durch eine längere Trainingssequenz mit vier zusätzlichen Videos ergänzt wurde. Nach dieser Schulung bewerteten die Teilnehmer die neun ursprünglichen Videos erneut in zufälliger Reihenfolge. Diese Anlage erlaubte es, den Einfluss des Trainings auf die Qualität des Bewertungsprozesses und auf psychometrische Kennwerte zu analysieren.

Die Autoren diskutieren die Ergebnisse entlang der Kategorien des Messick-Modells. Für den **response process** berichten sie, dass der Einfluss konstruktirrelevanter Varianz unter anderem dadurch reduziert wurde, dass die Bewerter auf typische Fehlerquellen und Verzerrungen hingewiesen wurden. Nach der Schulung gaben 18 von 19 Teilnehmern an, die Ratingskalen als ausreichend zu empfinden, und 17 Teilnehmer fühlten sich sicher im Umgang mit dem Instrument. Einzelne Rückmeldungen zeigten, dass die Kombination aus Kategorien-, Element- und Globalratings einerseits als hilfreich, andererseits aber auch als komplex wahrgenommen wurde. Zwei Personen sahen einen weiteren Trainingsbedarf. Mehrere Teilnehmer hoben hervor, dass ANTSdk differenziertes Feedback zu Kompetenzen ermögliche, die sonst nur schwer zu beurteilen seien. Die Mehrheit beurteilte auch die videobasierten Szenarien als realistisch. Insgesamt deuten diese Angaben darauf hin, dass ANTSdk im untersuchten Setting als **praktikabel und nutzbar** wahrgenommen wurde.

Im Bereich der **internen Struktur** wurden mehrere Kennwerte berichtet. Für die **Interrater-Reliabilität** wurden Intraklassenkorrelationen berechnet. Die Autoren definieren dabei Werte über 0,6 als ausreichend für klinische Feedbacksituationen und Werte über 0,7 als ausreichend für High-Stakes-Assessments. Für die Summenscores der Kategorien und Elemente lagen die **average measures ICCs** bereits vor dem Training bei mindestens **0,97** und nach dem Training bei mindestens **0,98**. Die entsprechenden **single measures ICCs** lagen bei mindestens **0,70** vor und **0,71** nach der Schulung. Für spezifische Kategorien, Elemente und den Global Rating Score lagen die average measures bei mindestens **0,96** vor dem Training und

mindestens **0,97** nach dem Training, während die single measures für die meisten Kategorien und Elemente vor dem Training bei mindestens **0,68** lagen. Die Autoren interpretieren dies als Hinweis auf eine insgesamt gute Interrater-Reliabilität, selbst bereits vor der Schulung. Besonders hervorgehoben wird im Diskussionsteil, dass die Kategorie **Situation Awareness** eine hohe Übereinstimmung aufwies, was im Vergleich zu früheren Untersuchungen anderer Instrumente bemerkenswert sei. Gleichzeitig werden einzelne Elemente wie **Supporting others**, **Demonstrating self-awareness** und **Reassessing decisions** als schwieriger beobachtbar beschrieben, deren Übereinstimmung sich nach der Schulung jedoch verbesserte.

Zusätzlich wird die Kohärenz der verschiedenen Bewertungsebenen untersucht. Die **Korrelation zwischen Elementratings und Global Rating Score** betrug **0,93**, die **Korrelation zwischen Kategorieratings und Global Rating Score 0,92**. Diese hohen Zusammenhänge interpretieren die Autoren als Hinweis darauf, dass ANTSdk die wesentlichen Aspekte nichttechnischer Kompetenzen umfassend erfasst. Auch wenn der Artikel den Begriff der internen Konsistenz nicht im klassischen Sinn eines homogenen Skaleninstruments in den Vordergrund stellt, deuten diese Werte auf eine starke innere Stimmigkeit der verschiedenen Bewertungsebenen hin.

Ein weiterer psychometrischer Aspekt ist die **Accuracy** der Ratings im Vergleich zu den Referenzbewertungen. In **Tabelle 2 auf Seite 8** werden die prozentualen Anteile der Bewertungen dargestellt, die höchstens einen Skalenpunkt vom Referenzwert abwichen. Für die meisten Kategorien und Elemente lagen diese Genauigkeitswerte bereits vor der Schulung bei **76 % oder höher**. Besonders hohe Werte zeigten etwa **Situation Awareness** mit 97 % vor und 93 % nach dem Training, **Team Working** mit 93 % sowohl vor als auch nach dem Training sowie mehrere Elemente der Führung und Situationswahrnehmung. Schwieriger zu bewerten waren dagegen insbesondere die Elemente **Demonstrating self-awareness**, **Choosing, communicating and implementing decisions**, **Reassessing decisions** und vor allem **Assessing competencies**, das mit 55 % vor und 58 % nach dem Training den niedrigsten Wert aufwies. Das Dokument leitet daraus ab, dass diese Elemente in zukünftigen Schulungen stärker berücksichtigt werden sollten. Zugleich wird festgestellt, dass die Teilnehmer vor der Schulung tendenziell strengere Bewertungen vergaben als die Expertengruppe und sich ihre Ratings nach der Schulung stärker in Richtung der Referenzwerte verschoben.

Hinsichtlich des Anwendungsbereichs wird ANTSdk im Dokument vor allem als Instrument für die **anästhesiologische Weiterbildung** dargestellt. Es soll zur strukturierten Beobachtung und Beurteilung nicht-technischer Kompetenzen von Weiterzubildenden dienen und insbesondere **formatives Assessment** ermöglichen. Die Autoren betonen, dass das Instrument dabei

helfen könne, konkrete Verhaltensbeispiele zu dokumentieren und darauf bezogen Rückmeldungen zu geben. Im Unterschied zu 360°-Beurteilungen werde die Rückmeldung bei ANTSdk direkt auf beobachtete Verhaltensweisen in einer klar definierten Situation bezogen. Das Instrument ist damit sowohl als Assessmentinstrument als auch als didaktisches Werkzeug zur Förderung reflexiver Lernprozesse zu verstehen.

Gleichzeitig weist das Dokument auf mehrere Limitationen hin. Zunächst wird betont, dass es sich nur um eine **kleine Gelegenheitsstichprobe** von 19 Ratern handelte. Darüber hinaus war die Schulung kürzer als für solche Trainingsformate empfohlen, was die Aussagekraft zur Wirkung des Trainings begrenzt. Die Teilnehmer könnten überdurchschnittlich bildungsaffin und für nicht-technische Kompetenzen sensibilisiert gewesen sein, sodass die Ergebnisse nicht ohne Weiteres auf alle klinischen Anwender übertragbar sind. Auch die Verwendung derselben Videos vor und nach der Schulung könnte theoretisch Erinnerungseffekte erzeugt haben. Ferner wird darauf hingewiesen, dass die Videos relativ kurz waren und daher keine Aussagen über die Bewertung schwankender Leistungen über längere Zeiträume erlauben. Schließlich wurden die Beziehungen zu anderen Variablen sowie Konsequenzen des Instrumenteneinsatzes in dieser Studie nicht untersucht. Diese Bereiche der Validitätsevidenz bleiben somit offen.

Zusammenfassend zeigt das hochgeladene Dokument, dass ANTSdk ein differenziert strukturiertes, markerbasiertes Instrument zur Beurteilung nicht-technischer Kompetenzen in der Anästhesie darstellt. Es umfasst vier Kategorien und 16 Elemente, die sowohl auf Kategorien- als auch auf Elementebene mit einer fünfstufigen Skala sowie zusätzlich mit einem globalen Gesamtrating eingeschätzt werden. Die Ergebnisse der Studie sprechen für eine gute Praktikabilität, eine hohe Interrater-Reliabilität, eine starke Kohärenz der Bewertungsebenen und eine beachtliche Genauigkeit der Ratings im Vergleich zu Expertenurteilen. Zugleich wird deutlich, dass einzelne Elemente schwerer zu erfassen sind und dass weitere Forschung insbesondere zur Generalisierbarkeit, zur Anwendung in Echtzeit und zur Optimierung von Ratertrainings notwendig bleibt. Insgesamt wird ANTSdk im Dokument jedoch als ein **valides und praktikables Instrument für die formativen Beurteilungsprozesse in der anästhesiologischen Weiterbildung** eingeschätzt.

5.5 Anaesthesiology Students' Non-Technical Skills (AS-NTS)

Die frei verfügbare Fassung der Studie von Malec et al. (2007), in der die ursprüngliche Mayo High Performance Teamwork Scale entwickelt und validiert wurde, war nicht vollständig zu-

gänglich. Daher wurde auf die Arbeit von Moll-Khosrawi P, Kamphausen A, Hampe W, Schulte-Uentrop L, Zimmermann S, Kubitz JC. Anaesthesiology students' Non-Technical Skills: Development and evaluation of a behavioural marker system for students (AS-NTS) zurückgegriffen. BMC Medical Education. 2019;19:205. doi:10.1186/s12909-019-1609-8

Abbildung 8: Anaesthesiology Students' Non-Technical Skills (AS-NTS)

Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Moll-Khosrawi et al. (2019)

Das Instrument *Anaesthesiology Students' Non-Technical Skills* (AS-NTS) wurde zur strukturierten Erfassung nicht-technischer Fähigkeiten von Medizinstudenten im Bereich der Anästhesiologie und Notfallmedizin entwickelt. Ausgangspunkt für die Entwicklung war die im Beitrag dargestellte Annahme, dass neben technischen Fertigkeiten insbesondere nicht-technische Kompetenzen wesentlich zur sicheren und effizienten Patientenversorgung beitragen. Da Defizite in diesen Kompetenzen als bedeutsame Ursache für unsichere Versorgung, unerwünschte Ereignisse und Störungen der Teamarbeit beschrieben werden, wird im zugrunde liegenden Dokument die Notwendigkeit betont, NTS nicht erst in der Weiterbildung, sondern bereits im Medizinstudium gezielt zu vermitteln und zu erfassen. Vor diesem Hintergrund verfolgten die Autoren das Ziel, ein speziell für Studenten geeignetes, praktikables und valides

Bewertungsinstrument zu entwickeln, das in simulationsbasierten Lehrsettings der Anästhesiologie und Notfallmedizin eingesetzt werden kann.

Die Entwicklung des AS-NTS erfolgte in einem mehrstufigen, empirisch fundierten Verfahren. Zunächst wurde auf Grundlage einer Literaturrecherche eine Liste relevanter nicht-technischer Fähigkeiten erstellt. Daran anschließend wurden diese Inhalte in einer Expertengruppe sowie in einer Fokusgruppe mit anästhesiologischen Fachärzten diskutiert und hinsichtlich ihrer Relevanz für die Zielgruppe der Studenten eingeordnet. Im weiteren Verlauf wurden halbstrukturierte Interviews und Feldbeobachtungen genutzt, um zu prüfen, welche der identifizierten Kompetenzen in simulationsbasierten Lehrsituationen tatsächlich beobachtbar und praktikabel beurteilbar sind. Die Entwicklung mündete schließlich in eine Implementierungs- und Validierungsphase, in der das Instrument unter realen Bedingungen des undergraduate Trainings getestet wurde. Das Vorgehen verdeutlicht, dass die Konstruktion des AS-NTS nicht als bloße Übernahme bestehender Taxonomien erfolgte, sondern als gezielte Adaption etablierter NTS-Konzepte an die Anforderungen und den Entwicklungsstand von Medizinstudenten.

Die konzeptuelle Grundlage des Instruments bildeten verschiedene in der Literatur beschriebene nicht-technische Fähigkeiten, darunter Situational Awareness, Priorisierung, Entscheidungsfindung, Maintaining Standards, Koordination von Teammitgliedern und Aktivitäten, Kommunikation, Leadership, Autorität und Durchsetzungsfähigkeit, Teambildung, Teamorientierung beziehungsweise Teamarbeit, Konfliktlösung sowie die Unterstützung anderer. Im Entwicklungsprozess zeigte sich jedoch, dass nicht alle dieser Fähigkeiten für die Zielgruppe der Studenten gleichermaßen geeignet waren. Zum einen wurden einzelne Fähigkeiten in den Fokusgruppen nicht als vorrangig relevant für undergraduate Lerner eingeschätzt, zum anderen erwiesen sich einige Kompetenzen in der Feldbeobachtung als schwer oder nicht hinreichend beobachtbar. Das Dokument betont in diesem Zusammenhang, dass insbesondere Instrumente, die ursprünglich für Fachärzte oder fortgeschrittene Weiterbildungsassistenten konzipiert wurden, nicht ohne Weiteres auf Studenten übertragbar sind. Daher wurde im Rahmen der Instrumentenentwicklung eine Reduktion und Reorganisation der zugrunde liegenden NTS vorgenommen.

Das finale AS-NTS besteht aus drei Dimensionen, die die zentralen Beobachtungseinheiten des Instruments bilden. Die erste Dimension trägt die Bezeichnung *Planning tasks, prioritising and problem solving* und entstand vor allem aus der Zusammenführung der zuvor getrennt betrachteten Bereiche *Decision making* und *Task management*. Die Autoren begründen diese Fusion damit, dass komplexe Entscheidungsfindung in der differenzierten Form, wie sie in Instrumenten für erfahrene Fachkräfte beschrieben wird, für Studenten in frühen Ausbildungs-

phasen nur eingeschränkt angemessen sei. Stattdessen wurde der Fokus auf jene beobachtbaren Vorläuferprozesse gelegt, die sich im studentischen Simulationskontext erfassen lassen, insbesondere das strukturierte Bearbeiten von Problemen, das Setzen von Prioritäten und das planvolle Abarbeiten notwendiger Handlungsschritte. Die zweite Dimension, *Teamwork and leadership*, bündelt Kompetenzen wie Koordination, Kommunikation und Führung. Sie zielt damit auf jene Verhaltensweisen ab, die für die Leitung eines Teams und die gemeinsame Aufgabenbewältigung in einer Notfallsituation bedeutsam sind. Die dritte Dimension, *Team orientation*, ergänzt dieses aufgabenbezogene Verständnis von Teamarbeit um eine stärkere beziehungs- und teambezogene Perspektive. Während die zweite Dimension auf die kollaborativen Prozesse zur Zielerreichung fokussiert, richtet sich die dritte auf Verhaltensweisen, die den Aufbau, die Einbindung und die Aufrechterhaltung eines funktionierenden Teams unterstützen.

Die konkrete Struktur des Instruments ist in Form eines verhaltensbasierten Ratingsystems angelegt. Die Beurteilung erfolgt nicht auf Ebene zahlreicher Einzelitems, sondern auf Ebene der drei genannten Dimensionen. Für jede dieser Dimensionen steht eine fünfstufige Likert-Skala von „very poor“ bis „very good“ zur Verfügung. Zur Unterstützung einer möglichst einheitlichen Bewertung wurden jeder Dimension verhaltensverankerte Beispiele guter und schlechter Performanz zugeordnet. Für die erste Dimension wird eine gute Leistung etwa dadurch charakterisiert, dass die studentische Person den CPR-Algorithmus durchführt, alle erforderlichen Schritte ergreift sowie Aufgaben priorisiert und den Arbeitsablauf strukturiert. Eine schlechte Leistung wird demgegenüber dadurch beschrieben, dass sich die Person durch weniger wichtige Aufgaben ablenken lässt. In der zweiten Dimension gelten klare Instruktionen an das Team, die Vergabe von Verantwortlichkeiten, das Teilen neuer Erkenntnisse, das Herstellen eines gemeinsamen Situationsverständnisses sowie erkennbares Führen des Teams als Merkmale guter Performanz. Schlechte Leistung liegt hier beispielsweise dann vor, wenn Teammitglieder nicht einbezogen werden, Änderungen im Vorgehen nicht kommuniziert werden, unklar bleibt, ob Aufgaben verstanden wurden, oder wenn die Person das Team eher verwirrt als anleitet. Die dritte Dimension wird durch Verhaltensweisen operationalisiert, die eine aktive Einbindung der Teammitglieder in diagnostische und therapeutische Prozesse erkennen lassen. Positiv bewertet werden unter anderem das aktive Nachfragen nach Informationsbedarfen, das laute Denken über eigene Annahmen und die Schaffung einer guten Teamatmosphäre. Als negative Verhaltensweisen gelten hingegen das Nicht-Einholen von Meinungen aus dem Team, das Ignorieren von Vorschlägen, die Abwertung anderer Teammitglieder sowie das Unterlassen entlastender Maßnahmen, etwa eines Wechsels bei der kardiopulmonalen Reanimation.

Ein zentrales Merkmal des AS-NTS besteht darin, dass das Instrument ausdrücklich für den Einsatz in der undergraduate Lehre konzipiert wurde. Es wurde in simulationsbasierten Trainings der Anästhesiologie und Notfallmedizin an der Medizinischen Fakultät Hamburg eingesetzt. Die Validierung erfolgte in vier curricular verankerten Lehrformaten, nämlich in drei aufeinander aufbauenden Advanced-Cardiac-Life-Support-Kursen sowie in einer Operationsraumsimulation. Mit jedem Studienabschnitt nahmen die technischen und nicht-technischen Anforderungen der Szenarien zu. Die Studenten arbeiteten in Dreiergruppen, wobei jeweils eine Person die Rolle der ärztlichen Leitung übernahm und die anderen Gruppenmitglieder als paramedizinisches oder anästhesiologisches Assistenzpersonal agierten. Bewertet wurde ausschließlich die studentische Person in der leitenden ärztlichen Rolle. Damit ist das Instrument auf Situationen zugeschnitten, in denen Studenten unter simulierten Bedingungen Führungsverhalten, Koordination, Prioritätensetzung und Teaminteraktion zeigen.

Hinsichtlich seiner psychometrischen Eigenschaften weist das AS-NTS nach den im Dokument dargestellten Ergebnissen günstige Kennwerte auf. Die Inhaltsvalidität wurde mithilfe des Content Validity Index für jede der drei Dimensionen bestimmt. Dabei ergaben sich Werte von 0,90 für die erste, 0,95 für die zweite und 0,80 für die dritte Dimension. Da im Dokument ein Wert von 0,75 oder höher als exzellent eingeordnet wird, erreichten alle drei Dimensionen des Instruments eine exzellente Inhaltsvalidität. Dies spricht dafür, dass die im AS-NTS enthaltenen Dimensionen aus Sicht der Beurteiler Fachpersonen als inhaltlich relevant für die Erfassung studentischer nicht-technischer Fähigkeiten angesehen wurden. Gleichzeitig wird im Dokument jedoch darauf hingewiesen, dass diese Berechnung auf den Einschätzungen von 21 Anästhesiologen beruhte und die begrenzte Stichprobengröße daher als Einschränkung zu berücksichtigen ist, da bei kleineren Stichproben der Einfluss zufälliger Übereinstimmung steigt.

Besonders ausführlich wird im Dokument die Interrater-Reliabilität untersucht, die zu den zentralen Qualitätsmerkmalen des Instruments zählt. Insgesamt nahmen 21 Anästhesiologen als Rater teil, die sich hinsichtlich Geschlechts, Alter, Lehrerfahrung und Weiterbildungsstand unterschieden. Bemerkenswert ist, dass diese Beurteiler lediglich eine fünfminütige Einführung in das Instrument erhielten. Die Reliabilität wurde in einem zweistufigen Verfahren geprüft. Zunächst wurden sechs Raterpaare analysiert, die jeweils mindestens sechs identische Szenarien unabhängig voneinander beurteilt hatten. In einem zweiten Schritt wurde der Gesamtdatensatz von 98 Simulationsszenarien ausgewertet, wobei die Daten nach Ausbildungsstand der Rater aggregiert wurden, um zu prüfen, ob medizinische Erfahrung oder die Beteiligung an der Instrumentenentwicklung einen Einfluss auf die Übereinstimmung haben könnten. Die Auswertung erfolgte mittels Intraclass Correlation Coefficient für ordinal skalierte Daten sowie

mittels Cohen's Kappa. Die berichteten Werte zeigen überwiegend hohe bis exzellente Übereinstimmungen. Im Abstract wird eine mittlere Interrater-Reliabilität von 0,89 angegeben. Auch die differenzierten Analysen der Raterpaare und der aggregierten Ausbildungsgruppen zeigen mehrheitlich Werte im Bereich guter bis exzellenter Übereinstimmung. Lediglich in der Vergleichsgruppe von Ratern im dritten und fünften Weiterbildungsjahr fiel die Übereinstimmung in einer Dimension niedriger aus und wurde dort nur als fair beschrieben. Insgesamt interpretieren die Autoren die Befunde jedoch dahingehend, dass die Anwendung des AS-NTS nur in geringem Maße von der Lehrerfahrung, dem anästhesiologischen Ausbildungsstand oder einer besonderen Vertrautheit mit dem Instrument abhängt.

Neben Validität und Reliabilität wurde auch die Praktikabilität des Instruments untersucht. In diesem Zusammenhang wurde AS-NTS mit dem etablierten System *Anaesthetists' Non-Technical Skills* (ANTS) verglichen. Laut Dokument zeigte sich in Interviews mit acht Anästhesiologen im ersten Weiterbildungsjahr, die beide Instrumente in Simulationstrainings einschließlich Videoaufzeichnungen angewendet hatten, dass ANTS für Studenten sowie für Ärzte in den ersten beiden Weiterbildungsjahren als zu komplex wahrgenommen wurde. Darüber hinaus wurde beschrieben, dass ANTS ohne Videoaufzeichnungen kaum vollständig eingesetzt werden könne, was die Anwendung im regulären Unterricht erheblich erschwere und die Feedbackschleife verzögere. Demgegenüber wurde das AS-NTS als praktikabel und unmittelbar einsetzbar bewertet. Auch eine zusätzliche Evaluation durch 21 Anästhesiologen, die das Instrument mindestens dreimal in der studentischen Lehre verwendet hatten, bestätigte dessen Feasibility und praktische Nutzbarkeit. Damit zeigt sich, dass das AS-NTS nicht nur gute psychometrische Eigenschaften aufweist, sondern auch den Anforderungen des Lehralltags entgegenkommt.

Im Vergleich zu ANTS liegt die besondere Stärke des AS-NTS somit in seiner zielgruppenspezifischen Reduktion und Adaptation. Während ANTS für erfahrene Anästhesisten beziehungsweise fortgeschrittene Weiterbildungsstufen entwickelt wurde und dementsprechend komplexere Strukturen aufweist, ist AS-NTS auf den Kompetenzstand von Studenten zugeschnitten. Die Reduktion auf drei Dimensionen, die Orientierung an beobachtbaren Vorläuferfähigkeiten sowie die Möglichkeit des Einsatzes ohne Videoanalyse stellen wesentliche Unterschiede dar. Diese Vereinfachung ist im Dokument nicht als Verlust an fachlicher Substanz dargestellt, sondern als notwendige didaktische Anpassung an die Anforderungen und Möglichkeiten undergraduate Lerner.

Gleichwohl weist das Dokument auch auf Limitationen des Instruments und seiner Evaluation hin. So wurde das AS-NTS bislang ausschließlich in deutscher Sprache, an einer einzelnen

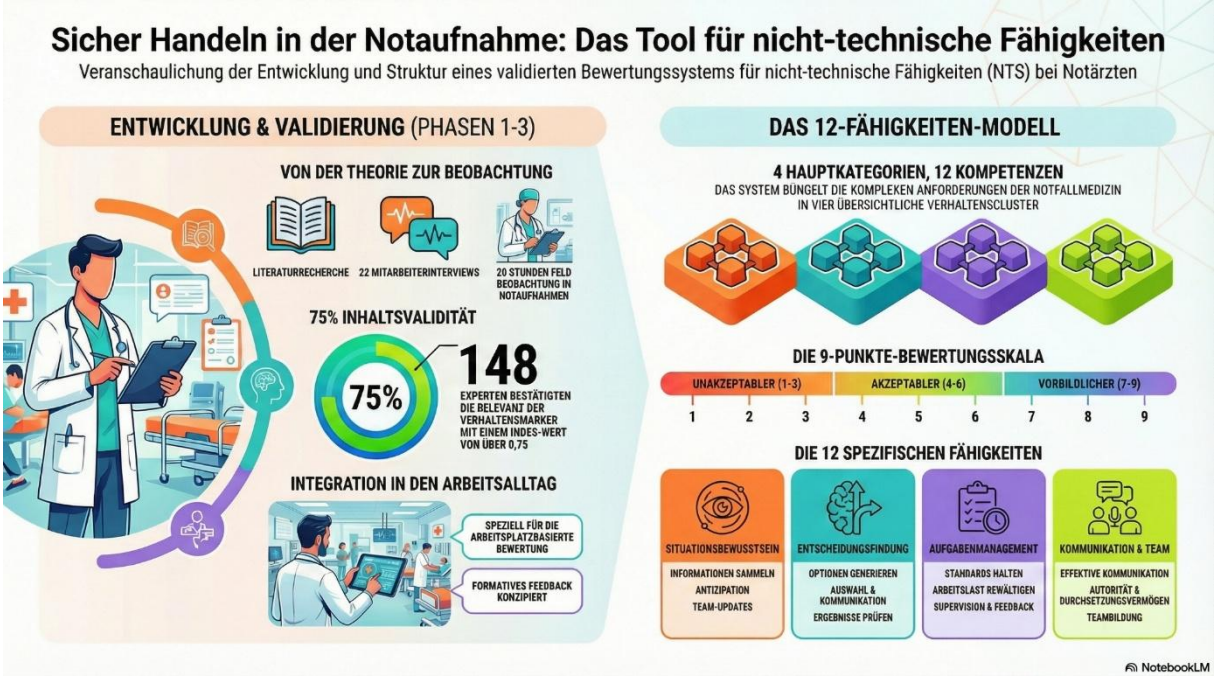
Institution und mit einer begrenzten Zahl von Lehrern untersucht. Die Autoren heben daher hervor, dass weitere Studien erforderlich sind, um Validität, Reliabilität und Praktikabilität der englischen Version zu prüfen. Zudem ist zu berücksichtigen, dass nicht alle in bestehenden Taxonomien beschriebenen nicht-technischen Fähigkeiten in das finale Instrument übernommen wurden. Einige Fähigkeiten wurden ausgeschlossen, weil sie im Simulationskontext nicht hinreichend beobachtbar waren oder für Studenten als noch nicht vollständig entwickelt galten. Ferner könnte kritisch angemerkt werden, dass die Bewertung auf Dimensionsebene eine geringere Differenzierung ermöglicht als stärker ausdifferenzierte Instrumente, die einzelne Teilkompetenzen separat erfassen. Das Dokument begegnet diesem Einwand jedoch mit dem Hinweis, dass die Zusammenführung mehrerer Teilbereiche, insbesondere von Entscheidungsfindung und Aufgabenmanagement, der Zielgruppe und dem Ausbildungsstand der Studenten angemessen sei.

Zusammenfassend kann das AS-NTS auf Grundlage des vorliegenden Dokuments als ein speziell für Medizinstudenten entwickeltes, verhaltensbasiertes und curricular gut integrierbares Instrument zur Erfassung nicht-technischer Fähigkeiten beschrieben werden. Es beruht auf einer systematischen, empirisch fundierten Entwicklung, umfasst drei inhaltlich klar begründete Dimensionen und weist exzellente Werte für die Inhaltsvalidität sowie überwiegend hohe bis exzellente Interrater-Reliabilität auf. Hinzu kommt eine hohe Praktikabilität im simulationsbasierten Unterricht, die insbesondere im Vergleich zu komplexeren Instrumenten wie ANTS als Vorteil hervortritt. Trotz der genannten Einschränkungen spricht die im Dokument dargestellte Evidenz dafür, dass AS-NTS eine geeignete Grundlage für die strukturierte Beurteilung und Rückmeldung nicht-technischer Fähigkeiten in der undergraduate anästhesiologischen und notfallmedizinischen Ausbildung darstellt.

5.6 Instrument zur Erfassung nontechnical skills von Emergency Physicians

Quelle: Flowerdew L, Brown R, Vincent C, Woloshynowych M. Development and validation of a tool to assess emergency physicians' nontechnical skills. Ann Emerg Med. (2012) 59:376–85.e4.

Abbildung 9: Instrument zur Erfassung nontechnical skills von Emergency Physicians



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Flowerdew et al. (2012)

Das von Flowerdew et al. Entwickelte Instrument zur Erfassung der nontechnical skills von Emergency Physicians stellt ein verhaltensbasiertes Beobachtungsinstrument dar, das speziell für die Anforderungen der Notaufnahme konzipiert wurde. Ziel der Entwicklung war es, ein strukturiertes Verfahren bereitzustellen, mit dem die nicht-technischen Kompetenzen von Notfallmedizinern unter realen Arbeitsbedingungen beobachtet und beurteilt werden können. Im Zentrum stand dabei die Annahme, dass nontechnical skills als kognitive, soziale und personale Ressourcen technische Fertigkeiten ergänzen und wesentlich zu einer sicheren sowie effizienten Aufgabenbewältigung beitragen. Im Dokument wird hervorgehoben, dass bislang vorhandene Instrumente häufig entweder auf Teamarbeit im Reanimationskontext oder auf den Simulationsbereich begrenzt waren und damit nur einen Teil der tatsächlichen Anforderungen des Arbeitsalltags von Emergency Physicians abbildeten. Vor diesem Hintergrund verfolgte die Studie das Ziel, ein Instrument zu entwickeln, das nicht nur die Versorgung eines einzelnen kritisch kranken Patienten, sondern das gesamte Spektrum notfallmedizinischer Tätigkeit in der Notaufnahme einschließlich Routinearbeit, Unterbrechungen, Koordination und Management mehrerer Patienten gleichzeitig erfasst.

Die Entwicklung des Instruments erfolgte in einem mehrphasigen Verfahren, das auf der Triangulation unterschiedlicher Datenquellen beruhte. In einer ersten Phase wurden einschlägige Literatur sowie relevante Curricula systematisch analysiert, um eine vorläufige Liste potenziell relevanter nontechnical skills zu erstellen. Die Analyse bezog sich nicht nur auf wissenschaftliche Arbeiten zu Sicherheit und Fehlern in der Notaufnahme, sondern auch auf bereits vorhandene Beobachtungsinstrumente aus anderen Bereichen des Gesundheitswesens. Ergänzend wurden das College of Emergency Medicine Generic Skills Curriculum sowie das Medical Leadership Curriculum der Academy of Royal Medical Colleges herangezogen, um sicherzustellen, dass das Instrument mit den in der Weiterbildung beschriebenen Kompetenzanforderungen übereinstimmt. Aus dieser Synthese entstand zunächst eine Liste von 13 potenziell relevanten nontechnical skills, zu denen unter anderem Maintaining Standards, Decisionmaking, Managing Workload, Communicating, Resolving Conflict, Team Building, Leadership, Supporting Others, Teaching and Providing Feedback, Using Authority and Assertiveness, Situational Awareness, Coordinating Team Members sowie Supervising/Assessing Capabilities gehörten. Im Anschluss wurde die vorläufige Version des Instruments in einer kurzen Machbarkeitsprüfung erprobt, indem leitende Notfallärzte während einstündiger Beobachtungen mit dem Tool begleitet wurden. Diese frühe Testung diente dazu, die grundsätzliche Akzeptanz des Beobachtungsprozesses sowie die Beobachtbarkeit der vorgesehenen Kompetenzen zu prüfen.

In einer zweiten Entwicklungsphase wurden semistrukturierte Interviews mit ärztlichem und pflegerischem Personal der Notaufnahme durchgeführt und zusätzlich Feldbeobachtungen in zwei Londoner Teaching Hospitals vorgenommen. Ziel dieser Phase war es, festzustellen, ob die in der ersten Phase entwickelte Liste relevante Auslassungen aufwies, ob die vorgesehenen Skills in der Praxis tatsächlich beobachtbar waren und welche konkreten Verhaltensweisen als geeignete Marker für die einzelnen Kompetenzen dienen konnten. Insgesamt wurden 22 Interviews mit Consultants, Registrars, Ärzten niedrigerer Hierarchiestufen und Pflegekräften geführt. Die Interviewten beschrieben positives und negatives Teamverhalten, den Einfluss dieser Verhaltensweisen auf die Teamfunktion sowie Fehler und mögliche Verbesserungsansätze für die Zusammenarbeit. Die Aussagen wurden transkribiert und mit den in der ersten Phase identifizierten 13 Skills kodiert. Parallel dazu wurden 20 etwa einstündige Feldbeobachtungen in zwei Londoner Notaufnahmen durchgeführt, bei denen vor allem Registrars in ihrer Tätigkeit begleitet wurden. Die erhobenen Feldnotizen dokumentierten beobachtbares Verhalten, Kommunikationsereignisse, wahrgenommene Fehler und suboptimale Praxisbeispiele und wurden ebenfalls anhand des entwickelten Kodierschemas analysiert. Die Kombination aus Interview- und Beobachtungsdaten ermöglichte es, die relative Bedeutung einzelner Skills

empirisch zu überprüfen und zu analysieren, welche Kompetenzen im ED-Alltag tatsächlich sichtbar werden.

Im Verlauf dieser zweiten Phase zeigte sich, dass die ursprünglich identifizierten 13 Skills nicht unverändert in das finale Instrument übernommen werden konnten. Besonders deutlich wurde, dass Leadership nur schwer als eigenständige Kategorie zu isolieren war, da Führung in der Notaufnahme verschiedene andere nontechnical skills integriert, etwa Situational Awareness, Managing Workload, Decisionmaking, Teaching und Team Building. Aus diesem Grund wurde Leadership nicht als separates Element in das finale Instrument aufgenommen, sondern in die Definitionen und Verhaltensmarker anderer Skills integriert. Auch weitere Kompetenzen wurden überarbeitet, zusammengeführt oder in andere Kategorien überführt. So wurde Resolving Conflict nicht als unabhängiger Skill beibehalten, sondern in die Dimension Using Authority and Assertiveness integriert. Team Building, Teaching and Providing Feedback sowie Supervising/Assessing Capabilities wurden ebenfalls kritisch geprüft, da sie entweder eher als allgemeiner Interaktionsstil als als klar abgrenzbarer Skill auftraten oder nur eingeschränkt direkt beobachtbar waren. Gleichwohl wurde insbesondere die Bedeutung von Supervision und Feedback aufgrund ihrer engen Verbindung zu Sicherheit und Fehlervermeidung betont, sodass Supervising und Providing Feedback schließlich zu einem gemeinsamen Element zusammengeführt wurden. Das Ergebnis dieses Revisionsprozesses war ein finales Instrument mit 12 emergency medicine-spezifischen nontechnical skills, die in vier übergeordnete Kategorien eingeordnet wurden.

Die Struktur des finalen Instruments ist in der im Dokument abgebildeten Assessmentvorlage dargestellt. Die vier Hauptkategorien lauten Management and Supervision, Teamwork and Cooperation, Decision-Making sowie Situational Awareness. Innerhalb der Kategorie Management and Supervision werden die drei Elemente Maintaining Standards, Managing Workload sowie Supervising and Providing Feedback erfasst. Maintaining Standards bezieht sich auf das Einhalten klinischer und sicherheitsbezogener Standards unter Berücksichtigung organisatorischer Vorgaben sowie auf die Überwachung der Einhaltung dieser Standards. Managing Workload beschreibt das Management der eigenen und fremden Arbeitsbelastung, insbesondere durch Priorisieren, Delegieren, Einfordern von Hilfe und Unterstützen anderer, um sowohl Überlastung als auch Unterauslastung zu vermeiden. Supervising and Providing Feedback umfasst die Einschätzung von Fähigkeiten, das Erkennen von Wissenslücken sowie das Schaffen von Möglichkeiten für Lehre und konstruktives Feedback.

Die zweite Hauptkategorie, Teamwork and Cooperation, setzt sich aus Team Building, Communicating Effectively sowie Authority and Assertiveness zusammen. Team Building bezeich-

net die Motivation und Unterstützung des Teams sowie ein freundliches und ansprechbares Auftreten. Communicating Effectively meint die präzise und effektive mündliche wie schriftliche Informationsweitergabe sowie das Zuhören, Bestätigen und Klären von Informationen. Authority and Assertiveness umfasst ein situationsangemessen durchsetzungsfähiges Verhalten, effektive Konfliktlösung und die Fähigkeit, auch unter Druck ruhig zu bleiben. Die dritte Kategorie, Decision-Making, besteht aus Generating Options, Selecting and Communicating Options sowie Reviewing Outcomes. Generating Options beschreibt die Nutzung aller verfügbaren Informationsquellen und Ressourcen zur Entwicklung geeigneter Handlungsoptionen sowie die Einbeziehung des Teams in den Entscheidungsprozess. Selecting and Communicating Options bezieht sich auf das Abwägen von Risiken verschiedener Optionen, deren Diskussion im Team sowie die klare Mitteilung und gegebenenfalls Begründung der gewählten Entscheidung. Reviewing Outcomes meint die fortlaufende Überprüfung der Angemessenheit getroffener Entscheidungen unter Berücksichtigung neuer Informationen und die Sicherstellung, dass delegierte oder prioritäre Maßnahmen tatsächlich erfolgt sind. Die vierte Kategorie, Situational Awareness, umfasst Gathering Information, Anticipating und Updating the Team. Gathering Information bezieht sich auf das Wahrnehmen relevanter Hinweise in der Umgebung und das aktive Einholen von Informationen anderer. Anticipating beschreibt das Vorwegnehmen möglicher Probleme, etwa im Hinblick auf Personalverfügbarkeit oder räumliche Kapazitäten, und das Planen von Handlungsalternativen. Updating the Team umfasst die Überprüfung der Zuverlässigkeit von Informationen sowie die fortlaufende Kommunikation der aktuellen Situation an das Team, damit dieses über relevante Entwicklungen informiert bleibt.

Das Instrument ist als behavioral marker system konzipiert und operiert damit nicht primär über isolierte Itemformulierungen im Sinne eines klassischen Fragebogens, sondern über definierte Kompetenzbereiche, die durch konkrete Verhaltensanker operationalisiert werden. Für jedes der zwölf Elemente enthält das Tool Beispiele für gutes und schlechtes Verhalten. So wird im Bereich Maintaining Standards positives Verhalten etwa daran festgemacht, dass unleserliche Dokumentation bemerkt und deren Bedeutung thematisiert wird, dass die Stabilität eines kranken Patienten vor einem Transfer sichergestellt wird oder dass Leitlinien und Formulare eingehalten werden. Negatives Verhalten zeigt sich hier beispielsweise in fehlender zeitnaher Dokumentation, fehlender Händehygiene oder Nichteinhaltung von Sicherheitsprozeduren. Im Bereich Managing Workload gelten das Erkennen überlasteter Kollegen, das Sicherstellen angemessener Pausen sowie der effektive Umgang mit Unterbrechungen als positive Marker, während das Nichtreagieren auf Überlastung, der Verlust des Überblicks über die Abteilung oder unzureichende Eskalation bei Überforderung als negative Marker beschrieben werden. Entsprechend differenzierte Verhaltensmarker liegen auch für die übrigen Elemente vor. Im

Bereich Team Building werden etwa freundliche Reaktionen auf Hilfegesuche und motivierende Verhaltensweisen positiv hervorgehoben, während abrupter, unhöflicher oder belastender Umgang mit Teammitgliedern negativ bewertet wird. Für Communicating Effectively gelten präzise Übergaben, das Sicherstellen korrekten Verstehens wichtiger Botschaften und klare Überweisungen an Fachabteilungen als gute Praxis. Authority and Assertiveness wird positiv über angemessene Durchsetzungsfähigkeit, das Ansprechen von Bedenken gegenüber senioren Kollegen sowie Ruhe unter Druck operationalisiert. Die Skalenanker in den Entscheidungs- und Situational-Awareness-Dimensionen umfassen beispielsweise das aktive Einholen von Teaminput, die Überprüfung von Behandlungseffekten, die Nutzung des Patient Tracking Systems, die Antizipation erhöhter Nachfrage oder die fortlaufende Information des Teams über Statusveränderungen und neue Probleme.

Die Bewertung erfolgt mittels einer neun Punkte umfassenden Ratingskala. Diese ist in drei Leistungsbereiche gegliedert und differenziert zwischen einem unacceptable standard im Bereich von 1 bis 3 Punkten, einem acceptable standard im Bereich von 4 bis 6 Punkten und einem exemplary standard im Bereich von 7 bis 9 Punkten. Ein unacceptable standard ist nach der Definition des Instruments durch mehrere Beispiele schlechten Verhaltens oder durch Verhalten gekennzeichnet, das die Patientensicherheit unmittelbar beeinträchtigt. Ein acceptable standard beschreibt eine insgesamt zufriedenstellende Leistung mit überwiegend gutem Verhalten und entspricht dem Niveau eines kompetenten Trainees. Ein exemplary standard steht für eine konsistent hohe Leistung, die als Modell für andere Teammitglieder gelten kann. Die Wahl einer neun Punkte umfassenden Skala wird im Dokument damit begründet, dass differenziertere Skalen theoretisch eine höhere Reliabilität aufweisen können und in anderen Beurteilungskontexten insbesondere für die Unterscheidung zwischen grenzwertig unzureichender, zufriedenstellender und überdurchschnittlicher Leistung Vorteile gezeigt haben. Darüber hinaus wurde im Forschungsteam der Nutzen einer differenzierten Skala betont, um zwischen leicht schwacher Performanz und deutlichem Förderbedarf unterscheiden zu können.

Hinsichtlich der psychometrischen Eigenschaften liegt der Schwerpunkt der vorliegenden Studie auf der Inhaltsvalidität des Instruments. Diese wurde mithilfe einer Expertenbefragung unter 148 Mitarbeiter der Notfallmedizin überprüft, darunter Consultants, trainee registrars, middle-grade physicians und senior ED nurses. Die Befragten bewerteten 36 exemplarische Verhaltensweisen auf einer fünfstufigen Relevanzskala hinsichtlich ihrer Bedeutung für Emergency Physicians. Der Content Validity Index wurde für jedes Verhalten als Anteil der Befragten berechnet, die ein Verhalten als sehr wichtig oder essenziell einschätzten. Als Kriterium für eine exzellente Inhaltsvalidität wurde ein Wert von mindestens 0,75 angesetzt. Die

Ergebnisse zeigten, dass 75 % der Items diesen Schwellenwert erreichten. Neun Items unterschritten den Grenzwert und wurden auf dieser Grundlage sprachlich oder inhaltlich überarbeitet. Zusätzlich wiesen zwei Items mehr als vier fehlende Antworten auf und wurden ebenfalls revidiert. Das Dokument macht anhand der ergänzenden Tabellen deutlich, dass die Revision der Verhaltensmarker nicht allein auf numerischen CVI-Werten beruhte, sondern auch qualitative Kommentare der Befragten einbezog. So wurde etwa eine als zu scharf empfundene Formulierung wie „reprimands doctor“ in eine weniger sanktionierende und stärker beobachtungsbasierte Formulierung geändert. Ebenso wurden Formulierungen angepasst, wenn Teilnehmer auf Mehrdeutigkeiten, zu starke Rollenspezifik oder unpassende Konnotationen hingewiesen hatten. Darüber hinaus konnten die Befragten zusätzliche Verhaltensweisen oder Skills benennen, die ihrer Ansicht nach relevant, aber nicht im Instrument enthalten waren. Aus insgesamt 101 Vorschlägen wurden 53 potenziell relevante Verhaltensweisen oder Skills abgeleitet. Nach Analyse durch das Forschungsteam kam man jedoch zu dem Ergebnis, dass sämtliche relevanten Ergänzungen in die bestehenden zwölf nontechnical skills eingeordnet werden konnten, sodass keine wesentliche inhaltliche Lücke des Instruments angenommen wurde. Damit liefert die Studie neben der Relevanzprüfung einzelner Verhaltensmarker auch Evidenz für die inhaltliche Vollständigkeit des Instruments.

Gleichzeitig wird deutlich, dass die psychometrische Prüfung des Instruments im vorliegenden Dokument noch nicht abgeschlossen ist. Zwar liegt eine ausführliche Untersuchung der Inhaltsvalidität vor, eine numerische Prüfung der Interrater-Reliabilität des finalen Instruments wird jedoch nicht berichtet. In der Diskussion wird vielmehr ausdrücklich darauf hingewiesen, dass bei workplace-based assessments üblicherweise mehrere Beobachtungen durch verschiedene Beurteiler erforderlich sind, um ein angemessenes Reliabilitätsniveau zu erzielen, und dass dies in zukünftigen Studien weiter untersucht werden müsse. Auch die Akzeptanz des Instruments im breiteren Einsatz sowie seine Übertragbarkeit auf unterschiedliche Kontexte und Rollen innerhalb der Notfallmedizin werden als Gegenstand weiterer Forschung benannt. In Bezug auf den Simulator wird dem Instrument zwar eine gute Face Validity zugeschrieben, doch wird auch hier betont, dass zusätzliche Untersuchungen notwendig seien, bevor ein solcher Einsatz umfassend empfohlen werden könne.

Der vorgesehene Anwendungsbereich des Instruments liegt primär in der formativen Beurteilung am Arbeitsplatz. Das Tool wurde für die direkte Beobachtung einer ärztlichen Person über etwa eine Stunde im Setting der Notaufnahme entwickelt und soll dabei Reflexion, Rückmeldung, zukünftiges Lernen und die Beobachtung individueller Fortschritte unterstützen. Im Dokument wird hervorgehoben, dass die Förderung von Reflexion über nontechnical skills gerade

im schnelllebigen ED-Kontext von besonderer Bedeutung ist. Da Ärzte ihre eigenen Kompetenzen, insbesondere im interpersonellen Bereich, nur begrenzt zutreffend einschätzen können, wird die strukturierte Fremdbeobachtung als wichtiger Bestandteil professioneller Entwicklung betrachtet. Idealerweise sollte jede Beobachtung mit einer ausführlichen Debriefing- und Feedbackphase verbunden werden, in der nicht nur Verhalten rückgemeldet, sondern auch die hinter dem beobachteten Handeln liegenden Überlegungen reflektiert werden. Darüber hinaus wird im Dokument darauf verwiesen, dass das Instrument potenziell auch für kürzere Beobachtungssequenzen, für spezifische Ereignisse wie die Beurteilung eines vom Junior vorgestellten Patientenmanagementplans, für Lehrsituationen und zur Strukturierung von Fall- oder Zwischenfalldiskussionen eingesetzt werden kann. Ferner wird eine mögliche Nutzung für Peer Assessment auf Registrar- und Consultant-Ebene in Aussicht gestellt.

Trotz seiner Stärken ist das Instrument mit mehreren Limitationen verbunden, die im Dokument ausdrücklich benannt werden. Zunächst beruht die Entwicklungsphase auf Daten aus lediglich zwei Londoner Teaching Hospitals, was die Generalisierbarkeit auf andere Notaufnahmen einschränken könnte. Zwar wurde die Validierungsbefragung landesweit durchgeführt und bezog Teilnehmer aus verschiedenen Regionen des Vereinigten Königreichs ein, dennoch bleibt der unmittelbare Entwicklungskontext begrenzt. Hinzu kommt, dass der Umfang der Feldbeobachtungen mit 20 Stunden relativ klein war, auch wenn die Autoren dies als ausreichend für die Identifikation zentraler nontechnical skills einschätzen. Eine weitere Limitation besteht darin, dass negative Verhaltensmarker aus Gründen der Praktikabilität nicht in die Inhaltsvalidierungsbefragung einbezogen wurden. Damit konnte die Relevanz negativer Marker in diesem Schritt nicht in gleicher Weise systematisch geprüft werden wie jene der positiven Marker. Zudem kann für die Befragung ein Sampling Bias nicht ausgeschlossen werden, da die Teilnehmer überwiegend selbstselektiert waren und aus Veranstaltungen zu akademischer Notfallmedizin und Patientensicherheit rekrutiert wurden. Diese Gruppe dürfte zwar besonders kompetent hinsichtlich der zu Beurteiler Konstrukte gewesen sein, repräsentiert aber möglicherweise nicht die gesamte notfallmedizinische Community. Schließlich bleibt als wesentliche Einschränkung festzuhalten, dass zum Zeitpunkt der Publikation noch keine umfassenden Daten zur Reliabilität des finalen Instruments vorlagen.

Insgesamt zeigt das Dokument, dass das von Flowerdew et al. Entwickelte Instrument einen wichtigen Schritt hin zu einer strukturierten, arbeitsplatzbasierten Erfassung nontechnical skills in der Notaufnahme darstellt. Die Entwicklung erfolgte auf Grundlage eines systematischen und methodisch breit angelegten Prozesses, der Literatur, Curricula, Interviews, Feldbeobachtungen und eine umfangreiche Expertenbefragung miteinander verknüpfte. Das Ergebnis ist

ein ED-spezifisches behavioral marker system mit zwölf klar definierten Kompetenzbereichen, die zentrale Anforderungen notfallmedizinischer Arbeit jenseits rein technischer Fertigkeiten abbilden. Besonders hervorzuheben sind die differenzierte Struktur mit vier Hauptkategorien, die konkrete Verhaltensverankerung der Bewertung sowie die empirisch überprüfte Inhaltsvalidität. Auch wenn weiterführende Untersuchungen zu Reliabilität, Akzeptanz und Übertragbarkeit erforderlich sind, bietet das Instrument eine fundierte Grundlage für formative Beobachtung, strukturiertes Feedback und die gezielte Förderung nontechnical skills von Emergency Physicians im klinischen Alltag.

5.7 Assessment of Obstetrical Team Performance (AOTP) und dem Global Assessment of Obstetrical Team Performance (GAOTP)

Quelle: Tregunno D, Pittini R, Haley M, Morgan PJ. Development and usability of a behavioural marking system for performance assessment of obstetrical teams. Qual Saf Health Care. (2009) 18:393–6. Doi: 10.1136/qshc.2007.026146

Abbildung 10: Assessment of Obstetrical Team Performance (AOTP) und dem Global Assessment of Obstetrical Team Performance (GAOTP)

Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Tregunno et al. (2009)

Mit dem *Assessment of Obstetrical Team Performance* (AOTP) und dem *Global Assessment of Obstetrical Team Performance* (GAOTP) wurden zwei verhaltensbasierte Bewertungsinstrumente für die Leistungsbeurteilung interdisziplinärer Teams in der Geburtshilfe entwickelt. Ausgangspunkt der Instrumentenentwicklung war die im Dokument beschriebene Problemlage, dass Teamarbeit und Kommunikation als zentrale Ursachen von Sentinel Events mit Säuglingstod und Verletzungen während der Geburt identifiziert wurden, gleichzeitig jedoch valide und reliable Marker zur Bewertung obstetrischer Teamleistung fehlten. Obwohl Teamtraining in der Geburtshilfe als wichtige Strategie zur Verbesserung mütterlicher und fetaler Sicherheit hervorgehoben wird, fehlten nach Darstellung der Autoren geeignete Instrumente, um die Wirksamkeit entsprechender curricularer Maßnahmen systematisch zu überprüfen. Vor diesem Hintergrund verfolgte die Studie das Ziel, zwei spezifisch auf die Geburtshilfe ausgerichtete behavioural marking systems zu entwickeln und ihre Usability im simulationsbasierten Einsatz zu prüfen.

Die Entwicklung der beiden Instrumente erfolgte mittels qualitativer Methoden und beruhte auf mehreren komplementären Datenquellen. Als Grundlage diente zunächst eine frühere Studie, in der zwölf interdisziplinäre obstetrische Teams in Teams mit jeweils fünf bis sechs Mitgliedern wiederholt einem von vier geburtshilflichen Hochrisikoszenarien zugeteilt worden waren. Nach Abschluss der Simulationen wurden die Teilnehmer gebeten, narrative Beschreibungen von Verhaltensweisen zu formulieren, die nach ihrer Einschätzung zur Funktionsfähigkeit der Teams beigetragen hatten. Ergänzend wurde eine Fokusgruppe mit Personen durchgeführt, die nicht an den Simulationen beteiligt waren, um Sicherheits- und Teamworkfragen in obstetrischen Settings weiter zu explorieren. Darüber hinaus erfolgte eine Literaturrecherche nach bereits publizierten behavioural marking systems für die Beurteilung medizinischer und obstetrischer Teams. Da nach Aussage des Dokuments kein spezifisches behavioural marking system für obstetrische Teams identifiziert werden konnte, war die Entwicklung von AOTP und GAOTP als genuine Neuentwicklung innerhalb dieses Fachgebiets angelegt.

Ein wesentlicher Entwicklungsschritt bestand in der Sichtung der Videobänder aus den Hochrisiko-Simulationen durch 13 Reviewer, die unterschiedlichen Perspektiven in den Entwicklungsprozess einbrachten. Diese umfassten sowohl teambezogen erfahrene als auch inhaltlich naive Personen sowie inhaltlich ausgewiesene Experten. Die Reviewer sichteten die Videomitschnitte der obstetrischen Teams und generierten eine Liste von Verhaltensaspekten, die aus ihrer Sicht die Teamleistung positiv oder negativ beeinflussten. Simulationsteilnehmer, Fokusgruppenmitglieder und Videoreviewer identifizierten auf diese Weise insgesamt 1294 Verhaltensitems, die für die obstetrische Teamleistung relevant erschienen. Diese breite em-

pirische Basis verdeutlicht, dass die Instrumente nicht lediglich aus theoretischen Überlegungen abgeleitet, sondern stark an beobachteten und berichteten Verhaltensweisen orientiert entwickelt wurden.

Die qualitative Auswertung der erhobenen Daten erfolgte mithilfe von NVivo. Der Erstautor analysierte das Material Zeile für Zeile, um wiederkehrende Themen und Subthemen zu identifizieren. Die Co-Investigatoren prüften diese Ergebnisse ebenfalls und verglichen ihre Einschätzungen in einem iterativen Konsensprozess. Auf dieser Grundlage wurden schließlich sechs Themen und 18 Subthemen der obstetrischen Teamleistung herausgearbeitet. Im nächsten Schritt entwickelten die Forscher verhaltensverankerte Beschreibungen für exzellente und schlechte Teamleistung innerhalb der jeweiligen Kategorien. Diese anchored descriptors bildeten die Grundlage für die Konstruktion der beiden Instrumente. Das AOTP wurde als detailliertes Instrument mit Themen und Subthemen konzipiert und mit einer fünfstufigen Likert-Skala versehen, wobei der Wert 1 eine schlechte und der Wert 5 eine exzellente Leistung abbildet. Das GAOTP wurde demgegenüber als globaleres Instrument entwickelt, das ausschließlich die sechs Themen und nicht die differenzierten Subthemen umfasst. Bereits aus dieser Struktur geht hervor, dass das AOTP stärker auf differenziertes, formatives Feedback und das GAOTP eher auf globalere, summative Gesamtbeurteilungen ausgerichtet ist.

Die inhaltliche Struktur des AOTP und damit zugleich die Grundlage des GAOTP wird im Dokument in Form von sechs Themen mit insgesamt 18 Subthemen dargestellt. Das erste Thema, *Communication with patient and partner*, umfasst die Subthemen *Information sharing*, *Reassuring attitude* und *Partner management*. Diese Kategorie macht deutlich, dass die Instrumente nicht allein auf die Interaktion zwischen professionellen Teammitgliedern fokussieren, sondern auch patienten- und familienbezogene Teamverhaltensweisen systematisch berücksichtigen. Das zweite Thema, *Task/case management*, beinhaltet die Subthemen *Plan of action*, *Problem solving* und *Resource utilisation* und bildet somit zentrale Aspekte der fallbezogenen Steuerung und Ressourcennutzung ab. Das dritte Thema, *Teamwork*, umfasst *Leadership*, *Role assignment* und *Team interaction* und adressiert damit klassische Kernelemente der Teamkoordination. Das vierte Thema, *Situational awareness*, ist mit *Anticipation*, *Realising limitations*, *Fixation*, *Responsiveness* und *Vigilance* die umfangreichste Kategorie und spiegelt die hohe Bedeutung des Lagebewusstseins in dynamischen obstetrischen Notfallsituationen wider. Das fünfte Thema, *Communication with team members*, besteht aus *Focussed communication* und *Closing the loop* und fokussiert die Qualität der teaminternen Informationsweitergabe. Das sechste Thema, *Environment in the room*, umfasst die Subthemen *Management of*

disruptive behaviour und *Atmosphere of the room* und verweist damit auf die Relevanz des sozialen und situativen Klimas im Versorgungsraum.

Als behavioural marking systems operieren die Instrumente nicht primär über knappe Itemformulierungen im Sinne klassischer Testinventare, sondern über definierte Themen und Subthemen, die mit konkreten Verhaltensankern versehen sind. Diese Verhaltensanker sollen den Beurteiler eine möglichst konsistente und inhaltlich nachvollziehbare Einschätzung des beobachteten Teamverhaltens ermöglichen. Exemplarisch werden im Dokument für das Thema *Communication with patient and partner* sowohl Marker schlechter als auch Marker exzellenter Teamleistung dargestellt. Im Subthema *Information sharing* wird schlechte Teamleistung etwa dadurch beschrieben, dass keine Teammitglieder eine Einführung vornehmen, in den Raum gesprochen wird und Patient oder Partner selbst nachfragen müssen, was geschieht und wie es dem Kind geht. Exzellente Leistung liegt dagegen vor, wenn sich Teammitglieder bei Betreten des Raumes vorstellen und ihre Rollen benennen, Patient und Partner direkt ansprechen und während des gesamten Ereignisses fortlaufend kommunizieren. Im Subthema *Reassuring attitude* wird schlechte Leistung dadurch charakterisiert, dass niemand Verantwortung für die emotionale Begleitung von Patient oder Partner übernimmt, sodass deren Angst im Verlauf zunimmt. Exzellente Leistung wird demgegenüber dann angenommen, wenn mindestens ein Teammitglied diese Verantwortung übernimmt und kontinuierlich Information sowie Beruhigung anbietet. Im Bereich *Partner management* besteht schlechte Teamleistung darin, dass Verhaltensänderungen des Partners unbeachtet bleiben und frühe Anzeichen störenden Verhaltens nicht adressiert werden, während exzellente Leistung dann vorliegt, wenn das Team potenziell störendes Verhalten antizipiert und frühzeitig eingreift, um eine Eskalation zu verhindern. Diese Beispiele verdeutlichen, dass das AOTP und das GAOTP nicht nur technische Abläufe oder interne Koordination erfassen, sondern eine ausgesprochen patientenzentrierte und interaktive Konzeption von Teamleistung zugrunde legen.

Im Hinblick auf die Funktion der beiden Instrumente beschreibt das Dokument eine klare Differenzierung. Das AOTP soll als Vorlage für formative Rückmeldungen zur Teamleistung dienen und erlaubt aufgrund seiner feineren Ausdifferenzierung eine detaillierte Erfassung einzelner Dimensionen der Teamarbeit. Es ist damit besonders geeignet, Verbesserungsbedarfe in spezifischen Bereichen zu identifizieren und Entwicklungen der Teamleistung über die Zeit nachzuzeichnen. Das GAOTP soll demgegenüber eine globalere, eher summative Einschätzung der Teamleistung ermöglichen. Im Discussion-Teil wird hervorgehoben, dass das AOTP im Unterschied zur Mayo High Performance Teamwork Scale wesentlich feingliedriger ist. Während die genannte Vergleichsskala als kurzes Selbstbeurteilungsinstrument mit 16 Ver-

haltenselementen für Lerner konzipiert wurde, umfasst das AOTP 18 Subthemen in sechs Themen und ist darauf angelegt, beobachtete Teamleistung differenziert zu erfassen. Dies macht deutlich, dass die beiden entwickelten Instrumente nicht primär als schnelle Selbsteinschätzung, sondern als strukturierte Werkzeuge für externe Beobachtung und Leistungsbeurteilung vorgesehen sind.

Ein zentrales Ergebnis der Studie betrifft die Usability des AOTP und des GAOTP. Nachdem die Prototypen entwickelt worden waren, wurden sie von 14 weiteren Reviewern erprobt, die nicht an der Generierung der Verhaltensitems beteiligt gewesen waren. Diese Gruppe bestand aus drei Pflegefachpersonen, sechs Ärzten und fünf Universitätsstudenten. Die Reviewer sichteten die Videobänder von zwölf Teams, die jeweils eines von vier Szenarien in einem High-Fidelity-Simulationszentrum bearbeitet hatten, und bewerteten die Teamleistung unmittelbar im Anschluss mithilfe beider Instrumente. Danach füllten sie einen Fragebogen zur Nutzbarkeit und zur benötigten Zeit aus. Die Ergebnisse zeigen eine insgesamt hohe Akzeptanz und gute Handhabbarkeit des AOTP. Die mediane Bearbeitungszeit lag bei 7,5 Minuten, wobei die Spannweite von 1,5 bis 50 Minuten reichte. Drei Viertel der Reviewer beschrieben den Zeitaufwand als moderat und handhabbar. In der detaillierten Usability-Befragung stimmten alle Reviewer zu, dass die Liste der im AOTP enthaltenen Verhaltensweisen umfassend sei. Ebenfalls einhellig wurde das Rasterformat als leicht handhabbar bewertet. Die große Mehrheit der Befragten verneinte, dass das Instrument zu viele oder zu wenige Verhaltensweisen enthalte, und gab an, die Themen, Subthemen sowie die Indikatoren guter und schlechter Leistung gut zu verstehen. Darüber hinaus stimmte die Mehrheit der Aussage zu, dass das AOTP ihre Einschätzung der Teamleistung zutreffend widerspiegele. Insgesamt schlussfolgern die Autoren aus diesen Ergebnissen, dass das AOTP umfassend, schnell und einfach anzuwenden sei und eine akkurate Reflexion der Beurteilungen der Reviewer erlaube.

Bezüglich der psychometrischen Eigenschaften ist hervorzuheben, dass die vorliegende Studie nicht primär der abschließenden Validierung oder Reliabilitätsprüfung der Instrumente diente. Dies wird im Dokument ausdrücklich betont. Gleichwohl werden erste Kennwerte berichtet, die Hinweise auf die Qualität der Instrumente geben. Zur Prüfung des Einflusses von Training wurden drei naive Rater gebeten, drei zufällig ausgewählte Szenarien zu bewerten. Anschließend sahen diese gemeinsam mit einem erfahrenen Moderator ein weiteres Szenario, wobei das Video angehalten und problematische Stellen gemeinsam diskutiert wurden, bis ein gemeinsames Verständnis der Leistungserwartungen erreicht war. Danach bewerteten die drei Rater dieselben drei Szenarien erneut unabhängig voneinander. Die interne Konsistenz der 18 Subthemen des AOTP verbesserte sich dabei von einem Cronbach’s Alpha von 0,829 vor

dem Training auf 0,914 nach dem Training. Für die sechs Themen des GAOTP stieg Cronbach's Alpha von 0,679 auf 0,868. Diese Befunde werden im Dokument dahingehend interpretiert, dass insbesondere nach Training eine gute bis sehr gute interne Konsistenz erreicht wurde und dass die zusammengefassten Bewertungsdimensionen offenbar ein kohärentes Konstrukt erfassen. Zugleich wird daraus geschlossen, dass das Training der Rater einen deutlichen positiven Einfluss auf die instrumentelle Anwendung hatte.

Auch die Interrater-Reliabilität wurde in diesem kleinen Trainingssample überprüft. Gemessen mittels Intraklassenkoeffizienten verbesserte sie sich von 0,54 vor dem Training auf 0,94 nach dem Training. Diese erhebliche Steigerung verdeutlicht, dass Ratertraining für die konsistente Anwendung beider Instrumente von zentraler Bedeutung ist. Die Autoren leiten daraus ab, dass eine intensive Schulung und gezieltes Feedback vor dem Einsatz der Instrumente in der Leistungsbewertung obstetrischer Teams notwendig erscheinen. Diese Interpretation wird auch durch die Rückmeldungen der Reviewer gestützt, die anregten, das Training durch Basisszenarien mit guter und schlechter Teamleistung zu ergänzen. Insbesondere content-naive Reviewer äußerten den Wunsch nach klareren Informationen über die Leistungserwartungen an obstetrische Teams, um ihre Aufmerksamkeit besser auf jene Verhaltensweisen richten zu können, die zu Verzögerungen oder Abweichungen von erwarteten Handlungsweisen führen.

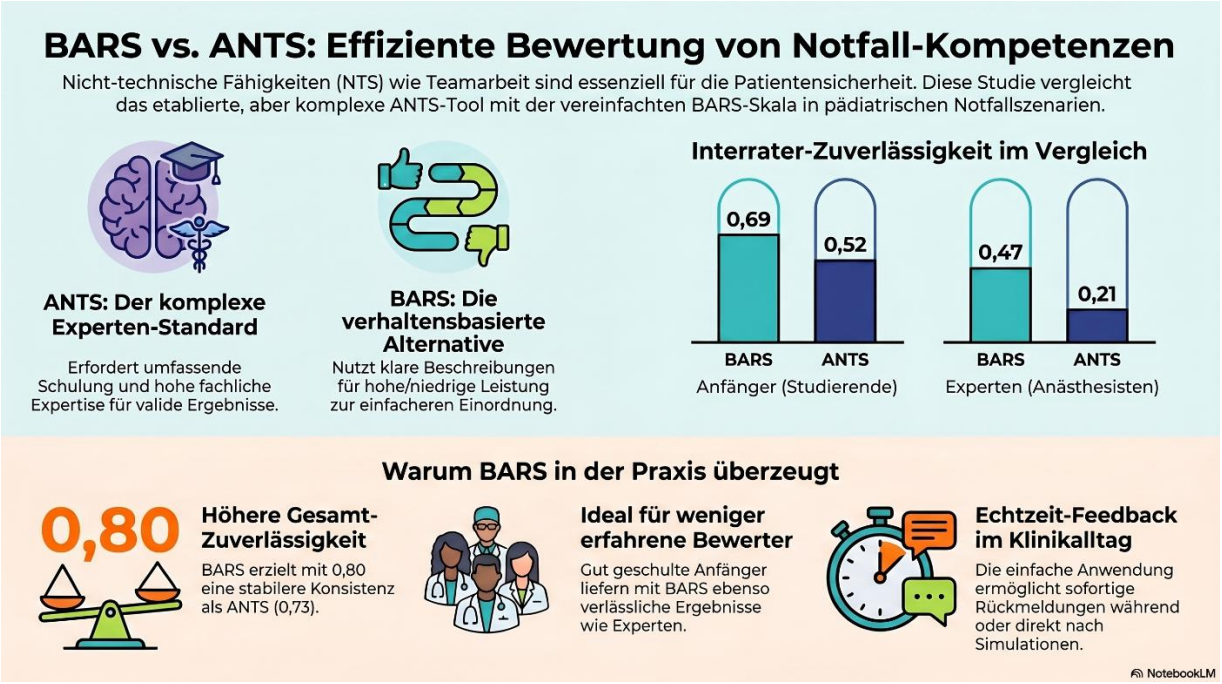
Trotz der insgesamt positiven Ergebnisse verweist das Dokument auf mehrere Limitationen. Zunächst ist zu betonen, dass Validität und Reliabilität der Instrumente nach Aussage der Autoren noch nicht abschließend bestimmt sind. Die vorliegenden Daten sind daher eher als erste Hinweise auf günstige psychometrische Eigenschaften zu verstehen, denn als endgültiger Nachweis der Messqualität. Hinzu kommt, dass die Analysen zur internen Konsistenz und Interrater-Reliabilität lediglich auf einer sehr kleinen Zahl von drei trainierten Ratern beruhen, sodass die Ergebnisse vorsichtig interpretiert werden müssen. Eine weitere, eher konzeptionelle Herausforderung betrifft die Abgrenzung zwischen Individual- und Teamleistung. In den Rückmeldungen der Reviewer wurde deutlich, dass Schwierigkeiten auftreten, wenn die Leistung eines einzelnen Teammitglieds problematisch ist, das Team als Ganzes diese Schwäche jedoch kompensiert. Das Dokument betont in diesem Zusammenhang, dass die Beurteiler nicht zwischen Individual- und Teamleistung abwägen sollen, sondern die Beiträge aller Teammitglieder integrieren und die Leistung des Teams als Ganzes einschätzen müssen. Gleichwohl wird hieran die grundsätzliche methodische Komplexität der Entwicklung von Teamleistungsinstrumenten sichtbar. Darüber hinaus zeigt die Studie, dass die Anwendung der Instrumente ohne ausreichendes Training mit Verständnisschwierigkeiten und geringerer Übereinstimmung einhergehen kann, was die Relevanz standardisierter Schulung unterstreicht.

Zusammenfassend lassen sich das AOTP und das GAOTP auf Grundlage des vorliegenden Dokuments als spezifisch für die Geburtshilfe entwickelte, beobachtungsbasierte behavioural marking systems charakterisieren, die auf einer breiten qualitativen Datengrundlage beruhen und die Leistung interdisziplinärer obstetrischer Teams in hochrealistischen Simulationen erfassen sollen. Das AOTP stellt dabei das differenziertere Instrument mit sechs Themen und 18 Subthemen dar und ist vor allem für formative Rückmeldung und die Beobachtung von Veränderungen der Teamleistung geeignet. Das GAOTP bildet dieselben sechs Themen in globalerer Form ab und ist eher auf summative Gesamteinschätzungen ausgerichtet. Beide Instrumente zeichnen sich durch eine starke Verhaltensorientierung, patientenzentrierte Dimensionen und eine hohe wahrgenommene Benutzerfreundlichkeit aus. Erste Befunde zur internen Konsistenz und Interrater-Reliabilität, insbesondere nach Training, fallen günstig aus, auch wenn die vollständige psychometrische Prüfung noch aussteht. Insgesamt liefern die Ergebnisse des Dokuments eine fundierte Grundlage für weiterführende Untersuchungen zur Validität, Reliabilität und zum Einsatz der Instrumente in der Evaluation obstetrischen Teamtrainings.

5.8 Anaesthetists' Nontechnical Skills Scale (ANTS) und Behaviorally Anchored Rating Scale Tool (BARS)

Quelle: Watkins SC, Roberts DA, Boulet JR, McEvoy MD, Weinger MB. Evaluation of a simpler tool to assess nontechnical skills during simulated critical events. Sim Healthcare. (2017) 12:69–75. doi: 10.1097/SIH.0000000000000199

Abbildung 11: Anaesthetists' Nontechnical Skills Scale (ANTS) und Behaviorally Anchored Rating Scale Tool (BARS)



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. (Watkins et al. (2017)

Im vorliegenden Dokument werden zwei Instrumente zur Erfassung nicht-technischer Fähigkeiten im anästhesiologischen Kontext behandelt und miteinander verglichen, nämlich die *Anaesthetists' Nontechnical Skills Scale* (ANTS) und ein vereinfachtes *Behaviorally Anchored Rating Scale* Tool (BARS). Ausgangspunkt der Untersuchung ist die Beobachtung, dass nontechnical skills wie Teamarbeit, Kommunikation, Entscheidungsfindung und Vigilanz eine wesentliche Rolle für die sichere Patientenversorgung spielen, in der medizinischen Ausbildung und Leistungsbeurteilung jedoch lange Zeit gegenüber technischen Fertigkeiten und Fachwissen in den Hintergrund getreten sind. Obwohl im Bereich der Anästhesiologie mit dem ANTS bereits ein etabliertes Instrument zur Erfassung dieser Kompetenzen existiert, wird im Dokument zugleich hervorgehoben, dass dessen Anwendung komplex ist und eine erhebliche fachliche Expertise sowie umfangreiche Schulung voraussetzt. Vor diesem Hintergrund verfolgten die Autoren das Ziel, die Reliabilität eines einfacheren Instruments, des BARS, zu untersuchen, erste Evidenz für dessen Validität bereitzustellen und die damit erhobenen Bewertungen mit den Ergebnissen des ANTS zu vergleichen.

Das ANTS wird im Dokument als das im Bereich der Anästhesiologie am intensivsten untersuchtes Instrument zur Erfassung nicht-technischer Fähigkeiten beschrieben. Es wurde entwickelt, um die NTS von Anästhesieanbietern systematisch zu bewerten, und erzielt nach Darstellung der Autoren dann reliabel nutzbare Ergebnisse, wenn es von fachlich erfahrenen Anästhesisten verwendet wird, die in seiner Anwendung angemessen geschult wurden. Gleichzeitig betont das Dokument, dass die Komplexität dieses Instruments seine Verbreitung in der Praxis einschränkt. Selbst erfahrene Kliniker oder Simulationsdozenten können Schwierigkeiten haben, ANTS konsistent anzuwenden. Daher eignet sich das Instrument nach Einschätzung der Autoren eher für forschungsbezogene oder summativere Kontexte als für formative Beurteilungssituationen, in denen zeitnahes Feedback gegeben werden soll. Das BARS wurde demgegenüber als vereinfachte Alternative konzipiert, die ähnliche Konstrukte wie das ANTS erfassen, jedoch mit geringerem Schulungsaufwand auskommen und im Idealfall auch von geschulten Nichtexperten angewendet werden kann. Die Entwicklung des BARS erfolgte durch ein Team aus Simulations-, Bildungs- und Domänenexperten im Rahmen einer großen multizentrischen Studie zur simulationsbasierten Leistungsbeurteilung. Nach einer umfassenden Literaturrecherche und der Prüfung vorhandener Bewertungsinstrumente wurde entschieden, ein neues, einfacheres Instrument zu entwickeln, da das ANTS sowohl hinsichtlich seiner Komplexität als auch im Hinblick auf spezifische Studienanforderungen als unzureichend praktikabel erschien. Das BARS wurde anschließend über einen Zeitraum von sechs Monaten iterativ weiterentwickelt und an acht Simulationszentren pilotiert.

Hinsichtlich ihrer Struktur weisen die beiden Instrumente sowohl Gemeinsamkeiten als auch deutliche Unterschiede auf. Das ANTS besteht aus vier Hauptkategorien, nämlich *Task Management*, *Team Working*, *Situation Awareness* und *Decision Making*. Diese Hauptkategorien sind in mehrere Unterelemente gegliedert. Im Bereich des *Task Management* gehören dazu etwa *Planning and Preparing*, *Prioritizing*, *Providing and Maintaining Standards* sowie *Identifying and Utilizing Resources*. Die Kategorie *Team Working* umfasst unter anderem die Koordination von Aktivitäten im Team, den Informationsaustausch, den Einsatz von Autorität und Durchsetzungsfähigkeit, das Einschätzen von Fähigkeiten und die Unterstützung anderer. *Situation Awareness* beinhaltet das Sammeln von Informationen, das Erkennen und Verstehen sowie das Antizipieren, während *Decision Making* die Identifikation von Optionen, das Abwägen von Risiken und das Re-Evaluieren umfasst. Charakteristisch für das ANTS ist damit ein hierarchischer und vergleichsweise feingliedriger Aufbau, bei dem einzelne Komponenten innerhalb der vier Hauptkategorien separat bewertet und anschließend zu Kategoriescores zusammengeführt werden. Ursprünglich verwendet das Instrument eine kategoriale Skala mit den Werten gut, akzeptabel, grenzwertig, schlecht und nicht beobachtet. Für die Zwecke der

vorliegenden Studie wurde das ANTS jedoch leicht modifiziert, indem Zwischenwerte von 0,5 eingeführt wurden, um eine feinere Abstufung und eine bessere Vergleichbarkeit mit dem BARS zu ermöglichen. Dadurch entstand für jede Kategorie eine achtstufige Bewertungsskala mit einem maximalen Gesamtscore von 32 Punkten.

Das BARS weist ebenfalls vier Hauptkategorien auf, die jedoch anders organisiert und insgesamt kompakter angelegt sind. Es umfasst die Bereiche *Vigilance/Awareness*, *Decision Making and Task Management*, *Communication* und *Teamwork*. Im Unterschied zum ANTS sind diese Bereiche nicht in zahlreiche Unterelemente untergliedert, sondern werden jeweils als Gesamtkategorien bewertet. Jede Kategorie ist mit zwei Sätzen verhaltensbezogener Beschreibungen versehen, die jeweils Beispiele für hohe beziehungsweise niedrige Leistung enthalten. Diese Behavior Descriptors dienen als Ankerpunkte für die Bewertung. Die Rater sollen das beobachtete Verhalten anhand dieser Anker einschätzen und der jeweiligen Kategorie einen Wert auf einer neunstufigen Skala zuweisen. Damit ergibt sich ein maximaler Gesamtscore von 36 Punkten. Zusätzlich enthält das BARS ein holistisches Globalrating, bei dem die Beurteiler nach Abschluss der vier Domänenbewertungen eine zusammenfassende Einschätzung der gesamten behavioral beziehungsweise nontechnical performance vornehmen. Das BARS ist somit deutlich kompakter und stärker verhaltensverankert als das ANTS. Es verzichtet auf die ausführliche Untergliederung einzelner Teilaspekte und fokussiert stattdessen auf globalere, an beobachtbaren Hoch- und Niedrigleistungsbeispielen orientierte Kategorien. Gerade diese strukturelle Vereinfachung bildet die Grundlage seiner intendierten leichteren Anwendbarkeit.

Die Untersuchung im Dokument fand im Kontext simulierter pädiatrischer kritischer Ereignisse statt. Bewertet wurden Anästhesie-Residents und Student Nurse Anesthetists, die in drei standardisierten Szenarien, nämlich Hypoxämie, supraventrikulärer Tachykardie sowie ventrikulärer Tachykardie beziehungsweise Kammerflimmern, die Rolle des primären Anästhesieproviders übernahmen. Die Szenarien hatten eine Dauer von acht bis zwölf Minuten. Sechs Rater, darunter vier Novizen und zwei Experten, bewerteten die Videoaufzeichnungen dieser Szenarien mit beiden Instrumenten. Die Novizen waren Medizinstudenten im dritten Jahr mit nur geringer Kenntnis der anästhesiologischen Praxis und wenig Simulationserfahrung. Die Experten waren board-zertifizierte pädiatrische Anästhesisten mit mehr als fünf Jahren klinischer Erfahrung und ausgewiesener Erfahrung als Simulationslehrer. Alle Rater erhielten vor Beginn der eigentlichen Bewertung ein Training in beiden Instrumenten. Die Schulung umfasste die Lektüre des ANTS User Manual und relevanter Literatur. Die Novizen erhielten darüber hinaus eine zusätzliche Einführung in Anästhesie-Crisis-Resource-Management, Human Factors und

nontechnical skills. Im Anschluss nahmen alle Rater an einem vierstündigen formalen Training teil, in dem beide Instrumente gemeinsam besprochen und anhand von Beispielvideos geübt wurden. Die Effektivität der Schulung wurde durch unabhängige Bewertungen zweier weiterer Trainingsvideos überprüft, wobei nach den a priori festgelegten Kriterien ausreichende Übereinstimmung zwischen den Ratern erzielt wurde.

Im Zentrum der Studie standen die psychometrischen Eigenschaften der mit beiden Instrumenten erzielten Bewertungen, insbesondere deren Reliabilität und die Beziehungen zwischen den aus beiden Verfahren gewonnenen Scores. Das Dokument ordnet diese Prüfung in ein modernes Validitätsverständnis ein und verweist auf das unified model of validity, nach der Evidenz aus verschiedenen Quellen erforderlich ist, um die Angemessenheit von Interpretationen und Inferenzschlüssen aus Assessmentdaten zu stützen. Für das BARS sollte insbesondere Evidenz zur internen Struktur der erhobenen Scores, zu ihren Beziehungen mit anderen Variablen sowie ergänzend zur Inhaltsvalidität gewonnen werden. Die Inhaltsvalidität des BARS wird im Dokument unter anderem dadurch gestützt, dass das Instrument von einer Gruppe erfahrener akademischer Anästhesisten mit umfassender Expertise in simulationsbasierter NTS-Bewertung entwickelt wurde. Eine numerische Bestimmung der Inhaltsvalidität im Sinne eines Content Validity Index wird im Dokument jedoch nicht berichtet. Im Vordergrund stehen vielmehr Reliabilitätsanalysen und die Korrelation mit dem etablierten ANTS.

Die Intrarater-Reliabilität wurde durch Korrelation der ersten und zweiten Bewertung desselben Raters bestimmt. Für das ANTS ergab sich ein Gesamtwert von 0,73. Die entsprechenden Werte unterschieden sich jedoch deutlich zwischen Novizen und Experten. Während die Novizen für den Overall Score des ANTS eine Intrarater-Reliabilität von 0,84 erreichten, lag der Wert bei den Experten lediglich bei 0,57. Auf Ebene der Kategorien war die Intrarater-Reliabilität im ANTS am höchsten für *Task Management* mit 0,74 und am niedrigsten für *Situational Awareness* mit 0,58. Das BARS zeigte insgesamt eine etwas höhere Intrarater-Reliabilität mit einem Overall Score von 0,79. Dabei lagen die Werte für Novizen mit 0,81 und für Experten mit 0,79 nahezu gleichauf. Auf Ebene der Einzelkategorien war die Intrarater-Reliabilität im BARS am höchsten für *Decision Making* mit 0,75 und am niedrigsten für *Vigilance* mit 0,63. Zusätzlich wurde für das holistische Globalrating ein Wert von 0,76 berichtet. Insgesamt legen diese Ergebnisse nahe, dass das BARS in Bezug auf die zeitliche Stabilität der Bewertungen mindestens ebenso günstig, tendenziell sogar günstiger abschneidet als das ANTS und dabei insbesondere weniger stark von der Ratergruppe abhängig ist.

Auch hinsichtlich der Interrater-Reliabilität zeigte das BARS günstigere Werte als das ANTS. Für das ANTS lag der Overall Score in der Interrater-Reliabilität bei 0,49, wobei sich erneut

deutliche Unterschiede zwischen Novizen und Experten zeigten. Die Novizen erreichten einen Wert von 0,57, die Experten lediglich 0,22. Die kategoriespezifischen Werte des ANTS fielen besonders niedrig für *Situational Awareness* aus, wo insgesamt nur ein Wert von 0,24 berichtet wurde. Für das BARS lag die Interrater-Reliabilität des Overall Score bei 0,62. Die Werte betrugen 0,67 für die Novizen und 0,52 für die Experten. Besonders hoch fiel die Übereinstimmung im Bereich *Communication* mit 0,70 aus; auch der holistische Score zeigte mit 0,66 eine vergleichsweise gute Interrater-Reliabilität. Diese Befunde wurden durch szenarienspezifische Analysen weiter differenziert. Dabei zeigte sich, dass die Reliabilität des ANTS je nach Szenario stark schwankte und bei Experten in einzelnen Szenarien sogar extrem niedrig ausfiel, während das BARS in allen drei Szenarien moderatere und stabilere Werte aufwies. Die Autoren interpretieren dies dahingehend, dass das BARS gegenüber dem ANTS weniger anfällig für raterbedingte Messfehler und möglicherweise auch weniger kontextspezifisch ist.

Ein besonders bemerkenswerter Befund der Studie besteht darin, dass die Novizen in beiden Instrumenten reliablere Bewertungen erzielten als die Experten. Das Dokument diskutiert hierfür mehrere Erklärungen. Zum einen wird darauf verwiesen, dass das ANTS in seiner Struktur und in einzelnen Verhaltensankern eng mit technischen und klinischen Aspekten verknüpft ist. So verweist das ANTS-Handbuch beispielsweise darauf, dass das Einhalten von Standards und Leitlinien ein Marker guter Praxis sei. Experten könnten daher stärker durch ihre klinischen Erfahrungen und ihr Wissen über angemessenes oder unangemessenes Patientenmanagement beeinflusst worden sein, obwohl sie angewiesen waren, ausschließlich die nontechnical skills zu bewerten. Die Novizen verfügten demgegenüber über deutlich weniger klinische Vorerfahrung und waren daher möglicherweise eher in der Lage, ihre Aufmerksamkeit konsequent auf beobachtbare nicht-technische Verhaltensweisen zu richten. Das Dokument schließt daraus, dass die Komplexität des ANTS und seine stärkere Nähe zu fachlich-technischen Aspekten eine größere Anforderung an die Bewerter darstellen als das BARS, dessen vereinfachte und stärker verhaltensverankerte Struktur offenbar konsistentere Urteile ermöglicht.

Zur Prüfung der Beziehungen zwischen beiden Instrumenten wurden Pearson-Korrelationen zwischen den Scores von ANTS und BARS berechnet. Der Overall Score des BARS korrelierte mit dem Overall Score des ANTS mit 0,74. Nach Darstellung des Dokuments bedeutet dies, dass 55 % der Varianz des ANTS-Gesamtscores durch den BARS-Gesamtscore erklärt werden können. Auch zwischen inhaltlich vergleichbaren Subdimensionen beider Verfahren ergaben sich substanziell positive Zusammenhänge, etwa zwischen der Vigilance/Awareness-Dimension des BARS und der Situational-Awareness-Kategorie des ANTS, zwischen den jeweiligen Decision-Making-Dimensionen sowie zwischen den Teamwork-bezogenen Bereichen.

Darüber hinaus wird im Dokument hervorgehoben, dass die Rangordnung der Performances mit beiden Instrumenten ähnlich ausfällt. Diese Befunde werden als Evidenz dafür interpretiert, dass das BARS ähnliche Konstrukte wie das ANTS misst. Damit liefert die Studie wichtige Hinweise darauf, dass die Vereinfachung des Instruments nicht mit einem vollständigen Verlust der inhaltlichen Nähe zum etablierten Referenzinstrument einhergeht.

Die Autoren leiten aus den Ergebnissen ab, dass das BARS insbesondere für formative Zwecke eine praktikable Alternative zum ANTS darstellen kann. Da das Instrument einfacher aufgebaut ist, weniger komplexe Urteilsbildung verlangt und mit geringerem Schulungsaufwand genutzt werden kann, erscheint es besonders geeignet für Assessments „for learning“, in denen rasches, gegenwartsnahes Feedback erforderlich ist. Das Dokument hebt hervor, dass BARS-Ratings potenziell nahezu in Echtzeit vorgenommen werden könnten und dadurch die Wahrscheinlichkeit steigt, dass nontechnical skills durch unmittelbare Rückmeldung verbessert werden. Darüber hinaus eröffnet die Möglichkeit, auch geschulte Nichtexperten als Rater einzusetzen, zusätzliche praktische Anwendungsperspektiven. Das ANTS behält demgegenüber seine Stärke in einer detaillierten und differenzierten Analyse einzelner Komponenten nicht-technischer Leistung, ist jedoch aufgrund seines Schulungsbedarfs und seiner Komplexität eher für spezialisierte, forschungsbezogene oder formalere Beurteilungskontexte geeignet. Das Dokument betont außerdem, dass für summative Leistungsbewertungen bei beiden Instrumenten wahrscheinlich mehr Beobachtungen beziehungsweise Ratings erforderlich wären, um ein ausreichend hohes Reliabilitätsniveau zu erzielen.

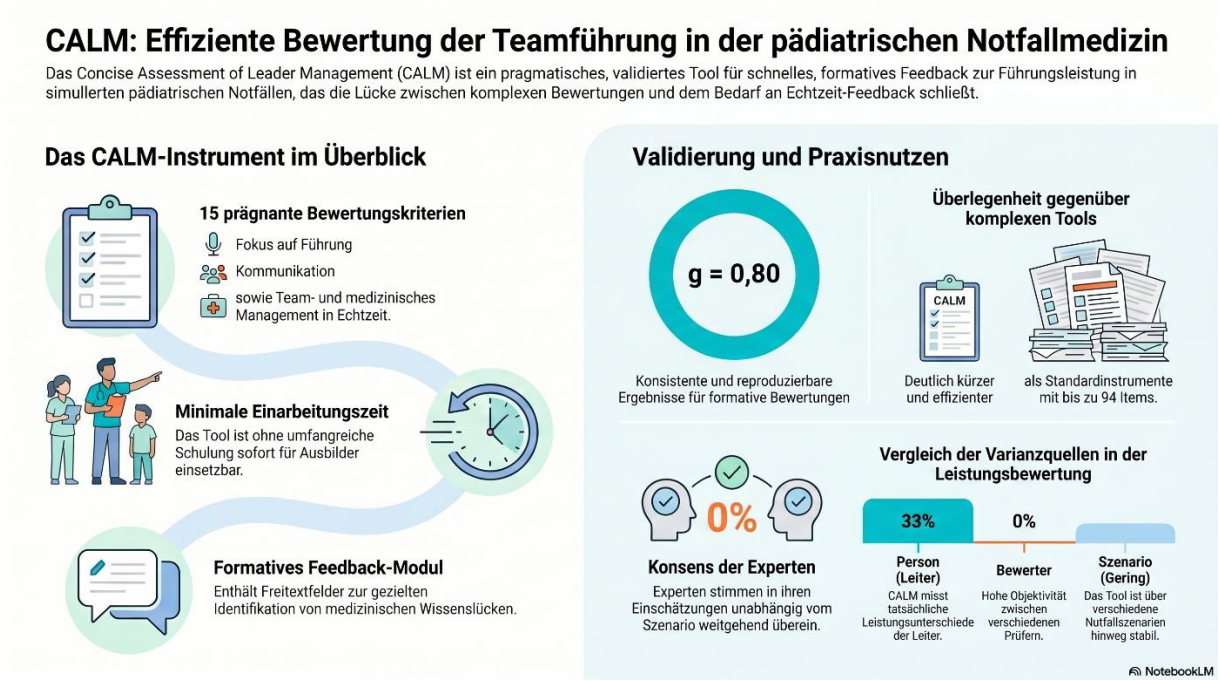
Gleichzeitig weist die Studie mehrere Limitationen auf. Die Zahl der Experten als Rater war mit zwei Personen sehr klein, sodass Verallgemeinerungen über Unterschiede zwischen Novizen und Experten nur eingeschränkt möglich sind. Eine größere Zahl fachkundiger Beurteilerr hätte möglicherweise andere Ergebnisse erbracht. Zudem wurden lediglich drei klinische Szenarien untersucht. Da bestimmte nontechnical skills je nach Situation leichter oder schwieriger zu beurteilen sein können, bleibt offen, inwieweit die Ergebnisse auf andere klinische Kontexte übertragbar sind. Ein weiterer methodischer Schwachpunkt besteht darin, dass das Zeitintervall zwischen Erst- und Zweitbewertung nicht kontrolliert wurde, sodass Unterschiede in der Intrarater-Reliabilität teilweise auch durch zeitliche Effekte beeinflusst worden sein könnten. Darüber hinaus wurde der sogenannte Task-Sampling-Error nicht bestimmt, also nicht untersucht, wie viele Szenarien notwendig wären, um ein hinreichend reliables Maß der tatsächlichen Fähigkeit zu erhalten. Schließlich weisen die Autoren selbst darauf hin, dass zwar die Novizenscores reliabel erschienen, ihre Validität aber nicht abschließend gesichert ist, da sie nicht mit einem objektiven Standardscore für jedes einzelne Szenario verglichen wurden.

Zusammenfassend lässt sich festhalten, dass das im Dokument untersuchte BARS als vereinfachtes, verhaltensverankertes Instrument zur Erfassung nicht-technischer Fähigkeiten im anästhesiologischen Simulationskontext konzipiert wurde und im Vergleich zum etablierten ANTS mehrere praktische und psychometrische Vorteile zeigt. Während das ANTS durch seinen differenzierten Aufbau eine präzisere Analyse einzelner NTS-Komponenten ermöglicht, ist es zugleich komplex, trainingsintensiv und in seiner Reliabilität stärker von der Ratergruppe und vom Szenario abhängig. Das BARS reduziert die Komplexität auf vier globalere Kategorien mit klaren Hoch- und Niedrigleistungsankern und erreicht dabei insgesamt günstigere Intra- und Interrater-Reliabilitäten. Gleichzeitig korrelieren die damit erzielten Scores deutlich mit den Ergebnissen des ANTS, was auf eine substanzielle inhaltliche Nähe beider Instrumente hinweist. Vor dem Hintergrund der im Dokument dargestellten Ergebnisse kann das BARS daher als praktikable Alternative zum ANTS für formative Beurteilungen nicht-technischer Fähigkeiten von Anästhesieanbietern in simulierten kritischen Ereignissen eingeordnet werden, auch wenn für weitergehende Generalisierungen und insbesondere für summative Verwendungszwecke zusätzliche Validierungsstudien erforderlich bleiben.

5.9 Concise Assessment of Leader Management (CALM)

Quelle: Nadkarni LD, Roskind CG, Auerbach MA, Calhoun AW, Adler MD, Kessler DO. The development and validation of a concise instrument for formative assessment of team leader performance during simulated pediatric resuscitations. Sim Healthcare. (2018) 13:77–82.

Abbildung 12: Concise Assessment of Leader Management (CALM)



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Nadkarni et al. (2018)

Das *Concise Assessment of Leader Management* (CALM) wurde als kompaktes Instrument zur formativen Beurteilung der Leistung von Teamleitern in simulierten pädiatrischen Reanimationssituationen entwickelt. Ausgangspunkt für die Entwicklung war die im Dokument beschriebene Problematik, dass pädiatrische Resuscitationen zwar selten, zugleich aber hochkritische Ereignisse sind und Lerner daher nur begrenzte Möglichkeiten haben, echte Führungserfahrung in diesen Situationen zu erwerben. Da effektive Führung in Reanimationen als zentral für Teamleistung und Patientenversorgung hervorgehoben wird und zugleich standardisierte Beurteilungsinstrumente für die Leistung individueller Teamleiter fehlen, bestand ein deutlicher Bedarf an einem pragmatischen, leicht einsetzbaren Werkzeug, das in simulationsbasierten Ausbildungssettings unmittelbares formatives Feedback unterstützen kann. Die Autoren betonen, dass viele existierende Instrumente entweder die Leistung des gesamten Teams und nicht spezifisch die des Teamleiters adressieren oder zwar individuelle Teamleiterleistung erfassen, aber so komplex oder trainingsintensiv sind, dass ihre praktische Nutzung im klinischen und edukativen Alltag eingeschränkt bleibt. Vor diesem Hintergrund wurde das CALM als „off the shelf"-Instrument konzipiert, das mit minimalem Schulungsaufwand unmittelbar für die formative Beurteilung eingesetzt werden kann.

Die Entwicklung des CALM erfolgte in einem systematischen und iterativen Prozess, der auf bestehender Literatur, beruflicher Erfahrung und Expertenkonsens beruhte. Drei Experten aus den Bereichen pädiatrische Notfallmedizin und graduate medical education trafen sich über einen Zeitraum von drei Monaten regelmäßig, um vorhandene Assessmentinstrumente und Publikationen mit Validitätsdaten zur Beurteilung von Führungs- und Teamleistung zu sichten. Aus diesen Quellen wurden relevante Fragen, Elemente und Themen zunächst in ihrer ursprünglichen Formulierung extrahiert. Anschließend wurden Duplikate zusammengeführt, Formulierungen im Hinblick auf Verständlichkeit und Einfachheit überarbeitet und die Inhalte mithilfe eines modifizierten Delphi-Prozesses priorisiert. Das Ergebnis dieses Prozesses war zunächst ein 18-Item-Instrument. Diese erste Version wurde in einem nächsten Schritt durch pädiatrische Notfallmediziner an einer Institution über einen Zeitraum von drei Monaten pilotiert, indem damit die Leistung von Resuscitation-Leitern während Mock-Resuscitations in der Notaufnahme bewertet wurde. Bemerkenswert ist in diesem Zusammenhang, dass die Pilotanwender keine spezifische Schulung in der Nutzung des Instruments erhielten. Dies geschah bewusst, da ein zentrales Entwicklungsziel des CALM darin bestand, ein benutzerfreundliches Instrument zu schaffen, das ohne aufwendiges Training auskommt. Das im Rahmen der Pilotierung erhobene Feedback wurde in die Revision einbezogen und führte schließlich zur finalen Version des CALM.

Die Struktur des finalen Instruments ist im Dokument anhand der abgebildeten Bewertungsform dargestellt. Das CALM umfasst 15 vierstufige Likert-Items sowie ein dichotomes Verhaltensitem und ist in vier übergeordnete Domänen gegliedert, die nach Auffassung der Autoren die vier wesentlichen Elemente der Führung in akuten Reanimationssituationen repräsentieren. Diese Domänen sind Leadership, Communication, Team Management und Medical Management. Innerhalb der Domäne Leadership werden sowohl die Übernahme der Führungsrolle als auch deren Klarheit und Angemessenheit bewertet. So wird unter anderem erfasst, ob die führende Person ihre Rolle als Leader explizit ankündigt, ob die Führungsrolle über den gesamten Fall hinweg erkennbar bleibt und ob der gewählte Führungsstil für die Situation angemessen und effektiv ist. Die Domäne Communication umfasst Aspekte wie die Lautstärke und Klarheit der Stimme, die direkte Ansprache von Teammitgliedern sowie die Unterstützung geschlossener Kommunikationskreisläufe im Sinne von Closed-Loop-Communication. Im Bereich Team Management werden die Zuweisung beziehungsweise Bestätigung von Rollen, die effektive Steuerung des Teams, die Verteilung der Arbeitslast, die Einbeziehung von Teammitgliedern in die Entscheidungsfindung und die regelmäßige Zusammenfassung des Falls beurteilt. Die vierte Domäne, Medical Management, fokussiert auf die medizinisch-organisatorische Steuerung des Szenarios. Dazu zählen die Priorisierung der Aufgabenreihenfolge, die Auf-

rechterhaltung des Gesamtüberblicks unter Vermeidung von Fixierungsfehlern, die wiederholte Re-Evaluation der Patientensituation, das Benennen der nächsten Versorgungsschritte sowie die Fähigkeit, eigene Grenzen zu erkennen und bei Bedarf Hilfe einzuholen.

Neben diesen in den CALM-Score eingehenden Kernelementen enthält das Instrument zwei zusätzliche Sektionen, die explizit der formativen Rückmeldung dienen, jedoch nicht in die Summenbewertung einfließen. Zum einen gibt es einen Bereich *Medical Knowledge*, in dem ein Aktionsplan zur Bearbeitung identifizierter Wissenslücken aus dem jeweiligen Szenario formuliert werden soll. Zum anderen enthält das Instrument eine globale Einschätzung der Leistung im Vergleich zu Peers, bei der die führende Person als unter dem erwarteten Niveau, auf dem erwarteten Niveau, über den Erwartungen oder in den besten fünf Prozent eingeordnet werden kann. Ferner sind in jeder Domäne Felder für spezifische Beispiele und Kommentare vorgesehen. Damit wird deutlich, dass das CALM nicht allein als numerisches Beurteilungsinstrument gedacht ist, sondern zugleich eine strukturierende Funktion für qualitativ gehaltvolles Feedback erfüllen soll.

Das Antwortformat des Instruments ist überwiegend vierstufig und verwendet die Kategorien „rarely“, „sometimes“, „mostly“ und „always“. Ein Item ist dichotom mit „yes“ beziehungsweise „no“ angelegt. Laut Dokument beträgt der maximal erreichbare CALM-Score 74 Punkte. Die beiden zusätzlichen Sektionen zu Medical Knowledge und Global Assessment werden bei dieser Summenbildung nicht berücksichtigt. Dadurch bleibt die eigentliche Skala auf die vier zentralen Führungsdomänen fokussiert. Die Entscheidung für ein kompaktes Format mit wenigen, klar formulierten Antwortoptionen steht in enger Beziehung zum Grundgedanken des Instruments, nämlich eine unkomplizierte, schnelle und realitätsnahe Anwendung zu ermöglichen.

Der Anwendungsbereich des CALM liegt spezifisch in der Beurteilung von Teamleiterleistung in simulierten pädiatrischen Reanimationssituationen. Im Rahmen der Validierungsstudie wurden vier unterschiedliche Resuscitation-Leader beurteilt, die jeweils vier verschiedene pädiatrische Reanimationsszenarien bearbeiteten. Die insgesamt 16 Videoaufzeichnungen stammten aus dem Archiv der *Improving Pediatric Acute Care Through Simulation* Gruppe und zeigten interprofessionelle Teams, die einen kindlichen Herzstillstand nach Ertrinkungsereignis, einen Säugling mit Atemstillstand durch Fremdkörper, einen Säugling mit hypoglykämiebedingtem Krampfanfall sowie einen Säugling mit Sepsis infolge Bakteriämie versorgten. Die Teams bestanden aus klinisch tätigen Fachkräften, darunter ein oder zwei ärztliche Teammitglieder mit Facharztanerkennung in pädiatrischer Notfallmedizin oder Notfallmedizin, mehrere Pflegekräfte sowie weitere Mitarbeiter wie certified nursing assistants oder emergency medicine technicians. Die Szenarien unterschieden sich hinsichtlich ihres Teamarbeitsbedarfs und

ihrer Komplexität, was die Generalisierbarkeit des Instruments über unterschiedliche Arten pädiatrischer Resuscitationen hinweg unterstützen sollte.

Die eigentliche Bewertung der Videos erfolgte durch vier unabhängige Rater, die aus dem Netzwerk INSPIRE rekrutiert wurden und als PEM-Fellowship-Directors an unterschiedlichen akademischen Institutionen tätig waren. Die Reihenfolge der Videos wurde für jede Beurteiler Person randomisiert. Die Rater sahen jedes Video genau einmal und durften das Material während des Abspielens weder pausieren noch zurückspulen. Auffällig ist, dass sie ausdrücklich keine weiteren spezifischen Anweisungen zur Anwendung des Instruments erhielten, sondern das CALM „nach bestem Vermögen“ nutzen sollten. Damit wurde die Validierung in einem Anwendungskontext durchgeführt, der dem intendierten Einsatz des Instruments möglichst nahekommt, nämlich einer unmittelbaren, ohne aufwendige Vorbereitung durchführbaren Beurteilung zur anschließenden formativen Rückmeldung.

Hinsichtlich der psychometrischen Eigenschaften berichtet das Dokument Evidenz zur Inhaltsvalidität und zur internen Strukturvalidität beziehungsweise Reliabilität. Die Inhaltsvalidität wird vor allem durch den Entwicklungsprozess des Instruments gestützt. Das CALM wurde von Experten in pädiatrischer Notfallmedizin und medizinischer Weiterbildung entwickelt, basierte auf bereits existierenden Resuscitation-Leader-Assessmentinstrumenten, wurde über einen modifizierten Delphi-Prozess strukturiert und anschließend pilotgetestet. Die Autoren werten diesen mehrstufigen, theorie- und konsensbasierten Prozess als inhaltliche Absicherung dafür, dass das Instrument tatsächlich die intendierten Konstrukte, nämlich zentrale Dimensionen der Führung in akuten pädiatrischen Resuscitationen, erfasst. Eine numerische Berechnung der Inhaltsvalidität wird im Dokument allerdings nicht berichtet; vielmehr wird diese Validitätsevidenz aus der systematischen und expertengeleiteten Entwicklung abgeleitet.

Die interne Strukturvalidität des CALM wurde mithilfe einer Generalisierbarkeitsanalyse untersucht. Dieses Vorgehen ermöglicht eine differenzierte Betrachtung verschiedener Varianzquellen der Messwerte und geht damit über einfache Reliabilitätskoeffizienten hinaus. Das Studiendesign war vollständig gekreuzt, sodass alle vier Rater alle vier Szenarien für jede der vier Führungspersonen bewerteten. Aus der Generalisierbarkeitsanalyse ergab sich, dass der größte Anteil der Varianz auf die Person, also die jeweilige Führungsperson, entfiel. Dieser Anteil lag bei 32,9 %. Das bedeutet, dass die Unterschiede in den CALM-Scores in erster Linie tatsächliche Unterschiede zwischen den beurteilten Teamleitern widerspiegelten. Genau dies ist aus Sicht eines Leistungsbewertungsinstruments wünschenswert, weil damit die gemessenen Unterschiede primär auf die Zielperson und nicht auf methodische Störfaktoren zurückgehen. Die Interaktion zwischen Szenario und Rater trug 16,4 % zur Gesamtvarianz bei, die

Interaktion zwischen Person und Rater 14,0 %. Der Szenarioeffekt allein war mit 1,6 % gering, was darauf hindeutet, dass das Instrument nicht primär von der Art des Szenarios gesteuert wurde. Besonders hervorzuheben ist der Befund, dass der Raterfaktor selbst einen Varianzanteil von 0 % aufwies. Das Dokument interpretiert dies als Ausdruck einer sehr hohen Übereinstimmung zwischen den Ratern und damit als starkes Argument für eine hohe Interrater-Reliabilität.

Auf Basis der Generalisierbarkeitsanalyse wurde für das Studiendesign mit vier Ratern und vier Szenarien ein absoluter Generalisierbarkeitskoeffizient von 0,80 berechnet. Im Dokument wird hervorgehoben, dass dieser Wert über dem als akzeptabel beschriebenen Bereich von 0,70 bis 0,79 für formative Assessments liegt und damit im Einklang mit der einschlägigen Literatur zu Performance Assessments steht. Die Autoren werten dies als wichtigen Hinweis darauf, dass das CALM für seine intendierte Verwendung, nämlich die formative Beurteilung und Rückmeldung, ausreichend zuverlässig ist. Ergänzend wurde eine D-Study durchgeführt, die den theoretischen Einfluss einer Veränderung der Anzahl der Rater oder Szenarien auf den Generalisierbarkeitskoeffizienten simuliert. Die im Dokument abgebildete Kurve zeigt, dass mit steigender Anzahl von Szenarien und Ratern die Zuverlässigkeit erwartungsgemäß zunimmt. Dies unterstreicht, dass die Güte des Instruments auch von den Rahmenbedingungen seiner Anwendung beeinflusst wird.

Trotz der insgesamt positiven Ergebnisse benennt das Dokument mehrere Limitationen. Die bedeutsamste Einschränkung besteht in der Nutzung von Videomaterial statt direkter Beobachtung vor Ort. Auch wenn dies aus Gründen der Durchführbarkeit notwendig war und dem Echtzeitkontext möglichst nahekommen sollte, könnten bestimmte Verhaltensweisen durch Kamerawinkel, Tonqualität oder Beginn und Ende der Aufnahme nicht vollständig erfasst worden sein. So weisen die Autoren darauf hin, dass eine Führungsperson ihre Rolle möglicherweise bereits vor Beginn der Videoaufzeichnung erklärt hatte, was im Rating dann nicht mehr sichtbar gewesen wäre. Dies könnte zu hohen Fehlerkomponente in der Generalisierbarkeitsanalyse beigetragen haben. Eine weitere Limitation sehen die Autoren in der Kürze und Offenheit mancher Formulierungen. Gerade die intendierte Knappheit des Instruments könnte dazu geführt haben, dass die Bedeutung einzelner Antwortoptionen unterschiedlich interpretiert wurde. Als Beispiel wird das Item genannt, das die Einbeziehung von Teammitgliedern in die Entscheidungsfindung erfasst. Wenn eine Führungsperson das Team einmal aktiv in eine Entscheidung einbindet, bleibt unklar, ob dies bereits als „always“, eher als „mostly“ oder nur als „sometimes“ bewertet werden sollte. Das Dokument schlägt daher vor, künftige Versionen

durch kurze Erläuterungen der Antwortkategorien zu ergänzen, um eine einheitlichere Interpretation zu fördern.

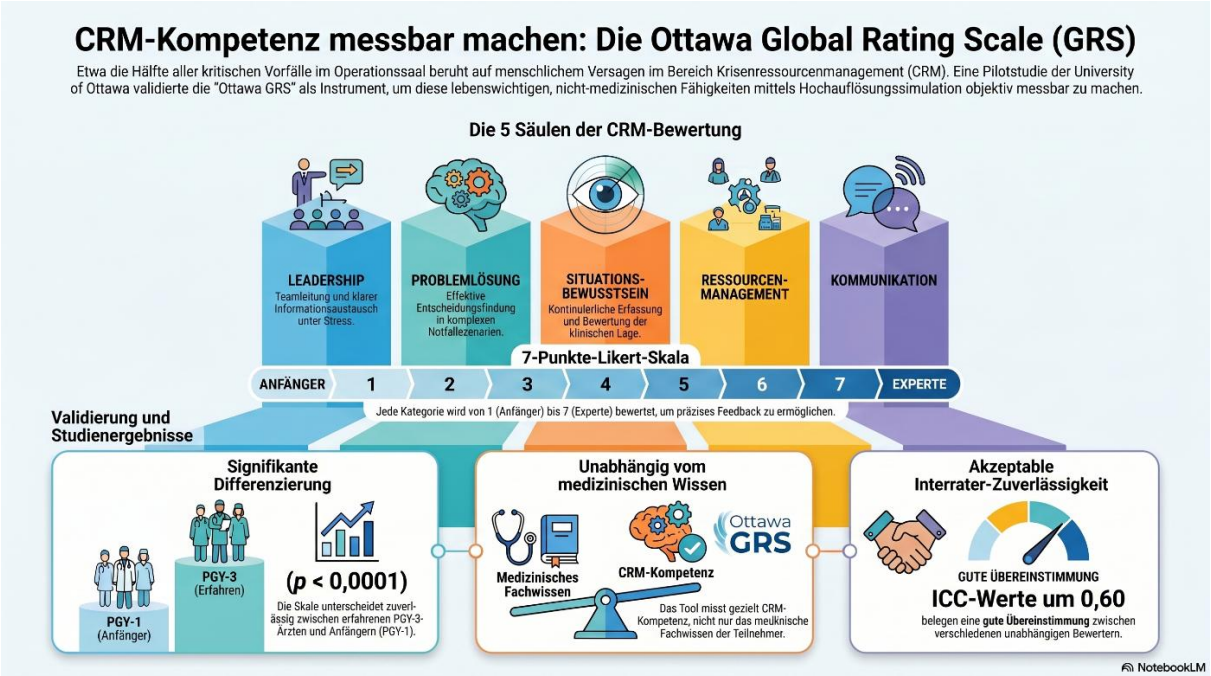
Eine weitere Einschränkung liegt darin, dass alle Rater als PEM-Fellowship-Directors zugleich Führungsexperten waren. Dies könnte die Generalisierbarkeit der Ergebnisse einschränken, da unklar bleibt, ob Personen ohne diese besondere Expertise das Instrument in vergleichbarer Weise anwenden würden. Hinzu kommt die geringe Stichprobengröße. Zwar war das Design vollständig gekreuzt, dennoch basieren die Befunde nur auf 16 Videos von vier Führungspersonen in vier Szenarien. Die Autoren weisen ausdrücklich darauf hin, dass größere Stichproben zu anderen Generalisierbarkeits- und Phi-Koeffizienten führen könnten. Schließlich wurde keine Rückmeldung der Lerner zur Nützlichkeit der mit dem CALM generierten formativen Daten erhoben. Da das Instrument primär dazu dienen soll, die Qualität unmittelbarer Rückmeldung zu unterstützen, stellt dies eine relevante Lücke dar, die in zukünftigen Studien adressiert werden sollte. Darüber hinaus wird betont, dass für höherstufige oder gar high-stakes Anwendungen zusätzliche Untersuchungen erforderlich wären, insbesondere im Hinblick auf die Beziehung zwischen CALM-Scores und langfristiger klinischer Leistung.

Zusammenfassend lässt sich das CALM auf Basis des Dokuments als ein knappes, pragmatisches und auf die Leistung des Resuscitation-Leaders fokussiertes Bewertungsinstrument charakterisieren, das speziell für den Einsatz in simulierten pädiatrischen Reanimationssituationen entwickelt wurde. Seine Entwicklung erfolgte strukturiert und expertenbasiert unter Einbezug vorhandener Instrumente, professioneller Erfahrung, eines modifizierten Delphi-Prozesses und einer Pilotierung im intendierten Anwendungskontext. Inhaltlich erfasst das Instrument vier zentrale Führungsdomänen, nämlich Leadership, Communication, Team Management und Medical Management, und ergänzt diese um qualitative Feedbackbereiche, die seine formative Funktion unterstreichen. Die berichteten psychometrischen Befunde liefern erste überzeugende Evidenz für die Inhalts- und interne Strukturvalidität des Instruments. Insbesondere der Generalisierbarkeitskoeffizient von 0,80 sowie der nahezu nicht vorhandene Rateranteil an der Gesamtvarianz sprechen für eine gute Eignung des CALM im formativen Einsatz. Trotz der genannten Limitationen deutet das Dokument somit darauf hin, dass das CALM ein reliables und praktikables Instrument für die strukturierte formative Rückmeldung an Teamleiter in simulierten pädiatrischen Resuscitationen darstellt.

5.10 Ottawa Crisis Resource Management Global Rating Scale (Ottawa GRS)

Quelle: Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: the university of ottawa critical care medicine, high-fidelity simulation, and crisis resource management. Crit Care Med. (2006) 34:2167–74. doi: 10.1097/01.CCM.0000229877.45125.CC

Abbildung 13: Ottawa Crisis Resource Management Global Rating Scale (Ottawa GRS)



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Kim et al. (2006)

Die *Ottawa Crisis Resource Management Global Rating Scale* (Ottawa GRS) wurde entwickelt, um die nicht-technischen Fähigkeiten von Ärzten in akuten Resuscitationssituationen strukturiert zu erfassen. Im Mittelpunkt stehen dabei jene Kompetenzen, die im Dokument unter dem Begriff *Crisis Resource Management* (CRM) zusammengefasst werden. Ausgehend von der Annahme, dass die Versorgung kritisch kranker Patienten nicht nur medizinisches Wissen und technische Fertigkeiten, sondern in hohem Maße auch Leadership, Problemlösefähigkeit, Situational Awareness, Kommunikationsfähigkeit und Ressourcenmanagement erfordert, verfolgte die Studie zwei zentrale Zielsetzungen. Zum einen sollte geprüft werden, ob *high-fidelity simulation* ein geeignetes Medium zur Bewertung von CRM-Leistung darstellt. Zum anderen sollte, da für die Messung von CRM zu diesem Zeitpunkt kein Goldstandard

verfügbar war, die Konstruktvalidität der Ottawa GRS untersucht werden. Das Instrument ist damit nicht nur als reines Bewertungswerkzeug, sondern auch als Teil eines größeren Versuchs zu verstehen, CRM-Leistung in simulierten medizinischen Krisensituationen überhaupt formalisierbar und systematisch beurteilbar zu machen.

Die Entwicklung der Ottawa GRS basierte auf den von Gaba beschriebenen Kernelementen effektiven Crisis Resource Managements. Diese theoretische Grundlage wurde im Instrument in fünf CRM-spezifische Kategorien übersetzt, die die Struktur der Skala bilden. Hinzu kommt eine zusätzliche globale Gesamtkategorie zur Beurteilung der allgemeinen CRM-Leistung. Die fünf inhaltlichen Teilbereiche sind *Leadership Skills*, *Problem Solving Skills*, *Situational Awareness Skills*, *Resource Utilization Skills* und *Communication Skills*. Jeder dieser Bereiche wird anhand einer siebenstufigen Likert-Skala bewertet, deren Skalenpunkte durch deskriptive Anker unterstützt werden. Im Dokument wird erläutert, dass ein Wert von 1 die Leistung eines vollständigen Novizen repräsentiert, ein Wert von 3 die Leistung einer Person mit begrenzter CRM- und Resuscitation-Erfahrung, ein Wert von 5 die Leistung eines Arztes, die beziehungsweise die kritischen Ereignisse kompetent bewältigen kann, und ein Wert von 7 eine expertengleiche Leistung. Damit ist die Skala nicht nur ordinal, sondern zugleich normativ auf eine Kompetenzentwicklung bezogen. Die beschreibenden Anker wurden so gestaltet, dass sie zusätzlich den Grad der notwendigen Hilfestellung beziehungsweise des Cueings berücksichtigen, der erforderlich war, damit die bewertete Person in der Simulation angemessen handelt. Dies stellt ein besonderes Merkmal des Instruments dar, da es die Bewertung an die Bedingungen des simulatorgestützten Kontextes anpasst und nicht lediglich beobachtbares Verhalten ohne Bezug auf erforderliche Unterstützung erfasst.

Die inhaltliche Ausgestaltung der fünf Hauptkategorien wird im Appendix des Dokuments detailliert beschrieben. Die Kategorie *Leadership Skills* umfasst Aspekte wie das Bewahren von Ruhe und Kontrolle in der Krise, zügige und entschlossene Entscheidungsfindung sowie das Aufrechterhalten einer globalen Perspektive im Sinne eines Überblicks über das Gesamtgeschehen. Die Kategorie *Problem Solving Skills* bezieht sich auf eine organisierte und effiziente Problemlösestrategie nach den grundlegenden Prinzipien der ABC-Struktur, die rasche Umsetzung von Maßnahmen im Sinne eines gleichzeitigen Managements mehrerer Anforderungen sowie die Berücksichtigung alternativer Handlungsoptionen. Die Kategorie *Situational Awareness Skills* erfasst die Fähigkeit, Fixierungsfehler zu vermeiden, die Situation fortlaufend neu zu bewerten und wahrscheinliche Ereignisse zu antizipieren. Unter *Resource Utilization Skills* werden das situationsgerechte Einholen von Hilfe, der angemessene Einsatz verfügbarer Ressourcen und eine adäquate Priorisierung von Aufgaben zusammengefasst. Die Kate-

gorie *Communication Skills* fokussiert schließlich auf klare und prägnante Kommunikation, den zielgerichteten Einsatz verbaler und nonverbaler Kommunikation sowie die Fähigkeit, auf Teaminput zu hören. Ergänzend zu diesen fünf Kategorien ermöglicht die globale Skala eine zusammenfassende Einschätzung der gesamten CRM-Leistung. Durch diese Struktur bildet das Ottawa GRS zentrale, inhaltlich klar voneinander abgegrenzte Domänen nicht-technischer Leistung in medizinischen Krisensituationen ab und verbindet diese mit einer globalen Gesamtbeurteilung.

Die Entwicklung der Skala und ihrer deskriptiven Anker erfolgte in mehreren Schritten. Zunächst wurden Ärzte der University of Ottawa mit Erfahrung in Resuscitation und Critical Care Medicine in die Ausarbeitung des Instruments einbezogen. Anschließend wurde das Bewertungssystem gemeinsam mit Simulationsinstruktoren und in Resuscitation oder Intensivmedizin erfahrenen Ärzten aus ganz Kanada überarbeitet. Ein weiterer Entwicklungsschritt erfolgte im Rahmen eines Delphi-Prozesses des Ratertrainings, in dem die Skala und ihre deskriptiven Anker nochmals modifiziert wurden. Diese mehrstufige, expertenbasierte Entwicklung wird im Dokument als wesentliche Stütze der Inhaltsvalidität interpretiert. Zusätzlich wurden auch die beiden in der Studie eingesetzten Simulationsszenarien einem Peer-Review hinsichtlich Realismus und inhaltlicher Plausibilität unterzogen, sodass nicht nur das Instrument selbst, sondern auch der Kontext seiner Anwendung inhaltlich abgesichert werden sollte.

Die Anwendung der Ottawa GRS erfolgte im Rahmen einer simulatorbasierten Pilotstudie mit Residents unterschiedlicher Fachrichtungen an der University of Ottawa. Teilgenommen haben 32 Residents im ersten Weiterbildungsjahr und 28 Residents im dritten Weiterbildungsjahr. Ausgeschlossen wurden Personen mit vorheriger Simulationserfahrung innerhalb der Residency sowie alle, die bereits formales CRM-Training erhalten hatten. Die Studie fand in einer hochrealistischen Simulationsumgebung statt, die Elemente eines Operationssaals sowie einer Intensivstation nachbildete und ein computergesteuertes Simulationsmannequin einsetzte. Zusätzlich waren ein standardisiert agierender ICU-Registered Nurse und ein Respiratory Therapist anwesend, deren Reaktionen auf die Handlungen der Teilnehmer vorab trainiert und skriptbasiert standardisiert wurden. Alle Residents nahmen zunächst an einer Tutorial-Sitzung teil, in der sie an die Simulationsumgebung, den Simulationsraum und das Mannequin herangeführt wurden. Dort wurden Inhalte zur Behandlung akuter respiratorischer Insuffizienz, grundlegendes Atemwegsmanagement und Schockmanagement wiederholt, ohne jedoch CRM-spezifische Instruktionen zu erhalten. Auf diese Weise sollte gewährleistet werden, dass Unterschiede in der späteren Leistung nicht bloß auf fehlende Vertrautheit mit der Umgebung oder grobe Wissensdefizite zurückzuführen waren.

Die Residents absolvierten zwei standardisierte Simulationsszenarien in derselben Reihenfolge. Das erste Szenario betraf kardiale Ereignisse bei einem postoperativen Patienten, das zweite einen Patienten mit akutem Schock und respiratorischem Versagen nach schwerem Trauma infolge eines Sturzes. Beide Szenarien dauerten jeweils 20 Minuten und waren so programmiert, dass die Teilnehmer fortlaufend neue Probleme erkennen, re-evaluieren und adäquat beantworten mussten. Alle Leistungen wurden videografiert, anonymisiert und in ein einheitliches Format überführt. Drei Rater, die jeweils als Akutmediziner und CRM-Instruktoren qualifiziert waren, beurteilten die Videos im Anschluss mithilfe der Ottawa GRS. Der Simulationsinstruktor, der bei allen Sitzungen anwesend war, wurde bewusst von der Bewertungsrolle ausgeschlossen, um die Integrität der Verblindung zu sichern. Besonders hervorzuheben ist, dass vor der eigentlichen Ratingphase ein umfassender Delphi-basierter Standardisierungsprozess mit den Ratern durchgeführt wurde. Hierbei wurden zunächst Beispielvideos eines unterdurchschnittlichen, durchschnittlichen und nahezu expertischen Leistungsniveaus gemeinsam bewertet, Unterschiede in der Bewertung diskutiert und die Skala sowie ihre Anker nochmals angepasst. Erst nachdem Konsens über die Interpretation der Skala erzielt worden war, erfolgte die eigentliche Bewertung aller Resident-Sitzungen. Dieses Vorgehen stellt einen zentralen Bestandteil des im Dokument beschriebenen *response process* dar und sollte Fehler in der Testadministration und der Bewertung minimieren.

Die Validierung der Ottawa GRS folgt dem im Dokument herangezogenen Konstruktvaliditätsverständnis nach Downing und den Standards for Educational and Psychological Testing. Untersucht wurden die Bereiche *content*, *response process*, *internal structure* und *relationship to other variables*, während *consequential validity* ausdrücklich nicht Gegenstand der Studie war. Die Inhaltsvalidität wird insbesondere durch die inhaltliche Orientierung an Gabas CRM-Domänen, die mehrstufige Expertenkonsultation sowie die Überarbeitung durch Simulations- und CRM-Instruktoren gestützt. Auch die Simulationsfälle selbst wurden durch Experten begutachtet, was aus Sicht der Autoren zusätzlich zur Repräsentativität der Messdomäne beiträgt. Der *response process* wurde durch eine Vielzahl an Standardisierungsmaßnahmen unterstützt. Dazu gehörten die Voraborientierung der Residents, die Standardisierung der Simulationsumgebung, das Training der unterstützenden Schauspiel- beziehungsweise Supportrollen, die identische Fallabfolge für alle Teilnehmer, die uniforme Aufbereitung der Videodaten sowie der ausführliche Ratertrainings- und Konsensprozess. Im Dokument wird daraus abgeleitet, dass die Datenerhebung und Bewertung weitgehend unter Bedingungen erfolgten, die eine Minimierung von Fehlern bei der Testadministration ermöglichen.

Die interne Struktur der Ottawa GRS wurde im Hinblick auf ihre diskriminative Fähigkeit und ihre Reliabilität untersucht. Ein zentrales Ergebnis der Studie besteht darin, dass die Skala deutlich zwischen Residents des ersten und dritten Weiterbildungsjahres differenzieren konnte. Der mittlere Overall-CRM-Score lag bei den PGY-1-Residents bei 4,13 Punkten und bei den PGY-3-Residents bei 5,54 Punkten, wobei dieser Unterschied hochsignifikant war. Derselbe Effekt zeigte sich in beiden Szenarien getrennt betrachtet. Auch auf Ebene der fünf Einzelkategorien unterschieden sich die PGY-1- und PGY-3-Gruppen in beiden Sitzungen jeweils signifikant voneinander. So erzielten die PGY-3-Residents in Leadership, Problem Solving, Situational Awareness, Resource Utilization und Communication durchweg höhere Werte. Diese durchgängig vorhandene Differenzierungsfähigkeit wird im Dokument als wichtiger Hinweis auf die diskriminative Eignung des Instruments interpretiert. Zugleich wird hervorgehoben, dass der Unterschied nicht allein durch Fachwissen erklärt werden könne. Da nur ein kleiner Teil der PGY-3-Residents aus chirurgischen oder notfallmedizinischen Programmen stammte und Anästhesiologie-PGY-3 mit früherer Simulationserfahrung ausgeschlossen wurden, argumentieren die Autoren, dass das Instrument nicht bloß medizinisches Spezialwissen, sondern tatsächlich CRM-bezogene Leistungsunterschiede erfasst.

Hinsichtlich der Reliabilität wurde die Interrater-Reliabilität mithilfe eines Typ-III-ICC bestimmt. Für die globale CRM-Gesamtbewertung ergaben sich Werte von 0,590 für das erste und 0,613 für das zweite Szenario. Diese Werte werden im Dokument als akzeptabel interpretiert, wenngleich ausdrücklich betont wird, dass sie für summative oder high-stakes Anwendungen nicht ausreichen. Auf Ebene der Einzelkategorien zeigten Leadership, Problem Solving und Situational Awareness ähnliche, mäßige ICC-Werte im Bereich zwischen etwa 0,475 und 0,626. Deutlich schwächer fielen die Werte für Resource Utilization und Communication aus, die nur Werte zwischen 0,236 und 0,384 erreichten. Dies deutet darauf hin, dass insbesondere diese beiden Kategorien im damaligen Entwicklungsstand des Instruments hinsichtlich ihrer Bewertungsgenauigkeit und ihrer Ankerformulierungen überarbeitungsbedürftig waren. Ein weiterer wichtiger Befund betrifft die Analyse einzelner Raterurteile. Auch wenn alle drei Rater die PGY-3-Gruppe im Vergleich zur PGY-1-Gruppe konsistent als leistungsstärker einstuften, zeigte sich ein sogenannter Dove-Hawk-Effekt, indem ein Rater systematisch höhere Werte vergab als die beiden anderen. Das Dokument interpretiert dies als Hinweis darauf, dass ein Teil der verbleibenden Varianz vermutlich nicht nur auf das Instrument selbst, sondern auch auf die Qualität und Kalibrierung des Ratertrainings zurückzuführen ist.

Die Studie untersuchte darüber hinaus die Stabilität der Leistungen über die beiden Szenarien hinweg. Es zeigte sich kein signifikanter Unterschied zwischen den Gesamtwerten in Szenario

1 und Szenario 2, weder insgesamt noch getrennt nach Weiterbildungsstufe. Die Autoren interpretieren dies nicht als Schwäche des Instruments, sondern eher als erwartbares Ergebnis, da zwischen den beiden Sitzungen kein formales CRM-Training stattfand und der zeitliche Abstand kurz war. Vielmehr deute das Fehlen eines Trainingseffekts darauf hin, dass die Teilnahme am ersten Szenario die Leistung im zweiten nicht nennenswert beeinflusst habe. Dies wird als potenzieller Hinweis darauf gewertet, dass high-fidelity simulation in diesem Kontext tatsächlich eine stabile Leistungserfassung ermöglicht. Allerdings wird zugleich eingeräumt, dass die Studie nicht darauf ausgelegt war, Lern- oder Partizipationseffekte systematisch zu untersuchen.

Trotz der insgesamt positiven Befunde benennt das Dokument mehrere Limitationen. Ein erster Kritikpunkt betrifft die Querschnittslogik des Gruppenvergleichs. Zwar konnten Unterschiede zwischen PGY-1- und PGY-3-Residents gezeigt werden, doch erlaubt dieses Design keinen echten intraindividuellen Entwicklungsnachweis. Die Autoren schlagen daher vor, zukünftige Studien mit denselben Residents über mehrere Weiterbildungsjahre hinweg durchzuführen, sodass jede Person ihre eigene Kontrollbedingung darstellt. Eine zweite Limitation liegt in der unklaren klinischen Relevanz der gemessenen Unterschiede. Auch wenn die PGY-3-Gruppe im Mittel über dem als kompetent definierten Schwellenwert von 5 lag und die PGY-1-Gruppe darunter, war die Studie nicht dazu konzipiert, Cutoffs für summative Entscheidungen oder klinisch relevante Leistungsniveaus belastbar festzulegen. Eine weitere Einschränkung betrifft die Interrater-Reliabilität. Zwar wird ein ICC um 0,60 als akzeptabel für erste instrumentelle Anwendungen angesehen, für summative oder hochrelevante Prüfungen ist dieses Niveau jedoch unzureichend. Besonders problematisch erscheinen die geringen Reliabilitätswerte in den Kategorien Communication und Resource Utilization. Hieraus leiten die Autoren die Notwendigkeit ab, sowohl die Skalenanker als auch das Ratertraining weiter zu überarbeiten.

Neben den psychometrischen Grenzen wird auch die Durchführbarkeit als relevante Herausforderung diskutiert. Jede Simulationssitzung erforderte mindestens 20 Minuten, hinzu kamen etwa 20 Minuten für die videobasierte Bewertung. Zudem waren pro Szenario mindestens ein Instruktor sowie zwei unterstützende Akteure erforderlich. Das Verfahren ist damit ressourcenintensiv. Darüber hinaus wird die Frage der Übertragbarkeit des Instruments auf andere Standorte und Kontexte offengelassen. Auch die Rolle der sogenannten *content specificity* wird kritisch reflektiert. Das Dokument verweist auf die medizinpädagogische Literatur, nach der Problemlöseleistung stark inhaltsspezifisch sein kann und für verlässliche Beurteilungen oft mehrere Fälle notwendig sind. Die Autoren diskutieren zwar, dass die Ottawa GRS in zwei sehr

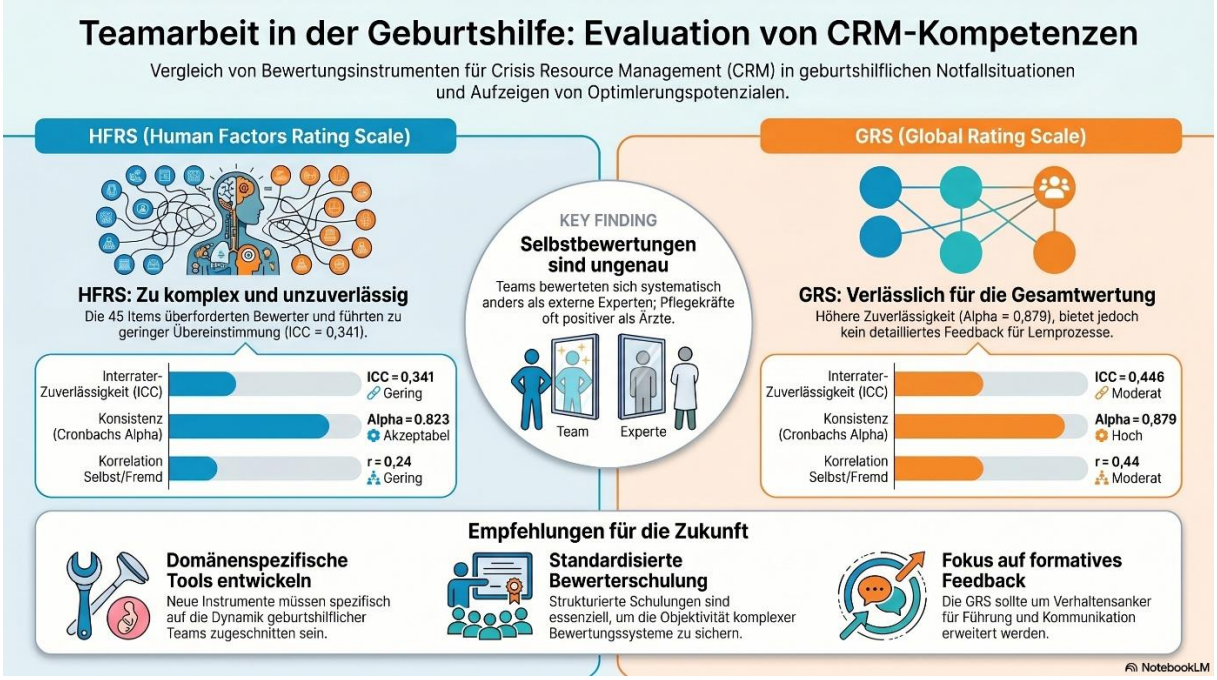
unterschiedlichen Szenarien konsistent zwischen PGY-1- und PGY-3-Leistungen unterscheiden konnte, wollen daraus jedoch nicht vorschnell ableiten, dass Content Specificity für CRM-Bewertungen bedeutungslos sei. Vielmehr sehen sie weiteren Forschungsbedarf, um das Verhältnis von Szenarioinhalt und CRM-Leistungsmessung genauer zu bestimmen.

Zusammenfassend kann die Ottawa GRS auf Grundlage des vorliegenden Dokuments als ein theoriegeleitet entwickeltes, domänenspezifisches Globalrating-Instrument zur Erfassung von Crisis Resource Management in simulatorbasierten Akutsituationen beschrieben werden. Es basiert auf fünf CRM-Kategorien sowie einer globalen Gesamtbewertung, verwendet eine siebenstufige Likert-Skala mit deskriptiven Ankern und wurde unter Einbezug umfassender Expertenrückmeldungen entwickelt und trainiert. Die vorliegenden Daten liefern erste Evidenz für die Konstruktvalidität der Skala, insbesondere im Hinblick auf Inhaltsvalidität, standardisierten Response Process, interne Struktur und die Beziehung zu einer externen Variable in Form des Ausbildungsstands. Besonders hervorzuheben ist die Fähigkeit des Instruments, konsistent zwischen unterschiedlichen Erfahrungsniveaus zu differenzieren. Die Reliabilitätswerte der Gesamtbewertung können als akzeptabel, jedoch noch nicht als ausreichend für summative Prüfungszwecke betrachtet werden. Vor allem in den Kategorien Kommunikation und Ressourcenmanagement besteht Überarbeitungsbedarf. Insgesamt erscheint die Ottawa GRS nach den im Dokument dargestellten Befunden als vielversprechender erster Ansatz zur formalen Bewertung von CRM-Leistung in high-fidelity Simulationsszenarien, der insbesondere für weiterführende Validierungsarbeiten und für formative oder explorative Anwendungen von Bedeutung ist.

5.11 Human Factors Rating Scale (HFRS) und eine Global Rating Scale (GRS)

Quelle: Morgan PJ, Pittini R, Regehr G, Marrs C, Haley MF. Evaluating teamwork in a simulated obstetric environment. Anesthesiology. (2007) 106:907–15.

Abbildung 14: Human Factors Rating Scale (HFRS) und eine Global Rating Scale (GRS)



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Morgan et al. (2007)

Im Dokument *Evaluating Teamwork in a Simulated Obstetric Environment* werden zwei Instrumente zur Bewertung von Teamleistung in simulierten geburtshilflichen Krisensituationen untersucht, nämlich die *Human Factors Rating Scale* (HFRS) und eine *Global Rating Scale* (GRS). Ausgangspunkt der Untersuchung war die Beobachtung, dass Kommunikations- und Teamprobleme in der Geburtshilfe wesentlich zu substandarder Versorgung und zu schwerwiegenden unerwünschten Ereignissen beitragen können. Vor diesem Hintergrund verfolgte die Studie das Ziel, zu prüfen, ob eine für den obstetrischen Kontext adaptierte Human-Factors-Skala sowie eine globale Rating-Skala geeignet sind, die Leistung obstetrischer Teams reliabel zu erfassen. Damit stand nicht nur die Beurteilung von Teamarbeit im Mittelpunkt, sondern zugleich die Frage, ob die eingesetzten Instrumente psychometrisch ausreichend tragfähig sind, um im Bereich der Teamleistungsbeurteilung sinnvoll eingesetzt zu werden.

Die HFRS wurde im Dokument als minimal für den obstetrischen Kontext adaptierte Version des *Operating Room Management Attitudes Questionnaire* (ORMAQ) beschrieben. Das ORMAQ selbst war aus der Luftfahrtforschung hervorgegangen und wurde ursprünglich entwickelt, um Sicherheits- und Teamhaltungen in Operationssaalteams abzubilden. Es war zuvor bereits vielfach weiterentwickelt und in verschiedenen Varianten genutzt worden, insbesondere auch im Kontext des späteren *Safety Attitudes Questionnaire*. Die HFRS stellt in diesem

Zusammenhang den Versuch dar, ein primär auf Teamhaltungen und allgemeine Sicherheitsaspekte ausgerichtetes Instrument in eine verhaltensorientierte Bewertungsform für die Leistung obstetrischer Teams zu überführen. Die GRS ist demgegenüber wesentlich einfacher konzipiert und dient der globalen Gesamtbeurteilung der Teamleistung im jeweiligen Szenario. Im Unterschied zur HFRS verfolgt sie keinen mehrdimensionalen, checklistenartigen Zugriff, sondern arbeitet mit einer einzigen globalen Fünf-Punkte-Skala, deren Bewertungsstufen durch inhaltliche Anker beschrieben werden.

Strukturell unterscheidet sich die HFRS deutlich von der GRS. Die HFRS ist ein umfangreiches Instrument mit 45 Items, die fünf Themenbereichen zugeordnet sind. Diese Themen umfassen Leadership–Structure, Confidence–Assertion, Information Sharing, Teamwork und Error. Bewertet wird jedes Item auf einer fünfstufigen Likert-Skala von „strongly disagree“ bis „strongly agree“. Im Dokument wird darauf hingewiesen, dass Items, die auf ein bestimmtes Szenario nicht zutreffen, leer gelassen werden können. Die HFRS ist somit als umfangreiche Checkliste angelegt, die unterschiedliche Aspekte teambezogenen Verhaltens in relativ hoher Breite erfasst. Innerhalb der Domäne Leadership–Structure finden sich beispielsweise Items dazu, ob Fachärzte Fragen von Residents ermutigten, ob Pflegekräfte angemessen konsultiert wurden oder ob der Erfolg der Fallbewältigung im Wesentlichen auf die Expertise einzelner Ärzte zurückzuführen war. Die Domäne Confidence–Assertion adressiert die Hierarchie und Durchsetzungsfähigkeit im Team, etwa durch Items zur Frage, ob Residents oder Pflegekräfte ärztliche Entscheidungen in kritischen Situationen hinterfragten oder ob Unsicherheiten offen angesprochen wurden. Information Sharing bezieht sich auf das klare Verbalisieren von Handlungsplänen, die Sicherstellung der Bestätigung von Aufträgen sowie die Effizienz des Informationsaustauschs zwischen und innerhalb der Berufsgruppen. Die Teamwork-Domäne erfasst beispielsweise Feedbackprozesse zwischen den Professionen, Koordinationsleistungen durch verschiedene Teammitglieder, Priorisierung von Aktivitäten, Konfliktlösung und das allgemeine Funktionieren der Zusammenarbeit. Die Error-Domäne nimmt eine besondere Stellung ein, da sie weniger direktes Teamverhalten als vielmehr Fehler und deren vermutete Ursachen adressiert, etwa Mängel in Wissen, Kommunikation, Ausrüstung, Technik, Erfahrung oder Ressourcen.

Demgegenüber ist die GRS als globale Leistungsbewertung deutlich kompakter. Sie basiert auf einer einzigen fünfstufigen Bewertungsskala, die von „unacceptable performance“ über „borderline“ und „acceptable“ bis zu „good“ und „superior performance“ reicht. Im Dokument wird auf Seite 9 ausgeführt, dass diese Bewertungsstufen mit spezifischen Deskriptoren hinterlegt sind, die insbesondere Fehlerzahl, Geschwindigkeit der Reaktion auf kritische Ereig-

nisse, Qualität der Teamkommunikation und Auswirkungen auf die Patientensicherheit einbeziehen. Eine Stufe von 1 beschreibt ein Team, das multiple Fehler begeht, kritische Ereignisse nicht oder nur mit Hilfe erkennt und praktisch keine Teamkommunikation zeigt. Eine Stufe von 3 repräsentiert eine akzeptable Leistung mit einigen Fehlern, die jedoch nicht zu irreversiblem Schaden geführt hätten, sowie einer zufriedenstellenden, aber führungsschwachen Kommunikation. Eine Stufe von 5 beschreibt schließlich eine überlegene Leistung mit nur wenigen geringfügigen Fehlern, promptem Erkennen und Managen kritischer Ereignisse und exzellenter Führung mit klarer, prägnanter Teamkommunikation. Die GRS ist somit weniger differenziert, integriert aber in ihrer Globalbewertung mehrere sicherheitsrelevante Aspekte, insbesondere Teamkommunikation, Fehlerkontrolle und Führungsqualität.

Beide Instrumente wurden in einem hochrealistischen Simulationssetting erprobt. Das Simulationszentrum war als geburtshilflicher Operationssaal eingerichtet und mit chirurgischem Material, fetalem Monitoring, Videotechnik sowie einem adaptieren Simulationsmodell ausgestattet, das einen termingerechten Uterus simulierte und operative Eingriffe sowie Blutverlust realitätsnah abbilden konnte. Entwickelt wurden vier Szenarien, die auf den häufigsten Ereignissen beruhten, die im Rahmen des *National Confidential Enquiry into Maternal Deaths* mit maternaler Mortalität assoziiert worden waren und zusätzlich durch Befragungen von Obstetrikern, Anästhesisten sowie obstetrischen Pflegekräften in ihrer Relevanz für das Simulationstraining bestätigt wurden. Die Szenarien umfassten eine schwierige Intubation mit Hypoxämie und Kreislaufstillstand bei einer morbid adipösen Schwangeren mit pathologischem fetalem Herzfrequenzmuster, eine Notfallsituation bei sich verschlechternder Präeklampsie mit schwerer Hypertonie und Lungenödem, eine Fruchtwasserembolie nach Notsectio bei Nabelschnurvorfall sowie eine massive Blutung bei okkulter Abruptio placentae mit fetaler Bradykardie. Diese Szenarien zeichneten sich durch ihre Komplexität und ihre hohen Anforderungen an interprofessionelle Teamarbeit aus.

An der Studie nahmen 34 Personen teil, darunter 16 Pflegekräfte, 6 Obstetriker, 6 Anästhesisten sowie 6 Residents. Die Teilnehmer repräsentierten laut Dokument mehr als 70 % der an der obstetrischen Versorgung eines akademischen Zentrums beteiligten Ärzte sowie eine Querschnittsstichprobe der dort tätigen Pflegekräfte. In drei unabhängigen Sitzungen wurden aus diesen Personen Teams von jeweils fünf oder sechs Mitgliedern gebildet, wobei die Teamzusammensetzungen für die verschiedenen Szenarien computerbasiert randomisiert wurden. Dadurch nahm jede Person an zwei oder drei Szenarien in jeweils unterschiedlich zusammengesetzten Teams teil. Nach jeder etwa 20-minütigen Simulation füllten die Teammitglieder sowohl die HFRS als auch die GRS zur Selbsteinschätzung der Teamleistung aus. Zusätzlich

wurden alle zwölf aufgezeichneten Teamperformances von neun externen Ratern bewertet, die entweder über Expertise im obstetrischen Bereich oder im Bereich Human Factors verfügten. Die beiden Instrumente wurden damit sowohl in Selbst- als auch in Fremdbewertungsformaten geprüft.

Die psychometrische Analyse der Instrumente konzentrierte sich auf ihre Reliabilität sowie auf die Frage, inwieweit sie zwischen unterschiedlichen Teamperformances und Szenarien unterscheiden können. Für die externen Bewertungen ergaben sich deutliche Unterschiede zwischen HFRS und GRS. Die Interrater-Reliabilität der HFRS auf Ebene eines einzelnen Raters war mit einem Intraklassenkorrelationskoeffizienten von 0,341 niedrig. Der Mittelwert über neun externe Rater erreichte jedoch ein Cronbachs Alpha von 0,823, was darauf hindeutet, dass erst die Aggregation einer großen Zahl von Beurteilungen eine einigermaßen stabile Bewertung ermöglicht. Für die GRS zeigte sich ein etwas besseres Bild. Hier lag die Einzelrater-ICC bei 0,446 und der Mittelwert über neun Rater erreichte ein Cronbachs Alpha von 0,879. Auch wenn die Reliabilität einzelner Urteile damit ebenfalls begrenzt war, erwies sich die GRS als insgesamt stabiler als die HFRS. Gleichzeitig korrelierten die über alle externen Rater gemittelten HFRS- und GRS-Werte über die zwölf Szenarien mit 0,934 sehr hoch miteinander. Das Dokument interpretiert dies dahingehend, dass beide Verfahren im Wesentlichen dasselbe Konstrukt der Teamleistung erfassen, auch wenn sie dies mit unterschiedlicher psychometrischer Qualität tun.

Auch im Hinblick auf die Fähigkeit, Unterschiede zwischen Szenarien abzubilden, zeigen sich Unterschiede zwischen beiden Instrumenten. In den externen Bewertungen wiesen sowohl HFRS als auch GRS signifikante Unterschiede zwischen den vier Szenarien auf, was nahelegt, dass einige Szenarien von den Ratern konsistent als schwieriger oder leistungskritischer erlebt wurden als andere. Bei den Selbstbewertungen der Teammitglieder ergab sich für die HFRS jedoch kein signifikanter Unterschied zwischen den Szenarien, während die GRS zumindest grenzwertig Unterschiede abbilden konnte. Diese Befunde deuten darauf hin, dass die HFRS in der Selbstanwendung nicht in der Lage war, Unterschiede in der Schwierigkeit oder Qualität von Teamperformances ausreichend sensibel zu erfassen, während die GRS hier etwas leistungsfähiger erschien.

Besonders kritisch fällt im Dokument die Bewertung der Selbstbeurteilungen aus. Für die HFRS lag das Cronbachs Alpha der sechs Teammitglieder, deren Einschätzungen zu einem Teamscore aggregiert wurden, bei nur 0,15. Dies deutet auf eine extrem geringe Übereinstimmung hin und macht deutlich, dass die HFRS für teaminterne Selbstbewertungen im vorliegenden Kontext ungeeignet ist. Die GRS zeigte in den Selbstbewertungen deutlich bessere

Werte mit einem Cronbachs Alpha von 0,74, was als moderat stabil eingeordnet werden kann. Zudem ergab sich bei den HFRS-Selbstbewertungen ein signifikanter Einfluss der Profession des Beurteilers. Pflegekräfte vergaben im Mittel höhere Teamwerte als Ärzte. Für die GRS bestand ein solcher berufsgruppenabhängiger Unterschied nicht. Gleichzeitig zeigte eine professionsspezifische Subanalyse der HFRS, dass keine Interaktion zwischen der Profession der Bewerter und der bewerteten Personen vorlag. Das heißt, Pflegekräfte und Ärzte unterschieden sich zwar in ihrer generellen Großzügigkeit, bevorzugten aber nicht systematisch die eigene Berufsgruppe in ihren Urteilen. Der Zusammenhang zwischen Selbst- und Fremdbewertung fiel insgesamt gering aus. Für die HFRS betrug die Korrelation zwischen selbstgenerierten und externen Bewertungen 0,24, für die GRS 0,44. Das Dokument interpretiert diese Werte als weiteren Hinweis darauf, dass Selbstbewertungen nur in begrenztem Maße geeignet sind, tatsächliche Teamleistung valide abzubilden, und bestätigt damit die Notwendigkeit externer Beurteilungen.

Im Diskussionsteil zieht das Dokument eine deutlich kritische Bilanz hinsichtlich der HFRS. Zwar seien die inhaltlichen Themen des Instruments wie Leadership, Assertiveness, Information Sharing und Teamwork auf den ersten Blick sinnvolle Verhaltensaspekte von Teamleistung, es bestehe jedoch der Verdacht, dass die Vielzahl der Items innerhalb der Kategorien die HFRS als praktikables und reliables Performanceinstrument schwächt. Diese Interpretation wird mit Literatur zu Checklisteninstrumenten in anderen Kontexten in Verbindung gebracht, die ebenfalls darauf hinweist, dass umfangreiche Checklisten mit zunehmender Komplexität an Reliabilität verlieren können. Darüber hinaus betonen die Autoren, dass aus der Luftfahrt oder dem Operationssaal adaptierte Instrumente nicht ohne Weiteres in einen obstetrischen Teamkontext übertragen werden können. Obstetrische Krisen könnten sich in ihrer Struktur so deutlich von anderen Hochrisikosettings unterscheiden, dass ein spezifisch für diesen Bereich entwickeltes Bewertungssystem erforderlich sei. Das Dokument kommt daher zu dem Schluss, dass die vorliegenden Ergebnisse die Verwendung der HFRS zur Beurteilung obstetrischer Teams nicht unterstützen.

Die GRS wird demgegenüber deutlich positiver beurteilt. Sie war sowohl in der Fremd- als auch in der Selbstbewertung besser in der Lage, Teamperformances voneinander zu unterscheiden, unterschiedliche Szenarioschwierigkeiten zu erkennen und professionenspezifische Bewertungsverzerrungen zu vermeiden. Die Autoren führen dies unter anderem darauf zurück, dass bei der GRS nur ein einziger Gesamteindruck zu bewerten ist und die Rater das Resultat des Szenarios leichter als Gesamturteil integrieren können. Gleichzeitig wird aber auch betont, dass die Einfachheit der GRS ihre zentrale Limitation darstellt. Sie erlaubt zwar eine globale

Beurteilung, benennt aber keine spezifischen Stärken oder Schwächen eines Teams und liefert daher nur begrenzt nutzbare Informationen für formative Rückmeldungen und für die gezielte Gestaltung von Trainingsmaßnahmen. Gerade weil Debriefing und Rückmeldung in der simulationsbasierten Lehre als besonders bedeutsam angesehen werden, wäre für einen formativen Einsatz eine differenziertere Struktur notwendig. Das Dokument folgert daher, dass die GRS eher Potenzial als summatives Instrument besitzt, während sie für formative Zwecke um einige gezielte Subkategorien erweitert werden müsste, die differenziertere Rückmeldungen erlauben.

Neben den instrumentenspezifischen Befunden diskutiert das Dokument auch grundsätzliche Implikationen für die Entwicklung künftiger Teamleistungsinstrumente in der Geburtshilfe. Die Ergebnisse werden als Hinweis darauf gewertet, dass es sinnvoller sein könnte, ein domänenspezifisches Verhaltenserfassungssystem für obstetrische Teams von Grund auf neu zu entwickeln, anstatt ein bestehendes Instrument aus anderen Kontexten lediglich zu adaptieren. Dabei sollen qualitative Analysen von Sicherheitsattitüden aus Fokusgruppen sowie Experten- und Nichtexperteneinschätzungen zu beobachtbaren Verhaltensmarkern in simulierten obstetrischen Krisen genutzt werden. Das Dokument verweist darauf, dass ähnliche Vorgehensweisen bereits erfolgreich für die Entwicklung behavioraler Markierungssysteme bei Anästhesisten sowie Chirurgen eingesetzt wurden. Die vorliegende Studie wird somit explizit als Ausgangspunkt für die Entwicklung eines spezifischen, obstetrischen Teambewertungsinstruments verstanden.

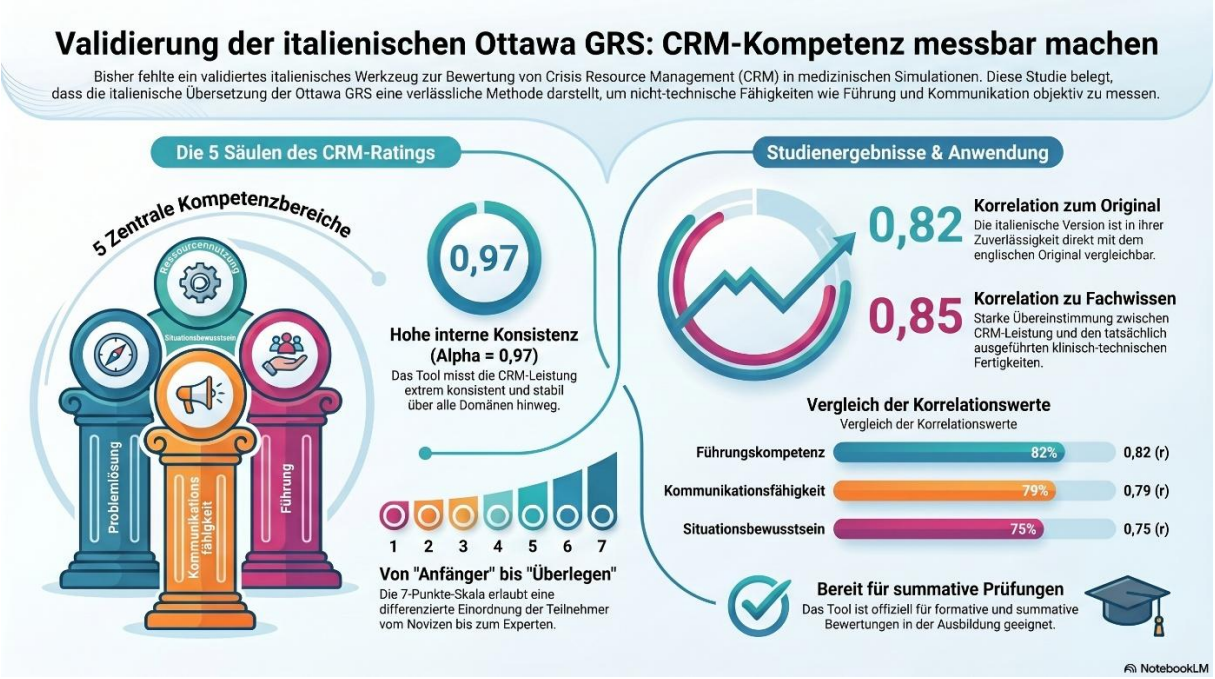
Zusammenfassend lässt sich auf Basis des Dokuments festhalten, dass die HFRS und die GRS zwei deutlich unterschiedliche Ansätze zur Erfassung von Teamleistung in simulierten obstetrischen Krisensituationen repräsentieren. Die HFRS ist ein umfangreiches, thematisch breit angelegtes, aus einem bestehenden Haltungstool adaptiertes Instrument mit 45 Items in fünf Domänen. Trotz ihrer inhaltlichen Breite erwies sie sich in der vorliegenden Studie insbesondere in Selbstbewertungen als psychometrisch unzureichend und auch in externen Bewertungen erst nach Aggregation einer großen Zahl von Ratern als einigermaßen stabil. Die GRS ist demgegenüber ein einfaches Globalurteil mit fünf verankerten Leistungsstufen und zeigte insgesamt günstigere psychometrische Eigenschaften. Sie differenzierte besser zwischen unterschiedlichen Teamleistungen und Szenarien und war weniger anfällig für professionsbezogene Verzerrungen. Gleichzeitig ist ihre diagnostische Tiefe begrenzt, sodass sie für formative Rückmeldungen nur eingeschränkt geeignet erscheint. Insgesamt stützt das Dokument daher die Schlussfolgerung, dass die HFRS in der untersuchten Form keine geeignete Grundlage für die Bewertung obstetrischer Teamleistung darstellt, während die GRS zwar Potenzial für

summative Bewertungen erkennen lässt, langfristig jedoch ein spezifisch entwickeltes, verhaltensorientiertes Instrument für obstetrische Teams erforderlich ist.

5.12 Die italienische Version der Ottawa Crisis Resource Management Global Rating Scale

Quelle: Franc JM, Verde M, Gallardo AR, Carenzo L, Ingrassia PL. An Italian version of the Ottawa crisis resource management global rating scale: a reliable and valid tool for assessment of simulation performance. Intern Emerg Med. (2017) 12:651–6.

Abbildung 15: Die italienische Version der Ottawa Crisis Resource Management Global Rating Scale



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Franc et al. (2017)

Die italienische Version der Ottawa Crisis Resource Management Global Rating Scale wurde mit dem Ziel untersucht, ein sprachlich angepasstes Instrument zur objektiven Beurteilung von Simulationsleistung im italienischsprachigen Raum bereitzustellen. Ausgangspunkt der Studie war die Feststellung, dass im italienischen Sprachraum kein publiziertes Instrument zur Bewertung simulationsbasierter Leistung verfügbar war, dessen Reliabilität und Validität doku-

mentiert worden wären. Vor diesem Hintergrund sollte die Übersetzung eines bereits etablierten englischsprachigen Instruments, des Ottawa GRS, eine praktikable Lösung darstellen. Das Dokument ordnet dieses Vorhaben in den breiteren Wandel medizinischer Ausbildung ein, der durch einen Übergang von der traditionellen, stark beobachtungsorientierten Ausbildung hin zu aktiveren, kompetenzbasierten und simulationsgestützten Lernformaten gekennzeichnet ist. Simulation wird dabei als besonders geeignet für kurze, intensive Lernsituationen mit unmittelbarer Rückmeldung, Reflexion und Verhaltenskorrektur beschrieben. Gerade in diesem Kontext kommt der Qualität von Bewertungsinstrumenten eine zentrale Bedeutung zu, da nach Auffassung der Autoren Assessmentdaten nur dann sinnvoll interpretiert werden können, wenn sie hinreichend reliabel und valide sind.

Die Studie versteht die italienische Version des Ottawa GRS nicht als bloße sprachliche Kopie des Originals, sondern als zu prüfende Neufassung eines bestehenden Instruments. Im Dokument wird ausdrücklich hervorgehoben, dass selbst eine sorgfältige Übersetzung eines reliablen und validen englischsprachigen Instruments keineswegs automatisch ein ebenfalls reliables und valides Instrument in einer anderen Sprache hervorbringt. Vielmehr können sprachliche Nuancen, insbesondere im Bereich der Bewertung menschlicher Leistung, erhebliche Auswirkungen auf die Interpretierbarkeit und Verlässlichkeit eines Beurteilungsinstruments haben. Daher verfolgte die Studie zwei Ziele. Primär sollte die Reliabilität der italienischen Fassung im Vergleich zum englischsprachigen Original geprüft werden. Sekundär sollte die Validität der italienischen Version durch den Vergleich mit einem unabhängig erhobenen technischen Skills-Score abgeschätzt werden. Das methodische Grundprinzip bestand somit darin, die Stärke der Assoziation zwischen der italienischen GRS-Version und zwei etablierten beziehungsweise ergänzenden Vergleichsmaßen zu untersuchen.

Die Entwicklung der italienischen Version erfolgte in einem iterativen Gruppenprozess, an dem alle Autoren beteiligt waren. Dieser Prozess sollte die face validity des neuen Instruments sicherstellen. Die Übersetzung von Englisch nach Italienisch wurde nicht durch eine Einzelperson vorgenommen, sondern gemeinsam von mehrsprachigen Autoren erarbeitet, von denen drei Italienisch und eine Person Englisch als Erstsprache hatten. Das englischsprachige Ottawa GRS wurde im Rahmen der Studie unverändert in seiner publizierten Originalfassung verwendet, um einen möglichst direkten Vergleich mit der italienischen Fassung zu ermöglichen. Das Dokument verweist darauf, dass das Originalinstrument bereits gründlich hinsichtlich seiner Validität und Reliabilität untersucht worden war und deshalb als geeignete Referenz dienen konnte. Insbesondere wird hervorgehoben, dass das Ottawa GRS eine siebenstufige,

halb verankerte Skala für die Gesamtleistung sowie fünf domänenspezifische Felder umfasst, die ebenfalls auf derselben Skala beurteilt werden.

Strukturell folgt die italienische Version vollständig dem Aufbau des englischsprachigen Originals. Das Instrument umfasst ein Feld zur globalen Gesamtbewertung der Leistung sowie fünf spezifische Domänen, nämlich Leadership, Problem Solving, Situational Awareness, Resource Utilization und Communication. Alle Bereiche werden auf einer Skala von 1 bis 7 bewertet. Der Wert 1 steht für eine Leistung auf dem Niveau eines Novizen, während der Wert 7 eine deutlich überlegene Leistung bezeichnet. Die Skala ist als semi-anchored scale angelegt, da für die Stufen 1, 3, 5 und 7 textliche Anker vorhanden sind, die die Anforderungen für das Erreichen der jeweiligen Stufe genauer beschreiben. Dieses Prinzip gilt sowohl für die globale Gesamtbewertung als auch für die fünf Domänen. Bei dem Instrument handelt es sich damit nicht um eine umfangreiche Checkliste mit vielen Einzelitems, sondern um ein kompaktes Globalrating-Instrument mit klar definierten Leistungsdimensionen. Die Struktur ist auf die Beurteilung nicht-technischer Leistungen in Simulationssituationen zugeschnitten und verbindet eine globale Einschätzung mit domänenspezifischen Teilbewertungen.

Zusätzlich zum italienischen und englischen GRS wurde in der Studie ein technischer Skills-Score eingesetzt. Dabei handelte es sich um szenariospezifisch entwickelte technische Checklisten, in denen einzelne Handlungen mit 0, 1 oder 2 Punkten bewertet wurden, je nachdem, ob die Handlung nicht, unvollständig oder vollständig ausgeführt wurde. Dieser Skills-Score diente nicht als alternatives CRM-Instrument, sondern als unabhängiges Vergleichsmaß für die Beurteilung der Validität der italienischen GRS-Fassung. Seine Einbeziehung ermöglicht es, den Zusammenhang zwischen globaler CRM-Leistungsbewertung und technisch-operativer Leistung zu untersuchen.

Der Anwendungsbereich der italienischen Version des Ottawa GRS lag in der Beurteilung von Simulationsperformances im Rahmen eines Simulationsformats nach dem sogenannten Sim-War-Modell. Dieses Format ist als kompetitiver Teamansatz aufgebaut, bei dem Teams auf Basis vorab definierter Scoringtools in weitere Runden eines Turnierformats aufsteigen. Die Teilnehmer wurden über Mailinglisten an verschiedene Weiterbildungsprogramme in Italien rekrutiert und registrierten sich freiwillig in Teams zu je vier Personen. Insgesamt nahmen 28 Residents aus acht italienischen Universitäten teil. Die Teilnehmer stammten aus unterschiedlichen Fachrichtungen, darunter überwiegend Anästhesiologie und Notfallmedizin, aber auch Kardiologie, Innere Medizin, Pneumologie und Geriatrie. Das mediane Alter lag bei 30 Jahren, und die mediane Weiterbildungsdauer betrug drei Jahre. Die Simulationsszenarien umfassten ein breites Spektrum von Einsatzlagen, darunter eine neonatale Reanimation bei Geburtsas-

phyxie, ein internistisches Szenario mit akutem Lungenödem, ein obstetrisches Szenario mit Eklampsie, eine kardiologische Reanimationssituation mit Hyperkaliämie-bedingtem Herzstillstand sowie ein Virtual-Reality-Szenario zur Triage von zehn Verletzten nach einem Verkehrsunfall. Darüber hinaus umfasste das Gesamttournamentformat ein Halbfinale mit Schädel-Hirn-Trauma-Management und ein Finale mit Herzstillstand, peri-mortem Sectio und neonataler Reanimation. Alle Simulationen wurden auf Italienisch durchgeführt.

Während der Simulationen wurden die Teams von jeweils drei unabhängigen Ratern beurteilt, die jeweils nur eines der drei Beurteilungssysteme verwendeten, nämlich das italienische GRS, das englische GRS oder den technischen Skills-Score. Alle Rater waren Fakultätsmitglieder der Università del Piemonte Orientale und hatten keine Vorerfahrung mit dem Ottawa GRS. Die italienischsprachigen GRS-Bewertungen wurden von Muttersprachlern im Italienischen vorgenommen, während die Rater des englischen GRS sowohl Englisch als auch Italienisch fließend beherrschten. Obwohl sich alle Rater gleichzeitig im selben Raum befanden, wurden sie angewiesen, die Bewertung unabhängig voneinander und ohne Austausch über Instrumente oder Beobachtungen vorzunehmen. Für jedes Szenario bewerteten dieselben Rater alle Teams, sodass systematische Unterschiede zwischen Ratern zumindest innerhalb eines Szenariokontextes konstant blieben. Die Studie umfasste insgesamt 41 bewertete Simulationen und 123 vollständige Scoring-Beobachtungen

Die psychometrische Analyse konzentrierte sich auf die Stärke der Assoziation zwischen den Messwerten. Zur Bewertung der Reliabilität der italienischen Version wurde die Korrelation zwischen dem italienischen Overall GRS und dem englischen Overall GRS berechnet. Das Ergebnis zeigte einen Korrelationskoeffizienten von 0,82 bei einem adjustierten 95 %-Konfidenzintervall von 0,62 bis 0,92. Der entsprechende p-Wert lag unter 0,000001. Die im Dokument dargestellte Streudiagrammabbildung zeigt, dass sich die Bewertungen über den gesamten Skalenbereich von 1 bis 6 hinweg gut verteilten und die positive Beziehung zwischen beiden Sprachversionen über das gesamte Leistungsspektrum erhalten blieb. Auch auf Ebene der fünf Domänen ergaben sich hohe Korrelationen zwischen italienischer und englischer Fassung. Für Leadership lag die Korrelation bei 0,82, für Problem Solving bei 0,83, für Situational Awareness bei 0,81, für Resource Utilization bei 0,73 und für Communication bei 0,71. Alle Korrelationen waren hochsignifikant. Diese Ergebnisse sprechen dafür, dass die italienische Version in ihrer Struktur und ihrem Bewertungsverhalten dem englischen Original weitgehend entspricht.

Zur Validitätsprüfung wurde die Korrelation zwischen dem italienischen Overall GRS und dem technischen Skills-Score berechnet. Dabei ergab sich ein Korrelationskoeffizient von 0,85 mit

einem adjustierten 95 %-Konfidenzintervall von 0,68 bis 0,94. Auch hier lag der p-Wert unter 0,000001. Das zugehörige Streudiagramm zeigt, dass die technischen Skill-Werte ebenfalls über einen breiten Wertebereich verteilt waren und die positive Beziehung über niedrige und hohe Leistungsniveaus hinweg bestehen blieb. Das Dokument interpretiert diese Korrelation als Hinweis darauf, dass eine höhere CRM-bezogene Gesamtleistung mit einer besseren technischen Durchführung einhergeht. Diese Interpretation wird mit früherer Literatur in Verbindung gebracht, die ebenfalls positive Zusammenhänge zwischen technischen und nicht-technischen Leistungen im Simulationskontext gezeigt hatte. Die Studie versteht diese Befunde somit als erste Evidenz für die Validität der italienischen GRS-Version.

Zusätzlich berichtet das Dokument über eine post hoc berechnete interne Konsistenz in Form von Cronbachs Alpha. Dieser Wert lag bei 0,97, war allerdings nicht Teil des ursprünglich festgelegten Analyseplans. Die Autoren betonen daher ausdrücklich, dass dieser Befund nur hypothesengenerierend interpretiert werden dürfe. Gleichwohl wird diskutiert, dass ein so hoher Wert möglicherweise auf eine exzellente interne Konsistenz der italienischen Version hinweist. Im Dokument wird dieser Befund in Beziehung zu den von Downing vorgeschlagenen Reliabilitätsanforderungen gesetzt. Nach dieser Einordnung sollten Reliabilitätswerte über 0,9 für high-stakes Prüfungen wie Lizensierungsentscheidungen vorliegen, Werte zwischen 0,8 und 0,89 für summative Prüfungen wie Jahresabschlussbeurteilungen und Werte zwischen 0,7 und 0,79 für informelle formative oder classroom-type Assessments. Vor diesem Hintergrund kommen die Autoren zu dem Schluss, dass die italienische Version des Ottawa GRS zumindest für informelle und formative Einsätze sowie möglicherweise auch für niedrigschwellige summative Verwendungen eine ausreichende Reliabilität aufweist, während für high-stakes Anwendungen noch weitere Evidenz erforderlich ist.

Im Diskussionsteil werden die Ergebnisse jedoch mit der nötigen methodischen Vorsicht interpretiert. Ein zentrales Problem besteht darin, dass für jede Sprachversion jeweils nur ein Rater pro Szenario eingesetzt wurde. Unterschiede zwischen italienischer und englischer Bewertung lassen sich daher nicht eindeutig auf die Sprache des Instruments oder auf normale Interrater-Unterschiede zurückführen. Das Dokument begegnet diesem Problem, indem es die beobachteten Korrelationen mit den in der ursprünglichen Ottawa-GRS-Studie berichteten Interrater-Reliabilitäten des englischen Instruments vergleicht. Dort lagen die Intraklassenkorrelationen für die Gesamtbewertung zwischen 0,590 und 0,613 und für einzelne Domänen teils noch deutlich niedriger. Vor diesem Hintergrund argumentieren die Autoren, dass die Korrelationen zwischen italienischer und englischer Version des Instruments zumindest ebenso hoch oder sogar höher seien als die Korrelationen zwischen zwei englischsprachigen Beobachtern. Dies

werde als Hinweis gewertet, dass die italienische Version mindestens so reliabel sein könnte wie das Original.

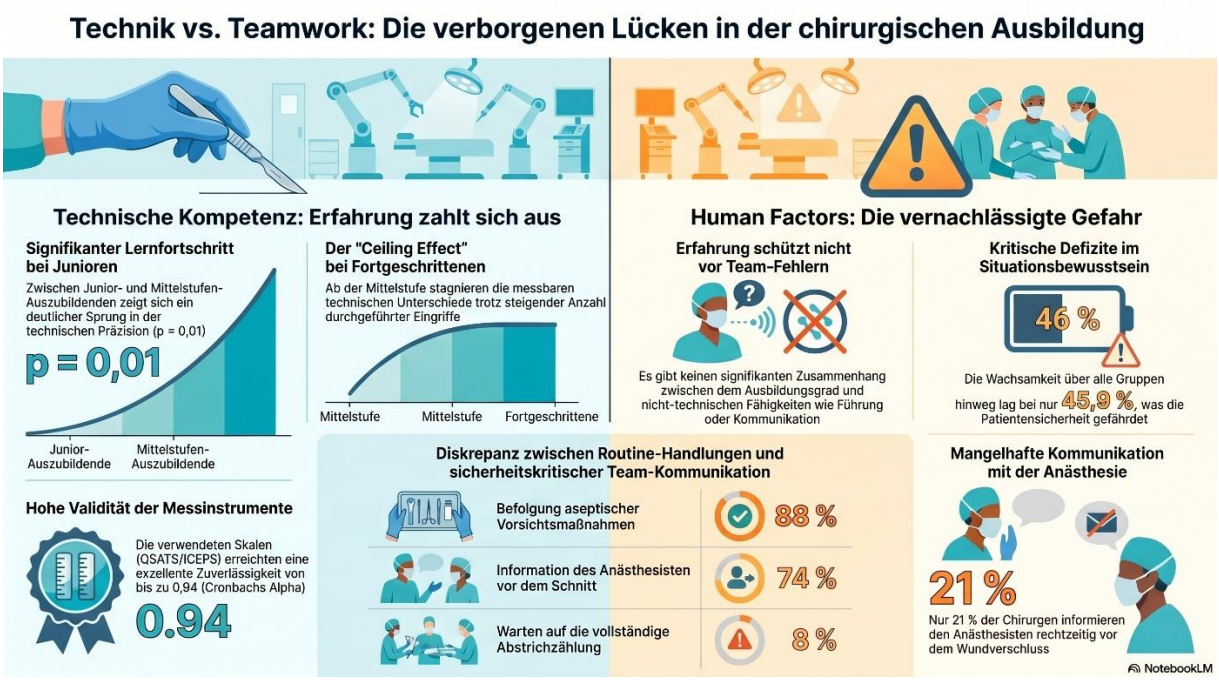
Gleichzeitig benennt das Dokument mehrere Limitationen. Die Konfidenzintervalle der Korrelationen waren trotz hochsignifikanter p-Werte relativ breit, und ihre unteren Grenzen lagen durchweg unter dem als wünschenswert definierten Wert von 0,7. Dies deutet darauf hin, dass die Stichprobe für eine präzise Schätzung der Korrelationen möglicherweise zu klein war. Außerdem erfolgte die Bewertung auf Teamebene. Da Teams aus mehreren Personen mit möglicherweise unterschiedlichem Kompetenzniveau bestanden, kann zusätzliche Varianz entstanden sein, die in einer Individualbewertung möglicherweise geringer ausgefallen wäre. Die Autoren schlagen daher vor, zukünftige Studien könnten dieselbe Fragestellung mit Fokus auf die Bewertung einzelner Personen statt ganzer Teams wiederholen. Eine weitere Einschränkung besteht darin, dass die italienische GRS-Version in dieser Studie nicht vollständig direkt validiert wurde, sondern primär über die Beziehung zum englischen Original und zum Skills-Score untersucht wurde. Auch wenn dies als pragmatischer und nachvollziehbarer Ansatz beschrieben wird, werden weitere direkte Studien zur Reliabilität und Validität ausdrücklich gefordert, bevor das Instrument in hochrelevanten Prüfungssettings eingesetzt werden kann.

Zusammenfassend lässt sich festhalten, dass die italienische Version des Ottawa GRS im vorliegenden Dokument als sprachlich adaptierte Fassung eines etablierten Globalrating-Instruments zur Bewertung simulationsbezogener CRM-Leistung untersucht wurde. Das Instrument übernimmt die Struktur des Originals mit einer globalen Gesamtbewertung und fünf Domänen, die auf einer siebenstufigen halb verankerten Skala bewertet werden. Die Übersetzung erfolgte in einem iterativen, bilingualen Gruppenprozess und wurde anschließend in einem simulationsbasierten Wettbewerbsformat mit multiprofessionellen Notfallszenarien geprüft. Die berichteten Korrelationen zwischen italienischer und englischer Version sowie zwischen italienischer Version und technischem Skills-Score liefern erste Hinweise auf eine angemessene Reliabilität und Validität. Auch wenn der zusätzliche post hoc berichtete Alpha-Wert auf eine sehr hohe interne Konsistenz hindeutet, bleibt die Evidenz für high-stakes Anwendungen noch unzureichend. Nach Maßgabe des Dokuments erscheint die italienische Version des Ottawa GRS daher vor allem für informelle und formative Beurteilungszwecke geeignet und stellt einen wichtigen ersten Schritt zur Etablierung eines objektiven italienischsprachigen Instruments für die Bewertung von Simulationsleistung dar.

5.13 Line Operations Safety Audit (LOSA)

Quelle: Moorthy K, Munz Y, Adams S, Pandey V, Darzi A. A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. Ann Surg. (2005) 242:631–9.

Abbildung 16: Line Operations Safety Audit (LOSA)



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Moorthy et al. (2005)

Im Dokument *A Human Factors Analysis of Technical and Team Skills Among Surgical Trainees During Procedural Simulations in a Simulated Operating Theatre* wird unter mehreren Bewertungsverfahren auch ein Instrument zur Erfassung nicht-technischer Fähigkeiten beschrieben, das auf ausgewählten Elementen der LOSA-Checkliste basiert. Die Studie verfolgte insgesamt das Ziel, technische und teambezogene Kompetenzen chirurgischer Weiterbildungsassistenten in einer simulierten Operationsumgebung gemeinsam zu erfassen. Vor dem Hintergrund, dass chirurgische Ausbildung traditionell vor allem technische und klinische Fertigkeiten fokussiert und nicht-technische Kompetenzen wie Kommunikation, Vigilanz oder Führung in der Regel weder systematisch trainiert noch strukturiert zurückgemeldet werden, sollte ein Simulationsansatz entwickelt werden, der eine ganzheitlichere Betrachtung operativer Kompetenz ermöglicht. Das LOSA-basierte Verfahren war dabei ausdrücklich nicht als voll-

ständig etabliertes Standardinstrument angelegt, sondern als exploratives, pilotiertes Assessment zur Beobachtung chirurgisch relevanter Human-Factors-Aspekte.

Die Grundlage für dieses Verfahren bildeten einzelne Elemente der LOSA-Checkliste, die ursprünglich für die Beurteilung nicht-technischer Fähigkeiten in der Luftfahrt entwickelt worden war. Im Dokument wird ausgeführt, dass aus dieser Checkliste jene Aspekte ausgewählt wurden, die für den chirurgischen Kontext als relevant erschienen. Ergänzt wurde diese Auswahl durch die Ergebnisse eines Pilotfragebogens, der an 35 beratende Chirurgen sowie höhere chirurgische Trainees versendet worden war. Auf dieser Grundlage entstand ein Verfahren, das vier zentrale Verhaltensbereiche abbildet, nämlich *preoperative preparation*, *communication and interaction*, *vigilance/situation awareness* und *leadership*. Bereits aus diesem Entwicklungsprozess wird deutlich, dass die Konstruktion des Instruments zwar theoriegeleitet und inhaltlich begründet erfolgte, zugleich aber explorativen Charakter hatte. Im Diskussionsteil wird ausdrücklich eingeräumt, dass die inhaltliche Fundierung auf diesem Weg nur in begrenztem Umfang als gesichert gelten könne und künftig systematischere Verfahren, etwa im Sinne einer kognitiven Aufgabenanalyse, notwendig wären, um die Inhaltsvalidität eines solchen Instruments robuster abzusichern.

Die Struktur des LOSA-basierten Instruments ist im Dokument tabellarisch dargestellt. Die Domäne *preoperative preparation* umfasst drei Elemente, nämlich die Vorstellung gegenüber den Teammitgliedern, die präoperative Kontrolle von Instrumenten und Equipment sowie ein Briefing. Die Domäne *communication and interaction* beinhaltet vier Elemente, die auf die Interaktion mit Assistenz- und Instrumentierpersonal abzielen. Bewertet wird, ob Anweisungen klar und höflich formuliert werden, ob eine Bestätigung der Anweisungen abgewartet wird, ob aktiv Unterstützung durch Teammitglieder gesucht wird und ob Hilfe oder Ratschläge aus dem Team anerkannt werden. Die dritte Domäne, *vigilance/situation awareness*, erfasst, ob die chirurgisch handelnde Person die Vitalparameter des Patienten während des gesamten Eingriffs im Blick behält, sich des Anästhesisten bewusst zeigt und aktiv die Kommunikation mit dem Anästhesisten initiiert. Die vierte Domäne *leadership* umfasst die Einhaltung bewährter Verfahrensweisen, den angemessenen Einsatz von Ressourcen im Sinne einer geeigneten Aufgabenverteilung und Delegation sowie Autorität beziehungsweise Assertivität. Jedes dieser Elemente wurde auf einer fünfstufigen Likert-Skala bewertet, wobei die Extremwerte durch Anker beschrieben waren. Der resultierende Gesamtscore wurde anschließend als Prozentwert ausgewiesen. Im Unterschied zu umfangreichen Checklisteninstrumenten mit vielen Einzelfragen handelt es sich hier somit um ein relativ kompaktes, verhaltensorientiertes Ratingsystem, das

eine begrenzte Zahl chirurgisch relevanter Human-Factors-Dimensionen strukturiert zusammenfasst.

Die Anwendung des Instruments erfolgte im Rahmen eines Simulated Operating Theatre, dass möglichst realitätsnah als Operationssaal konzipiert worden war. Ein standardisiertes OP-Team aus Anästhesisten, Operating Department Assistant, Scrub Nurse, Circulating Nurse und Assistenzperson war in allen Szenarien präsent. Die chirurgischen Teilnehmer wurden über die Kontrollzone in das Setting eingeführt, über den Fall informiert und führten anschließend einen simulierten Eingriff an einem synthetischen Modell der saphenofemoralen Junktion aus. Während der Simulation war zusätzlich ein vorprogrammiertes Hypoxieszenario integriert, dass gezielt die Aufmerksamkeit des chirurgischen Teilnehmers für den Zustand des Patienten und die Interaktion mit dem Anästhesisten prüfen sollte. Die Bewertung der nicht-technischen Fähigkeiten erfolgte anhand der Videoaufzeichnungen durch zwei unabhängige Beurteiler, nämlich einen Human-Factors-Experten mit Simulationserfahrung aus der Nuklearindustrie und einen Forschungsmitarbeiter, der durch diesen Experten geschult worden war. Um eine einheitliche Anwendung des Instruments sicherzustellen, wurden die ersten fünf Bewertungen gemeinsam durchgeführt; anschließend erfolgte die weitere Bewertung unabhängig voneinander. Die auf diese Weise strukturierte Einführung der Rater diente offensichtlich dazu, die Beurteilungskonsistenz zu verbessern und einen einheitlichen Bewertungsmaßstab zu etablieren.

Hinsichtlich der psychometrischen Eigenschaften berichtet das Dokument für dieses LOSA-basierte Instrument zunächst eine hohe Interrater-Reliabilität. Der Cronbach-Alpha-Koeffizient zwischen den beiden unabhängigen Beobachtern betrug 0,84. Dies wird im Dokument als hohe Übereinstimmung interpretiert und positiv hervorgehoben. Im Diskussionsteil wird diese gute Reliabilität unter anderem damit erklärt, dass die Skala nur eine begrenzte Zahl von Bewertungsdimensionen umfasste und damit die Wahrscheinlichkeit divergierender Einschätzungen reduziert worden sei. In diesem Zusammenhang wird auch auf frühere Beobachtungsstudien in der Anästhesiesimulation verwiesen, in denen eine größere Zahl von Bewertungsmaßen möglicherweise zu geringerer Interrater-Reliabilität beigetragen habe. Für das vorliegende Instrument scheint die kompakte Struktur also ein Vorteil gewesen zu sein.

Weniger überzeugend fallen die Befunde zur Konstruktvalidität aus. Das Dokument prüfte, ob das Instrument zwischen chirurgischen Trainees unterschiedlicher Erfahrungsstufen differenzieren kann. Die Teilnehmer waren in drei Gruppen eingeteilt worden, nämlich Junior-, Intermediate- und Senior-Trainees, definiert über die Anzahl bereits durchgeführter Saphenofemoral-junction-high-tie-Eingriffe. Für den Gesamtwert der nicht-technischen Fähigkeiten ergab

sich jedoch kein signifikanter Unterschied zwischen den drei Erfahrungsgruppen. Auch für die meisten Einzeldimensionen wurden keine signifikanten Unterschiede festgestellt. Lediglich für den Bereich *leadership* zeigte sich ein signifikanter Unterschied zwischen den Gruppen, wobei insbesondere Junior- und Intermediate-Trainees voneinander differenziert werden konnten. Zwischen Intermediate- und Senior-Trainees bestand wiederum kein signifikanter Unterschied. Diese Ergebnisse deuten darauf hin, dass das Instrument in seiner damaligen Form nur eingeschränkt in der Lage war, Erfahrungsunterschiede über die Gesamtheit der nicht-technischen Fähigkeiten hinweg abzubilden.

Das Dokument diskutiert mehrere mögliche Ursachen für diese eingeschränkte Konstruktvalidität. Zum einen wird darauf hingewiesen, dass das gesamte Untersuchungskollektiv ausschließlich aus chirurgischen Trainees bestand. Es sei denkbar, dass deutlich erfahrenere Chirurgen höhere Werte erreicht und dadurch Unterschiede klarer sichtbar gemacht hätten. Zum anderen wird eingeräumt, dass einzelne gewählte Bewertungsdimensionen möglicherweise nur begrenzt zur Chirurgie passen oder die operative Realität nicht hinreichend präzise abbilden. Trotz des vorgelagerten Fragebogens mit chirurgischen Fachpersonen könne daher nicht ausgeschlossen werden, dass die inhaltliche Auswahl der Maße nur eingeschränkt geeignet war. Als dritte Erklärung nennen die Autoren einen eher grundsätzlichen Aspekt: Nicht-technische Fähigkeiten seien in der chirurgischen Ausbildung traditionell nie ein expliziter Fokus gewesen, während technische Fertigkeiten stark trainiert und selektiv gefördert würden. Wenn nicht-technische Kompetenzen nicht systematisch unterrichtet und eingefordert werden, ist es auch schwerer zu erwarten, dass sie entlang klassischer Erfahrungsstufen in ähnlicher Weise ansteigen wie technische Fertigkeiten.

Aufschlussreich sind in diesem Zusammenhang die Befunde zur Gesamtleistung der Kohorte. Über alle Teilnehmer hinweg zeigten sich besonders niedrige Werte in den Bereichen *preoperative preparation* und *vigilance*. Die durchschnittlichen Prozentwerte lagen hier bei 35,8 % beziehungsweise 45,9 %. Auch der Bereich *leadership* erreichte mit 56,6 % nur ein moderates Niveau. Diese Ergebnisse machen deutlich, dass das Instrument trotz begrenzter Differenzierungsfähigkeit über Erfahrungsgruppen hinweg in der Lage war, konkrete Schwächen des Gesamtkollektivs sichtbar zu machen. Im Diskussionsteil wird insbesondere der Befund der niedrigen Vigilanz hervorgehoben. Das Dokument verweist darauf, dass Probleme mit Vigilanz und Situation Awareness in anderen Hochrisikobereichen, etwa in der Anästhesie, eine bedeutsame Rolle bei unerwünschten Ereignissen spielen. Im chirurgischen Bereich sei diese Frage hingegen bislang kaum untersucht worden. Die niedrigen Werte könnten darauf hinweisen, dass viele Operateure ihre Aufmerksamkeit fast ausschließlich auf die operative Aufgabe rich-

ten und den Zustand des Patienten nur unzureichend mitverfolgen oder sich stark darauf verlassen, durch den Anästhesisten informiert zu werden. Dies wird als möglicher kultureller Normeffekt im Operationssaal diskutiert und als wichtiger Ansatzpunkt für weitere Forschung betrachtet.

Ein weiterer Befund betrifft das Verhältnis technischer und nicht-technischer Leistung. Das Dokument berichtet nur eine geringe Korrelation zwischen technischem und nicht-technischem Gesamtscore. Diese niedrige Beziehung wird als Hinweis darauf interpretiert, dass beide Leistungsdimensionen unterschiedliche Kompetenzbereiche repräsentieren. Bemerkenswert ist zudem, dass der Zusammenhang bei den Senior-Trainees tendenziell höher war als bei den weniger erfahrenen Gruppen. Daraus leiten die Autoren ab, dass technische und nicht-technische Kompetenz möglicherweise erst auf höheren Erfahrungsstufen stärker zusammenfallen, während jüngere Trainees in beiden Bereichen sehr unterschiedliche Profile aufweisen können. Zugleich wird hervorgehoben, dass innerhalb aller Gruppen eine erhebliche Leistungsvariabilität bestand und einzelne Senior-Trainees in nicht-technischen Bereichen schwächer abschnitten als manche Junior-Trainees. Dies unterstreicht die Annahme, dass nicht-technische Kompetenz nicht automatisch mit zunehmender Operationszahl ansteigt, sondern durch komplexe Faktoren wie Mentoring, Kultur, Persönlichkeit und Vorbilder beeinflusst wird.

Inhaltlich besonders eng mit dem LOSA-basierten Instrument verknüpft ist die ergänzende Kommunikationsanalyse über die *utterance frequency*. Zwar stellt diese kein Bestandteil des eigentlichen Ratingsystems dar, doch sie ergänzt insbesondere die Kommunikationsdomäne um ein objektives Maß. Das Dokument berichtet eine nur geringe Korrelation zwischen der im nicht-technischen Instrument bewerteten Kommunikationsqualität und der quantitativen utterance frequency. Daraus schließen die Autoren, dass beide Maße unterschiedliche Facetten von Kommunikation erfassen. Während die utterance frequency die Häufigkeit verbaler Austauschakte abbildet, fokussiert das LOSA-basierte Instrument auf die Qualität der Interaktion, etwa Klarheit, Höflichkeit, Bestätigung von Anweisungen und das aktive Einholen von Hilfe. Gerade diese Unterscheidung verdeutlicht eine Stärke des Instruments, da es Kommunikation nicht bloß als Anzahl verbaler Äußerungen versteht, sondern als qualitative Teamkompetenz einordnet.

Trotz dieser Stärken benennt das Dokument mehrere Limitationen des Instruments. Die wichtigste Einschränkung betrifft, wie bereits ausgeführt, die noch unzureichend gesicherte Inhaltsvalidität. Die Auswahl der beobachteten Verhaltensdimensionen erfolgte zwar theoriegeleitet und unter Rückgriff auf eine Expertenbefragung, doch wird selbstkritisch betont, dass ein robusteres Verfahren zur Entwicklung eines solchen Instruments nötig wäre. Zudem bleibt die

Konstruktvalidität beschränkt, da das Instrument mit Ausnahme von Leadership kaum in der Lage war, unterschiedliche Erfahrungsstufen zu differenzieren. Ferner handelt es sich um ein Pilotinstrument in einem sehr spezifischen Kontext, nämlich einer simulierten Operation an einem synthetischen Modell mit standardisiertem Team und integriertem Hypoxieereignis. Die Übertragbarkeit auf andere chirurgische Eingriffe und reale Operationssituationen bleibt offen. Hinzu kommt, dass die Face Validity des Gesamtsettings zwar hoch war, die Realitätsnähe des eingesetzten synthetischen Modells aber nur von etwa der Hälfte der Teilnehmer als überzeugend eingestuft wurde. Auch wenn dies nicht unmittelbar eine Eigenschaft des Instruments selbst ist, beeinflusst die Realitätsnähe des Kontextes die Aussagekraft eines beobachtungsbasierten Verhaltensratings erheblich.

Zusammenfassend lässt sich das im Dokument beschriebene LOSA-basierte Instrument als früher, explorativer Versuch charakterisieren, nicht-technische Fähigkeiten chirurgischer Trainees in einem simulationsgestützten Operationssaal-Setting strukturiert zu erfassen. Es basiert auf ausgewählten LOSA-Elementen, wurde durch chirurgische Experteneinschätzungen ergänzt und umfasst die vier Bereiche präoperative Vorbereitung, Kommunikation und Interaktion, Vigilanz beziehungsweise Situation Awareness sowie Leadership. Die Skala ist kompakt aufgebaut, weist eine gute Interrater-Reliabilität auf und ermöglicht die Sichtbarmachung kollektiver Schwächen, insbesondere in den Bereichen Vorbereitung und Vigilanz. Gleichzeitig bleibt ihre Konstruktvalidität begrenzt, da sie Unterschiede zwischen Erfahrungsstufen nur unzureichend differenzieren konnte. Das Dokument kommt daher zu dem Schluss, dass weitere Forschung notwendig ist, um ein inhaltlich stärker fundiertes und psychometrisch robusteres Instrument für die Erfassung chirurgischer nicht-technischer Fähigkeiten zu entwickeln. Damit besitzt das LOSA-basierte Verfahren vor allem Bedeutung als pilotierter Ausgangspunkt für die spätere Entwicklung strukturierterer behavioral marker systems im chirurgischen Kontext.

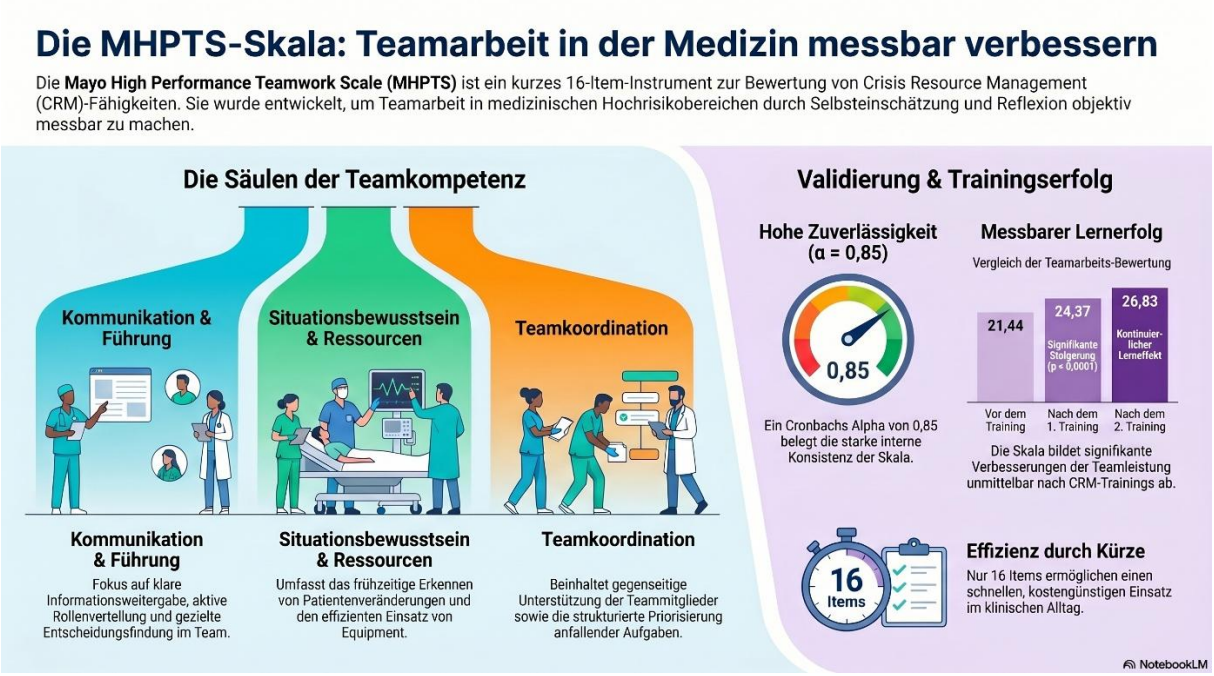
5.14 Mayo High Performance Teamwork Scale (MHPTS)

Quelle: Malec JF, Torsher LC, Dunn WF,Wiegmann DA, Arnold JJ, Brown DA, et al. The mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. Simul Healthc. (2007) 2:4–10.

Vergleich Studie französisch: Émilie Gosselin, Mélanie Marceau, Christian Vincelette, Charles-Olivier Daneau, Stéphan Lavoie & Isabelle Ledoux. (2019). French Translation and Validation of the Mayo High Performance Teamwork Scale for Nursing Students in a High-

Fidelity Simulation Context. Clinical Simulation in Nursing, 30, 25–33. https://doi.org/10.1016/j.ecns.2019.03.002

Abbildung 17: Mayo High Performance Teamwork Scale (MHPTS)



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Émilie Gosselin et al. (2019)

Die *Mayo High Performance Teamwork Scale* (MHPTS) ist ein Instrument zur Erfassung von Teamarbeitskompetenzen in simulationsbasierten Gesundheitskontexten und wurde entwickelt, um hochleistungsbezogene Teamfähigkeiten in kritischen Versorgungssituationen zu beurteilen. Im hochgeladenen Material wird die Skala vor allem über zwei Dokumente erschlossen. Zum einen liegt die tabellarische Originalform der MHPTS vor, zum anderen beschreibt ein Artikel die französisch-kanadische Übersetzung und Validierung der Skala für Pflegestudenten im Kontext hochrealistischer Notfallsimulationen. Aus beiden Dokumenten ergibt sich, dass die MHPTS als CRM-orientiertes Instrument konzipiert wurde, um Teamverhalten in Situationen zu erfassen, in denen Gesundheitsfachpersonen interdisziplinär zusammenarbeiten und unter Zeitdruck präzise und koordiniert handeln müssen. Der Hintergrund dieses Instruments ist die Annahme, dass nicht-technische Fertigkeiten wie Führung, Kommunikation, Rollenklärung und Fehlerprävention für die Patientensicherheit von zentraler Bedeutung sind und daher in der Simulation ebenso beurteilt werden sollten wie technische Fertigkeiten.

Die konzeptuelle Einbettung der MHPTS erfolgt im Dokument über das *crisis resource management* (CRM). Dieses Modell wird als theoretische Grundlage der Skala dargestellt und umfasst mehrere Kernpunkte wirksamen Krisenmanagements, darunter effektive Kommunikation, Situational Awareness, Antizipation und Planung, das Benennen einer Führungsperson, Rollenklärung, Lastenverteilung, Ressourcennutzung und den Einsatz kognitiver Hilfen. Das Dokument führt aus, dass CRM in vielen Ausbildungsprogrammen der Gesundheitsberufe verwendet wird, um Teamleistung zu verbessern und nicht-technische Fehler durch Simulationstraining zu reduzieren. Die MHPTS erscheint vor diesem Hintergrund als ein Instrument, das zentrale CRM-bezogene Teamverhaltensweisen in beobachtbarer Form operationalisiert und damit ein Bindeglied zwischen theoretischem Krisenmanagementmodell und praktischer Teamleistungsbeurteilung darstellt.

Die ursprüngliche Entwicklung der MHPTS wird im französisch-kanadischen Validierungsartikel auf frühere Arbeiten von Malec et al. zurückgeführt. Dort wird berichtet, dass die Skala ursprünglich in einer quasiexperimentellen Studie mit medizinischen Residents und amerikanischen Pflegefachpersonen eingesetzt wurde, die an einem CRM-basierten Trainingsprogramm mit simulierten kritischen Versorgungsszenarien teilnahmen. Bereits in dieser ursprünglichen Untersuchung habe die Skala gute psychometrische Eigenschaften gezeigt, insbesondere eine gute interne Konsistenz mit einem Cronbachs Alpha von 0,85 sowie eine Sensitivität für Veränderung nach Training. Diese Angaben sind für die Bewertung der Skala wesentlich, weil sie darauf hindeuten, dass die MHPTS von Beginn an nicht nur als beschreibendes Instrument, sondern auch als Instrument zur Erfassung trainierbarer Teamarbeitsleistungen konzipiert war.

Die Struktur der MHPTS ist im hochgeladenen Originalbogen vollständig dargestellt. Die Skala umfasst insgesamt 16 Items, die auf einer dreistufigen Antwortskala von 0 bis 2 bewertet werden. Die Antwortkategorien lauten „never or rarely“, „inconsistently“ und „consistently“. Zusätzlich enthält das Instrument die explizite Anweisung, konservativ zu bewerten, da Teams, die noch nicht lange miteinander gearbeitet haben, viele der beschriebenen Qualitäten typischerweise nicht durchgängig zeigen. Dies ist ein bemerkenswertes Merkmal der Skala, da es die Raterhaltung im Sinne einer zurückhaltenden und realistischen Beurteilung steuern soll. Im Dokument wird weiter erläutert, dass die MHPTS aus zwei Sektionen mit jeweils acht Items besteht. Die ersten acht Items müssen immer bewertet werden, während die Items 9 bis 16 bei Bedarf mit „not applicable“ versehen werden können, wenn im Szenario keine Situationen auftraten, die das entsprechende Verhalten erforderten. Diese Zweiteilung ist für die Anwendung besonders bedeutsam, da sie einerseits grundlegende Teamarbeitsaspekte erfasst und

andererseits komplexere, nur unter bestimmten Bedingungen beobachtbare Verhaltensweisen berücksichtigt.

Die ersten acht Items der MHPTS fokussieren auf grundlegende, nahezu in jeder Simulation beobachtbare Aspekte effektiver Teamarbeit. Erfasst wird zunächst, ob von allen Teammitgliedern eine Führungsperson klar erkannt wird und ob diese Führungsperson ein angemessenes Gleichgewicht zwischen Kommandogewalt und Beteiligung der Teammitglieder wahrt. Ein weiterer Schwerpunkt liegt auf der Rollenklarheit im Team sowie auf der wechselseitigen Unterstützung bei der Überwachung relevanter klinischer Indikatoren. Zudem wird erhoben, ob Teammitglieder ihre Handlungen am Patienten laut verbalisieren, ob Anweisungen und Klärungen durch Wiederholung oder Paraphrasierung abgesichert werden, ob etablierte Protokolle und Checklisten genutzt werden und ob alle Teammitglieder sich angemessen beteiligen. Diese erste Hälfte der Skala operationalisiert damit zentrale Basiskomponenten von Teamkoordination, Kommunikation und gemeinsamer Situationsorientierung.

Die zweite Hälfte der MHPTS erfasst demgegenüber komplexere und stärker situationsabhängige Teamreaktionen. Dazu gehören die Bearbeitung von Konflikten und Meinungsverschiedenheiten, ohne dabei die Situation Awareness zu verlieren, eine flexible Umverteilung von Rollen bei dringlichen Ereignissen, das Einfordern von Wiederholung und Klärung bei unklaren Anweisungen sowie der positive Umgang mit Rückmeldungen, die Fehler vermeiden oder eindämmen sollen. Darüber hinaus wird bewertet, ob Teammitglieder auf potenzielle Fehlerquellen aufmerksam machen, auf mögliche Komplikationen mit fehlervermeidenden Maßnahmen reagieren, sicherheitsbezogene Hinweise auch dann weiterverfolgen, wenn sie zunächst keine Reaktion auslösen, und ob Teammitglieder einander bei Überlastung aktiv um Unterstützung bitten. Diese zweite Sektion erweitert die Skala um jene Teamprozesse, die für fortgeschrittenes Krisenmanagement und Fehlerprävention charakteristisch sind, aber nicht in jeder Simulation in gleicher Weise sichtbar werden.

Der Anwendungsbereich der MHPTS ist eng an simulationsbasierte Ausbildungs- und Bewertungssituationen gebunden. Im französisch-kanadischen Validierungsdokument wurde die Skala mit Pflegestudenten in hochrealistischen Notfallsimulationen eingesetzt. Zielpopulation waren Abschlussstudenten pflegewissenschaftlicher Studiengänge, die an zwei hochrealistischen Szenarien, nämlich einem anaphylaktischen Schock und einer symptomatischen Bradykardie, teilnahmen. Eine Woche vor der Simulation erhielten die Studenten eine kurze Einführung in das CRM. Da es sich um Anfänger im Bereich Critical Care handelte, konzentrierte sich diese Einführung auf vier ausgewählte CRM-Elemente, nämlich effektive Kommunikation, Rollenklärung, Leadership und Antizipation beziehungsweise Planung. In den Si-

mulationen konnten die Studenten entweder aktiv an der Versorgung beteiligt sein oder als Beobachter teilnehmen. Nach Abschluss der beiden Szenarien füllten sie die französisch-kanadische Version der MHPTS aus. Das Dokument verdeutlicht damit, dass die Skala sowohl in aktiver Teilnahme als auch in beobachtungsbasierten Settings eingesetzt wurde, auch wenn letzteres nicht dem ursprünglichen primären Verwendungszweck des Instruments entsprach.

Ein zentraler Schwerpunkt des hochgeladenen Dokuments liegt auf der Übersetzung und kulturadaptiven Übertragung der MHPTS ins kanadische Französisch. Dieser Prozess wurde nach der Methodik von Beaton et al. durchgeführt und umfasste vier aufeinanderfolgende Schritte: Übersetzung, Rückübersetzung, Expertenreview und Pretest. In einem ersten Schritt erstellten zwei voneinander unabhängige Übersetzer, deren Erstsprache kanadisches Französisch war, zwei französische Fassungen des Originals. Anschließend wurden diese Versionen von zwei weiteren, vom Original geblindeten Übersetzern wieder zurück ins Englische übertragen. Ziel dieses Rückübersetzungsschrittes war es, Inkonsistenzen und semantische Verschiebungen sichtbar zu machen. Danach wurde ein Expertengremium aus sieben bilingualen Pflegefachpersonen zusammengestellt, die über unterschiedliche Expertisen in Methodologie, Forschung, Simulation, Statistik und Critical Care verfügten. Dieses Gremium verglich Original, Übersetzungen und Rückübersetzungen und erarbeitete durch Konsens eine präfinale Version. Dabei wurde nicht nur auf semantische und idiomatische Äquivalenz geachtet, sondern auch auf kulturelle Passung und konzeptuelle Übereinstimmung mit dem CRM-Modell. Anschließend wurde diese präfinale Fassung mit 44 Pflegestudenten vorgetestet. Insgesamt erwies sich die Verständlichkeit der Items und Antwortkategorien als gut. Lediglich ein Item aus der zweiten Sektion wurde leicht modifiziert, da fast die Hälfte der Befragten Schwierigkeiten hatte, den zugrunde liegenden Begriff sicher zu verstehen. Insgesamt wird im Dokument betont, dass der Übersetzungsprozess nur zu minimalen Änderungen führte und damit für die strukturelle Stabilität der Skala spricht.

Hinsichtlich der psychometrischen Eigenschaften berichtet das Dokument zunächst zur Originalversion zufriedenstellende Kennwerte. Die MHPTS wies in der ursprünglichen Studie eine gute interne Konsistenz auf und war sensibel gegenüber Veränderungen nach CRM-basiertem Training. In der französisch-kanadischen Adaptation wurde vor allem die Reliabilität in Form der internen Konsistenz untersucht. Da viele Teilnehmer die zweite Hälfte der Skala nicht vollständig ausfüllten, wurde die Reliabilitätsanalyse ausschließlich auf die ersten acht Items beschränkt. Für diese ergab sich ein Cronbachs Alpha von 0,74, was im Dokument als akzeptabel interpretiert wird. Die Interitem-Korrelationen reichten von schwach bis moderat, was darauf hindeutet, dass die Items zwar zusammenhängen, aber nicht redundant sind. Das Doku-

ment stellt fest, dass diese ersten acht Items wesentliche Facetten von Teamarbeit im Sinne des CRM abbilden und gemeinsam auf ein gemeinsames Konstrukt verweisen. Gleichwohl lag der Reliabilitätswert, unter den in der Originalversion berichteten Alphas. Als mögliche Erklärung wird genannt, dass die Studenten die Skala nicht nach jedem einzelnen Szenario, sondern erst nach zwei unterschiedlichen Simulationen ausfüllten. Es blieb somit unklar, ob sie sich bei ihrer Beurteilung auf das erste, das zweite oder auf beide Szenarien zugleich bezogen. Dieser Umstand könnte zu zusätzlichen Schwankungen in den Antworten geführt haben.

Neben der Reliabilität liefert das Dokument auch Hinweise zur Inhaltsvalidität der französischen Version. Diese wird nicht über quantitative Kennwerte, sondern qualitativ über den Expertenreview gestützt. Das Expertengremium kam zu dem Schluss, dass die übersetzten Inhalte klar, umfassend und für den vorgesehenen Einsatzkontext geeignet seien. Zusätzlich zeigte der Pretest, dass die überwiegende Mehrheit der Studenten die Items und Antwortoptionen verstand. Der Artikel weist allerdings darauf hin, dass ein Teil der Verständnisschwierigkeiten darauf zurückzuführen war, dass die Pretest-Teilnehmer zum Zeitpunkt der Erhebung noch keine formale Einführung in CRM erhalten hatten. Daher wurden einige Rückmeldungen bewusst nicht in inhaltliche Änderungen übersetzt, um die Äquivalenz zum Original nicht zu gefährden. Die Autoren betonen in diesem Zusammenhang, dass der Pretest zwar keine Konstruktvalidität belegt, aber einen wichtigen Beitrag zum Verständnis leistet, wie die Zielpopulation die Items interpretiert.

Trotz der insgesamt positiven Befunde weist das Dokument auch auf mehrere Limitationen der Skala und ihrer Validierung hin. Eine zentrale Einschränkung besteht darin, dass die zweite Hälfte der Skala in den gewählten Szenarien nur eingeschränkt zur Anwendung kam. Viele der dort beschriebenen Teamverhaltensweisen, etwa Konfliktbearbeitung, Rollenwechsel oder hartnäckiges Nachfassen bei Sicherheitswarnungen, traten in den eher niedrigschwelligen Lernszenarien nicht auf. Deshalb konnten diese Items nicht angemessen in die Reliabilitätsanalyse einbezogen werden. Das Dokument formuliert daraus die Hypothese, dass möglicherweise nur die erste Hälfte der Skala für novice learners uneingeschränkt relevant ist. Eine weitere Limitation betrifft den Umstand, dass sowohl aktive Teilnehmer als auch Beobachter die Skala ausfüllten. Da die MHPTS ursprünglich nicht für Heteroevaluation durch Beobachter entwickelt worden war, bleibt offen, ob beide Gruppen Teamarbeit in gleicher Weise wahrnehmen und beurteilen. Hier sieht das Dokument ausdrücklich weiteren Forschungsbedarf. Hinzu kommt die bereits erwähnte Möglichkeit eines Halo-Effekts, weil die Skala erst nach Abschluss von zwei Simulationen ausgefüllt wurde. Künftige Studien sollten sie deshalb unmittelbar nach jedem einzelnen Szenario einsetzen, um präzisere und weniger überlagerte Einschätzungen

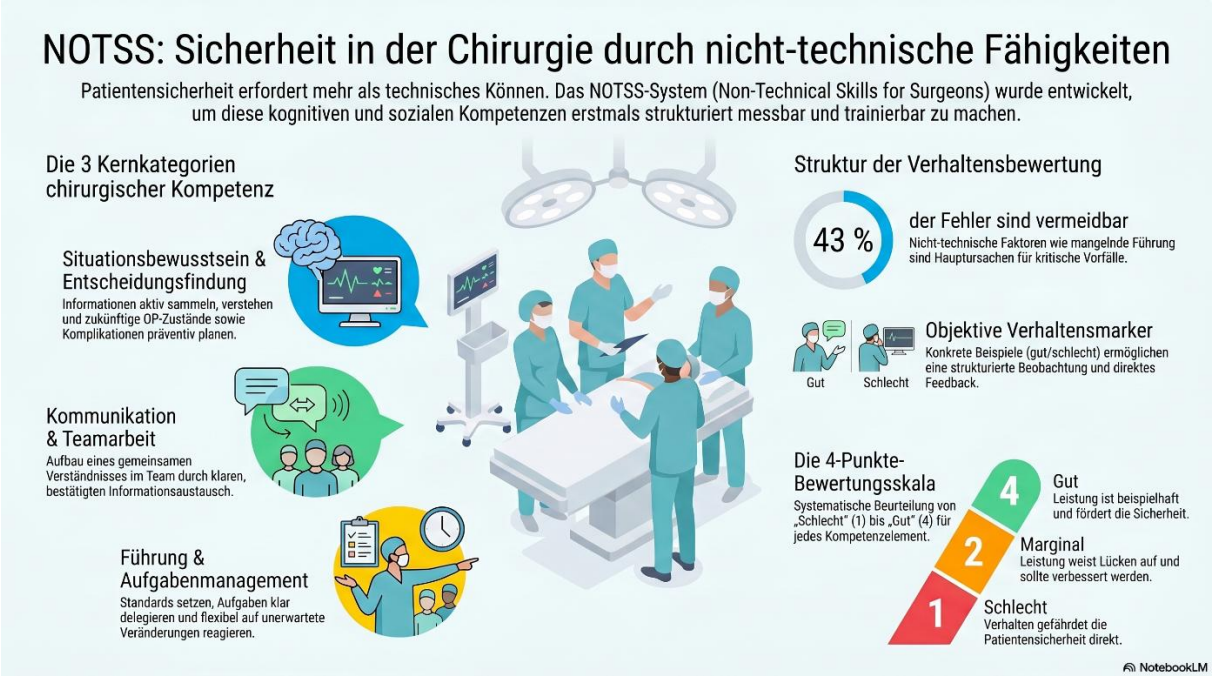
zu erhalten. Schließlich weisen die Autoren darauf hin, dass die französisch-kanadische Version von einem pflegewissenschaftlich geprägten Expertengremium entwickelt wurde. Für die Zielpopulation der Pflegestudenten sei dies zwar angemessen gewesen, doch könnte bei der Anwendung auf andere Gesundheitsberufe eine erneute Anpassung oder Prüfung erforderlich sein.

Zusammenfassend lässt sich die Mayo High Performance Teamwork Scale auf Basis der hochgeladenen Dokumente als ein kompaktes, CRM-orientiertes Instrument zur Erfassung beobachtbarer Teamarbeitsqualitäten in Simulationssituationen charakterisieren. Sie umfasst 16 Items, die über eine dreistufige Antwortskala bewertet werden und sowohl grundlegende als auch komplexere Aspekte effektiver Teamarbeit abbilden. Die ersten acht Items fokussieren auf Leadership, Rollenklarheit, Kommunikation, gemeinsame Überwachung und Beteiligung, während die zweite Hälfte stärker auf Konfliktbearbeitung, Fehlerprävention, flexible Rollenübernahme und Hilfesuche bei Überlastung ausgerichtet ist. Die Originalversion der Skala weist laut Dokument gute psychometrische Eigenschaften auf. Die französisch-kanadische Version wurde in einem methodisch strukturierten Übersetzungs- und Adaptationsprozess entwickelt und zeigt für die erste Hälfte der Skala eine akzeptable interne Konsistenz. Zugleich macht das Dokument deutlich, dass die Anwendbarkeit der MHPTS stark vom Kontext, von der Komplexität der Szenarien und vom Erfahrungsniveau der Zielgruppe abhängt. Für Pflegestudenten in hochrealistischen Simulationssettings stellt die MHPTS-F nach den vorliegenden Befunden ein brauchbares Instrument zur Erfassung von Teamarbeit dar, dessen weitere psychometrische Prüfung insbesondere für die zweite Hälfte der Skala sowie für Beobachterbewertungen jedoch noch aussteht.

5.15 Non-Technical Skills for Surgeons (NOTSS): Entwicklung eines Bewertungssystems für die Chirurgie 2006

Quelle: Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. Med Educ. (2006) 40:1098–104

Abbildung 18: Non-Technical Skills for Surgeons (NOTSS) 2006



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Yule et al. (2006)

Das Instrument **NOTSS** (*Non-Technical Skills for Surgeons*) wurde entwickelt, um die nicht-technischen Fähigkeiten von Chirurgen während der intraoperativen Phase systematisch zu erfassen und für Beobachtung, Beurteilung und Rückmeldung nutzbar zu machen. Ausgangspunkt seiner Entwicklung war die Erkenntnis, dass unerwünschte Ereignisse in der Chirurgie häufig nicht primär auf Defizite technischer Fertigkeiten zurückzuführen sind, sondern auf Probleme im Bereich der kognitiven und interpersonellen Leistung. Im zugrunde liegenden Dokument wird in diesem Zusammenhang hervorgehoben, dass Kommunikationsprobleme in einer Studie bei 43 % der chirurgischen Fehler eine ursächliche Rolle spielten. Vor diesem Hintergrund wird argumentiert, dass technische Fertigkeiten zwar eine notwendige Voraussetzung chirurgischer Kompetenz darstellen, für die Gewährleistung von Patientensicherheit jedoch nicht ausreichen. Vielmehr bedarf es ergänzend eines strukturierten Zugangs zu nicht-technischen Fähigkeiten wie Situationsbewusstsein, Entscheidungsfindung, Aufgabenmanagement, Führung sowie Kommunikation und Teamarbeit.

Das Dokument ordnet NOTSS in die Tradition der Behavioural Marker Systems ein, also jener verhaltensorientierten Beobachtungs- und Bewertungssysteme, die in Hochrisikobereichen entwickelt wurden, um nicht-technische Leistungen strukturiert zu erfassen. Solche Systeme basieren auf Fähigkeits- oder Kompetenz-Taxonomien und übersetzen diese in beobachtbare

Verhaltensmarker, die auf gute oder defizitäre Leistung hinweisen. Nach Darstellung des Artikels müssen derartige Instrumente explizit, transparent, reliabel und valide sein, um für die Beurteilung nicht-technischer Fähigkeiten sinnvoll eingesetzt werden zu können. Vor diesem Hintergrund entstand NOTSS als kontextspezifisches Instrument für die Chirurgie, das sich ausdrücklich auf die intraoperative Phase konzentriert. Zwar wird im Dokument eingeräumt, dass auch perioperative Faktoren einen Einfluss auf das Verhalten im Operationssaal ausüben, der Fokus des Projekts lag jedoch bewusst auf der chirurgischen Leistung während des eigentlichen Eingriffs.

Die Entwicklung des Instruments erfolgte systematisch auf Grundlage von Gordon's Modell des Systemdesigns, das den Weg von der Aufgabenanalyse über die Systementwicklung bis zur Evaluation beschreibt. Eine zentrale empirische Grundlage bildeten kognitive Aufgabenanalysen in Form von Critical-Incident-Interviews mit 27 beratenden Chirurgen aus elf Krankenhäusern in Schottland. Die Stichprobe umfasste Fachärzte aus der Allgemeinchirurgie, der Orthopädie und der Herzchirurgie. Ziel dieser Interviews war es, diejenigen nicht-technischen Fähigkeiten sichtbar zu machen, die Chirurgen in herausfordernden intraoperativen Situationen tatsächlich benötigen. Das Verfahren wurde gewählt, weil es sich nach Darstellung des Dokuments besonders dazu eignet, implizites Expertenwissen zu erschließen, das mit anderen Methoden nur schwer zugänglich ist. Die interviewten Chirurgen wurden gebeten, sich an anspruchsvolle, nicht-routinemäßige Fälle im Operationssaal zu erinnern und diese wiederholt zu rekonstruieren. Dabei wurden sowohl soziale und interpersonelle Aspekte als auch kognitive Prozesse thematisiert. Das im Anhang des Dokuments wiedergegebene Interviewprotokoll zeigt, dass unter anderem nach Führungsrollen, Kommunikationsanforderungen, Teamarbeit mit dem chirurgischen Team und der Anästhesie, Ressourcenmanagement, Zielbildung, Informationsverarbeitung, Entscheidungsstrategien, Situationsverständnis und Zukunftsprojektionen gefragt wurde. Durch diese Anlage des Interviews sollten nicht nur offensichtliche Verhaltensaspekte, sondern auch zugrunde liegende kognitive Prozesse und das „big picture" chirurgischen Handelns zugänglich gemacht werden.

Die Interviews wurden digital aufgezeichnet, vollständig transkribiert und anschließend mit einem line-by-line-coding im Sinne der Grounded Theory analysiert. Dies geschah durch erfahrene Psychologen, die zunächst unabhängig voneinander Teile des Materials kodierten, bis ein akzeptables Maß an Übereinstimmung erreicht war. Aus diesem Prozess entstand zunächst eine Liste von 150 unsortierten nicht-technischen Fähigkeiten. Diese Rohsammlung wurde in einem mehrstufigen Verfahren weiter verdichtet. In der ersten Phase der Systementwicklung reduzierte und präzisierte eine multidisziplinäre Forschungsgruppe die identifizierten

Fähigkeiten. In diesen Prozess flossen zusätzlich die Ergebnisse einer Literaturrecherche, einer Befragung von OP-Personal zu Teamarbeit, Fehlern und Sicherheit sowie Beobachtungen im Operationssaal ein. In einer zweiten Phase wurden die reduzierten Inhalte thematisch geordnet, sodass übergeordnete Kategorien mit zugehörigen Elementen entstehen konnten. Diese Struktur wurde von vier unabhängigen Panels beratender Chirurgen aus vier Krankenhäusern überprüft und sprachlich sowie inhaltlich an die Erfordernisse des chirurgischen Alltags angepasst. In der dritten Phase entwickelten 16 beratende Chirurgen für jedes Element beobachtbare Verhaltensmarker für gute und schlechte Leistung. Diese Marker wurden in mehreren interdisziplinären Review-Sitzungen weiter überarbeitet und konsequent als aktive Verben formuliert, um eine klare Beobachtungs- und Bewertungsgrundlage zu schaffen.

Das Ergebnis dieses Entwicklungsprozesses ist eine dreistufige, hierarchisch aufgebaute Taxonomie, die aus Kategorien, Elementen und Verhaltensmarkern besteht. Dieser Aufbau wurde aus bestehenden Systemen der Anästhesie und der Luftfahrt übernommen, jedoch an die Anforderungen der Chirurgie angepasst. Die Version NOTSS v1.1 umfasst fünf Hauptkategorien mit insgesamt 14 Elementen. Die Kategorie **Situation awareness** beinhaltet die Elemente *Gathering information*, *Understanding information* und *Projecting and anticipating future state*. Die Kategorie **Decision making** besteht aus *Considering options*, *Selecting and communicating option* sowie *Implementing and reviewing decisions*. Unter **Task management** werden *Planning and preparation* sowie *Flexibility/responding to change* zusammengefasst. Die Kategorie **Leadership** umfasst *Setting and maintaining standards*, *Supporting others* und *Coping with pressure*. Die fünfte Kategorie **Communication and teamwork** enthält die Elemente *Exchanging information*, *Establishing a shared understanding* und *Co-ordinating team activities*. Damit integriert das Instrument kognitive, organisatorische, kommunikative und führungsbezogene Dimensionen der chirurgischen Leistung.

Ein zentrales Merkmal von NOTSS ist die Verknüpfung dieser Fähigkeitsbereiche mit konkreten Verhaltensmarkern. Das Dokument erläutert diesen Aspekt exemplarisch anhand der Kategorie *Situation awareness*. Für das Element *Gathering information* gelten etwa das Sicherstellen, dass relevante Untersuchungen wie Bildgebung vorliegen und geprüft wurden, oder die Abstimmung mit dem Anästhesisten über den Anästhesieplan als positive Verhaltensbeispiele. Negativ bewertet würden dagegen verspätetes Erscheinen im Operationssaal oder das Einholen wichtiger Befunde erst im letzten Moment. Für *Understanding information* werden das aktive Betrachten relevanter Befunde und die Diskussion ihrer Bedeutung als gutes Verhalten beschrieben, während das Übersehen wichtiger Ergebnisse oder Fragen, die mangelndes Verständnis erkennen lassen, als problematisch gelten. Für *Projecting and anticipating*

future state gelten vorausschauende Planung, etwa unter Berücksichtigung möglicher Verzögerungen, und das Verbalisieren zukünftiger Bedarfe als positive Marker, wohingegen das Hineingeraten in vorhersehbare Komplikationen ohne frühzeitige Kommunikation oder das Operieren jenseits des eigenen Erfahrungsniveaus als negatives Verhalten gewertet werden. Diese Beispiele verdeutlichen, dass das Instrument nicht mit abstrakten Persönlichkeitsmerkmalen arbeitet, sondern mit spezifischen, kontextbezogenen Verhaltensindikatoren.

Die Konstruktion des Instruments folgte dabei klar definierten Designkriterien. Nach dem Dokument sollten die erfassten Fertigkeiten auf das Verhalten des Chirurgen in der intraoperativen Phase anwendbar sein, aus möglichst spezifischen und beobachtbaren Verhaltensweisen bestehen, kognitive und soziale Fähigkeiten gleichermaßen abbilden und zugleich ökonomisch gestaltet sein. Besonders hervorgehoben wird, dass das Instrument parsimonisch sein und auf eine A4-Seite passen sollte, um im Operationssaal oder in einer hochrealistischen Simulationsumgebung praktisch nutzbar zu bleiben. Zudem sollten die Kategorien und Elemente möglichst trennscharf sein, auch wenn das Dokument anerkennt, dass aufgrund der Interdependenz nicht-technischer Fähigkeiten keine vollständige gegenseitige Exklusivität erreichbar sei. Schließlich wurde großer Wert darauf gelegt, auf psychologischen Fachjargon zu verzichten und stattdessen eine alltagsnahe, domänenspezifische Sprache zu verwenden, die für Chirurgen unmittelbar anschlussfähig ist.

Die Bewertung im Rahmen von NOTSS erfolgt auf einer vierstufigen Ratingskala mit den Ausprägungen *good*, *acceptable*, *marginal* und *poor*. Zusätzlich ist die Kategorie *not observed* vorgesehen. Diese Zusatzkategorie wird genutzt, wenn ein bestimmtes Verhalten in einer gegebenen Situation nicht gezeigt wurde, weil für die entsprechende Fertigkeit keine Anforderung bestand. Falls eine Fertigkeit jedoch hätte demonstriert werden müssen und das Verhalten ausblieb, soll dies ausdrücklich mit *poor* bewertet werden. Gleiches gilt für Verhaltensweisen, die potenziell die Patientensicherheit gefährden. Das Dokument weist außerdem darauf hin, dass es sich in der Anwendung als sinnvoll erweisen könne, zunächst die einzelnen Elemente und erst anschließend die globaleren Kategorien zu bewerten. Die Wahl einer vierstufigen Skala wird damit begründet, dass sie eine feinere Differenzierung als zwei- oder dreistufige Skalen erlaubt und zugleich mit im Vereinigten Königreich etablierten chirurgischen Beurteilungsverfahren kompatibel ist.

Hinsichtlich der psychometrischen Eigenschaften zeigt das Dokument ein differenziertes Bild. Es macht deutlich, dass der vorliegende Beitrag primär die Entwicklung des Instruments beschreibt und noch keine abgeschlossene psychometrische Evaluation vorlegt. Konkrete Kennwerte zu Interrater-Reliabilität, interner Konsistenz oder Kriteriumsvalidität werden nicht be-

richtet. Stattdessen wird mehrfach betont, dass die Reliabilität des Systems zum Zeitpunkt der Veröffentlichung noch anhand standardisierter Szenarien überprüft werde. In der Diskussion wird ausgeführt, dass hierzu standardisierte, in einer simulierten OP-Umgebung gefilmte Szenarien verwendet werden, die von beratenden Chirurgen mit Hilfe des Ratingsystems analysiert werden. Erst wenn sich das Instrument in dieser experimentellen Evaluation als reliabel und reproduzierbar erweise, solle es im nächsten Schritt in einer realen OP-Umgebung erprobt werden. Insofern erlaubt das Dokument keine Aussage über bereits nachgewiesene Reliabilitätskennwerte, sondern lediglich über den vorgesehenen Prüfpfad.

Dennoch liefert der Artikel deutliche Hinweise auf die angestrebte Inhalts- und Konstruktvalidität des Systems. Erstens ist die Taxonomie empirisch in der chirurgischen Praxis verankert, da sie aus Interviews mit erfahrenen Chirurgen sowie aus ergänzenden Datenquellen wie Literaturreview, Beobachtungen im OP und Einstellungen von Theaterpersonal zu Teamarbeit und Sicherheit abgeleitet wurde. Zweitens wurden Domänenexperten auf allen Entwicklungsebenen einbezogen, was die fachliche Passung und Relevanz der Kategorien und Marker absichern soll. Drittens wird im Fazit ausdrücklich betont, dass das System mit dem Ziel entwickelt wurde, explizit und transparent zu sein und ein akzeptables Maß an Konstruktvalidität zu besitzen. Die inhaltliche Nähe zur chirurgischen Praxis, die alltagsnahe Sprache sowie die Orientierung an direkt beobachtbaren oder aus Kommunikation inferierbaren Verhaltensweisen sprechen ebenfalls für eine hohe Inhaltsvalidität des Instruments. Gleichwohl bleibt festzuhalten, dass die formale psychometrische Prüfung im engeren Sinne im hier vorliegenden Dokument noch nicht abgeschlossen ist.

Aus dem Dokument ergeben sich mehrere potenzielle Anwendungsbereiche von NOTSS. Zunächst soll das Instrument dazu dienen, Beobachtungen im Operationssaal systematisch zu strukturieren und Rückmeldungen an Weiterbildungsassistenten wie auch an erfahrene Chirurgen zu standardisieren. Darüber hinaus wird NOTSS als Grundlage für formative Beurteilungen verstanden, bei denen Stärken und Entwicklungsbedarfe in nicht-technischen Fähigkeiten sichtbar gemacht werden können. Ferner kann das Instrument dazu beitragen, Trainingsbedarfe zu identifizieren und nicht-technische Fertigkeiten explizit in die chirurgische Aus- und Weiterbildung zu integrieren. In diesem Zusammenhang verweist das Dokument auf bereits eingeführte oder begonnene Schulungsangebote der Royal Colleges of Surgeons zu Crew Resource Management und nicht-technischen Fertigkeiten. Darüber hinaus wird die Entwicklung eines standardisierten Beobachtungssystems als Forschungsfortschritt bewertet, weil sie neue Möglichkeiten eröffnet, intraoperative Leistung systematisch mit Sicherheits- und Qualitätsfragen zu verknüpfen.

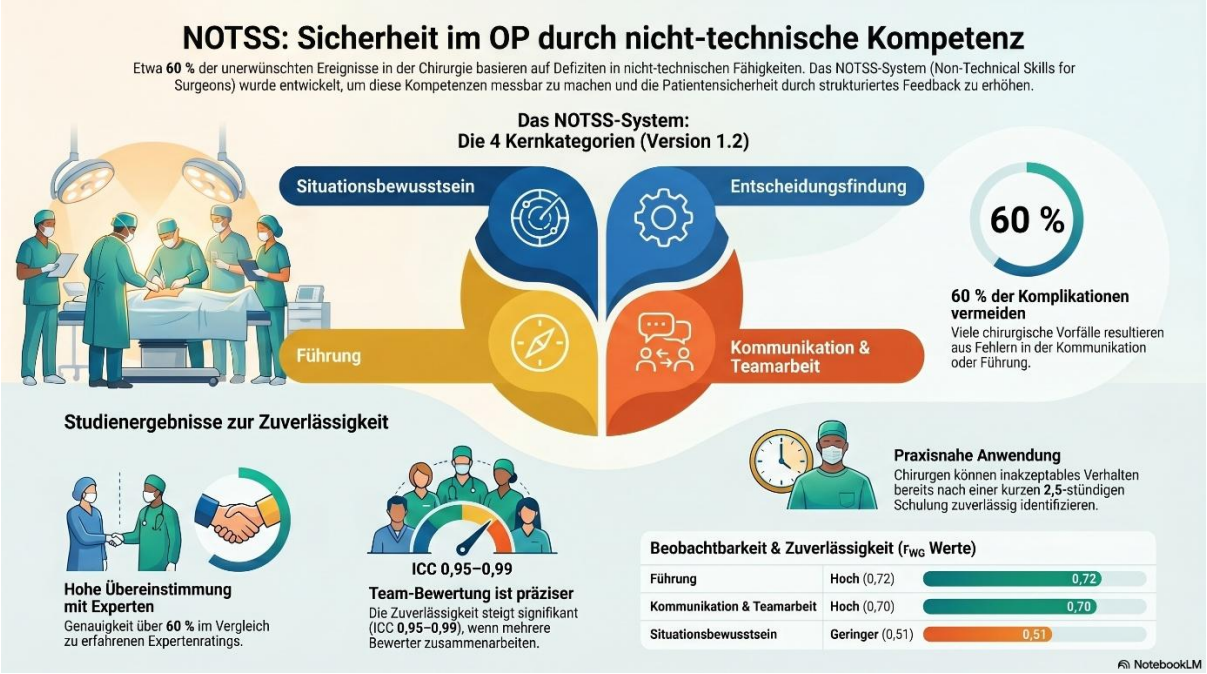
Neben diesen Potenzialen benennt das Dokument auch mehrere Begrenzungen. Eine erste Limitation liegt in der bewussten Konzentration auf die intraoperative Phase. Damit wird zwar ein klar umrissener und hochrelevanter Handlungsbereich fokussiert, gleichzeitig bleiben perioperative Einflussfaktoren außen vor, obwohl diese das intraoperative Verhalten mitprägen können. Eine zweite Limitation besteht darin, dass das Dokument vor allem den Entwicklungsprozess und nicht die abgeschlossene Evaluation des Systems beschreibt. Aussagen zu Reliabilität, Usability und Kriteriumsvalidität bleiben daher vorläufig. Drittens weist die Anlage des Instruments auf eine grundsätzliche methodische Schwierigkeit hin, nämlich die Erfassung kognitiver Fähigkeiten wie Situationsbewusstsein oder Entscheidungsfindung, die häufig nicht direkt beobachtbar sind, sondern nur über kommunikative Hinweise oder indirekte Verhaltensindikatoren erschlossen werden können. Schließlich wird deutlich, dass die praktische Nutzung des Instruments eine angemessene Schulung der Anwender voraussetzen wird, da nur so eine konsistente und sachgerechte Bewertung gewährleistet werden kann.

Zusammenfassend lässt sich festhalten, dass NOTSS im vorliegenden Dokument als systematisch entwickeltes, empirisch fundiertes und domänenspezifisches Verhaltensmarkierungssystem für die intraoperativen nicht-technischen Fähigkeiten von Chirurgen beschrieben wird. Das Instrument basiert auf einer mehrstufigen Entwicklungslogik, die kognitive Aufgabenanalyse, qualitative Interviewauswertung, Expertenrevision und die Formulierung konkreter Verhaltensmarker miteinander verbindet. Seine Struktur umfasst fünf Kategorien, 14 Elemente und eine vierstufige Ratingskala mit zusätzlicher Option „not observed“. Besonders hervorzuheben sind die starke Verankerung in der chirurgischen Praxis, die Orientierung an beobachtbarem Verhalten, die ökonomische und sprachlich alltagsnahe Gestaltung sowie die klare Ausrichtung auf Beobachtung, Feedback und Training. Gleichzeitig ist zu betonen, dass das Dokument noch keine vollständige psychometrische Absicherung des Instruments berichtet. NOTSS erscheint damit als konzeptionell überzeugendes und für Ausbildung, Feedback und Forschung sehr relevantes Instrument, dessen endgültige psychometrische Qualität jedoch erst durch nachfolgende Evaluationsstudien gesichert werden muss.

5.16 Non-technical Skills for Surgeons (NOTSS) 2008

Quelle: Yule S, Flin R, Maran N, Rowley D, Youngson G, Paterson-Brown S. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. World J Surg. (2008) 32:548–56.

Abbildung 19: Non-technical Skills for Surgeons (NOTSS) 2008



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Yule et al. (2008)

Das Instrument „Non-technical Skills for Surgeons“ (NOTSS) wurde zur systematischen Erfassung nichttechnischer Kompetenzen von Chirurgen im Operationssaal entwickelt. Die Entwicklung des Verfahrens basiert auf der Annahme, dass chirurgische Versorgungsqualität nicht ausschließlich durch technisches Können bestimmt wird, sondern in erheblichem Maß auch von kognitiven und interpersonellen Fähigkeiten abhängt. Im zugrunde liegenden Beitrag wird ausgeführt, dass Defizite in Bereichen wie Kommunikation, Teamarbeit, Führung, Situationsbewusstsein und Entscheidungsfindung mit Fehlern, ungünstigen Outcomes und Sicherheitsrisiken in der Chirurgie assoziiert sind. Vor diesem Hintergrund wurde NOTSS als verhaltensorientiertes Bewertungssystem konzipiert, das beobachtbare nichttechnische Verhaltensweisen im intraoperativen Setting strukturiert erfassen und beurteilbar machen soll.

Die Entwicklung des Instruments erfolgte auf Grundlage mehrerer Methoden der Aufgabenanalyse unter Einbezug erfahrener Konsiliar- beziehungsweise Consultant-Chirurgen. Ziel war es, jene nichttechnischen Fertigkeiten zu identifizieren, die für eine sichere chirurgische Praxis als zentral erachtet werden. Das daraus hervorgegangene System umfasste zunächst fünf Kategorien, nämlich Situation Awareness, Decision Making, Task Management, Communication & Teamwork sowie Leadership. Diese fünf Hauptkategorien wurden in insgesamt 14 Elemente untergliedert, denen jeweils beispielhafte gute und schlechte Verhaltensweisen zuge-

ordnet waren. Die Autoren führen aus, dass die inhaltliche Validität des Instruments aus diesem systematischen Entwicklungsprozess mit Fachexperten abgeleitet werden kann. NOTSS wurde somit von Beginn an als taxonomisch fundiertes Beobachtungs- und Bewertungssystem verstanden, das nichttechnische Kompetenzen über konkrete Verhaltensindikatoren operationalisiert.

In seiner ursprünglichen Struktur erfasste die Kategorie Situation Awareness die Elemente „Gathering information“, „Understanding information“ und „Projecting and anticipating future state“. Damit wurden die Aufnahme relevanter Informationen, deren Interpretation sowie die antizipierende Einschätzung des weiteren Verlaufs abgebildet. Die Kategorie Decision Making setzte sich aus den Elementen „Considering options“, „Selecting and communicating options“ sowie „Implementing and reviewing decisions“ zusammen und bezog sich damit auf das Abwägen von Handlungsalternativen, die Auswahl und Kommunikation einer Entscheidung sowie deren Umsetzung und nachfolgende Überprüfung. Task Management umfasste „Planning and preparation“ sowie „Flexibility/responding to change“ und sollte planerische sowie adaptive Aspekte des Handelns erfassen. Leadership beinhaltete die Elemente „Setting and maintaining standards“, „Supporting others“ und „Coping with pressure“, während Communication & Teamwork über die Elemente „Exchanging information“, „Establishing a shared understanding“ und „Co-ordinating team activities“ strukturiert wurde. Das Instrument war damit so angelegt, dass sowohl kognitive als auch soziale und führungsbezogene Dimensionen chirurgischer Leistung differenziert beobachtet werden konnten.

Für die Bewertung wurde eine vierstufige Ratingskala verwendet. Die Ausprägungen reichten von 4 für gutes Verhalten über 3 für akzeptables Verhalten und 2 für grenzwertiges Verhalten bis 1 für schlechtes Verhalten. Zusätzlich stand die Kategorie „N/A“ zur Verfügung, wenn eine bestimmte Fertigkeit in der vorliegenden klinischen Situation nicht erforderlich oder nicht erwartbar war. Die Beurteilung konnte sowohl auf Ebene der Oberkategorien als auch auf Ebene der zugehörigen Elemente erfolgen. Dies ermöglichte einerseits globale Einschätzungen über zentrale Kompetenzbereiche und andererseits eine feinere Differenzierung einzelner beobachtbarer Verhaltensaspekte.

Zur empirischen Prüfung des Instruments wurden standardisierte videobasierte Operationssaalszenarien eingesetzt. Im Dokument wird beschrieben, dass insgesamt elf Szenarien mit simulierten, aber realitätsnah gestalteten klinischen Situationen erstellt wurden, von denen sechs für die eigentliche Evaluation ausgewählt wurden. Diese Szenarien waren zwischen 2:30 und 5:40 Minuten lang und deckten sowohl allgemein- als auch orthopädisch-chirurgische Situationen ab. Drei weitere Szenarien dienten dem Training der Beurteiler. Die Szenarien

wurden in Operationssälen unter Einsatz eines Patientensimulators sowie mit praktizierenden Chirurgen, Anästhesisten und Pflegekräften in den jeweiligen Rollen gefilmt. Durch diese Gestaltung sollte gewährleistet werden, dass die Reliabilität des Systems anhand verschiedener klinischer Konstellationen geprüft werden konnte.

Die Stichprobe bestand aus 44 Consultant-Chirurgen aus fünf schottischen Krankenhäusern, die an insgesamt sechs experimentellen Sitzungen teilnahmen. Der größte Teil der Teilnehmer stammte aus der Allgemein- und Orthopädiechirurgie; daneben waren auch einzelne Vertreter anderer chirurgischer Disziplinen beteiligt. Die durchschnittliche Erfahrung auf Consultant-Ebene betrug 8,9 Jahre. Obwohl ein Großteil der Teilnehmer angab, an der Beurteilung chirurgischer Weiterzubildender beteiligt zu sein, hatte nur ein Teil eine formale Schulung im Bereich der Leistungsbeurteilung erhalten. Vor der eigentlichen Evaluation absolvierten alle Teilnehmer ein 2,5-stündiges Training zum NOTSS-System. Dieses umfasste Grundlagen zu Human Factors und nichttechnischen Fertigkeiten, eine Einführung in das System und die Verhaltensbeurteilung sowie praktische Übungen anhand von drei Trainingsszenarien. Eine formale Kalibrierung der Beurteiler fand jedoch nicht statt; stattdessen wurden die Beobachtungen und Ratings nach den Übungsszenarien gemeinsam besprochen.

Die psychometrische Analyse konzentrierte sich auf die Bereiche Sensitivität, Interrater-Reliabilität und interne Struktur. Zur Prüfung der Sensitivität wurden die Urteile der Teilnehmer mit sogenannten Referenzratings verglichen. Diese Referenzbewertungen wurden von den Szenarioentwicklern vergeben, die zugleich als besonders erfahrene Fachpersonen in der Verhaltensbeobachtung sowie in der Bewertung technischer und nichttechnischer Leistungen beschrieben werden. Die Sensitivität wurde als mittlere absolute Differenz zwischen den von den Teilnehmern vergebenen Kategorienratings und den Referenzratings berechnet. Niedrigere Werte standen dabei für eine höhere Genauigkeit. Im Mittel lag die Sensitivität des Systems auf Kategorienebene bei 0,67 Skalenpunkten. Die höchste Sensitivität zeigte sich für die Kategorie Task Management, während Situation Awareness die geringste Sensitivität aufwies. Decision Making wurde sensibler bewertet als Situation Awareness, und die Kategorien Leadership sowie Communication & Teamwork lagen in einem ähnlichen Bereich mit einer mittleren Abweichung von etwa 0,8 Skalenpunkten von den Referenzurteilen.

Darüber hinaus wurde untersucht, inwieweit die Teilnehmer in der Lage waren, zwischen akzeptablem und inakzeptablem Verhalten zu unterscheiden. Hierzu wurden die Skalenwerte 1 und 2 als „unakzeptabel“ und die Werte 3 und 4 als „akzeptabel“ zusammengefasst. Die Ergebnisse zeigten, dass die Übereinstimmung mit den Referenzbewertungen in allen Kategorien über 60 % lag. Besonders hoch war sie in den Bereichen Decision Making und Commu-

nication & Teamwork, in denen jeweils durchschnittlich 82 % der Teilnehmer mit den Referenzratings übereinstimmten. Für Leadership lag die Übereinstimmung bei 69 %, während Situation Awareness und Task Management jeweils von 63 % der Rater in Übereinstimmung mit der Experteneinschätzung beurteilt wurden. Diese Befunde deuten darauf hin, dass das Instrument insbesondere auf einer dichotomen Ebene, also bei der Unterscheidung zwischen akzeptabler und nicht akzeptabler Leistung, bereits mit begrenztem Training eine brauchbare Sensitivität aufweist.

Die Interrater-Reliabilität wurde mit dem Within-group-agreement-Koeffizienten rwg sowie ergänzend mit Intraklassenkorrelationskoeffizienten des Typs ICC(2) untersucht. Für den rwg wurde im Beitrag ein akzeptabler Bereich von mehr als 0,7 bis 0,8 zugrunde gelegt. Auf Kategorienebene erreichten lediglich Leadership mit einem mittleren rwg von 0,72 und Communication & Teamwork mit 0,70 diesen Bereich. Decision Making lag mit 0,68 knapp darunter, Task Management mit 0,66 ebenfalls unterhalb des Kriteriums, und Situation Awareness zeigte mit 0,51 die geringste Übereinstimmung. Auf Ebene der einzelnen Elemente überschritt nur das Leadership-Element „Supporting others“ mit einem mittleren rwg von 0,74 den akzeptablen Schwellenwert. Die übrigen Elemente blieben teilweise deutlich darunter, was auf begrenzte Übereinstimmung zwischen den einzelnen Beurteiler hinweist.

Die ergänzend berechneten Intraklassenkorrelationskoeffizienten zeigten ein differenziertes Bild. Die ICC(2)-Werte für Einzelbeurteilungen erreichten in keiner Kategorie das im Beitrag genannte Kriterium von mehr als 0,7. Die höchsten Werte fanden sich für Leadership mit 0,66, Communication & Teamwork mit 0,63 und Decision Making mit 0,60. Besonders niedrig lagen die Werte für Situation Awareness mit 0,29 und Task Management mit 0,39. Demgegenüber fielen die ICC(2)-Werte für gemittelte Beurteilungen über alle Kategorien hinweg sehr hoch aus und lagen zwischen 0,95 und 0,99. Dies spricht dafür, dass das Instrument bei Aggregation mehrerer Urteile eine hohe Zuverlässigkeit aufweisen kann, während Einzelbewertungen deutlich variabler ausfallen. Diese Ergebnisse sind insbesondere für den praktischen Einsatz relevant, da sie nahelegen, dass NOTSS eher in Kontexten mit mehreren Beobachtern oder wiederholten Beobachtungen robust eingesetzt werden, kann als isoliertes Einzelrating.

Die Analyse der internen Struktur des Instruments ergab eine hohe Konsistenz zwischen den Bewertungen auf Kategorien- und Elementebene. Da die Elemente konzeptionell den übergeordneten Kategorien zugeordnet sind, wurde erwartet, dass die jeweiligen Ratings eng miteinander korrespondieren. Diese Annahme wurde geprüft, indem die mittlere absolute Differenz zwischen den Elementbewertungen und dem zugehörigen Kategorienrating berechnet wurde. Für alle Kategorien zeigte sich eine sehr geringe mittlere Differenz von weniger als 0,25 Ska-

lenpunkten auf der vierstufigen Skala. Damit weist das Instrument nach den Angaben des Artikels eine konsistente interne Struktur auf. Die enge Beziehung zwischen Kategorien und Elementen kann als Hinweis darauf verstanden werden, dass die strukturelle Logik des Instruments nachvollziehbar ist und die untergeordneten Verhaltensaspekte die übergeordneten Kompetenzbereiche kohärent repräsentieren.

Die Ergebnisse zeigten darüber hinaus, dass die Interrater-Reliabilität zwischen den einzelnen Szenarien variierte. Eine höhere Übereinstimmung wurde für die Szenarien 1 bis 3 berichtet als für die Szenarien 4 bis 6. Im Beitrag wird dies damit erklärt, dass bestimmte Verhaltensweisen in einzelnen Szenarien möglicherweise extremer dargestellt waren und dadurch leichter identifiziert und bewertet werden konnten. Zudem wurde untersucht, ob die fachliche Spezialisierung der Beurteiler einen Einfluss auf die Übereinstimmung hatte. Dabei zeigte sich, dass orthopädische Chirurgen signifikant homogener urteilten als Allgemeinchirurgen, und zwar unabhängig davon, ob im jeweiligen Szenario allgemeinchirurgische oder orthopädische Situationen dargestellt wurden. Dieser Befund deutet darauf hin, dass nicht nur das Instrument selbst, sondern auch Merkmale der Beurteiler Einfluss auf die Reliabilität der Bewertungen ausüben.

Auf Grundlage der Ergebnisse wurde das Instrument weiterentwickelt. Insbesondere wurde die Kategorie Task Management aus der ursprünglichen Taxonomie entfernt und relevante Verhaltensweisen wurden, soweit angemessen, in andere Kategorien integriert. Diese Entscheidung wurde damit begründet, dass die Interrater-Reliabilität für Task Management vergleichsweise schwach ausfiel, dass Teilnehmer angaben, entsprechende Verhaltensweisen intraoperativ häufig nicht sichtbar auszuüben, da viele dieser Aufgaben bereits präoperativ delegiert würden, und dass mehrere Inhalte dieser Kategorie konzeptionell ohnehin Aspekte des Situationsbewusstseins widerspiegelten. Darüber hinaus wurde auf Rückmeldungen verwiesen, wonach die Verringerung der Kategorienzahl das Instrument übersichtlicher und ökonomischer mache und die kognitive Belastung der Nutzer reduziere. Die revidierte Version NOTSS v1.2 umfasste daher nur noch vier Kategorien, nämlich Situation Awareness, Decision Making, Communication & Teamwork sowie Leadership. Die zugehörigen Elemente wurden entsprechend angepasst und auf zwölf reduziert.

Hinsichtlich seiner Anwendungsbereiche wurde NOTSS primär als pädagogisches Instrument konzipiert. Es soll Chirurgen eine Struktur und eine geeignete Fachsprache bereitstellen, um Verhalten während routinemäßiger operativer Eingriffe zu beobachten, zu bewerten und zu besprechen. Der Fokus liegt somit auf formativer Beurteilung, Feedback und Training. Im Artikel wird darüber hinaus berichtet, dass sich weitere Einsatzmöglichkeiten ergeben haben, da-

runter der Umgang mit leistungsschwachen Chirurgen, die Analyse chirurgischer Zwischenfälle sowie die Strukturierung nichttechnischer Fertigkeitstrainings. Langfristig wird das Ziel formuliert, NOTSS in die Curricula aller chirurgischen Fachgebiete zu integrieren. Zugleich wird betont, dass eine Anpassung an bestehende Formate zur Bewertung technischer Fertigkeiten sinnvoll sein könnte, um den Aufwand zusätzlicher Schulungs- und Bewertungsverfahren im klinischen Alltag zu begrenzen. Die Studienergebnisse legen nahe, dass Chirurgen mit minimalem Training in der Lage sind, das Instrument zumindest für eine formative Einschätzung auf der Ebene „akzeptabel“ versus „unakzeptabel“ einzusetzen.

Trotz der grundsätzlich positiven Befunde weist die Studie mehrere Limitationen auf. Als zentrale Einschränkung benennen die Autoren die geringe Dauer der Rater-Schulung. Für die Anwendung verhaltensbasierter Bewertungssysteme dieser Art wird im Beitrag eine Schulungsdauer von mindestens zwei Tagen als empfehlenswert genannt, während im vorliegenden Projekt lediglich 2,5 Stunden realisiert werden konnten. Hinzu kommt, dass die Beurteiler nicht formal kalibriert wurden, sodass keine gemeinsamen Bewertungsstandards etabliert wurden. Dies dürfte wesentlich zur beobachteten Variabilität der Urteile beigetragen haben. Eine weitere Limitation betrifft die Verwendung kurzer, videobasierter Szenarien. Obwohl diese als realistisch und klinisch plausibel beschrieben werden, bilden sie nicht die Komplexität längerer intraoperativer Beobachtungen ab, bei denen Beobachter über einen längeren Zeitraum zusätzliche Kontextinformationen sammeln können. Im Artikel wird zudem darauf hingewiesen, dass manche Rater sich mehr Hintergrundinformationen gewünscht hätten, etwa zum Kompetenzniveau des assistierenden Personals, um das Verhalten des konsultierenden Chirurgen angemessener bewerten zu können. Schließlich wird hervorgehoben, dass die chirurgische Praxis im Vergleich zu stark standardisierten Hochrisikobranchen wie Luftfahrt oder Kernenergie größere Interpretationsspielräume zulässt, was die Herstellung hoher Übereinstimmung zwischen Beurteiler zusätzlich erschweren kann.

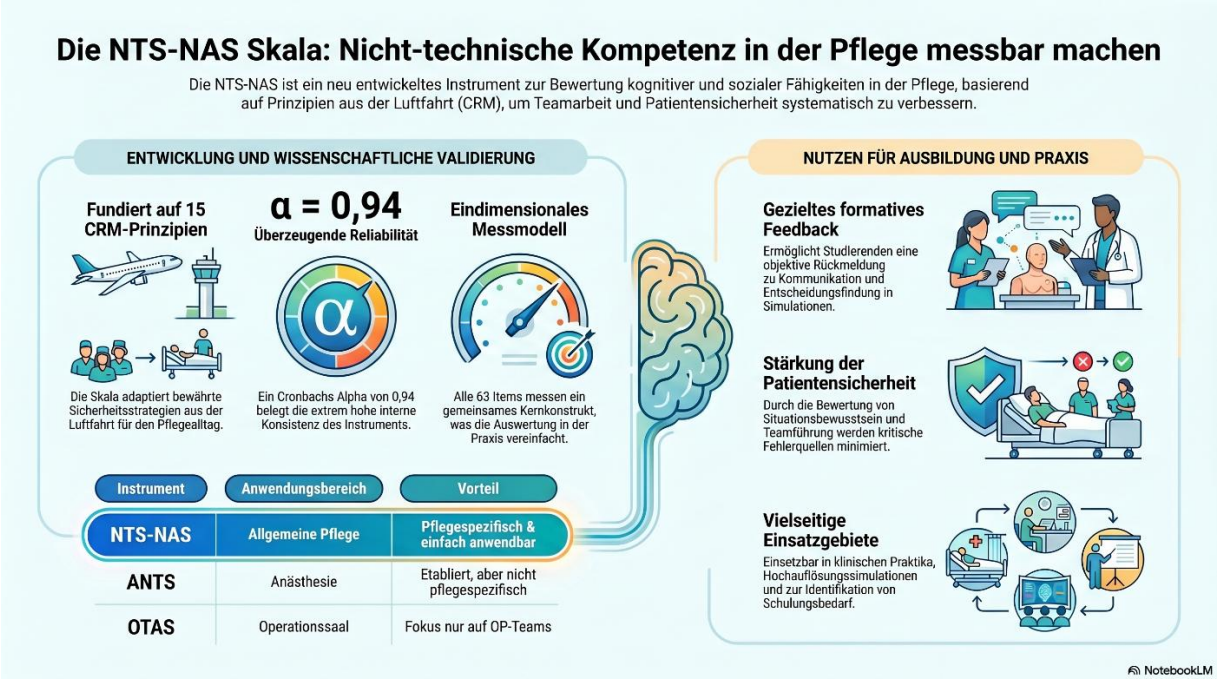
Als Stärke der Untersuchung ist hervorzuheben, dass das Instrument nicht anhand eines einzelnen Falls, sondern über mehrere unterschiedlichen Szenarien hinweg geprüft wurde. Die Verwendung standardisierter Videos ermöglichte stabile Beobachtungsbedingungen und eine systematische Überprüfung der Bewertungen über verschiedene klinische Konstellationen hinweg. Dadurch konnte gezeigt werden, dass das Instrument grundsätzlich in unterschiedlichen Operationssaalsituationen anwendbar ist. Gleichwohl leiten die Autoren aus ihren Ergebnissen ab, dass vor einem breiten Einsatz eine formale Usability-Prüfung im realen Operationssaal erforderlich ist.

Zusammenfassend lässt sich festhalten, dass NOTSS ein systematisch entwickeltes und inhaltlich fundiertes Instrument zur Erfassung nichttechnischer Kompetenzen in der Chirurgie darstellt. Die Struktur des Instruments ist klar taxonomisch aufgebaut, die interne Konsistenz zwischen Kategorien und Elementen fällt hoch aus, und die Sensitivität gegenüber akzeptablem beziehungsweise inakzeptablem Verhalten ist insgesamt zufriedenstellend. Die Reliabilitätsbefunde zeigen jedoch zugleich, dass insbesondere auf Ebene einzelner Beurteilungen und bei bestimmten kognitiven Domänen, vor allem Situation Awareness, noch Einschränkungen bestehen. Die Ergebnisse sprechen dafür, dass NOTSS vor allem bei ausreichender Schulung, zunehmender Vertrautheit der Anwender und gegebenenfalls unter Einbezug mehrerer Beurteilungen ein vielversprechendes Instrument für Training, Beobachtung und formative Leistungsrückmeldung im chirurgischen Kontext sein kann.

5.17 Non-Technical Skills – Nursing Assessment Scale (NTS-NAS)

Quelle: Pires SMP, Monteiro SOM, Pereira AMS, Stocker JNM, Chaló DM, Melo EMOP. Non-technical skills assessment scale in nursing: construction, development and validation. Rev Lat Am Enfermagem. (2018) 26:e3042. doi: 10. 1590/1518-8345.2383.3042

Abbildung 20: Non-Technical Skills – Nursing Assessment Scale (NTS-NAS)



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Pires et al. (2018)

Die Non-Technical Skills – Nursing Assessment Scale (NTS-NAS) wurde mit dem Ziel entwickelt, ein spezifisches Instrument zur Erfassung nichttechnischer Fertigkeiten im pflegerischen Ausbildungskontext bereitzustellen. Ausgangspunkt der Instrumentenentwicklung war die im zugrunde liegenden Beitrag formulierte Feststellung, dass nichttechnische Kompetenzen wie Kommunikation, Teamarbeit, Führung, Entscheidungsfindung und Situationsbewusstsein für die Sicherheit von Patienten sowie für erfolgreiches klinisches Handeln von zentraler Bedeutung sind, jedoch bislang kein theoretisch fundiertes und zugleich leicht anwendbares Verfahren zur spezifischen Beurteilung dieser Kompetenzen in der Pflegeausbildung vorlag. Vor diesem Hintergrund wurde die NTS-NAS als methodisches Forschungsprojekt konstruiert, entwickelt und validiert, um nichttechnische Fertigkeiten von Pflegestudenten systematisch erfassen zu können.

Die konzeptionelle Grundlage des Instruments bildet das aus der Luftfahrt stammende Konzept des Crew Resource Management, das im Gesundheitswesen als Crisis Resource Management adaptiert wurde. Dieses Modell dient der simulationsbasierten Schulung nichttechnischer Fertigkeiten und basiert auf 15 Handlungsprinzipien, darunter das Kennen der Umgebung, Antizipation und Planung, frühes Einholen von Hilfe, Führung und Followership, Arbeits-

verteilung, effektive Kommunikation, Nutzung verfügbarer Informationen, Fehlervermeidung, Gegenkontrolle, der Einsatz kognitiver Hilfsmittel, wiederholte Reevaluation, Teamarbeit, Aufmerksamkeitssteuerung und dynamische Prioritätensetzung. Auf Grundlage einer Literaturübersicht, der fachlichen Erfahrung des Forschungsteams im Bereich nichttechnischer Fertigkeiten im Gesundheitswesen sowie des Wissens über die Prinzipien des Crisis Resource Management wurde zunächst ein umfangreicher Itempool erstellt. An diesem Entwicklungsprozess war ein interprofessionelles Team beteiligt, das sich aus Pflegefachpersonen, Pflegelehrern, einem Anästhesisten sowie drei Psychologen zusammensetzte. Für jedes der 15 CRM-Prinzipien wurden Aussagen formuliert, die die jeweilige Kompetenz im pflegerischen Kontext abbilden sollten. Dieser erste Entwicklungsschritt führte zu einer Liste von 64 Items mit einer fünfstufigen Likert-Skala von „totally disagree“ bis „totally agree“ sowie der zusätzlichen Antwortmöglichkeit „not applicable“.

Im Anschluss daran wurde die erste Instrumentenversion einer inhaltlichen Prüfung durch ein Expertengremium unterzogen. Drei Pflegeexperten diskutierten gemeinsam mit dem Forschungsteam sämtliche Items im Hinblick auf ihre Klarheit, Repräsentativität, Beobachtbarkeit, Verständlichkeit und Passung zu den Kompetenzen von Pflegestudenten. Ziel dieses Schrittes war die Sicherung der Inhaltsvalidität des Instruments. Darüber hinaus prüfte das Panel, ob die Aussagen für hoch- und niedrig-fidelische klinische Simulationskontexte geeignet sind. Im Verlauf dieses Überarbeitungsprozesses wurden einzelne Formulierungen verändert, Items gestrichen, neue Items aufgenommen und einzelne Aussagen anderen Kompetenzbereichen zugeordnet. Besonders bedeutsam war die Entscheidung, das CRM-Prinzip „Mobilize all available resources“ aus dem Instrument zu entfernen. Dies wurde damit begründet, dass dieser Bereich im pflegerischen Ausbildungskontext schwer zu messen sei und Pflegestudenten noch nicht über die notwendige Autonomie verfügten, um diese Kompetenz in vollem Umfang auszuüben. Dadurch reduzierte sich die Zahl der theoretisch zugrunde gelegten Dimensionen von 15 auf 14.

Zur weiteren Überprüfung der Verständlichkeit wurde anschließend ein Pretest mit sechs fortgeschrittenen Pflegestudenten durchgeführt. Dieser diente dazu, mögliche Unklarheiten in den Instruktionen und Items aufzudecken. Infolge dieses Pretests wurden insbesondere die Instruktionen präzisiert. Die Teilnehmer wurden nun aufgefordert, den Fragebogen auf Grundlage ihrer letzten Erfahrung in einem Pflegeteam und mit Blick auf ihre übliche Leistung auszufüllen. Zusätzlich wurden zentrale Begriffe erläutert, etwa der Begriff „Scenarios“ als unterschiedliche diagnostische Hypothesen oder Ausgangspunkte vor Entscheidungen sowie der Begriff „Leader“ als die für das Versorgungsteam verantwortliche Person. Nach Abschluss die-

ses Prozesses lag die Studienversion der NTS-NAS mit 63 Items vor. Das Instrument wurde in portugiesischer Sprache entwickelt und ausdrücklich für Ausbildungs- und Simulationskontexte in der Pflege konzipiert.

In seiner ursprünglichen Struktur umfasste die NTS-NAS 14 Dimensionen, die den verbliebenen CRM-Prinzipien entsprachen. Diese Dimensionen lauteten „Know the environment“, „Anticipate and plan“, „Call for help early“, „Exercise leadership and followership“, „Distribute the workload“, „Communicate effectively“, „Use all available information“, „Prevent and manage fixation errors“, „Cross (double) check“, „Use cognitive aids“, „Re-evaluate repeatedly“, „Have a good teamwork“, „Allocate attention wisely“ und „Set priorities dynamically“. Die Zahl der zugeordneten Items variierte zwischen den einzelnen Dimensionen erheblich. So bestanden „Know the environment“ und „Anticipate and plan“ jeweils aus acht Items, „Call for help early“ aus fünf Items, „Exercise leadership and followership“ aus elf Items, „Distribute the workload“ aus zwei Items, „Communicate effectively“ aus sechs Items, „Use all available information“ und „Prevent and manage fixation errors“ jeweils aus einem Item, „Cross (double) check“ aus fünf Items, „Use cognitive aids“ aus zwei Items, „Re-evaluate repeatedly“ aus vier Items, „Have a good teamwork“ aus sieben Items, „Allocate attention wisely“ aus zwei Items und „Set priorities dynamically“ wiederum aus einem Item. Diese ungleiche Verteilung der Items auf die Dimensionen erwies sich in der psychometrischen Analyse als relevant, da insbesondere Ein-Item-Dimensionen und sehr kurze Skalen nur eingeschränkt hinsichtlich ihrer internen Konsistenz überprüfbar waren.

Die Validierung des Instruments erfolgte an einer Stichprobe von 177 Pflegestudenten der School of Health Sciences der Universität Aveiro in Portugal. In die Untersuchung wurden ausschließlich Studenten des zweiten, dritten und vierten Ausbildungsjahres einbezogen, da diese bereits über klinische Erfahrungen verfügten, während Erstsemester aufgrund fehlender praktischer Erfahrungen ausgeschlossen wurden. Die Stichprobe umfasste sowohl weibliche als auch männliche Teilnehmer, wobei der Anteil weiblicher Studenten deutlich überwog. Alle Teilnehmer verfügten über Erfahrungen in klinischen Praktika, jedoch nicht über Vorerfahrungen im Crisis Resource Management oder in hoch-fidelischer Simulation. Zur Prüfung der psychometrischen Eigenschaften wurden deskriptive Statistiken, Korrelationsanalysen, Analysen der internen Konsistenz mittels Cronbachs Alpha sowie explorative Faktorenanalysen durchgeführt.

Die Analyse der Sensitivität zeigte zunächst auf Dimensionsebene günstige Ergebnisse. Die Mittelwerte der Dimensionen wurden nach Angaben der Autoren nicht wesentlich durch Ausreißer beeinflusst. Zugleich lagen die Schiefe- und Kurtosiskoeffizienten in den meisten Fällen

nahe an der Einheit, was als Hinweis auf keine oder nur geringe Abweichungen von der Normalverteilung interpretiert wurde. Die Minimal- und Maximalwerte lagen deutlich auseinander, sodass von einer guten Verteilung der Antworten über die verschiedenen Antwortoptionen ausgegangen wurde. Auch auf Itemebene ergab sich ein insgesamt vergleichbares Bild. Für die meisten Items lagen Mittelwert und Median nahe beieinander, und die Antworten verteilten sich gut über die vorhandenen Antwortkategorien. Einzelne Items zeigten jedoch auffällige Werte hinsichtlich Schiefe und Kurtosis, was im Artikel als möglicher Hinweis auf sozial erwünschtes Antwortverhalten interpretiert wird. Die Autoren vermuten, dass Studenten möglicherweise eher solche Antworten wählten, die als erwartbar oder professionell gelten, als dass sie ihr tatsächliches Verhalten berichteten.

Die Korrelationsanalysen ergaben auf Dimensionsebene überwiegend signifikante und positive Zusammenhänge. Dies wurde als Hinweis darauf verstanden, dass höhere Ausprägungen in einer nichttechnischen Kompetenz tendenziell mit höheren Ausprägungen in anderen Kompetenzbereichen einhergehen. Besonders hohe Zusammenhänge zeigten sich zwischen „Know the environment“ und „Exercise leadership and followership“ mit einem Korrelationskoeffizienten von 0,64, zwischen „Call for help early“ und „Allocate attention wisely“ mit 0,60, zwischen „Exercise leadership and followership“ und „Distribute the workload“ mit 0,60 sowie zwischen „Use all available information“ und „Prevent and manage fixation errors“ mit 0,62. Niedrigere Zusammenhänge bestanden etwa zwischen „Exercise leadership and followership“ und „Use all available information“ mit 0,21, zwischen „Prevent and manage fixation errors“ und „Have a good teamwork“ mit 0,19 sowie zwischen „Use cognitive aids“ und „Have a good teamwork“ mit 0,22. Auf Itemebene bestätigte sich dieses Muster überwiegend. Die meisten Items standen in signifikant positiven Beziehungen zueinander. Eine Ausnahme bildete Item 52, das negativ formuliert war und daher eine signifikante negative Korrelation aufwies. Nach Auffassung der Autoren ist dies sachlogisch erklärbar, da dieses Item eine Konfliktbeteiligung beschreibt, während die übrigen Items positiv formulierte nichttechnische Verhaltensweisen erfassen.

Die interne Konsistenz wurde zunächst für die 14 theoretisch angenommenen Dimensionen separat untersucht. Als Referenzwert wurde ein Cronbach-Alpha von 0,70 zugrunde gelegt. Für mehrere Dimensionen ergaben sich akzeptable bis gute Werte. So erreichte „Know the environment“ ein Alpha von 0,77, „Anticipate and plan“ 0,73, „Call for help early“ 0,85, „Exercise leadership and followership“ 0,88, „Communicate effectively“ 0,74, „Re-evaluate repeatedly“ 0,71 und „Allocate attention wisely“ ebenfalls 0,71. Demgegenüber blieben einige Dimensionen unterhalb des gewünschten Wertes. „Cross (double) check“ erreichte 0,68, „Distri-

bute the workload“ 0,54, „Use cognitive aids“ 0,42 und „Have a good teamwork“ lediglich 0,36. Für die Dimensionen „Use all available information“, „Prevent and manage fixation errors“ sowie „Set priorities dynamically“ konnte keine interne Konsistenz berechnet werden, da diese jeweils nur ein Item enthielten. Diese Befunde deuten darauf hin, dass die ursprünglich angenommene 14-dimensionale Struktur nur teilweise durch eine zufriedenstellende interne Homogenität gestützt wird.

Zusätzlich wurden korrigierte Item-Gesamt-Korrelationen analysiert, um zu prüfen, inwieweit einzelne Items mit dem Gesamtscore ihrer jeweiligen Dimension zusammenhängen. Werte unter 0,30 wurden als problematisch angesehen. Besonders auffällig war die Dimension „Have a good teamwork“, innerhalb derer mehrere Items sehr niedrige oder sogar negative korrigierte Item-Gesamt-Korrelationen aufwiesen. Im Artikel werden für diese Dimension unter anderem Werte von -0,02 für Item 50, 0,27 für Item 51, -0,01 für Item 52 und 0,28 für Item 56 berichtet. Auch innerhalb der Dimension „Use cognitive aids“ fielen die Items 44 und 45 mit Korrelationen von jeweils 0,27 auf. Diese Befunde stützen die Annahme, dass einzelne Items die ihnen zugedachten Dimensionen nicht hinreichend konsistent abbilden. Entsprechend wird im Beitrag darauf hingewiesen, dass der Ausschluss bestimmter Items die interne Konsistenz der jeweiligen Skalen verbessern könnte.

Besondere Bedeutung kommt der Untersuchung der faktoriellen Struktur zu. Auf Grundlage der theoretischen Herleitung wurde zunächst eine explorative Faktorenanalyse mit 14 fixierten Faktoren durchgeführt. Die Voraussetzungen für eine Faktorenanalyse erwiesen sich als günstig. Der Kaiser-Meyer-Olkin-Wert lag bei 0,849, was auf eine gute Eignung der Daten für eine Faktorenanalyse hinweist, und der Bartlett-Test auf Sphärizität ergab mit $\chi^2 = 6483{,}998$ und $p = 0{,}000$ ein signifikantes Ergebnis, das ausreichende Interkorrelationen zwischen den Variablen belegt. Trotz dieser grundsätzlich guten Voraussetzungen bestätigte die Faktorenanalyse die theoretisch postulierte 14-dimensionale Struktur nicht. Die Auswertung der Komponentenmatrix und des Scree-Plots zeigte vielmehr eine deutliche Dominanz des ersten Faktors, auf den sämtliche 63 Items luden. Die Autoren leiten daraus ab, dass die NTS-NAS empirisch eher ein gemeinsames Gesamtkonstrukt nichttechnischer Fertigkeiten als klar unterscheidbare Einzeldimensionen erfasst.

Aufgrund dieses Befundes wurde das Instrument erneut unter Annahme einer unidimensionalen Struktur psychometrisch geprüft. Für diese eindimensionale Lösung ergaben sich insgesamt günstigere Ergebnisse. Die interne Konsistenz der Gesamtskala war mit einem Cronbach-Alpha von 0,94 sehr hoch. Lediglich wenige Items wiesen problematische korrigierte Item-Gesamt-Korrelationen unterhalb von 0,30 auf, darunter Item 13 mit 0,29, Item 40 mit 0,28,

Item 52 mit -0,02 und Item 53 mit 0,12. Auch die erneute Faktorenanalyse bestätigte die Eignung der unidimensionalen Lösung. Der Kaiser-Meyer-Olkin-Wert blieb mit 0,849 unverändert hoch, der Bartlett-Test zeigte weiterhin ein signifikantes Ergebnis, und das Ein-Faktor-Modell erklärte 26 % der Gesamtvarianz. Die Faktorladungen lagen überwiegend zwischen 0,37 und 0,73, während die Kommunalitäten Werte zwischen 0,24 und 0,53 erreichten. Diese Ergebnisse sprechen dafür, dass die Items in relevanter Weise auf ein gemeinsames zugrunde liegendes Konstrukt bezogen sind. Die Autoren kommen daher zu dem Schluss, dass die finale Version der NTS-NAS nicht als multidimensionales, sondern als unidimensionales Instrument mit 63 Items zur Erfassung allgemeiner nichttechnischer Fertigkeiten in der Pflege verstanden werden sollte.

Im Hinblick auf ihre Merkmale ist die NTS-NAS als Selbstbeurteilungsinstrument zu charakterisieren. Die Befragten schätzen ihre übliche Leistung in Bezug auf nichttechnische Fertigkeiten selbst ein, bezogen auf ihre Erfahrungen im Teamkontext. Daraus ergibt sich, dass das Instrument insbesondere für reflexive, formative und ausbildungsbezogene Zwecke geeignet ist. Die Bearbeitungsdauer wird mit etwa fünf bis fünfzehn Minuten angegeben, und während der Datenerhebung wurden keine nennenswerten Verständnisschwierigkeiten berichtet. Das Instrument wurde für den Einsatz in hoch- und niedrig-fidelischen Simulationen sowie in klinischen Lernumgebungen entwickelt und soll dazu beitragen, nichttechnische Fertigkeiten im pflegerischen Ausbildungsprozess sichtbarer und systematischer beurteilbar zu machen.

Die Anwendungsbereiche der NTS-NAS liegen nach Angaben des Artikels vor allem in Trainings- und Ausbildungskontexten. Das Instrument kann in curricularen Praktika, in spezifischen Workshops oder in Interventionsprogrammen eingesetzt werden, die neben technischen Fertigkeiten auch die Entwicklung nichttechnischer Kompetenzen adressieren. Nach Auffassung der Autoren kann die Erfassung solcher Fähigkeiten die Leistung, das Vertrauen und die Selbstwirksamkeit von Pflegestudenten stärken und ihnen helfen, sich besser an die Anforderungen komplexer klinischer Kontexte anzupassen. Darüber hinaus wird ein Einsatz in postgradualen Bildungsangeboten anderer Gesundheitsberufe als möglich angesehen. Ebenso könne die Skala verwendet werden, um Entwicklungsbedarfe und Verbesserungsmöglichkeiten in Versorgungseinrichtungen wie Krankenhäusern oder Privatpraxen zu identifizieren. Die NTS-NAS besitzt damit nicht nur eine diagnostische, sondern auch eine didaktische Funktion, da sie als Grundlage für Reflexion, Feedback und gezielte Förderung dienen kann.

Gleichzeitig weist die Studie mehrere Limitationen auf. Zunächst ist hervorzuheben, dass die theoretisch erwartete 14-dimensionale Struktur empirisch nicht bestätigt werden konnte. Dies

begrenzt die Möglichkeit, differenzierte Subskalen für einzelne Kompetenzbereiche zuverlässig zu interpretieren. Hinzu kommt, dass mehrere Dimensionen aufgrund zu geringer Itemzahlen oder unzureichender interner Konsistenz psychometrisch problematisch waren. Eine weitere Einschränkung besteht darin, dass die Validierung ausschließlich an einer Stichprobe von Pflegestudenten einer einzelnen Hochschule durchgeführt wurde, was die Generalisierbarkeit der Befunde einschränkt. Die Autoren weisen deshalb ausdrücklich darauf hin, dass weitere Untersuchungen mit größeren und repräsentativeren Stichproben aus unterschiedlichen Versorgungskontexten erforderlich sind. Darüber hinaus lässt die Beobachtung auffälliger Schiefe- und Kurtosiswerte bei einzelnen Items vermuten, dass sozial erwünschtes Antwortverhalten die Ergebnisse beeinflusst haben könnte. Schließlich ist zu berücksichtigen, dass es sich um ein Selbstbeurteilungsinstrument handelt, sodass die erhobenen Werte die subjektive Selbsteinschätzung der Befragten und nicht unmittelbar beobachtetes Verhalten widerspiegeln.

Zusammenfassend lässt sich festhalten, dass die NTS-NAS ein theoriegeleitet entwickeltes Instrument zur Erfassung nichttechnischer Fertigkeiten im pflegerischen Ausbildungskontext darstellt, das auf den Prinzipien des Crisis Resource Management basiert und in mehreren Entwicklungsschritten inhaltlich geprüft und überarbeitet wurde. Die empirischen Analysen zeigen, dass die ursprünglich angenommene 14-dimensionale Struktur nicht tragfähig ist, während eine unidimensionale Gesamtstruktur psychometrisch günstiger erscheint. Insbesondere die hohe interne Konsistenz der Gesamtskala, die guten Voraussetzungen für die Faktorenanalyse sowie die insgesamt zufriedenstellenden Sensitivitäts- und Korrelationsbefunde sprechen dafür, dass die NTS-NAS als globales Maß nichttechnischer Fertigkeiten in der Pflegeausbildung ein vielversprechendes Instrument darstellt. Zugleich machen die Ergebnisse deutlich, dass weitere Validierungsstudien erforderlich sind, um die Anwendbarkeit und Aussagekraft des Instruments in unterschiedlichen Bildungs- und Versorgungskontexten weiter abzusichern.

5.18 Objective Structured Assessment of Nontechnical Skills (OSANTS)

Quelle: Dedy NJ, Szasz P, Louridas M, Bonrath EM, Husslein H, Grantcharov TP. Objective structured assessment of nontechnical skills: reliability of a global rating scale for the in-training assessment in the operating room. Surgery. (2015) 157:1002–13. doi: 10.1016/j.surg.2014.12.023

Abbildung 21: Objective Structured Assessment of Nontechnical Skills (OSANTS)

Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Dedy et al. (2015)

Das Instrument Objective Structured Assessment of Nontechnical Skills (OSANTS) wurde entwickelt, um ein evidenzbasiertes und reliables Verfahren zur arbeitsplatznahen Beurteilung nichttechnischer Leistungen chirurgischer Weiterbildungsassistenten im Operationssaal bereitzustellen. Hintergrund dieser Entwicklung ist die im Beitrag dargestellte zunehmende Bedeutung nichttechnischer Kompetenzen wie Teamarbeit, Kommunikation und Führung für die Patientensicherheit sowie deren verpflichtende Berücksichtigung in chirurgischen Weiterbildungsprogrammen. Trotz dieser bildungspolitischen und klinischen Relevanz bestand nach Auffassung der Autoren bislang kein allgemein akzeptierter Ansatz für die routinemäßige In-Training-Beurteilung individueller nichttechnischer Leistungen im Operationssaal. Vorhandene Verfahren fokussierten überwiegend auf Teamleistungen oder waren nicht spezifisch auf die Leistungsbeurteilung von chirurgischen Weiterzubildenden ausgerichtet. Vor diesem Hintergrund zielte die Entwicklung von OSANTS darauf ab, ein praktikables, mit begrenztem Schulungsaufwand einsetzbares und sowohl im Simulationssetting als auch im realen Operationssaal anwendbares Instrument bereitzustellen.

Die Entwicklung des Instruments erfolgte in einem mehrstufigen Verfahren. Zunächst wurden die relevanten Inhalte auf Basis bestehender evidenzbasierter Rahmenmodelle und Bewer-

tungssysteme nichttechnischer Fertigkeiten im Operationssaal sowie unter Berücksichtigung von Weiterbildungsanforderungen verschiedener chirurgischer Fachgesellschaften und Akkreditierungsinstitutionen identifiziert. In die Entwicklung flossen unter anderem Anforderungen des Accreditation Council for Graduate Medical Education, des Royal College of Physicians and Surgeons of Canada, des Intercollegiate Surgical Curriculum Programme sowie des Royal Australasian College of Surgeons ein. Ein zentrales Auswahlkriterium für die spätere Skalenstruktur bestand darin, dass die zu erfassenden Kompetenzen sowohl für die chirurgische Weiterbildung relevant als auch bei chirurgischen Weiterbildungsassistenten im Operationssaal tatsächlich beobachtbar sein mussten. Anschließend wurde eine vorläufige Version des Instruments durch zwei erfahrene chirurgische Weiterbildungsassistenten in einem simulierten Operationssaal anhand geskripteter Videos pilotiert. Ziel dieser Pilotierung war die Überprüfung der Klarheit von Definitionen und Ankerbeschreibungen sowie die Identifikation möglicher Mehrdeutigkeiten. Auf dieser Basis wurden Formulierungen und Beschreibungen in einem iterativen Prozess überarbeitet. Danach erfolgte ein formales Ratertraining mit einer Einführung in die Konzepte nichttechnischer Fertigkeiten, einer Besprechung der Skalenitems und der Analyse beobachtbarer Verhaltensbeispiele. Daran schloss sich eine Kalibrierungsphase an, in der die Rater unabhängig voneinander zwölf ungeskriptete Videoszenarien aus einem simulierten Operationssaal bewerteten und ihre Urteile im Anschluss diskutierten. Der gesamte Schulungs- und Kalibrierungsaufwand belief sich nach Angaben der Autoren auf etwa sechs Stunden.

OSANTS ist als globale Ratingskala mit sieben Items konzipiert. Die finale Version des Instruments umfasst die Domänen Situation Awareness, Decision Making, Teamwork, Communication, Leading and Directing, Professionalism sowie Managing and Coordinating. Jede dieser Domänen wird auf einer fünfstufigen ordinalen Skala bewertet. Charakteristisch für die Struktur des Instruments ist, dass für jedes Item eigene deskriptive Anker für die Skalenwerte 1, 3 und 5 formuliert wurden. Diese spezifischen Verhaltensanker gelten im Artikel als zentrales Gestaltungsmerkmal des Instruments, da sie eine kriterienreferenzierte Bewertung ermöglichen sollen. Im Unterschied zu generischen Beurteilungsrastern oder Skalen, die primär die Auswirkungen von Verhalten auf Patientensicherheit oder Teamfunktion beschreiben, soll OSANTS durch itembezogene Verhaltensanker eine stärker objektivierte Bewertung beobachtbaren Verhaltens erlauben, ohne dass die Bewerter die Bedeutung eines Verhaltens zunächst umfassend interpretieren müssen. Dadurch soll sowohl die Durchführbarkeit als auch die Objektivität der Bewertung verbessert werden.

Die Domäne Situation Awareness beschreibt die Bereitschaft für eine Operation, die Fähigkeit zur Wahrnehmung und Sammlung von Informationen aus der Umgebung, deren Einordnung in den aktuellen Kontext sowie die Antizipation möglicher zukünftiger Entwicklungen. Auf der höchsten Bewertungsstufe wird eine gut vorbereitete Person beschrieben, die ihre Umgebung fortlaufend überwacht, Informationen angemessen interpretiert und künftige Ereignisse sowie Materialbedarfe routinemäßig antizipiert. Die Domäne Decision Making umfasst die Fähigkeit, Probleme zu erkennen, Handlungsoptionen zu entwickeln, Entscheidungen zu treffen und umzusetzen sowie deren Ergebnis zu überprüfen und gegebenenfalls den Handlungsplan zu ändern. Teamwork wird im Instrument als Fähigkeit verstanden, ein gemeinsames Verständnis im Operationsteam herzustellen und aufrechtzuerhalten, beispielsweise durch Briefings, operative Pausen, die rechtzeitige Weitergabe neuer Informationen, die aktive Einbeziehung anderer Teammitglieder sowie die Bereitschaft, Unterstützung anzubieten. Die Domäne Communication fokussiert auf die wirksame Übermittlung relevanter Informationen, etwa durch klare Botschaften, angemessene Lautstärkeanpassung an Umgebungsgeräusche sowie durch direkte Adressierung anderer Personen mittels Namen, Funktion oder Blickkontakt. Leading and Directing bezieht sich auf die Bereitschaft und Fähigkeit, im Operationssaal Führungsverantwortung zu übernehmen, wenn dies in der jeweiligen Situation angemessen ist. Dazu gehört auch, bei Bedarf entschlossen die Leitung zu übernehmen und Autorität sowie Durchsetzungsfähigkeit zu zeigen. Professionalism erfasst Verhaltensweisen wie Verantwortungsbewusstsein, Respekt gegenüber Teammitgliedern und Patienten, die Einhaltung professioneller Standards und guter klinischer Praxis sowie die Aufrechterhaltung professioneller Haltung auch unter Stress. Managing and Coordinating schließlich beschreibt die Fähigkeit, Abläufe im Operationssaal effizient und effektiv zu organisieren, Aufgaben zu delegieren und vorhandene personelle, materielle und informationelle Ressourcen zielorientiert zu nutzen.

Die Struktur des Instruments wurde bewusst auf beobachtbare Verhaltensweisen zugeschnitten, die für chirurgische Weiterzubildende relevant sind. Die Autoren betonen, dass komplexe Konstrukte möglichst auf zentrale, beobachtbare und für den Ausbildungskontext bedeutsame Merkmale reduziert wurden, um die Anwendbarkeit zu erhöhen und die Beurteilung auch für Personen mit begrenzter Erfahrung in der Bewertung nichttechnischer Fertigkeiten praktikabel zu machen. Dies erklärt auch, weshalb im Instrument Kommunikationsverhalten als eigenständige Domäne behandelt wird und nicht lediglich als Bestandteil von Teamarbeit oder Führung verstanden wird. Ebenso wurden breitere Führungskonzepte in die drei spezifischen Bereiche Leading and Directing, Professionalism und Managing and Coordinating aufgeteilt, um trennschärfere und leichter beobachtbare Kompetenzbereiche zu schaffen.

Die psychometrische Prüfung des Instruments erfolgte zunächst in einer simulierten Operationssaalumgebung. Zwei Rater bewerteten unabhängig voneinander sechs videografierte ungeskriptete Krisenszenarien mit chirurgischen Weiterbildungsassistenten der Allgemeinchirurgie. Die Interrater-Reliabilität des Gesamtmittelwerts erwies sich dabei als hoch. Für den Average-Measure-Intraclass-Correlation-Coefficient wurde ein Wert von 0,95 berichtet, für den Single-Measure-Koeffizienten ein Wert von 0,90. Auch auf Ebene der Einzelitems zeigten sich überwiegend gute Übereinstimmungen. Fünf der sieben Domänen erreichten Average-Measure-ICC-Werte zwischen 0,79 und 1,00. Die Domäne Professionalism erreichte lediglich eine moderate Übereinstimmung mit einem Wert von 0,62. Für die Domäne Communication konnte aufgrund fehlender Varianz in den Bewertungen kein ICC berechnet werden. Insgesamt deuten diese Befunde darauf hin, dass OSANTS im Simulationskontext insbesondere auf Ebene des Gesamtscores eine hohe Beurteilerübereinstimmung ermöglicht.

Die Anwendbarkeit des Instruments im realen Operationssaal wurde anhand von zehn Live-Beobachtungen chirurgischer Eingriffe geprüft, an denen Weiterbildungsassistenten unterschiedlicher Ausbildungsjahre beteiligt waren. Die Beobachtungen wurden durch dieselben beiden Rater durchgeführt, die sich während des Eingriffs unauffällig im Operationssaal aufhielten. Auch in diesem Setting erwies sich die Interrater-Reliabilität des Gesamtmittelwerts als hoch. Der Average-Measure-ICC lag wiederum bei 0,95 und der Single-Measure-ICC bei 0,90. Auf Ebene der Einzelitems erreichten fünf der sieben Domänen gute Übereinstimmungswerte zwischen 0,75 und 0,95. Für Teamwork ergab sich mit einem Average-Measure-ICC von 0,70 eine nur moderate Übereinstimmung. Für Professionalism konnte aufgrund einer sehr geringen Varianz in den vergebenen Bewertungen erneut kein ICC berechnet werden. Die Ergebnisse sprechen damit dafür, dass OSANTS nicht nur in standardisierten Simulationsumgebungen, sondern auch bei direkten Beobachtungen im klinischen Alltag zuverlässig eingesetzt werden kann.

Zur weiteren Prüfung der Reliabilität und internen Struktur wurde ein zusätzlicher Datensatz aus 31 weiteren videobasierten Krisensimulationen einbezogen. Auch in diesem erweiterten Datensatz zeigte sich eine hohe Interrater-Reliabilität des Gesamtscores mit einem Average-Measure-ICC von 0,95 und einem Single-Measure-ICC von 0,90. Die Einzelitems wiesen überwiegend gute Reliabilitätswerte auf. Situation Awareness erreichte einen Average-Measure-ICC von 0,82, Decision Making von 0,89, Teamwork von 0,85, Communication von 0,71, Leading and Directing von 0,95, Professionalism von 0,65 und Managing and Coordinating von 0,84. Die Werte zeigen, dass insbesondere Leading and Directing, Decision Making und Situation Awareness zuverlässig bewertet werden konnten, während Communication und insbe-

sondere Professionalism weniger hohe, aber noch akzeptable beziehungsweise moderate Übereinstimmungen aufwiesen.

Hinsichtlich der internen Konsistenz ergab sich für die Gesamtskala ein Cronbach-Alpha von 0,80, was im Artikel als hohe interne Konsistenz interpretiert wird. Die Item-Gesamt-Korrelationen lagen für fast alle Domänen zwischen 0,51 und 0,65, was auf substanziell positive Zusammenhänge der Einzelitems mit dem Gesamtscore hinweist. Eine Ausnahme bildete erneut die Domäne Professionalism, die lediglich eine schwache Korrelation mit dem Gesamtscore von 0,22 aufwies. Der Ausschluss dieses Items hätte den Alpha-Wert nur geringfügig auf 0,82 erhöht. Trotz dieser vergleichsweisen schwächeren psychometrischen Einbindung wurde Professionalism nicht aus dem Instrument entfernt, da die Autoren diesen Bereich als inhaltlich bedeutsam für die chirurgische Weiterbildung und für die kontinuierliche Beobachtung professionellen Handelns im Verlauf der Weiterbildung ansehen.

Ein weiterer Hinweis auf die Validität des Instruments ergibt sich aus seiner Beziehung zu einem bereits etablierten Verfahren, dem NOTSS-System. Einer der Rater, der Erfahrung in der Anwendung von NOTSS hatte, bewertete sowohl die Simulationsvideos als auch die Live-Beobachtungen zusätzlich mit diesem Instrument. Zwischen den NOTSS-Gesamtwerten dieses Raters und den OSANTS-Werten des zweiten Raters, der NOTSS nicht verwendete, zeigte sich eine starke positive Korrelation. Für die Simulationsvideos wurde ein Korrelationskoeffizient von 0,97 bei einem Signifikanzniveau von $p = 0,001$ berichtet, für die Beobachtungen im realen Operationssaal ein Korrelationskoeffizient von 0,82 bei $p = 0,004$. Diese Ergebnisse sprechen dafür, dass OSANTS inhaltlich in hohem Maße mit einem bestehenden Referenzinstrument zur Erfassung nichttechnischer chirurgischer Fertigkeiten übereinstimmt.

Die Autoren ordnen die Ergebnisse in einen validitätstheoretischen Rahmen nach Messick ein. Demnach lässt sich die Validität, der mit OSANTS gewonnenen Werte aus mehreren Perspektiven stützen. Die Inhaltsvalidität ergibt sich aus der Herleitung der Skaleninhalte aus evidenzbasierten Rahmenmodellen und curricularen Anforderungen chirurgischer Weiterbildung. Hinweise auf die Qualität des Response Process wurden durch Pilotierung, Instrumentenüberarbeitung sowie Ratertraining und Kalibrierung gewonnen. Die interne Struktur des Instruments wurde über Interrater-Reliabilität, interne Konsistenz und Item-Gesamt-Korrelationen untersucht und insgesamt als günstig bewertet. Die Beziehung zu anderen Variablen wurde durch die starke Korrelation mit den NOTSS-Werten gestützt. Insgesamt interpretieren die Autoren diese Befunde als Evidenz dafür, dass OSANTS valide und reliabel zur Beurteilung nichttechnischer Leistungen im chirurgischen Weiterbildungskontext eingesetzt werden kann.

Im Hinblick auf seine Anwendungsbereiche ist OSANTS primär für die In-Training-Beurteilung chirurgischer Weiterbildungsassistenten im Operationssaal vorgesehen. Das Instrument kann sowohl in simulationsbasierten Lehr-Lern-Umgebungen als auch im realen Operationssaal eingesetzt werden. Es soll in die alltäglichen klinischen Arbeitsabläufe integrierbar sein und vor allem formative Rückmeldungen ermöglichen, die zur gezielten Förderung nichttechnischer Kompetenzen genutzt werden können. Die Autoren heben hervor, dass insbesondere die hohen Single-Measure-ICC-Werte von praktischer Relevanz sind, da sie darauf hindeuten, dass auch die Bewertung durch einen einzelnen Beobachter, etwa einen supervidierenden Facharzt, bereits hinreichend reliabel sein könnte. Darüber hinaus wird das Instrument ausdrücklich als geeignet für Forschungszwecke beschrieben, da es in unterschiedlichen Kontexten stabile und konsistente Ergebnisse liefert.

Trotz der insgesamt positiven Befunde weist die Studie mehrere Limitationen auf. Erstens basieren die Untersuchungen auf Beobachtungen von Weiterbildungsassistenten aus einem einzelnen Ausbildungsprogramm, was die Generalisierbarkeit der Ergebnisse einschränken könnte. Zwar argumentieren die Autoren, dass die zugrunde liegenden Kompetenzbereiche fachübergreifend relevant seien und die Anwendbarkeit des Instruments durch den Einbezug unterschiedlicher chirurgischer Disziplinen gestützt werde, dennoch bleibt die empirische Basis institutionell begrenzt. Zweitens erfolgten die Bewertungen durch speziell geschulte Beobachter und nicht durch regulär supervidierende Fachärzte im normalen klinischen Alltag. Damit bleibt offen, inwieweit sich die gezeigte Reliabilität bei routinemäßiger Anwendung durch klinisches Lehrpersonal unter Alltagsbedingungen gleichermaßen erreichen lässt. Als weitere Einschränkung ist anzumerken, dass einzelne Domänen, insbesondere Professionalism und in Teilen auch Communication, psychometrisch weniger robust waren als die übrigen Bereiche. Im Artikel wird dies teilweise auf Deckeneffekte zurückgeführt, da viele Teilnehmer in diesen Bereichen hohe oder sehr hohe Bewertungen erhielten und dadurch nur geringe Varianz entstand. Zudem wird diskutiert, dass die Domäne Professionalism in ihrer Definition stärker an curricularen Anforderungen als an klassischen Rahmenmodellen nichttechnischer Fertigkeiten orientiert war, was ihre schwächere Einbindung in das Gesamtkonstrukt mit erklären könnte.

Zusammenfassend lässt sich festhalten, dass OSANTS ein spezifisch für die chirurgische Weiterbildung entwickeltes Beobachtungsinstrument zur Erfassung nichttechnischer Kompetenzen im Operationssaal darstellt. Seine Entwicklung basiert auf evidenzbasierten Modellen und curricularen Anforderungen und mündete in eine sieben Items umfassende globale Ratingskala mit domänenspezifischen Verhaltensankern. Die psychometrischen Ergebnisse weisen auf eine hohe Reliabilität des Gesamtscores in Simulations- und Realkontexten, eine gute

interne Konsistenz sowie eine starke inhaltliche Übereinstimmung mit dem NOTSS-System hin. Damit erscheint OSANTS als vielversprechendes Instrument für formative Leistungsbeurteilung und Forschung im chirurgischen Weiterbildungskontext. Gleichzeitig machen die Befunde deutlich, dass weitere Untersuchungen zur Generalisierbarkeit und zur routinemäßigen Anwendung durch klinisches Lehrpersonal notwendig sind, um die Einsatzmöglichkeiten des Instruments weiter abzusichern.

5.19 Observational Skill-Based Clinical Assessment Tool for Resuscitation (OSCAR): Ein Instrument für die Notfallmedizin

Quelle: Walker S, Brett S,McKay A, Lambden S, Vincent C, Sevdalis N. Observational skill-based clinical assessment tool for resuscitation (OSCAR): development and validation. Resuscitation. (2011) 82:835–44. doi: 10.1016/j.resuscitation.2011.03.009

Abbildung 22: Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR)

Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Walker et al. (2011)

Das Instrument **OSCAR** (*Observational Skill-based Clinical Assessment tool for Resuscitation*) wurde entwickelt, um nicht-technische Fähigkeiten in Reanimationsteams systematisch zu er-

fassen und einer strukturierten Beobachtung, Bewertung und Rückmeldung zugänglich zu machen. Ausgangspunkt seiner Entwicklung war die Erkenntnis, dass erfolgreiche Reanimationen nicht allein von technischen Fertigkeiten abhängen, sondern in hohem Maße auch von nicht-technischen Kompetenzen wie Kommunikation, Entscheidungsfindung, Führung, Aufgabenkoordination und Monitoring beziehungsweise Situationsbewusstsein. Das zugrunde liegende Dokument betont, dass insbesondere Notfall- und Reanimationssituationen aufgrund ihrer Zeitkritik, der hohen Interventionsdichte, der oft unvollständigen Informationslage und der spontanen Zusammensetzung der beteiligten Teams besonders anfällig für Fehler und unerwünschte Ereignisse sind. Vor diesem Hintergrund bestand ein klarer Bedarf an einem Instrument, das die Leistung von Teammitgliedern in Reanimationssituationen nicht nur technisch, sondern auch im Hinblick auf ihr nicht-technisches Verhalten differenziert erfassen kann. Während in anderen Bereichen der Akutmedizin bereits Instrumente zur Bewertung nicht-technischer Fähigkeiten entwickelt worden waren, fehlte bislang nach Darstellung des Dokuments ein spezifisches Tool, das die Leistung einzelner Teammitglieder im Reanimationskontext systematisch abbildet.

Die Entwicklung von OSCAR erfolgte in einem methodisch klar strukturierten, dreiphasigen Prozess, der im Dokument sowohl textlich als auch schematisch dargestellt wird. Ziel dieses Prozesses war es, ein Instrument zu schaffen, das hinsichtlich Validität, Reliabilität und Praktikabilität abgesichert ist. In der ersten Phase wurden bestehende Instrumente aus anderen medizinischen Einsatzfeldern als Evidenzbasis herangezogen, um daraus eine erste Version des Instruments zu entwickeln. Als Ausgangssysteme dienten das **Observational Teamwork Assessment for Surgery (OTAS)**, das Instrument **Anaesthetists' Non-Technical Skills (ANTS)** sowie die revidierte **NOTECHS-Skala** für den OP-Kontext. Diese Instrumente wurden ausgewählt, weil sie bereits als validierte und in Echtzeit sowie in Simulationen einsetzbare Systeme zur Erfassung nicht-technischer Fertigkeiten vorlagen. Das Dokument betont, dass die mit ihnen erfassten Verhaltensweisen zwar teilweise unterschiedlich bezeichnet würden, inhaltlich jedoch weitgehend vergleichbare nicht-technische Kompetenzbereiche abdecken. Auf dieser Grundlage wurde OSCAR für den spezifischen Kontext der kardiopulmonalen Reanimation konzipiert.

Die erste Entwicklungsphase führte zu einem Instrument, das sechs Verhaltensdomänen für drei zentrale Rollen innerhalb eines typischen Reanimationsteams erfasst. Bei diesen drei Kernrollen handelt es sich erstens um den Atemwegs-, Beatmungs- und Gefäßzugangsspezialisten, der im Instrument als *Anaesthetist* bezeichnet wird, wobei das Dokument ausdrücklich darauf hinweist, dass diese Rolle lokal auch durch andere Berufsgruppen wie etwa Respiratory

Therapists oder OP-Personal ausgefüllt werden kann. Zweitens wird die Rolle des medizinisch führenden Arztes als *Physician* bezeichnet, wobei diese Funktion ebenfalls je nach lokaler Struktur durch andere Fachdisziplinen übernommen werden kann. Drittens wird die *Senior nurse* als pflegerische Schlüsselrolle innerhalb des Reanimationsteams erfasst. Diese Fokussierung auf drei Rollenbereiche zeigt, dass OSCAR nicht die Gesamtleistung des Teams pauschal bewertet, sondern die nicht-technischen Leistungen rollenspezifisch analysiert.

Die sechs durch OSCAR erfassten Domänen sind **Communication**, **Co-operation**, **Co-ordination**, **Monitoring**, **Leadership** und **Decision making**. Damit deckt das Instrument zentrale kommunikative, kooperative, organisatorische, führungsbezogene und kognitive Aspekte der Teamleistung ab. Ein zentrales Konstruktionsprinzip von OSCAR besteht darin, dass jede dieser Domänen nicht abstrakt, sondern durch konkrete Verhaltensbeispiele operationalisiert wird. Diese sogenannten *exemplar behaviours* wurden entwickelt, um optimale beziehungsweise problematische Formen des Teamverhaltens im Reanimationskontext sichtbar und bewertbar zu machen. Das Dokument verdeutlicht dies anhand eines Beispiels aus der Kommunikation: Als positives Verhalten wird etwa angesehen, wenn die primär betreuende Pflegekraft dem eintreffenden Reanimationsteam eine klare und knappe Schilderung der Situation liefert, idealerweise im Sinne des SBAR-Schemas. Als negatives Gegenbeispiel wird beschrieben, wenn beim Eintreffen des Teams keine hilfreichen Informationen bereitgestellt werden und dadurch die Reanimation behindert wird. Die Verhaltensbeispiele wurden auf Basis der bereits validierten OTAS-Exemplare entwickelt und für den Reanimationskontext angepasst.

In der ersten Version umfasste OSCAR insgesamt 54 Verhaltensbeispiele, da für jede der drei Rollen in jeder der sechs Domänen drei Verhaltensbeispiele formuliert wurden. Diese erste Fassung wurde in der zweiten Entwicklungsphase einer systematischen Face- und Content-Validation unterzogen. Hierzu bewerteten zehn Experten aus dem Bereich der Reanimationsversorgung die Relevanz der Verhaltensbeispiele. Um sowohl die fachspezifische als auch die professionsübergreifende Einschätzung zu berücksichtigen und potenzielle Disziplinenverzerrungen zu reduzieren, wurde jedes Set von Verhaltensbeispielen sowohl von fünf Experten der entsprechenden Disziplin als auch von fünf fachfremden Experten eingeschätzt. Die Bewertung erfolgte auf einer vierstufigen Skala hinsichtlich der Bedeutung des jeweiligen Verhaltens, wobei eins für geringe und vier für kritische Bedeutung stand. Zusätzlich konnten Formulierungen kommentiert, Erweiterungen vorgeschlagen oder Streichungen empfohlen werden. Auf Basis dieser Überprüfung wurden Mittelwerte und Standardabweichungen berechnet. Verhaltensbeispiele mit einem Mittelwert von drei oder darunter wurden durch das Entwicklungsteam detailliert überprüft. Das Dokument berichtet, dass 39 der 54 ursprünglichen Exemplare von

den Experten als „kritisch wichtig“ bewertet wurden. Fünfzehn Verhaltensbeispiele lagen im Grenzbereich und wurden deshalb erneut diskutiert. Im Ergebnis wurden sieben Exemplare sprachlich verändert, vier entfernt, vier zwar überprüft, aber beibehalten und zusätzlich ein neues Exemplar aufgenommen. Insgesamt wurden 18 Änderungen vorgenommen. Dieser Schritt zeigt, dass das Instrument nicht nur auf bestehender Literatur aufbaut, sondern in einem strukturierten Expertenprozess inhaltlich nachgeschärft wurde.

In der dritten Phase wurde die Reliabilität des Instruments überprüft. Dabei standen die interne Konsistenz sowie die Interrater-Reliabilität im Vordergrund. Zwei erfahrene klinische Beobachter bewerteten unabhängig voneinander acht Videos von Reanimationsteams. Vier dieser Videos stammten aus Simulationstrainings in einer Trainingsumgebung, vier weitere aus unangekündigten In-situ-Simulationen in realen klinischen Settings. Die Szenarien umfassten unterschiedliche Konstellationen, beispielsweise eine massive postpartale Blutung auf der Geburtsstation oder ein rupturiertes Bauchaortenaneurysma in der Radiologie. Dadurch wurde die Anwendbarkeit des Instruments in unterschiedlichen reanimationsbezogenen Akutsituationen geprüft.

Die interne Konsistenz wurde mit **Cronbachs Alpha** berechnet. Das Dokument führt aus, dass Werte im Bereich von 0,70 bis 0,90 üblicherweise als adäquat gelten. Die Ergebnisse zeigten insgesamt hohe bis sehr hohe interne Konsistenz. Für die Anästhesiegruppe lagen die Alpha-Werte zwischen 0,745 in der Domäne *Co-operation* und 0,965 in der Domäne *Decision making*. Die Domänen *Communication* und *Leadership* erreichten mit 0,951 beziehungsweise 0,952 ebenfalls sehr hohe Werte. Für die Physician-Gruppe lagen die Werte zwischen 0,855 für *Co-ordination* und 0,949 für *Monitoring*, wobei alle Domänen deutlich im guten bis sehr guten Bereich lagen. Auch bei der Pflegegruppe bewegten sich die Alpha-Werte auf hohem Niveau und reichten von 0,736 für *Monitoring* bis 0,948 für *Co-operation*. Insgesamt lagen die Alpha-Werte damit zwischen 0,736 und 0,965. Das Dokument hebt hervor, dass 15 der 18 Verhaltensskalen, also 83 %, eine sehr hohe interne Konsistenz mit Werten über 0,80 aufwiesen. In Folge der Analysen wurden drei Verhaltensbeispiele aus dem Instrument entfernt, da sie nicht konsistent messbar waren. Dabei handelte es sich nicht um klinisch irrelevante Verhaltensweisen, sondern um solche, die im Rahmen des Instruments nicht stabil genug erfasst werden konnten.

Die Interrater-Reliabilität wurde mit **Intraklassenkorrelationen (ICC)** bestimmt. Nach Angaben des Dokuments gelten Werte von 0,70 oder höher typischerweise als Hinweis auf eine adäquate Übereinstimmung zwischen unabhängigen Beobachtern. Die Ergebnisse zeigen auch hier überwiegend gute bis sehr gute Werte. Für die Anästhesiegruppe lagen die ICC-

Werte zwischen 0,664 für *Monitoring* und 0,876 für *Co-ordination*, der Gesamtwert betrug 0,767. Bei der Physician-Gruppe reichten die Werte von 0,743 für *Co-ordination* bis 0,895 für *Decision making*, mit einem Gesamtwert von 0,809. Für die Pflegegruppe wurden Werte zwischen 0,652 für *Co-operation* und 0,911 für *Decision making* erreicht, der Gesamtwert lag bei 0,807. Alle Koeffizienten waren hochsignifikant mit $p < 0,001$. Auch wenn einzelne Werte knapp unterhalb des im Methodenteil genannten Richtwertes von 0,70 lagen, bewertet das Dokument die Interrater-Reliabilität insgesamt als sehr gut. Zusammen mit der hohen internen Konsistenz bildet dies die Grundlage für die Einschätzung der Autoren, OSCAR sei psychometrisch robust, wissenschaftlich fundiert und klinisch relevant.

Die Struktur des fertigen Instruments ist in der im Dokument abgebildeten Endversion besonders anschaulich dargestellt. Dort wird deutlich, dass die Bewertung jedes einzelnen Verhaltens auf einer Skala von 0 bis 6 erfolgt. Die Anker reichen von „Team severely compromised“ bis „Highly effective in enhancing teamwork“. Zusätzlich ist pro Rollenbereich und Domäne ein globaler Verhaltensscore vorgesehen. Diese Skalenlogik erlaubt nicht nur die Einschätzung einzelner Verhaltensbeispiele, sondern auch eine zusammenfassende Beurteilung der jeweiligen Domäne für jede Kernrolle. In der Domäne *Communication* wird etwa für die Anästhesiegruppe erfasst, ob das Team über Atemanstrengung des Patienten, weitere klinische Zeichen oder eine geplante Intubation informiert wird. Für die Physician-Gruppe wird unter anderem bewertet, ob die Anamnese bzw. Aktenlage überprüft und relevante Informationen klar weitergegeben werden und ob die Kommunikation zwischen Subteams gefördert wird. Für die Pflegegruppe wird etwa erhoben, ob beim Eintreffen des Reanimationsteams klare Informationen zum Arrest gegeben werden und ob junioren Pflegekräften angemessene Anweisungen erteilt werden. In den weiteren Domänen finden sich ebenso konkrete Verhaltensitems, beispielsweise zur Unterstützung anderer Teamgruppen, zur Koordination des Materialeinsatzes, zur Führung bei Airway- oder Basic-Life-Support-Aufgaben, zum Monitoring des Patienten und des Teamzustands oder zur raschen Identifikation des Problems und zum Entwickeln eines Reanimationsplans.

Ein wesentliches Merkmal von OSCAR ist seine **rollenspezifische Differenzierung**. Anders als Instrumente, die ausschließlich die Gesamtleistung eines Teams bewerten, erlaubt OSCAR eine detaillierte Zuordnung von Stärken und Schwächen zu einzelnen Rollen innerhalb des Teams. Gerade dieser Aspekt wird im Dokument auch im Vergleich zu anderen Instrumenten besonders betont. In der Diskussion wird beispielsweise das Instrument **TEAM** als ein Tool beschrieben, das die Gesamtteamleistung anhand von zwölf Punkten bewertet. Im Unterschied dazu erfasst OSCAR jedes der drei Teammitglieder beziehungsweise Teamsegmente

separat und bewertet sechs Verhaltensbereiche detailliert innerhalb dieser Subgruppen, was insgesamt 48 bewertbare Punkte ergibt. Das Dokument leitet daraus ab, dass TEAM zwar vermutlich schneller anzuwenden sei, OSCAR aber eine deutlich differenziertere und einsichtsreichere Analyse des Reanimationsverhaltens ermögliche. Insbesondere erlaube OSCAR eine individuelle Rückmeldung an einzelne Teammitglieder zu ihren nicht-technischen Fertigkeiten.

Die vorgesehenen Anwendungsbereiche des Instruments sind breit gefasst. Das Dokument sieht den primären Einsatz von OSCAR sowohl in **simulation centre training** als auch in **realen klinischen Reanimationssituationen**. Bereits im Abstract wird vorgeschlagen, das Tool in simulierten wie in realen kardialen Arrestsituationen einzusetzen, um nicht-technische Fähigkeiten zu bewerten, anzuleiten und zu trainieren. In der Diskussion wird ausgeführt, dass das Instrument von einer Person mit Reanimationserfahrung genutzt werden könne, ohne dass zwingend Vorerfahrungen im Umgang mit Verhaltensbewertungssystemen erforderlich seien. Allerdings werde eine begrenzte Einführung in die Nutzung des Instruments vorausgesetzt. Die Autoren gehen davon aus, dass OSCAR im Rahmen realer wie simulierter Reanimationen helfen kann, Schwächen und Entwicklungsmöglichkeiten im nicht-technischen Verhalten von Teammitgliedern sichtbar zu machen. Dies soll wiederum die Grundlage für konstruktives Debriefing und gezieltes Training bilden. Langfristig erwarten die Autoren, dass dies zu einer allgemeinen Verbesserung der Teamleistung bei Notfällen, zu einer Reduktion von Fehlern und unerwünschten Ereignissen sowie indirekt zu einer Abflachung hierarchischer Strukturen beiträgt, was wiederum die Kultur der Patientensicherheit stärken könne.

Trotz dieser Stärken weist das Dokument auch auf mehrere Limitationen hin. Eine erste Einschränkung besteht in der **Komplexität des Instruments**. Gerade im Vergleich zu einfacheren teamorientierten Tools wie TEAM oder dichotomen Checklisten erscheint OSCAR aufwendiger, da es eine Vielzahl einzelner Verhaltensweisen und Rollen differenziert bewertet. Das Dokument erkennt an, dass dies die Anwendung anspruchsvoller machen könnte, hält den höheren Detaillierungsgrad jedoch für einen wesentlichen Vorteil, insbesondere im Hinblick auf individuelles Feedback. Eine weitere Einschränkung besteht darin, dass die vorgestellte Validierung auf **Erwachsenenreanimationen** beschränkt ist. Für pädiatrische Kontexte oder für komplexere Notfallsettings wie Major Trauma wären nach Aussage der Autoren weitere Anpassungen erforderlich, auch wenn die zugrunde liegenden Prinzipien wahrscheinlich übertragbar seien. Darüber hinaus beruhen die Reliabilitätsprüfungen auf acht Videofällen und zwei Beobachtern, was trotz überzeugender Kennwerte eine methodisch begrenzte Grundlage darstellt. Schließlich zeigt die Entfernung von drei ursprünglich enthaltenen Verhaltensbeispielen,

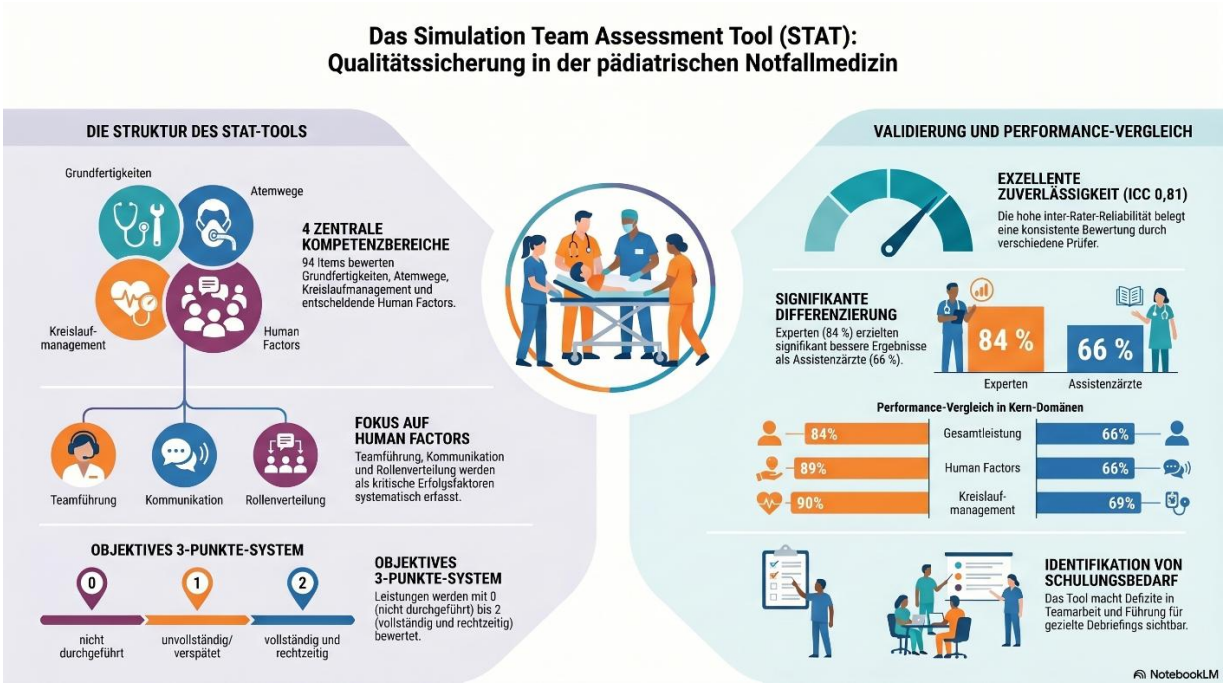
dass nicht jede klinisch sinnvolle Verhaltensweise zugleich hinreichend konsistent und reliabel messbar ist.

Zusammenfassend lässt sich festhalten, dass OSCAR im zugrunde liegenden Dokument als methodisch sorgfältig entwickeltes, rollenspezifisches und psychometrisch gut abgesichertes Instrument zur Erfassung nicht-technischer Fähigkeiten in Reanimationsteams beschrieben wird. Die Entwicklung erfolgte in drei aufeinander aufbauenden Phasen unter Rückgriff auf bestehende validierte Instrumente, strukturierte Expertenvalidierung und anschließende Reliabilitätsprüfung. Das Instrument erfasst sechs zentrale nicht-technische Domänen in drei Kernrollen des Reanimationsteams und verbindet diese mit konkreten Verhaltensitems, die auf einer siebenstufigen Skala von 0 bis 6 bewertet werden. Besonders hervorzuheben sind die hohe interne Konsistenz, die gute bis sehr gute Interrater-Reliabilität sowie die Möglichkeit, nicht-technische Fähigkeiten nicht nur global, sondern rollenspezifisch und verhaltensnah zu analysieren. Gerade diese Struktur macht OSCAR zu einem Instrument, das sich nicht nur für Forschung, sondern insbesondere auch für Simulation, Debriefing, individuelles Feedback und gezieltes Training eignet. Gleichzeitig ist zu berücksichtigen, dass das Tool in der Anwendung komplexer ist als einfachere Alternativen und dass seine bisherige Prüfung auf den Kontext der Erwachsenenreanimation beschränkt bleibt. Insgesamt erscheint OSCAR auf Basis des Dokuments als ein klinisch relevantes und wissenschaftlich fundiertes Instrument, das einen wichtigen Beitrag zur systematischen Erfassung und Weiterentwicklung nicht-technischer Fähigkeiten in Reanimationsteams leisten kann.

5.20 Simulation Team Assessment Tool (STAT)

Quelle: Reid J, Stone K, Brown J, Caglar D, Kobayashi A, Lewis-Newby M, et al. The simulation team assessment tool (STAT): development, reliability and validation. Resuscitation. (2012) 83:879–86. doi: 10.1016/j.resuscitation.2011.12.012

Abbildung 23: Simulation Team Assessment Tool (STAT



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Reid et al. (2012)

Der Simulation Team Assessment Tool (STAT) wurde mit dem Ziel entwickelt, die Gesamtleistung von Teams in simulierten pädiatrischen Reanimationssituationen umfassend zu erfassen. Der Ausgangspunkt für die Entwicklung des Instruments war die im Artikel formulierte Feststellung, dass zwar zahlreiche Simulationsformate zur Vorbereitung auf seltene, aber kritische pädiatrische Notfallsituationen eingesetzt werden, jedoch kein validiertes Instrument existierte, das die Gesamtleistung eines Reanimationsteams in ihrer Breite beurteilen kann. Bestehende Verfahren konzentrierten sich überwiegend auf einzelne Personen oder auf Teilbereiche wie technisches Handeln oder Human Factors. Der STAT sollte demgegenüber medizinische Entscheidungsfindung, technische Fertigkeiten und nichttechnische Teamkompetenzen in einem gemeinsamen Instrument zusammenführen und dadurch eine umfassendere Beurteilung der Teamleistung ermöglichen.

Die Entwicklung des Instruments stützte sich auf etablierte inhaltliche Referenzrahmen der pädiatrischen Reanimation. Als Grundlage dienten die Standards des Pediatric Advanced Life Support (PALS), das Tool for Resuscitation Assessment using Computerized Simulation (TRACS) sowie mehrere bereits publizierte Checklisten und Bewertungsinstrumente. Zur Festlegung der endgültigen Instrumentenstruktur wurde ein modified-Delphi-Verfahren eingesetzt.

In diesen Prozess waren Experten aus den Bereichen pädiatrische medizinische Weiterbildung, pädiatrische Notfallmedizin und pädiatrische Intensivmedizin eingebunden. In zwei aufeinanderfolgenden Überarbeitungsrunden prüften sie potenzielle Items, deren Bewertungskriterien sowie die formale Gestaltung des Instruments. Erst nach Erreichen eines Konsenses wurde die endgültige Fassung des STAT festgelegt. Dieses Entwicklungsverfahren diente der systematischen Auswahl relevanter und fachlich angemessener Inhalte und bildet zugleich die Grundlage für die im Beitrag beschriebene Inhaltsvalidität des Instruments.

Die finale Version des STAT umfasst 94 Einzelelemente, die auf vier Domänen verteilt sind. Diese Domänen sind basic assessment skills, airway and breathing, circulation sowie human factors. Nach den Angaben des Artikels entfallen 15 Elemente auf die basic skills, 28 auf die Domäne airway and breathing, 25 auf circulation und 26 auf human factors. Die detaillierte Struktur des Instruments ist in der Tabelle auf den Seiten 3 und 4 des Dokuments dargestellt und zeigt, dass der STAT in hohem Maße verhaltensnah und handlungsbezogen konstruiert ist. Im Bereich der basic skills werden grundlegende Aufgaben der pädiatrischen Reanimationsversorgung erfasst. Dazu zählen unter anderem das Erheben einer SAMPLE-Anamnese, die Durchführung eines Primary Survey nach dem ABCDE-Schema, die Durchführung eines Secondary Survey, die Gewichtsermittlung, die Anbringung geeigneter Monitore, die Sicherung eines Gefäßzugangs, die Veranlassung geeigneter Laboruntersuchungen und bildgebender Diagnostik, das angemessene Reagieren auf Untersuchungsergebnisse, die Erkennung akuter Gefahrensituationen, die Einhaltung von Schutzmaßnahmen, die Konsultation weiterer Fachdisziplinen sowie die Kommunikation mit Angehörigen. Die Domäne airway and breathing umfasst die Beurteilung von Atemweg und Atmung, basale Atemwegsinterventionen, die Gabe von Sauerstoff, die Anwendung geeigneter Atemwegshilfen, die Beutel-Masken-Beatmung, Aspekte der Rapid Sequence Intubation, die Durchführung und Sicherung der endotrachealen Intubation, die Überprüfung der Tubuslage sowie Maßnahmen der Magendekompression. Die Domäne circulation beinhaltet die Beurteilung zentraler Kreislaufparameter, Volumentherapie, Reanimationsmaßnahmen wie Herzdruckmassage und Defibrillation, das Erkennen und Behandeln von Rhythmusstörungen sowie das korrekte algorithmische Vorgehen nach PALS. Die Domäne human factors integriert teambezogene und nichttechnische Aspekte wie professionelles Verhalten, die Identifikation einer klaren Teamleitung, die Rollenverteilung, die Nutzung geschlossener Kommunikationsschleifen, die Konfliktlösung, die Einbeziehung des Teams in Entscheidungen, Prioritätensetzung, Vermeidung von Fixierungsfehlern, Zwischenzusammenfassungen sowie die Anpassung von Rollen und Arbeitsbelastung. Dadurch erfasst das Instrument sowohl technische Reanimationskompetenzen als auch koordinative, kommunikative und führungsbezogene Aspekte teambezogenen Handelns.

Die Bewertung jedes einzelnen Elements erfolgt anhand eines verhaltensverankerten Punktesystems. Ein Wert von zwei Punkten wird vergeben, wenn eine Handlung vollständig und zeitgerecht durchgeführt wurde. Ein Punkt wird vergeben, wenn die Maßnahme zwar durchgeführt, aber unvollständig oder verspätet war. Null Punkte erhält ein Element, wenn die entsprechende Maßnahme erforderlich gewesen wäre, jedoch nicht durchgeführt oder fehlerhaft umgesetzt wurde. Zusätzlich kann ein Item als nicht anwendbar bewertet werden, wenn es im jeweiligen Szenario nicht erforderlich war. Nicht beobachtbare Elemente konnten entsprechend leer bleiben. Für jedes Item wurden spezifische Kriterien zur Bestimmung von Vollständigkeit und Timeliness formuliert. Der Artikel nennt als Beispiel die Anlage eines intraossären Zugangs, die mit zwei Punkten bewertet wurde, wenn sie nach zwei erfolglosen intravenösen Punktionsversuchen oder innerhalb von 90 Sekunden erfolgte. Diese verhaltensverankerte Ausgestaltung des Instruments stellt sicher, dass die Beurteilung möglichst konkret und standardisiert an beobachtbare Handlungen gekoppelt wird. Für die Berechnung der Gesamt- und Domänenscores wurden nur numerisch bewertete Items herangezogen; eine Gewichtung der Items erfolgte nicht.

Die Überprüfung der Inhaltsvalidität erfolgte durch sieben externe Experten aus den Bereichen Notfallmedizin, Intensivmedizin und medizinische Weiterbildung. Diese prüften, ob sämtliche aufgenommenen Elemente notwendig und angemessen waren, ob wichtige Aspekte fehlten und ob die jeweiligen Handlungen realistischerweise von pädiatrischen Weiterbildungsassistenten erwartet werden können. Die Autoren interpretieren die Bestätigung dieser Punkte als Hinweis auf eine gute Inhaltsvalidität. Das Instrument wurde somit nicht nur aus theoretischen Referenzmodellen abgeleitet, sondern auch fachlich auf seine Passung zum Zielkontext überprüft.

Die Prüfung der Konstruktvalidität erfolgte über den Vergleich von zwei Expertenteams mit zwei Residententeams. Die Expertenteams bestanden jeweils aus einem pädiatrischen Notfallmediziner auf Attending-Niveau, einem Fellow der pädiatrischen Notfallmedizin sowie einem Fellow der pädiatrischen Intensivmedizin. Die Residententeams setzten sich aus Assistenzärzten des ersten, zweiten und dritten Ausbildungsjahres zusammen. Alle Teams absolvierten dasselbe standardisierte Simulationsszenario, das einen zweijährigen Patienten mit septischem Schock, nachfolgender Apnoe und Kammerflimmern umfasste. Das Szenario war vorprogrammiert, enthielt zeitlich definierte Trigger und standardisierte Reaktionen des Simulators sowie des Moderationsteams. Die Logik der Konstruktvalidierung bestand darin, dass ein valides Instrument in der Lage sein sollte, erwartbare Leistungsunterschiede zwischen erfahrenen Expertenteams und weniger erfahrenen Residententeams abzubilden.

Die Ergebnisse zeigten, dass der STAT diese Anforderung auf Ebene des Gesamtscores und in mehreren Domänen erfüllte. Der mittlere Gesamtscore lag bei den Expertenteams bei 0,84 und bei den Residententeams bei 0,66. Dieser Unterschied war mit einem p-Wert von 0,02 statistisch signifikant. Auch in den Domänen basic skills, circulation und human factors erzielten die Expertenteams signifikant höhere Werte als die Residententeams. Für basic skills lagen die Mittelwerte bei 0,73 gegenüber 0,55, für circulation bei 0,90 gegenüber 0,69 und für human factors bei 0,89 gegenüber 0,66. Lediglich in der Domäne airway and breathing ergab sich kein signifikanter Unterschied zwischen Expertenteams und Residententeams; hier lagen die Werte bei 0,80 beziehungsweise 0,75 bei einem p-Wert von 0,25. Die im Artikel abgebildete grafische Darstellung der Konstruktvalidität zeigt entsprechend, dass sich insbesondere in den Domänen Basics, Circulation, Team Management beziehungsweise Human Factors sowie im Gesamtscore klare Leistungsunterschiede zwischen beiden Gruppen abzeichnen, während die Werte im Atemwegsbereich stark überlappen. Die Autoren interpretieren diese Befunde dahingehend, dass das Instrument insgesamt gut zwischen unterschiedlichen Erfahrungsniveaus differenzieren kann, die Atemwegsdomäne jedoch weiterer Prüfung bedarf.

Die Interrater-Reliabilität wurde durch sechs Rater untersucht, die jede videografierte Simulation unabhängig voneinander mit dem STAT bewerteten. Vor Beginn der Bewertung nahmen alle Rater an einer vierstündigen Schulung teil. Diese umfasste eine Einführung in das Instrument, die Verwendung eines Referenzmanuals mit Definitionen und Timeliness-Kriterien für jedes Item sowie das eigenständige Scoring zweier Übungsvideos, die nicht Teil der Studie waren. Anschließend wurden die Bewertungen dieser Trainingsvideos gemeinsam diskutiert, um ein konvergentes Bewertungsverständnis herzustellen. In der eigentlichen Studienphase sahen die Rater die Teamleistungen jeweils nur einmal in Echtzeit, konnten dabei jedoch zwei Kameraperspektiven, den Monitor und eine Stoppuhr parallel einsehen. Nach der Beobachtung stand ihnen eine fünfminütige Nachbearbeitungszeit zur Verfügung, in der sie ihre Bewertungen vervollständigen und das Referenzmanual konsultieren konnten.

Die Interrater-Reliabilität des Gesamtscores erwies sich mit einem Intraklassenkorrelationskoeffizienten von 0,81 als gut. Auch mehrere Domänen erreichten zufriedenstellende Werte. Für basic skills lag der ICC bei 0,73, für circulation bei 0,76 und für human factors bei 0,68. Deutlich niedriger fiel der Wert für die Domäne airway and breathing mit 0,30 aus. Die Autoren führen dieses Ergebnis auf die geringe Varianz in dieser Domäne zurück, da sowohl die Expertenteams als auch die Residententeams in diesem Bereich vergleichsweise gut abschnitten. In einer solchen Situation können bereits wenige Unterschiede zwischen den Ratern einen erheblichen Einfluss auf den ICC haben. Die niedrigere Reliabilität der Atemwegsdomäne wird

deshalb eher als Folge der konkreten Datenstruktur, denn als grundsätzliche Schwäche des Konzepts interpretiert. Insgesamt sprechen die Ergebnisse jedoch dafür, dass der STAT auf Ebene des Gesamtscores und in mehreren zentralen Domänen eine gute Beurteilerübereinstimmung ermöglicht.

Ein wichtiges Merkmal des Instruments ist seine detaillierte und granulare Erfassung von Teamleistung. Die Autoren betonen, dass der STAT ursprünglich als Forschungsinstrument entwickelt wurde und die Möglichkeit bietet, eine große Zahl relevanter Einzelelemente pädiatrischer Reanimationsleistung differenziert zu analysieren. Diese Stärke ist jedoch zugleich mit praktischen Herausforderungen verbunden. Die hohe Anzahl an Items erfordert Training, damit Bewerter das Instrument verlässlich anwenden können. Nach Angaben des Artikels konnten die Rater mit entsprechender Schulung den STAT erfolgreich verwenden, obwohl jede Simulation nur einmal angesehen werden durfte und die Bewertung in einem eng begrenzten Zeitfenster erfolgte. Zugleich wird berichtet, dass die Autoren das Instrument nach Abschluss der Studie auch für Echtzeitbewertungen simulierter Reanimationen eingesetzt haben. Dabei zeigte sich, dass die große Zahl an Elementen sowie eingeschränkte Sichtverhältnisse die Genauigkeit der Bewertung beeinträchtigen können. Für eine valide Echtzeitanwendung sei deshalb eine weitere Entwicklungsphase erforderlich, in der besonders kritische und diskriminierende Elemente identifiziert oder das Instrument stärker generalisiert wird.

Hinsichtlich der Anwendungsbereiche ist der STAT in erster Linie für simulationsbasierte Trainings- und Forschungssettings in der pädiatrischen Reanimation vorgesehen. Das Instrument eignet sich insbesondere für standardisierte pädiatrische Notfallszenarien, in denen Teamleistung umfassend und differenziert beurteilt werden soll. Die Autoren sehen den Nutzen des Instruments darin, Teamkompetenz in pädiatrischen Reanimationssituationen sichtbar zu machen und die verschiedenen technischen und nichttechnischen Komponenten dieses Handelns systematisch zu analysieren. Der STAT kann damit sowohl in der Aus- und Weiterbildung als auch in wissenschaftlichen Untersuchungen eingesetzt werden. Zugleich wird deutlich, dass das Instrument aufgrund seiner Komplexität derzeit vor allem für analysierende und strukturierte Simulationskontexte geeignet ist und noch nicht ohne Weiteres als effizientes Echtzeitinstrument im Routineeinsatz betrachtet werden kann.

Die Studie benennt mehrere Limitationen. Eine wesentliche Einschränkung besteht in der geringen Anzahl der untersuchten Teams, da nur zwei Expertenteams und zwei Residententeams einbezogen wurden. Dies reduziert die Varianz der Daten und erschwert insbesondere die psychometrische Prüfung einzelner Domänen. Die Autoren führen dies auch als mögliche Erklärung für die schwächeren Ergebnisse im Bereich airway and breathing an und empfehlen

daher weitere Untersuchungen mit mehr Teams, mehreren Institutionen und einem breiteren Spektrum an Erfahrungshintergründen. Eine weitere Limitation liegt im Fokus auf den ärztlichen Teamanteil. Zwar wurden in allen Teams standardisierte Pflegekräfte eingesetzt, diese handelten jedoch nicht autonom. Auf diese Weise sollte der Einfluss unterschiedlicher pflegerischer Erfahrung kontrolliert und die Konstruktvalidität zwischen Ärzteteams besser geprüft werden. Gleichzeitig bleibt damit offen, wie gut der STAT die Leistung multiprofessioneller Reanimationsteams abbildet. Die Validierung in interdisziplinären Teams wird deshalb als wichtiger nächster Entwicklungsschritt benannt.

Darüber hinaus wurde die Studie an nur einer Institution durchgeführt. Die Rater waren nicht gegenüber den Teilnehmenden Personen verblindet und kannten die bewerteten Residents, Fellows und Attendings teilweise aus dem klinischen Alltag. Die Autoren weisen darauf hin, dass diese Einschränkung auch in anderen vergleichbaren Studien vorliegt. Zur Reduktion möglicher Verzerrungen wurden die Bewertungen jedoch pseudonymisiert erfasst, sodass individuelle Raterwerte nicht einzelnen Personen zugeordnet werden konnten. Dennoch bleibt eine potenzielle Verzerrung durch Vorwissen über die Teilnehmer bestehen. Eine weitere Limitation ergibt sich daraus, dass nur ein einziges medizinisches Simulationsszenario untersucht wurde. Obwohl dieses Szenario ein breites Spektrum grundlegender, atemwegsbezogener, zirkulatorischer und teambezogener Anforderungen enthielt, ist die Generalisierbarkeit auf andere pädiatrische Reanimationssituationen noch nicht gesichert. Der Artikel fordert deshalb eine Prüfung des Instruments in einer größeren Bandbreite pädiatrischer Notfallszenarien.

Auch die ungewichtete Behandlung aller 94 Items wird als Einschränkung diskutiert. In der klinischen Realität dürften einzelne Maßnahmen einen größeren Einfluss auf den Patientenausgang haben als andere. Im STAT werden jedoch alle Elemente gleichbehandelt. Zudem enthält das Instrument keine szenariospezifischen Elemente, etwa die Auswahl bestimmter Medikamente oder hochspezialisierter Maßnahmen, die in einzelnen Notfällen relevant sein könnten. Die Autoren schlagen deshalb vor, künftige Versionen des Instruments, um szenariospezifische Komponenten zu ergänzen, um die diagnostische Genauigkeit weiter zu erhöhen.

Ergänzend wurden auch die subjektiven Einschätzungen der Teilnehmer zur Simulationssitzung erhoben. Diese Ergebnisse zeigen, dass der Fall von den Teilnehmern als hoch relevant, realistisch und lernwirksam wahrgenommen wurde. Die Bewertungen der einzelnen Aussagen lagen im Mittel durchgehend zwischen 4,33 und 4,92 auf einer fünfstufigen Likert-Skala. Insbesondere der Nutzen des Debriefings sowie der wahrgenommene Lernwert für zukünftige Reanimationen wurden hoch eingeschätzt. Diese Befunde betreffen zwar nicht direkt die

psychometrische Qualität des Instruments selbst, stützen jedoch dessen Einbettung in ein als lehrreich und praxisnah empfundenes Simulationssetting.

Zusammenfassend lässt sich festhalten, dass der Simulation Team Assessment Tool ein detailliert strukturiertes, verhaltensverankertes und inhaltlich breit angelegtes Instrument zur Erfassung der Teamleistung in simulierten pädiatrischen Reanimationssituationen darstellt. Seine Entwicklung erfolgte theorie- und expertenbasiert, und die Skala integriert vier zentrale Leistungsbereiche, nämlich grundlegende Reanimationsfertigkeiten, Atemwegs- und Beatmungsmanagement, Kreislaufmanagement sowie Human Factors. Die psychometrischen Befunde sprechen für eine gute Inhaltsvalidität, eine überzeugende Konstruktvalidität des Gesamtscores sowie mehrerer Domänen und eine gute Interrater-Reliabilität des Gesamtscores. Einschränkungen bestehen insbesondere hinsichtlich der Atemwegsdomäne, der kleinen Stichprobe, der monozentrischen Durchführung, des fehlenden multiprofessionellen Fokus sowie der praktischen Komplexität des Instruments. Insgesamt erscheint der STAT jedoch als vielversprechendes Instrument zur differenzierten Analyse von Teamkompetenz in der pädiatrischen Reanimationssimulation und damit als wertvoller Beitrag zur simulationsbasierten medizinischen Ausbildung und Forschung.

5.21 Trauma Non-Technical Skills Scale (T-NOTECHS) 2012

Quelle: Steinemann S, Berg B, DiTullio A, Skinner A, Terada K, Anzelon K, et al. Assessing teamwork in the trauma bay: introduction of a modified “NOTECHS” scale for trauma. Am J Surg. (2012) 203:69–75. doi: 10.1016/j.amjsurg.2011.08.004

Abbildung 24: Trauma Non-Technical Skills Scale (T-NOTECHS) 2012

Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Steinemann et al. (2012)

Das Instrument Trauma NOTECHS (T-NOTECHS) wurde entwickelt, um die Teamarbeit multidisziplinärer Traumareanimationsteams systematisch zu erfassen und zu bewerten. Der Entwicklung lag die im Beitrag formulierte Annahme zugrunde, dass Traumareanimationen hochkomplexe und zeitkritische Versorgungssituationen darstellen, in denen die koordinierte Zusammenarbeit unterschiedlicher Berufsgruppen eine zentrale Voraussetzung für eine qualitativ hochwertige Patientenversorgung ist. Obwohl die Bedeutung von Teamarbeit in der Traumaversorgung seit Langem anerkannt ist, sind entsprechende nichttechnische Kompetenzen nach den Ausführungen der Autoren in etablierten Traumacurricula bislang unterrepräsentiert. Zugleich bestand ein Bedarf an einem Instrument, das die wesentlichen Human Factors der Traumaversorgung valide erfasst, sowohl in realen als auch in simulierten Kontexten einsetzbar ist, von unterschiedlichen Bewertern genutzt werden kann und sich darüber hinaus auch für Rückmeldung, Selbstbeurteilung und Debriefing eignet. Vor diesem Hintergrund wurde mit T-NOTECHS eine modifizierte Version der ursprünglichen NOTECHS-Skala speziell für den Kontext der Traumaversorgung entwickelt.

Die Entwicklung des Instruments erfolgte auf Grundlage bereits etablierter Konzepte zur Erfassung nichttechnischer Fertigkeiten. Ausgangspunkt war die aus der Luftfahrt stammende NOTECHS-Systematik, die zuvor bereits für den Operationssaal modifiziert worden war. Zur

Konzeption von T-NOTECHS wurde ein interprofessionelles Expertengremium aus zwei Institutionen gebildet, das aus zwei Trauma- und Intensivchirurgen, einem trauma- beziehungsweise intensivmedizinisch tätigen Internisten sowie zwei trauma- und intensivmedizinischen Pflegefachpersonen bestand. Die Autoren betonen, dass dieses Panel kumulativ über mehr als 80 Jahre klinische Erfahrung in der Traumaversorgung verfügte und alle ärztlichen Mitglieder als Advanced Trauma Life Support-Instruktoren sowie intensivmedizinisch qualifiziert waren. In die Entwicklung flossen publizierte Instrumente zur Traumateam-Evaluation, die Nomenklatur essenzieller Teamkompetenzen sowie die modifizierte NOTECHS-Version aus dem operativen Kontext ein. Im Zuge der inhaltlichen Überarbeitung kam das Panel zu dem Schluss, dass für den Traumakontext insbesondere die Aspekte Rollenübernahme und Aufgabenbewältigung sowie der Umgang mit Stress und Unterbrechungen stärker betont werden müssten als in der Ausgangsversion. Auf dieser Grundlage wurden die fünf ursprünglichen Verhaltensdomänen der NOTECHS-Skala inhaltlich angepasst, während die fünfstufige Likert-Skala beibehalten wurde.

Die Struktur des T-NOTECHS umfasst fünf Domänen, die als zentrale Verhaltensbereiche effektiver Teamarbeit in der Traumaversorgung verstanden werden. Diese Domänen sind Leadership, Cooperation and Resource Management, Communication and Interaction, Assessment and Decision Making sowie Situation Awareness/Coping with Stress. Jede Domäne wird auf einer Skala von 1 bis 5 bewertet, wobei die Extremwerte und der mittlere Wert durch verhaltensnahe Anker beschrieben werden. In der Domäne Leadership steht die höchste Ausprägung für eine jederzeit klar erkennbare Teamleitung, gutes Zeitmanagement und die vollständige Erledigung aller Aufgaben auf nicht-hierarchische Weise, während die niedrigste Ausprägung dadurch gekennzeichnet ist, dass die Identität der Teamleitung unklar bleibt. Cooperation and Resource Management bezieht sich auf die Klarheit von Rollen, die Verteilung und Erfüllung von Aufgaben sowie den situationsgerechten Einsatz personeller Ressourcen. Communication and Interaction beschreibt insbesondere die Klarheit, Organisation und Zentralität der Kommunikation über die Teamleitung, die Weitergabe kritischer Informationen und die Nutzung geschlossener Kommunikationsschleifen. Assessment and Decision Making fokussiert auf die geordnete und vollständige Durchführung der Primär- und Sekundäruntersuchung, die Zusammenfassung von Befunden sowie die Kommunikation des Behandlungsplans. Situation Awareness/Coping with Stress schließlich erfasst die Fähigkeit des Teams, trotz unerwarteter Befunde, externer Störungen oder sich verschlechternder Patientenbedingungen systematisch, ruhig und vorausschauend zu handeln.

Ergänzend zu diesen fünf Domänen wurden exemplarische Verhaltensweisen entwickelt, die die Anwendung der Skala unterstützen sollten. Zunächst wurden 26 exemplarische Verhal-

tensweisen formuliert, die jedoch im Zuge der klinischen Erprobung überarbeitet wurden. Zwei Beispiele wurden verworfen, drei neue aufgenommen, sodass die endgültige Version des Instruments 27 exemplarische Verhaltensweisen umfasste. Die Autoren betonen ausdrücklich, dass diese Verhaltensbeispiele nicht als eigenständige Checklistenitems gedacht waren, sondern der Veranschaulichung der jeweiligen Domänen dienen sollten. Auf diese Weise sollte die Skala schnell und praktikabel anwendbar bleiben. Die Beispiele verdeutlichen, welche konkreten Verhaltensweisen in der Bewertung der jeweiligen Domäne berücksichtigt werden sollen. Im Bereich Leadership zählen hierzu etwa das klare Erkennen der Teamleitung, die Delegation von Aufgaben, das Bewahren des Überblicks, die Einbindung anderer Teammitglieder und die Durchführung von Briefing und Debriefing. Cooperation and Resource Management wird unter anderem durch die eindeutige Zuordnung von Rollen, die Erfüllung zugewiesener Aufgaben sowie das flexible Umverteilen von Arbeitslast veranschaulicht. Communication and Interaction wird beispielsweise durch lautes Vorlesen des Rettungsdienstberichts, das Verbalisieren aller kritischen Informationen, klare und hörbare Kommunikation sowie geschlossene Kommunikationsschleifen beschrieben. Assessment and Decision Making wird durch die vollständige und geordnete Durchführung der Surveys, die Zusammenfassung von Verletzungen und physiologischen Befunden sowie das Kommunizieren des nächsten Schritts konkretisiert. Situation Awareness/Coping with Stress wird unter anderem über das ruhige Verhalten des Teams, die Antizipation systemischer Probleme und die angemessene Reaktion auf unerwartete Befunde oder Störungen abgebildet.

Ein wesentliches Merkmal des T-NOTECHS liegt darin, dass es als einheitliches Instrument in verschiedenen Anwendungskontexten eingesetzt wurde. Die Autoren beschreiben, dass dieselbe Skala sowohl in simulationsbasierten Trainings als auch in realen Traumareanimationen verwendet wurde. In den Simulationssettings diente T-NOTECHS sowohl der unmittelbaren anonymen Bewertung der Teamleistung vor der Videoreflexion als auch als Strukturierungsinstrument für das anschließende Debriefing. Im klinischen Alltag wurde die Skala in Echtzeit durch traumaerfahrene Pflegefachpersonen und Forschungsassistenten genutzt. Nach Abschluss der Teamtrainings wurde sie darüber hinaus auch von Teamleitungen und verantwortlichen Traumatologen zur Selbstbeurteilung eingesetzt. Zudem diente T-NOTECHS in der realen Versorgung als Grundlage für eine kurze Rückmeldung an die Teamleitung. Die Autoren heben hervor, dass der gesamte Prozess der Bewertung und Rückmeldung im klinischen Alltag in der Regel nur fünf bis zehn Minuten erforderte.

Die Einführung des Instruments wurde durch ein strukturiertes Ratertraining begleitet. Zwölf additional trauma- und intensivmedizinische Pflegefachpersonen, die im klinischen Alltag als Dokumentationskräfte während der Traumaversorgung tätig waren, wurden als primäre Rater

rekrutiert. Ergänzt wurde diese Gruppe durch drei medizinische Forschungsassistenten. Das Training umfasste vorbereitende Lektüre zu Teamkompetenzen sowie zu spezifischen Rollen und Aufgaben in Traumateams und einen einstündigen Workshop, in dem die Domänen des Instruments und ihre Bedeutung erläutert wurden. Anschließend bewerteten die Teilnehmer videografierte simulierte Traumareanimationen mit unterschiedlich ausgeprägter Teamleistung und nahmen an Debriefings teil. Im Verlauf der ersten 6,5 Monate der klinischen Erprobung wurden zusätzlich drei 20-minütige Trainingssitzungen durchgeführt, in denen klinische Beispiele diskutiert, Verhaltensbeispiele weiter geschärft und ungewöhnliche Bewertungen mit Blick auf normative Einschätzungen besprochen wurden. Dieser begrenzte Schulungsaufwand wird von den Autoren als Hinweis auf die relativ schnelle Adaption und intuitive Anwendbarkeit der Skala interpretiert.

Hinsichtlich der psychometrischen Eigenschaften wurde zunächst die interne Struktur des Instruments betrachtet. Dabei zeigte sich, dass die Werte der fünf Domänen stark miteinander zusammenhingen. Cronbachs Alpha lag für alle Rater über 0,90 und verbesserte sich nicht durch das Entfernen einzelner Domänen. Die Autoren schlussfolgern daraus, dass die fünf Teilbereiche des Instruments eng miteinander verbunden sind und sich gut zu einem Gesamtscore zusammenfassen lassen. Entsprechend wurde in den weiteren Analysen primär der Gesamtsummenscore verwendet, der Werte zwischen 5 und 25 Punkten annehmen konnte. Diese hohe interne Konsistenz kann als Hinweis auf eine homogene Erfassung des übergeordneten Konstrukts Teamarbeit verstanden werden.

Die Interrater-Reliabilität wurde in mehreren Settings untersucht. Für T-NOTECHS-Bewertungen in Echtzeit während simulierter Traumareanimationen mit drei Ratern ergab sich ein Intraklassenkorrelationskoeffizient von 0,44. Für Echtzeitbewertungen realer Traumareanimationen mit zwei Ratern lag der ICC bei 0,48. Diese Werte deuten auf eine lediglich moderate Übereinstimmung hin. Eine höhere Reliabilität wurde für videobasierte Nachbewertungen simulierter Traumareanimationen durch zwei ärztliche Debriefer berichtet. In diesem Fall lag der ICC für den Gesamtscore bei 0,71. Damit war die Übereinstimmung bei Videoreview deutlich höher als bei der unmittelbaren Echtzeitbeurteilung. Die Autoren führen dies auf mehrere Faktoren zurück, darunter geringere Ablenkung außerhalb des Schockraums, die Möglichkeit, sich ausschließlich auf Teamarbeit zu konzentrieren, sowie die größere Erfahrung der Bewerter/Debriefer. Zugleich wird berichtet, dass die größte Variabilität in der Domäne Assessment and Decision Making auftrat, für die ein ICC von 0,33 ermittelt wurde. Nach Einschätzung der Autoren könnte dies darauf zurückzuführen sein, dass in dieser Domäne schwer zu unterscheiden ist, ob die Teaminteraktion im Entscheidungsprozess oder die inhaltliche Qualität der klinischen Entscheidung selbst bewertet wird.

Ein weiterer wichtiger Aspekt der psychometrischen Analyse betrifft die Korrelation des Instruments mit objektiven klinischen Leistungsparametern. In den 33 simulierten Traumareanimationen zeigten höhere T-NOTECHS-Scores der erfahrensten Rater signifikante Zusammenhänge mit einer besseren klinischen Leistung. So korrelierten höhere Werte mit einer größeren Anzahl abgeschlossener Reanimationsaufgaben mit einem Korrelationskoeffizienten von 0,50 bei einem Signifikanzniveau von $p < 0,01$. Darüber hinaus bestand ein signifikanter Zusammenhang mit einer schnelleren Durchführung der drei in allen Szenarien gemeinsamen Schlüsselaufgaben, wobei der Korrelationskoeffizient 0,38 bei $p < 0,05$ betrug. Diese Befunde deuten darauf hin, dass höhere Bewertungen der Teamarbeit mit einer effizienteren und vollständigeren Abarbeitung zentraler Reanimationsschritte in der Simulation einhergingen. Demgegenüber zeigten Selbstbewertungen durch die verantwortlichen Traumachirurgen keine signifikanten Korrelationen mit Geschwindigkeit, Vollständigkeit oder Berichtsraten der Aufgaben, weder in Simulationen noch in realen Reanimationen. Dies spricht dafür, dass Fremdbewertungen durch geschulte Beobachter im vorliegenden Kontext aussagekräftiger waren als Selbsteinschätzungen.

Auch in der klinischen Praxis ergaben sich Hinweise auf die Relevanz der Skala. In insgesamt 244 realen Traumareanimationen verbesserten sich die von den traumaerfahrenen Pflegefachpersonen vergebenen T-NOTECHS-Scores nach einem simulationsbasierten Teamtraining signifikant. Der Mittelwert stieg von 16,3 auf 17,7 Punkte, wobei dieser Unterschied mit $p < 0,001$ hochsignifikant war. Besonders ausgeprägt war die Verbesserung in der Domäne Communication, deren Mittelwert von 3,05 auf 3,33 anstieg und ebenfalls statistisch signifikant war. Darüber hinaus korrelierten höhere T-NOTECHS-Werte mit einer kürzeren Gesamtzeit der Reanimation in der Notaufnahme mit einem Korrelationskoeffizienten von -0,13 bei $p < 0,05$ sowie mit einer geringeren Anzahl nicht berichteter Aufgaben aus Primär- und Sekundärsurvey mit einem Korrelationskoeffizienten von -0,16 bei $p < 0,05$. Obwohl diese Korrelationen relativ klein ausfielen, interpretieren die Autoren sie als Hinweis auf die klinische Relevanz der Skala, da bessere Teamarbeit mit einer effizienteren und vollständigeren Versorgung verbunden war.

Die Anwendungsbereiche des T-NOTECHS liegen damit sowohl im Bereich simulationsbasierter Teamtrainings als auch in der realen Traumaversorgung. Das Instrument kann zur Beobachtung und Bewertung von Teams in Echtzeit, zur videobasierten Nachanalyse, zur Strukturierung von Debriefings sowie zur Selbst- und Fremdbeurteilung eingesetzt werden. Besonders hervorgehoben wird seine Funktion als Feedbackinstrument im Rahmen von Teamtrainings und klinischen Nachbesprechungen. Durch seine domänenbezogene Struktur und die exemplarischen Verhaltensanker bietet es eine gemeinsame Sprache für die Reflexion von

Teamarbeit im Traumakontext. Gleichzeitig zeigt der Einsatz in der klinischen Praxis, dass das Instrument prinzipiell mit begrenztem Zeitaufwand in den Arbeitsalltag integriert werden kann. Trotz dieser positiven Befunde weisen die Autoren auf mehrere Limitationen hin. Eine wesentliche Einschränkung betrifft die Möglichkeit von Verzerrungen bei der Echtzeitbewertung. Es sei denkbar, dass Rater eine schnellere Aufgabenbearbeitung wahrnahmen und dies ihre Einschätzung der Teamarbeit in sämtlichen Domänen beeinflusste. Um diesen Effekt in den Simulationen zu reduzieren, wurden die Debriefer vor der Bewertung nicht mit den zeitgestempelten Protokollen der klinischen Interventionen konfrontiert. In der realen Versorgung wurden die Zeiten erst nachträglich aus dem Dokumentationsbogen extrahiert, sodass die primären Rater während der Versorgung nicht gezielt auf Zeitunterschiede fokussiert waren. Gleichwohl bleibt ein potenzieller Zusammenhang zwischen der Wahrnehmung zügiger Versorgung und einer positiveren globalen Teambeurteilung bestehen.

Eine weitere zentrale Einschränkung liegt in der nur moderaten Interrater-Reliabilität der Echtzeitbewertungen. Die Autoren verweisen darauf, dass die Übereinstimmung zwischen Ratern unterschiedlicher Hintergründe bestimmt wurde. Während die traumaerfahrenen Pflegefachpersonen den klinischen Ablauf gut kannten, verfügten die medizinischen Forschungsassistenten nur über begrenzte Kenntnisse der Traumaversorgung. Dies könnte die Nachvollziehbarkeit der Abläufe und damit die Vergleichbarkeit der Bewertungen beeinträchtigt haben. Hinzu kommt, dass die Bewerter Traumachirurgen in der Simulationsphase vor der Selbstbeurteilung keine spezifische Schulung im Umgang mit dem Instrument erhalten hatten. Darüber hinaus war die gesamte Raterqualifizierung mit einer einstündigen Einführungsveranstaltung und einigen kurzen Nachschulungen relativ knapp bemessen. Im Vergleich zu anderen Verfahren, für deren reliablen Einsatz eine deutlich intensivere Schulung erforderlich war, erscheint dieser Trainingsumfang begrenzt. Die Autoren sehen darin eine mögliche Erklärung für die suboptimalen Reliabilitätswerte und folgern, dass für hoch belastbare Bewertungen insbesondere im klinischen Echtzeiteinsatz ein längeres und stärker spezialisiertes Ratertraining erforderlich sein könnte.

Als zusätzliche Limitation wird die konzeptuelle Schwierigkeit der Domäne Assessment and Decision Making diskutiert. Da diese Domäne sowohl Aspekte der Teaminteraktion im Entscheidungsprozess als auch Merkmale der inhaltlichen Qualität der klinischen Entscheidungen berührt, könnte sie Bewerter dazu verleiten, nicht nur Teamarbeit, sondern auch die medizinische Korrektheit der Versorgung zu beurteilen. Diese Vermischung unterschiedlicher Bewertungsdimensionen könnte die geringere Reliabilität in diesem Bereich erklären. Ferner weist die Studie darauf hin, dass trotz der beobachteten Zusammenhänge mit Prozessparametern keine signifikanten Korrelationen mit klassischen Outcomeparametern wie Mortalität oder Ver-

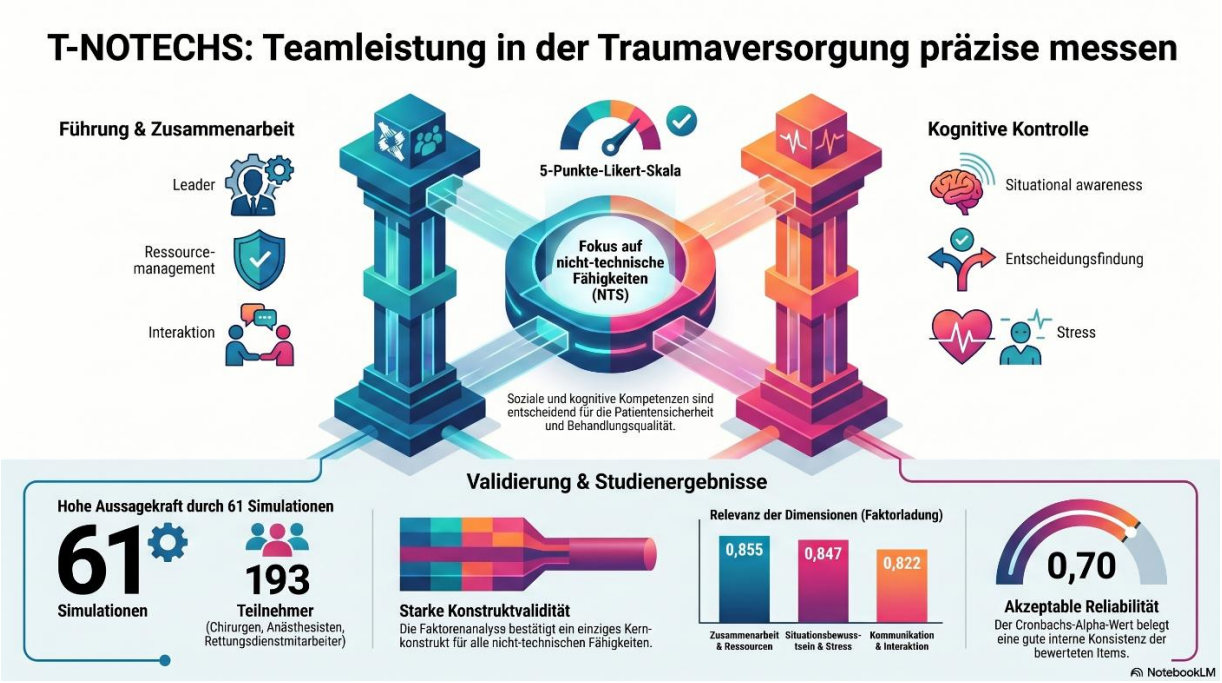
weildauer nachgewiesen werden konnten. Dies wird im Artikel vor allem mit der kleinen Fallzahl der beobachteten Reanimationen erklärt.

Zusammenfassend lässt sich festhalten, dass T-NOTECHS eine auf der ursprünglichen NOTECHS-Systematik basierende, an die Besonderheiten des Traumakontexts angepasste Skala zur Erfassung globaler Teamarbeit darstellt. Die fünf Domänen Leadership, Cooperation and Resource Management, Communication and Interaction, Assessment and Decision Making sowie Situation Awareness/Coping with Stress werden über eine fünfstufige Ratingskala beurteilt und durch exemplarische Verhaltensweisen konkretisiert. Das Instrument wurde in einem konsensbasierten interprofessionellen Entwicklungsprozess erstellt und in Simulations- wie Realkontexten erprobt. Die psychometrischen Ergebnisse zeigen eine hohe interne Konsistenz, Hinweise auf klinische Relevanz durch signifikante Zusammenhänge mit objektiven Prozessparametern sowie Verbesserungen nach Teamtraining. Gleichzeitig bleibt die Interrater-Reliabilität im klinischen Echtzeiteinsatz begrenzt, sodass weitere Untersuchungen und eine intensivere Schulung der Bewerter sinnvoll erscheinen. Insgesamt stellt T-NOTECHS jedoch ein vielversprechendes Instrument für Training, Beobachtung und Debriefing von Teamarbeit in der Traumaversorgung dar.

5.22 Trauma Non-Technical Skills Scale (T-NOTECHS) 2019

Quelle: Repo JP, Rosqvist E, Lauritsalo S, Paloneva J. Translatability and validation of non-technical skills scale for trauma (T-NOTECHS) for assessing simulated multi-professional trauma team resuscitations. BMC Med Educ. (2019) 19:40. doi: 10.1186/s12909-019-1474-5

Abbildung 25: Trauma Non-Technical Skills Scale (T-NOTECHS) 2019



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Repo et al. (2019)

Die Trauma Non-Technical Skills Scale (T-NOTECHS) wurde zur Erfassung nichttechnischer Fertigkeiten in Traumateam-Reanimationen entwickelt und in der vorliegenden Studie nicht neu konzipiert, sondern in eine nicht-angelsächsische Sprache übersetzt, kulturübergreifend adaptiert und psychometrisch geprüft. Im Mittelpunkt der Untersuchung stand somit die Frage, ob ein bereits validiertes englischsprachiges Instrument zur Beurteilung teambezogener Verhaltensaspekte in Traumareanimationen auch in einem sprachlich und kulturell deutlich unterschiedlichen Kontext anwendbar ist. Die Autoren begründen dieses Vorhaben mit der steigenden Bedeutung nichttechnischer Fertigkeiten in der Traumaversorgung sowie mit dem Bedarf an standardisierten Instrumenten, die internationale Vergleiche von Trainings- und Ausbildungseffekten ermöglichen. Da in Finnland trotz aktiver simulationsbasierter Traumateamtrainings kein geeignetes Instrument zur Erfassung dieser Kompetenzen verfügbar war, sollte die T-NOTECHS in eine finnische Version überführt und auf ihre psychometrische Eignung geprüft werden.

Die zugrunde liegende Ausgangsversion der T-NOTECHS wird im Beitrag als ein Instrument mit fünf Verhaltensdimensionen beschrieben. Diese fünf Domänen umfassen Leadership, Cooperation and Resource Management, Communication and Interaction, Assessment and De-

cision-Making sowie Situation Awareness/Coping with Stress. Jedes dieser Items wird auf einer fünfstufigen Skala bewertet, wobei ein Wert von eins für eine schlechte und ein Wert von fünf für eine exzellente Teamleistung steht. Die niedrigste Ausprägung bedeutet, dass das Team das angestrebte Verhalten nicht gezeigt hat, während die höchste Ausprägung eine fehlerfreie Ausführung des betreffenden Teamverhaltens kennzeichnet. Die englische Originalversion ist in der Abbildung auf Seite 2 des Artikels dargestellt, während die finnische Endfassung in der Abbildung auf Seite 4 dokumentiert ist. Die Struktur des Instruments blieb im Übersetzungsprozess unverändert erhalten, sodass auch die finnische Version dieselben fünf Domänen umfasst und somit als globales Instrument zur Beurteilung nichttechnischer Teamleistung in Traumareanimationen verstanden werden kann.

Die Übersetzung und kulturelle Adaptation erfolgte in einem mehrstufigen, strukturierten Verfahren, das sich an eigens von den Autoren entwickelten Leitlinien zur Übersetzung nichttechnischer Fertigkeitsskalen orientierte. Dieser Prozess ist in der Abbildung auf Seite 3 grafisch dargestellt. Zunächst wurden zwei voneinander unabhängige Vorwärtsübersetzungen aus dem Englischen ins Finnische angefertigt. Diese wurden von zwei im Gesundheitswesen tätigen Personen erstellt, deren Muttersprache Finnisch war und fließend Englisch beherrschten. Beide Übersetzer dokumentierten Begriffe, Formulierungen und mögliche kulturelle Bezüge, die mehrdeutig sein oder missverstanden werden könnten. Anschließend wurden die beiden Übersetzungen in einer Konsensusphase miteinander abgeglichen und zu einer gemeinsamen Version zusammengeführt. In einem nächsten Schritt erfolgte eine Rückübersetzung aus dem Finnischen ins Englische durch einen englischen Muttersprachler mit guten Finnischkenntnissen, jedoch ohne medizinischen Hintergrund und ohne Kenntnis des Originalinstruments. Die Rückübersetzung diente der inhaltlichen Kontrolle, indem geprüft werden sollte, ob die finnische Fassung den semantischen Gehalt der Ausgangsversion erhalten hatte. Im Anschluss wurden Vorwärts- und Rückübersetzung von einem Expertengremium beurteilt, das sich aus einer landesspezifischen Schlüsselperson, einer Projektleitung, einem Orthopäden, einem Facharzt für Anästhesiologie und einer Anästhesiepflegekraft zusammensetzte. Dieses Gremium erarbeitete auf der Grundlage der vorliegenden Berichte eine präfinale Version der finnischen T-NOTECHS.

Diese präfinale Fassung wurde anschließend einem Pretest unterzogen. Fünf Gesundheitsfachpersonen, darunter zwei Traumapfleger, eine Fachperson der Notfallmedizin, eine Leitungsfunktion im Rettungsdienst sowie eine Pflegefachperson aus der Primärversorgung, überprüften das Instrument in zwei multiprofessionellen High-Fidelity-In-situ-Traumasimulationen. Im Anschluss an diese Erprobung wurden kognitive Debriefings durchgeführt, um Prob-

leme im Verständnis der Items, der Erläuterungen und der Antwortoptionen zu identifizieren sowie mögliche sprachliche Umformulierungen zu erfassen. Die Ergebnisse des Pretests zeigten, dass die T-NOTECHS grundsätzlich gut verständlich und einfach auszufüllen war. Drei der fünf Testpersonen bezeichneten die Skala ausdrücklich als verständlich und leicht anwendbar. Zwei Testpersonen berichteten jedoch Schwierigkeiten mit den Begriffen im Zusammenhang mit der Domäne Leadership. Außerdem wurde im Item Communication and interaction das Wort „hub“ als im Finnischen missverständlich beurteilt und deshalb gestrichen, ohne den inhaltlichen Gehalt der Aussage zu verändern. Zusätzlich erfolgte eine geringfügige Umformulierung der dritten Antwortkategorie im Item Assessment and decision making. Nach Überarbeitung dieser Punkte bestätigten alle Mitglieder des Komitees sowie die Pretester, dass die finnische Version die Inhalte der Originalfassung angemessen abbilde. Auf dieser Grundlage leiten die Autoren eine gute Face Validity der finnischen T-NOTECHS ab.

Die empirische Prüfung des übersetzten Instruments erfolgte anhand von 61 In-situ-Simulationen multiprofessioneller Traumateams mit insgesamt 193 Teilnehmern. Die Untersuchung wurde in einem finnischen Zentralkrankenhaus durchgeführt, das für eine Bevölkerung von etwa 270.000 Menschen die Traumaversorgung sicherstellt. Im dortigen Traumateam arbeiten mindestens eine chirurgische Fachperson, eine anästhesiologische Fachperson, eine radiologische Fachperson, eine Traumapflegekraft sowie eine weitere Pflegekraft als Arbeitsassistenz für die Anästhesie zusammen. Die strukturierte Traumasimulation, in der das Instrument geprüft wurde, war in den klinischen Alltag eingebettet und wurde in der Notaufnahme des Krankenhauses in einer realitätsnahen Umgebung durchgeführt. Die Teilnehmer nahmen dabei ihre tatsächlichen professionellen Rollen ein. Das zweistündige Simulationsformat umfasste eine Einführung in die Methode, eine kurze Vorlesung, die Rollenverteilung, eine erste Simulation, ein Debriefing, eine zweite Simulation sowie ein abschließendes Debriefing. Insgesamt kamen vier unterschiedliche Traumaszenarien zum Einsatz, die auf Seite 5 in Tabelle 1 beschrieben werden. Diese umfassten unter anderem einen Patienten mit Stichverletzungen und Spannungspneumothorax, einen Sturz aus vier Metern mit Beckeninstabilität, einen Fahrradunfall sowie einen schwer verbrannten Patienten. Die Szenarien deckten verschiedene interventionelle Anforderungen wie Thorakozentese, intraossären Zugang, FAST, Blasenkatheter, Beckenschlinge, Crisis Resource Management und Escharotomie ab. An den Kursen nahmen insgesamt 193 Personen teil, darunter Anästhesiologen, Chirurgen, Pädiater, Notfallmedizin-Residents, Pflegefachpersonen und Pflegestudenten. Das mittlere Alter der Teilnehmer lag bei 37 Jahren, die durchschnittliche Berufserfahrung in der aktuellen Position bei 7,4 Jahren. Die durchschnittliche Teilnahme an Traumateamsimulationen lag bei fünf, die durchschnittliche Teilnahme an realen Traumateam-Reanimationen bei elf Einsätzen.

Die psychometrische Datenerhebung erfolgte durch zwei erfahrene Rater, nämlich einen Anästhesiologen und eine Traumapflegefachperson. Beide verfügten über umfangreiche Erfahrung in der Durchführung von Traumasimulationen in der untersuchten Klinik. Vor Beginn der Erhebung wurden sie umfassend in das Konzept nichttechnischer Fertigkeiten, in die Beobachtung und Bewertung koordinierten Teamverhaltens sowie in die spezifische Anwendung des T-NOTECHS eingeführt. Darüber hinaus hatten beide bereits vor Beginn der Studie Erfahrungen mit der finnischen T-NOTECHS gesammelt und ihre Bewertungen kalibriert. Die Beurteilung jeder Simulation erfolgte direkt nach dem Ende des Szenarios, unabhängig voneinander und ohne nachträgliche Diskussion mit den Teilnehmern oder untereinander.

Zur psychometrischen Prüfung wurden Floor- und Ceiling-Effekte, interne Konsistenz, Interrater-Reliabilität, absolute Reliabilität und Konstruktvalidität analysiert. Die Autoren formulierten im Vorfeld vier Hypothesen, wonach Floor- und Ceiling-Werte jeweils maximal 15 % betragen, Cronbachs Alpha zwischen 0,70 und 0,90 liegen, der Intraklassenkorrelationskoeffizient über 0,40 ausfallen und das Instrument auf einen Faktor laden sollte. Laut den Angaben auf Seite 6 wurden alle diese vordefinierten Hypothesen bestätigt.

Die Analyse der Floor- und Ceiling-Effekte ergab zunächst, dass weder auf Itemebene noch auf Gesamtscoreebene ein Floor-Effekt auftrat. Das bedeutet, dass keine der Bewertungen den minimal möglichen Wert erreichte. Auf Itemebene zeigten sich jedoch bei einzelnen Domänen Ceiling-Effekte. Für Rater 1 erreichte das Item Leadership bei 20 % der Bewertungen den Maximalwert, und Situation awareness/coping with stress lag bei 15 %. Bei Rater 2 zeigten sich höhere Maximalwertanteile, darunter 31 % bei Leadership, 18 % bei Cooperation and Resource Management, 18 % bei Communication and Interaction, 16 % bei Assessment and Decision Making sowie 49 % bei Situation awareness/coping with stress. Trotz dieser auffälligen Maximalwertverteilungen auf einzelner Itemebene zeigte der Gesamtscore keinen Ceiling-Effekt. Die Prozentsätze maximaler Gesamtwerte lagen nur bei 1,6 % für Rater 1 und 4,9 % für Rater 2. Die Autoren folgern daraus, dass das Instrument insgesamt nicht durch eine problematische Konzentration der Bewertungen am oberen Ende der Skala verzerrt wird, einzelne Domänen jedoch bei der Interpretation mit Vorsicht betrachtet werden sollten.

Die interne Konsistenz der finnischen T-NOTECHS erwies sich mit einem Cronbach-Alpha von 0,70 als akzeptabel. Dieser Wert wird von den Autoren als ausreichender Hinweis auf die Homogenität der Skala interpretiert. Die korrigierten Item-Gesamt-Korrelationen lagen bei Rater 1 zwischen 0,58 und 0,84 und bei Rater 2 zwischen 0,67 und 0,80. Die höchsten Zusammenhänge mit dem Gesamtscore zeigten bei Rater 1 die Domäne Assessment and decision making und bei Rater 2 die Domäne Leadership. Die Mediane der korrigierten Item-Gesamt-Kor-

relationen lagen bei 0,68 beziehungsweise 0,69. Nach Auffassung der Autoren weisen diese Ergebnisse darauf hin, dass die einzelnen Domänen in sinnvoller Weise zum übergeordneten Konstrukt nichttechnischer Teamfertigkeiten beitragen, ohne dabei eine übermäßige inhaltliche Redundanz aufzuweisen.

Die Interrater-Reliabilität wurde mit einem Intraklassenkorrelationskoeffizienten in einem Zwei-Wege-Random-Effects-Modell mit absoluter Übereinstimmung untersucht. Die mittleren T-NOTECHS-Werte der beiden Rater betrugen 3,76 beziehungsweise 4,01, sodass sich im Durchschnitt eine Differenz von 0,25 Punkten ergab. Der ICC lag bei 0,54 mit einem 95 %-Konfidenzintervall von 0,34 bis 0,70. Nach der im Artikel verwendeten Klassifikation entspricht dieser Wert einer fairen Reliabilität. Ergänzend wurde die absolute Reliabilität über den Coefficient of Repeatability bestimmt, der bei 1,53 lag. Dieser Kennwert beschreibt jenen Bereich, innerhalb dessen die absolute Differenz zwischen zwei Bewertungen mit einer Wahrscheinlichkeit von 95 % liegt. Die Autoren führen aus, dass der Coefficient of Repeatability eine differenzierte Einschätzung der absoluten Übereinstimmung erlaubt und in diesem Fall auf eine moderate Streuung zwischen den Beurteilungen hinweist. Im Vergleich zur ursprünglichen englischen Validierung von Steinemann et al., in der ein ICC von 0,44 für Echtzeitbewertungen simulierter Traumateams berichtet worden war, fiel die Reliabilität der finnischen Version etwas günstiger aus. Die Autoren führen dies möglicherweise auf Unterschiede im Setting, in der Zahl der Simulationen oder in der Zusammensetzung der Rater zurück.

Die Konstruktvalidität wurde mittels explorativer Faktorenanalyse untersucht. Die Autoren gingen von der Hypothese aus, dass die T-NOTECHS ein einziges latentes Konstrukt, nämlich nichttechnische Fertigkeiten, erfassen sollte. Hierzu wurden die Bewertungen beider Rater in einer gemeinsamen Datenreihe zusammengeführt, um die Datenbasis für die Analyse zu vergrößern. Die Hauptkomponentenanalyse mit Varimax-Rotation ergab, dass die Skala auf einen dominanten Faktor lud. Der erste Faktor wies einen Eigenwert von 2,84 auf, während der zweite Faktor lediglich einen Eigenwert von 0,93 erreichte. Nach dem im Artikel genannten Kriterium wurde deshalb nur der erste Faktor als bestätigt angesehen. Der auf Seite 7 dargestellte Scree-Plot zeigt einen deutlichen Knick nach dem ersten Faktor und stützt damit die Einfaktorstruktur. Die Faktorladungen der einzelnen Domänen lagen zwischen 0,383 für Leadership und 0,855 für Cooperation and resource management. Communication and interaction, Assessment and decision making sowie Situation awareness/coping with stress wiesen ebenfalls hohe Faktorladungen auf. Die Autoren interpretieren diese Ergebnisse dahingehend, dass die T-NOTECHS in ihrer finnischen Version ein gemeinsames latentes Konstrukt abbildet und die Gesamtsumme der fünf Items als Indexscore verwendet werden kann.

Im Hinblick auf ihre Anwendungsbereiche wird die finnische T-NOTECHS im Artikel vor allem als Instrument zur Bewertung der Teamleistung in simulierten multiprofessionellen Traumateam-Reanimationen verstanden. Die Autoren sehen in ihr insbesondere ein geeignetes Verfahren zur Beurteilung der Wirksamkeit simulationsbasierter In-situ-Traumateamtrainings. Durch ihre Übersetzung in eine nicht-angelsächsische Sprache und die strukturierte Adaptation soll die Skala zudem die Vergleichbarkeit von Trainingsergebnissen zwischen Ländern und Sprachräumen fördern. In den Schlussfolgerungen wird ausdrücklich betont, dass das Instrument für Simulationen mit Chirurgen, Anästhesiologen sowie Residents geeignet erscheint. Darüber hinaus wird sein Potenzial für Benchmarking und internationale Kooperation hervorgehoben.

Gleichzeitig weist die Studie auf mehrere Limitationen hin. Eine wesentliche Einschränkung besteht darin, dass keine externen Vergleichsinstrumente zur Verfügung standen, mit denen eine konvergente Validität hätte geprüft werden können. Die Aussagekraft der Validitätsprüfung stützt sich daher vor allem auf Face Validity und auf die Ergebnisse der Faktorenanalyse. Zudem wurde das Instrument ausschließlich in simulierten Situationen und nicht in realen Traumareanimationen untersucht. Aussagen über seine Reliabilität und Gültigkeit in der klinischen Praxis bleiben daher offen. Die Autoren weisen ferner darauf hin, dass weitere Untersuchungen mit Rasch-Measurement-Techniken zusätzliche Erkenntnisse zur Konstruktvalidität liefern könnten. Ebenso sollte geprüft werden, inwieweit das Instrument zwischen hoher und niedriger Teamleistung unterscheiden kann und ob es sich auch für High-Stakes-Assessments eignet. Als weiterer kritischer Punkt ist zu nennen, dass einzelne Domänen, insbesondere Leadership und Situation awareness/coping with stress, erhöhte Ceiling-Werte aufwiesen, was ihre Sensitivität für Leistungsunterschiede im oberen Bereich einschränken könnte.

Als Stärken der Studie werden die vergleichsweise große Zahl an Simulationen, die nach den COSMIN-Kriterien als gute Stichprobengröße eingestuft wird, die sorgfältige Schulung der Rater, die Durchführung im realitätsnahen Krankenhausumfeld sowie die Nutzung mehrerer unterschiedlicher Traumaszenarien hervorgehoben. Hinzu kommt die detaillierte Dokumentation der einzelnen Phasen des Übersetzungs- und Adaptationsprozesses, die den methodischen Anspruch der Studie unterstreicht und als Modell für zukünftige Übersetzungen ähnlicher Instrumente dienen kann.

Zusammenfassend lässt sich festhalten, dass die T-NOTECHS in ihrer finnischen Fassung erfolgreich in eine sprachlich und kulturell deutlich vom Englischen abweichende Umgebung übertragen werden konnte. Die Struktur des Instruments mit fünf Domänen blieb erhalten, kleinere sprachliche Anpassungen dienten der Verbesserung der Verständlichkeit und externen

Validität. Die psychometrischen Ergebnisse sprechen für eine akzeptable interne Konsistenz, eine faire Interrater-Reliabilität und eine gute Konstruktvalidität in Form einer überwiegend eindimensionalen Faktorstruktur. Die Skala erscheint damit als geeignetes Instrument zur Erfassung nichttechnischer Teamleistung in simulierten multiprofessionellen Traumareanimationen. Zugleich machen die Ergebnisse deutlich, dass weitere Studien erforderlich sind, um die Anwendbarkeit in realen Versorgungssituationen, die konvergente Validität und die Differenzierungsfähigkeit des Instruments weiter abzusichern.

5.23 Team Emergency Assessment Measure (TEAM)

Quelle: Cooper S, Cant R, Porter J, Sellick K, Somers G, Kinsman L, et al. Rating medical emergency teamwork performance: development of the team emergency assessment measure (TEAM). Resuscitation. (2010) 81:446–52. doi: 10.1016/j. resuscitation.2009.11.027

Abbildung 26: Team Emergency Assessment Measure (TEAM)

Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Cooper et al. (2010)

Das Team Emergency Assessment Measure (TEAM) wurde mit dem Ziel entwickelt, ein valides, reliables und praktikables Instrument zur Beobachtung und Bewertung der Teamleistung

bei medizinischen Notfallsituationen und Reanimationen bereitzustellen. Den Ausgangspunkt der Entwicklung bildete die im Beitrag beschriebene Problemlage, dass die Qualität kardiopulmonaler Reanimationen und die Leistung medizinischer Notfallteams trotz etablierter Leitlinien und Trainingsangebote weiterhin verbesserungsbedürftig sind. Die Autoren führen aus, dass effektive Teamleistung nicht allein von technischen Fertigkeiten abhängig ist, sondern in erheblichem Maße auch von nichttechnischen Kompetenzen wie Führung, Teamarbeit, Situationsbewusstsein, Entscheidungsfindung und Aufgabenmanagement beeinflusst wird. Obwohl bereits verschiedene generische und berufsspezifische Instrumente zur Beurteilung nichttechnischer Leistungen existierten, fehlte zum Zeitpunkt der Untersuchung ein Verfahren, das spezifisch auf die Anforderungen von medizinischen Reanimations- und Notfallteams zugeschnitten war. Vor diesem Hintergrund wurde TEAM entwickelt, um Teamleistung in simulierten und klinischen Notfallsituationen strukturiert erfassen und zugleich die Grundlage für konstruktives Feedback und Debriefing schaffen zu können.

Die Entwicklung des Instruments erfolgte in mehreren aufeinanderfolgenden Schritten. Zunächst wurde eine umfassende Literaturrecherche zu bereits bestehenden Instrumenten der Teamleistungsmessung durchgeführt. Hierzu nutzten die Autoren unter anderem die Datenbanken Medline, ProQUEST und PsycINFO sowie einschlägige Fachwebseiten. Insgesamt wurden 17 Instrumente identifiziert, von denen 14 für die Entwicklung von TEAM als relevant beurteilt wurden. Diese Instrumente umfassten unterschiedliche Zielgruppen und Anwendungsbereiche, darunter das Emergency Team Dynamics Scale, das Anaesthetists' Non-Technical Skills System, NOTSS, die Ottawa Crisis Resource Management Global Rating Scale sowie die Mayo High Performance Teamwork Scale. Die vorhandenen Instrumente wurden daraufhin in Bezug auf Kategorien, Elemente und Einzelitems analysiert. Aus dieser Analyse resultierte zunächst eine Liste von 15 Teamwork-Elementen und 57 Items. Diese wurden in einem weiteren Schritt durch einen akkreditierten Reanimationsexperten auf 27 Items reduziert. Anschließend verfeinerte das interprofessionelle Forschungsteam die Struktur auf acht Elemente mit insgesamt elf Items, die drei übergeordneten Kategorien zugeordnet wurden. Auf Grundlage weiterer Rückmeldungen wurde schließlich noch ein globales Gesamturteil ergänzt. Die Entwicklung des Instruments war damit sowohl literaturbasiert als auch iterativ-expertengeleitet.

Das Forschungsteam, das den Entwicklungsprozess trug, bestand aus erfahrenen Klinikern und Wissenschaftlern mit Berufserfahrung zwischen 19 und 41 Jahren. Beteiligt waren eine Resuscitation Officer-Funktion, zwei Pflegefachpersonen aus der Notfallversorgung, eine Person aus der Allgemeinmedizin sowie drei Fachpersonen aus Psychologie und medizinischer

Ausbildung. Vier der Teammitglieder verfügten über anerkannte Provider- oder Instruktorqualifikationen der Resuscitation Councils in Großbritannien oder Australien. Diese Zusammensetzung deutet darauf hin, dass TEAM in enger Anbindung an den klinischen Anwendungskontext und unter Berücksichtigung unterschiedlicher professioneller Perspektiven entwickelt wurde.

Die finale Version des TEAM umfasst zwölf Items. Davon sind elf als spezifische Verhaltensitems konzipiert und ein Item dient als globales Gesamturteil. Die elf spezifischen Items werden auf einer fünfstufigen Skala mit den Ausprägungen von null bis vier bewertet. Die verbalen Anker reichen von „never/hardly ever“ über „seldom“, „about as often as not“ und „often“ bis „always/nearly always“. Das zwölfte Item bildet eine globale Einschätzung der Gesamtleistung auf einer zehnstufigen Skala von eins bis zehn ab. Inhaltlich gliedert sich das Instrument in die drei Kategorien Leadership, Teamwork und Task Management. Die Kategorie Leadership umfasst zwei Items, nämlich die Frage, ob die Teamleitung dem Team klar gemacht hat, was erwartet wird, und ob die Teamleitung eine globale Übersicht bewahrt hat. Die Kategorie Teamwork enthält sieben Items und bezieht sich auf effektive Kommunikation, rechtzeitige gemeinsame Aufgabenerfüllung, ruhiges und kontrolliertes Handeln, positives Teamklima, Anpassungsfähigkeit an veränderte Situationen, Überwachung und Re-Evaluation der Situation sowie die Antizipation potenzieller Handlungen. Die Kategorie Task Management umfasst die Priorisierung von Aufgaben sowie die Orientierung an anerkannten Standards und Leitlinien. Im Anhang des Artikels werden diese Items zusätzlich durch kurze Beobachtungshinweise präzisiert, beispielsweise im Hinblick auf Delegation, Konfliktmanagement, Rollenflexibilität oder die antizipative Vorbereitung von Defibrillator, Medikamenten und Atemwegsmaterial. Das Instrument verbindet damit verhaltensnahe Einzelbeobachtungen mit einer globalen Gesamteinschätzung und erlaubt sowohl differenzierte Rückmeldungen zu einzelnen Aspekten als auch eine zusammenfassende Gesamtbewertung der Teamleistung.

Ein wesentliches Merkmal des TEAM ist die Kombination aus analytischen Einzelitems und einem globalen Rating. Die Autoren begründen diese Struktur damit, dass binäre Checklisten häufig die ganzheitlichen Aspekte klinischer Kompetenz nicht ausreichend erfassen, während globale Urteile eher geeignet seien, die Gesamtqualität von Teamleistung abzubilden. Zugleich ermöglichen die elf spezifischen Items eine strukturierte Rückmeldung zu einzelnen Facetten von Führung, Teaminteraktion und Aufgabenmanagement. Diese Konstruktion soll die Vorteile beider Bewertungsformen miteinander verbinden.

Die Inhaltsvalidität des TEAM wurde durch ein internationales Expertengremium geprüft. Dieses setzte sich aus sechs Reanimationsexperten aus Großbritannien, Australien und Neusee-

land zusammen, darunter zwei Ärzte sowie vier Pflegefachpersonen oder Resuscitation Officers mit 15 bis 29 Jahren Erfahrung in der Akutversorgung. Die Mitglieder dieses Panels bewerteten die Relevanz der zwölf TEAM-Items unabhängig voneinander auf einer fünfstufigen Skala. Zur Bestimmung der Inhaltsvalidität wurde ein Content Validity Index berechnet, der den Anteil der Experten erfasst, die ein Item mit mindestens drei Punkten bewerteten. Die Ergebnisse zeigten, dass alle Items einen Content Validity Index von über 0,83 erreichten. Der Gesamtindex der zwölf Items betrug 0,96 und lag damit deutlich über dem im Artikel genannten Akzeptanzwert von 0,90. Auf dieser Grundlage wurden sämtliche Items in der finalen Version des Instruments beibehalten. Die Autoren interpretieren diese Befunde als deutlichen Hinweis auf eine hohe inhaltliche Validität der Skala.

Die Konstruktvalidität wurde zunächst explorativ anhand einer Hauptkomponentenanalyse mit Varimax-Rotation untersucht. Datengrundlage waren die Bewertungen eines Reanimationsexperten zu 56 videografierten Reanimationsereignissen, darunter drei reale Krankenhausreanimationen und 53 simulierte Szenarien. Bei Anwendung eines Eigenwertkriteriums von größer als eins und einer Mindestladung von 0,4 ergab sich eine Einfaktorenlösung, die 80,27 % der Gesamtvarianz erklärte. Die Faktorladungen der Einzelitems lagen zwischen 0,64 und 0,88. Die Autoren werten dies als Hinweis darauf, dass TEAM ein gemeinsames übergeordnetes Konstrukt, nämlich Teamwork in Reanimationssituationen, erfasst. Unterstützt wird diese Interpretation durch die hohe uni-dimensionale Validität der Skala. Alle elf spezifischen Items korrelierten signifikant miteinander mit Spearman-Rho-Werten zwischen 0,621 und 1,0 bei einem Signifikanzniveau von $p < 0,01$. Darüber hinaus korrelierte jedes Item stark mit dem Gesamtscore der elf Items; die Korrelationskoeffizienten lagen zwischen 0,801 und 0,943. Ein weiterer Befund zugunsten der Konstruktvalidität besteht darin, dass sich die Ratings der drei realen Krankenhausereignisse nicht signifikant von jenen der simulierten Ereignisse unterschieden. Die fehlenden Unterschiede zwischen beiden Datenquellen wurden dahingehend interpretiert, dass beide Ereignistypen gemeinsam für die Entwicklung und Prüfung der Skala genutzt werden konnten.

Auch die interne Konsistenz der elf spezifischen TEAM-Items erwies sich als hoch. Auf Grundlage der 56 videografierten Reanimationsereignisse ergab sich für die Bewertungen eines Forschers ein Cronbach-Alpha von 0,97. In einer weiteren Analyse, die auf Bewertungen von drei Reanimationsinstruktoren aus der Echtzeittestung beruhte, lag Cronbachs Alpha bei 0,89. Die hohe interne Konsistenz zeigte sich darüber hinaus sowohl für reale Krankenhausereignisse mit einem Alpha von 0,98 als auch für simulierte Ereignisse mit einem Wert von 0,97. Diese

Ergebnisse deuten darauf hin, dass die Skala in hohem Maß homogene Aspekte eines gemeinsamen Konstrukts erfasst.

Zur Prüfung der konkurrenten Validität wurden die Bewertungen der elf Einzelitems sowie des Summenscores mit dem globalen Gesamturteil verglichen. Die Ergebnisse zeigen durchweg signifikante positive Zusammenhänge. Die einzelnen Items korrelierten mit dem globalen Rating zwischen 0,75 und 0,94. Den höchsten Zusammenhang zeigte das Führungsitem zur Klarheit von Richtung und Kommando durch die Teamleitung, den niedrigsten das Item zum positiven Teamklima. Besonders hervorzuheben ist, dass der Gesamtscore der elf Items mit dem globalen Gesamturteil mit einem Spearman-Rho von 0,95 bei $p < 0,01$ korrelierte. Diese sehr hohe Übereinstimmung wird im Artikel als starker Hinweis darauf gewertet, dass die differenzierten Einzelbewertungen inhaltlich mit der globalen Gesamtwahrnehmung der Teamleistung übereinstimmen.

Die Interrater-Reliabilität wurde anhand von sechs zufällig ausgewählten videografierten Reanimationsereignissen untersucht, was etwa elf Prozent des Gesamtdatensatzes entsprach. Zwei Experten bewerteten diese Sequenzen unabhängig voneinander. Die Auswertung mittels Cohen's Kappa ergab für die elf spezifischen Items einen Wert von 0,55, der nach der im Artikel zitierten Einordnung als faire Interrater-Übereinstimmung interpretiert wird. Ergänzend wurde ein mittlerer Intraklassenkorrelationskoeffizient von 0,60 berichtet. Die Autoren betonen, dass sich in den Fällen ohne exakte Übereinstimmung die Bewertungen jeweils nur um einen Punkt unterschieden. Dies deutet darauf hin, dass die Bewertungsdifferenzen eher gering ausfielen, auch wenn die Kappa-Werte noch nicht auf eine hohe Übereinstimmung hindeuten.

Die Stabilität der Bewertungen über die Zeit wurde über die Test-Retest-Reliabilität geprüft. Dazu wurden sechs der insgesamt 56 Videos nach einem Intervall von sechs Monaten erneut durch denselben Experten bewertet, ohne dass dieser Zugang zu seinen ursprünglichen Ratings hatte. Cohen's Kappa ergab in diesem Fall einen Wert von 0,53, was ebenfalls als faire Übereinstimmung gewertet wurde. Zugleich wurde ein mittlerer Intraklassenkorrelationskoeffizient von 0,80 berichtet. Die Autoren interpretieren diese Werte als insgesamt gute Stabilität, insbesondere wenn nicht nur exakte Identität, sondern auch die Distanz zwischen den Bewertungen berücksichtigt wird.

Die Praktikabilität und Feasibility des TEAM wurden in einer dritten Phase mittels Echtzeittestung geprüft. An dieser Untersuchung nahmen acht Medizinstudenten und sieben Pflegestudenten teil, die nach Absolvieren eines eintägigen Immediate Life Support-Kurses in multiprofessionellen Teams simulierte Reanimationsszenarien bearbeiteten. Drei erfahrene Reanima-

tionstrainer führten den Kurs jeweils mit einem Team durch und beurteilten anschließend ein anderes Team. Die TEAM-Ratings wurden unmittelbar nach Abschluss jedes Szenarios vergeben. Ergänzend bewerteten die drei Assessoren das Instrument mithilfe eines 24 Items umfassenden Fragebogens in Bezug auf Vollständigkeit, Beobachtbarkeit, Akzeptanz und Design. Die Ergebnisse zeigen, dass die TEAM-Scores in einem mittleren Bereich lagen. Der Mittelwert der elf spezifischen Items betrug 2,49 von 5, was für eine angemessene Ausnutzung der Skala spricht. Das am höchsten bewertetes Item war die effektive Kommunikation mit einem Mittelwert von 3,07, während die Orientierung an Standards und Leitlinien mit 1,79 den niedrigsten Mittelwert aufwies. Das globale Rating lag im Mittel bei 4,73 von 10. Die Einschätzungen der Praktikabilität fielen insgesamt sehr positiv aus. Mit einer Gesamtpunktzahl von 82 von 88 wurde das Instrument als weitgehend vollständig, akzeptabel und angemessen gestaltet beurteilt. Die erfassten Teamwork-Fähigkeiten galten als gut beobachtbar. Zugleich wurde angemerkt, dass das Item zum Teamklima beziehungsweise zur Teammoral schwerer zu beurteilen sei und dass Erfahrung im Bereich der Teamleistungsbeurteilung wesentlich für die adäquate Anwendung des Instruments sei.

In der Diskussion beschreiben die Autoren, dass TEAM inhaltlich auf drei Kategorien basiert, die sich wiederum auf neun Elemente beziehen. Diese Elemente sind Führungskontrolle, Kommunikation, Kooperation und Koordination, Teamklima, Anpassungsfähigkeit, Situationsbewusstsein in seiner wahrnehmenden und antizipierenden Dimension, Priorisierung sowie klinische Standards. Die elf Items der Skala bilden diese Elemente ab, während das globale Rating eine zusammenfassende intuitive Einschätzung der Gesamtleistung darstellen soll. Nach Auffassung der Autoren kann das Instrument sowohl über die Bildung von Kategoriensummen als auch über einen Gesamtscore genutzt werden. TEAM wird insbesondere für erfahrene klinische Beobachter als geeignet angesehen, um Teamleistung in Herz-Kreislauf- und Traumareanimationen zu erfassen und strukturiertes Feedback zu geben.

Die Anwendungsbereiche des TEAM liegen nach den Angaben des Artikels vor allem in simulierten Reanimations- und Notfallsituationen, perspektivisch aber auch in realen klinischen Kontexten. Es soll die Beobachtung und Messung von Teamleistung ebenso unterstützen wie die Rückmeldung im Rahmen von Debriefings. Die Autoren heben hervor, dass das Instrument in den Händen erfahrener Kliniker einen nützlichen Beitrag zur Leistungsbeurteilung und zur Förderung der Patientensicherheit leisten kann. Zudem wird angeregt, TEAM gegebenenfalls mit technischen Checklisten zu kombinieren, um sowohl technische als auch nichttechnische Leistungen in Reanimationsteams umfassender abzubilden.

Trotz der insgesamt günstigen Ergebnisse benennen die Autoren mehrere Limitationen. Eine wesentliche Einschränkung besteht in der Größe und Zusammensetzung der Stichprobe, die zunächst vor allem simulierte Ereignisse umfasste und nur in geringem Umfang reale Krankenhausreanimationen einschloss. Daher bleibt offen, inwieweit TEAM zuverlässig zwischen unterschiedlichen Kompetenzniveaus differenzieren kann und welches Leistungsniveau als kompetent einzustufen ist. Ebenso ist weitere Prüfung in realen klinischen Kontexten mit erfahrenen Fachpersonen erforderlich. Darüber hinaus beruhte die Reliabilitätsprüfung nur auf einem kleinen Teil des Gesamtdatensatzes, sodass größere Studien mit mehr geschulten Ratern notwendig erscheinen, um die Stabilität der Interrater-Reliabilität und internen Konsistenz weiter zu sichern. Auch die teilweise subjektive Natur einzelner Items, insbesondere des Teamklimas, wurde als potenzielle Schwäche benannt.

Zusammenfassend lässt sich festhalten, dass das Team Emergency Assessment Measure als spezifisch für medizinische Notfall- und Reanimationsteams entwickeltes Beobachtungsinstrument eine strukturierte Erfassung nichttechnischer Teamleistung ermöglicht. Die Entwicklung erfolgte auf Grundlage einer breiten Literaturrecherche, durch Expertenurteile und in einem mehrstufigen empirischen Prüfverfahren. Die finale Skala umfasst zwölf Items, davon elf spezifische Verhaltensitems in den Kategorien Leadership, Teamwork und Task Management sowie ein globales Gesamturteil. Die vorliegenden psychometrischen Ergebnisse sprechen für eine hohe Inhaltsvalidität, eine eindimensionale Konstruktstruktur, eine hohe interne Konsistenz, eine starke konkurrente Validität sowie faire bis gute Interrater- und Test-Retest-Reliabilität. Auch die Praktikabilität in der Echtzeitanwendung wurde positiv bewertet. Zugleich machen die begrenzte Stichprobe, die teilweise subjektive Bewertbarkeit einzelner Items und der bislang überwiegend simulationsbasierte Einsatz deutlich, dass weitere Untersuchungen notwendig sind, um die psychometrischen Eigenschaften und die klinische Anwendbarkeit des Instruments umfassend abzusichern.

5.24 Team Emergency Assessment Measure (TEAM): Vergleich von Novizen- und Expertenratings

Quelle: Freytag J, Stroben F, Hautz WE, Schauber SK, Kämmer JE. Rating the quality of teamwork—a comparison of novice and expert ratings using the team emergency assessment measure (TEAM) in simulated emergencies. Scand J Trauma Resusc Emerg Med. (2019) 27:12. doi: 10.1186/s13049-019-0591-9

Abbildung 27: Team Emergency Assessment Measure (TEAM): Vergleich von Novizen- und Expertenratings

Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Freytag et al. (2019)

Im vorliegenden Beitrag steht nicht die erstmalige Entwicklung des Team Emergency Assessment Measure (TEAM) im Mittelpunkt, sondern dessen Anwendung in einem erweiterten Validierungskontext, nämlich im Vergleich von Bewertungen durch Novizen einerseits und Experten andererseits in simulierten Notfallsituationen. Die Studie geht von der Annahme aus, dass Teamarbeit in hochakuten medizinischen Situationen wie Reanimationen und anderen Notfällen einen wesentlichen Beitrag zur Versorgungsqualität leistet, ihre Beobachtung und Beurteilung jedoch sowohl methodisch als auch organisatorisch herausfordernd ist. Insbesondere wird problematisiert, dass klinische Experten im realen Versorgungsgeschehen zumeist selbst in die Notfallversorgung eingebunden sind und deshalb nicht ohne Weiteres als externe Beobachter zur Verfügung stehen. Vor diesem Hintergrund untersuchten die Autoren, ob auch Personen mit geringerer klinischer Erfahrung, sofern sie entsprechend geschult werden und ein standardisiertes Beobachtungsinstrument verwenden, zu vergleichbaren Urteilen über Teamverhalten gelangen können. Das Ziel der Untersuchung lag damit sowohl in der weiteren psychometrischen Prüfung der deutschen TEAM-Version als auch in der Klärung der Frage, ob TEAM durch Novizen ebenso wie durch Experten sinnvoll eingesetzt werden kann.

Das TEAM wird im Artikel als bereits etabliertes Instrument beschrieben, das speziell zur Erfassung von Teamleistung in medizinischen Notfallsituationen entwickelt wurde. Es umfasst elf Items, die drei Subskalen zugeordnet sind, nämlich Leadership mit zwei Items, Teamwork mit sieben Items und Task Management mit zwei Items. Alle elf Items werden auf einer fünfstufigen Likert-Skala von null bis vier bewertet, wobei null für „never/hardly ever“ und vier für „always/nearly always“ steht. Aus diesen Einzelbewertungen kann ein Summenscore zwischen 0 und 44 Punkten gebildet werden. Ergänzend enthält das Instrument eine globale Ratingskala von eins bis zehn, mit der die Gesamtleistung des Teams eingeschätzt wird. TEAM kombiniert damit analytische Einzelbeobachtungen mit einem globalen Gesamturteil und erlaubt sowohl eine differenzierte Betrachtung einzelner Aspekte der Teamarbeit als auch eine übergreifende Bewertung des Teamhandelns. Nach den Angaben des Artikels war TEAM bereits in verschiedenen realen und simulierten Notfallkontexten eingesetzt worden und galt im Vergleich zu anderen Instrumenten als besonders geeignetes und psychometrisch überzeugendes Verfahren zur Beurteilung von Teamarbeit in medizinischen Notfällen.

Für die vorliegende Studie wurde TEAM in eine deutsche Version übertragen. Da zwar bereits eine französische Fassung vorlag, jedoch keine deutschsprachige Version verfügbar war, übersetzte das Forschungsteam das Instrument mit der TRAPD-Methode, die die Schritte translation, review, adjudication, pre-testing und documentation umfasst. Laut Artikel wurde zusätzlich eine Vorstudie durchgeführt, in der Praktikabilität und Interrater-Reliabilität geprüft und sehr gute Ergebnisse erzielt wurden. In der hier berichteten Untersuchung wurden nun die psychometrischen Eigenschaften der deutschen TEAM-Version systematisch analysiert und zugleich die Urteile von Novizen und Experten miteinander verglichen.

Die Datenerhebung fand im Rahmen einer notfallmedizinischen Simulation an der Charité Universitätsmedizin Berlin statt. Diese Simulation war für Medizinstudenten des letzten Studienjahres konzipiert und umfasste sechs verschiedene Fälle mit einer Dauer von jeweils etwa 30 Minuten. Die Szenarien deckten häufige Notfallsituationen ab und beinhalteten auch eine Reanimation. Die Umsetzung erfolgte mithilfe standardisierter Patienten sowie High-Fidelity-Simulation. Die Teilnehmer arbeiteten in Teams von fünf Personen, wobei für jeden Fall jeweils ein Teammitglied als Teamleitung bestimmt wurde. Diese Führungsrolle wechselte nach jedem Fall. Insgesamt wurden sieben Teams beobachtet, die durch zwölf Rater bewertet wurden. Dies führte zu insgesamt 84 Beobachtungen. Jede Teamleistung wurde jeweils von einem Novizen und einem Experten unabhängig voneinander direkt nach dem jeweiligen Szenario beurteilt.

Die Definition der beiden Ratergruppen erfolgte über deren klinische und inhaltliche Expertise. Die Novizengruppe bestand aus Tutoren des lokalen Skills Labs. Es handelte sich um fortgeschrittene Medizinstudenten mit notfallmedizinischer Vorerfahrung, die sie entweder durch klinische Praktika oder durch Tätigkeiten im Rettungsdienst erworben hatten. Die Expertengruppe setzte sich aus Ärzten sowie Psychologen zusammen, die umfangreiche Erfahrungen in der Notfallmedizin und/oder in der Beobachtung und Vermittlung von Teamarbeit in simulationsbasierten Lernsettings hatten. Die Novizen waren zwischen 20 und 33 Jahre alt, mit einem Median von 24 Jahren. Die Experten waren zwischen 26 und 37 Jahre alt, mit einem Median von 31,5 Jahren. Die Lehr- beziehungsweise Trainingserfahrung lag bei den Novizen zwischen einem und 2,5 Jahren im Bereich des studentisch assistierten Lernens, während die Expertengruppe über 3,5 bis 10 Jahre Erfahrung in klinischer Lehre, simulationsbasierter Ausbildung und Faculty Development verfügte. Hinsichtlich der klinischen Expertise reichte die Erfahrung der Novizen von Praktika mit bis zu 120 Tagen, während die Experten ein bis zehn Jahre klinische Erfahrung aufwiesen. Für die Bewertung mit TEAM war jedoch wesentlich, dass keine der beteiligten Personen vor Beginn der Untersuchung Erfahrungen mit diesem spezifischen Instrument hatte. Beide Gruppen erhielten vor der Datenerhebung ein identisches Ratertraining, das eine Einführung in TEAM, Informationen zu typischen Beurteilungsfehlern sowie eine Frame-of-Reference-Schulung mit videobasierten Beispielen umfasste. Damit wurde sichergestellt, dass Unterschiede zwischen den Gruppen nicht auf unterschiedliche Instrumentenerfahrung zurückgeführt werden konnten.

Die psychometrische Analyse der deutschen TEAM-Version umfasste zunächst die Prüfung der internen Konsistenz. Diese wurde für jeden Fall und getrennt nach Experten- und Novizengruppe berechnet. Für die Experten ergab sich ein mittleres Cronbach-Alpha von 0,89 bei einer Standardabweichung von 0,06. Für die Novizen lag der entsprechende Mittelwert bei 0,85 bei einer Standardabweichung von 0,19. Der niedrigste Alpha-Wert der Experten wurde im ersten Fall aus dem chirurgischen Kontext mit 0,79 beobachtet. Der niedrigste Alpha-Wert der Novizen trat im fünften Fall aus dem anästhesiologischen Bereich auf und lag bei 0,47. Trotz dieser fallbezogenen Streuung beurteilten die Autoren die interne Konsistenz insgesamt als hoch. In der vergleichenden Übersicht mit früheren Untersuchungen wird zudem berichtet, dass die Cronbach-Alpha-Werte der deutschen Version bei Experten 0,93 und bei Novizen 0,94 betrugen und damit im Bereich der Werte der englischen und französischen Versionen lagen. Diese Befunde deuten darauf hin, dass die deutsche TEAM-Version eine hinreichend homogene Erfassung der zugrunde liegenden Teamleistung ermöglicht.

Auch die diskriminativen Eigenschaften der Einzelitems wurden untersucht. Die mittleren Item-Gesamt-Korrelationen lagen bei den Experten bei 0,71 und bei den Novizen bei 0,69. In der tabellarischen Übersicht werden für die deutsche Version bei Experten Werte zwischen 0,59 und 0,81 und bei Novizen zwischen 0,38 und 0,81 angegeben. Diese Befunde sprechen dafür, dass die einzelnen TEAM-Items sinnvoll mit dem Gesamtscore zusammenhängen und damit einen substanziellen Beitrag zur Messung des gemeinsamen Konstrukts Teamarbeit leisten.

Zur weiteren Überprüfung der Konstruktvalidität wurde eine Hauptkomponentenanalyse durchgeführt. Ziel war es zu prüfen, ob die elf TEAM-Items zu einer übergeordneten allgemeinen Komponente zusammengefasst werden können. Die Voraussetzungen für diese Analyse wurden im Vorfeld statistisch überprüft. Die Inter-Item-Korrelationen lagen bei Experten zwischen 0,29 und 0,73 und bei Novizen zwischen 0,42 und 0,75. Das Kaiser-Meyer-Olkin-Kriterium betrug in beiden Gruppen 0,87 und lag damit deutlich über dem empfohlenen Mindestwert von 0,6. Der Bartlett-Test auf Sphärizität fiel für beide Gruppen hochsignifikant aus. Insgesamt deuteten diese Befunde darauf hin, dass die Items ausreichend miteinander zusammenhängen, um eine Faktorenanalyse durchzuführen. Die Hauptkomponentenanalyse ergab sowohl für Experten- als auch für Novizenratings einen dominanten ersten Faktor. Dieser erklärte 59 % der Varianz in den Expertenratings und 65 % der Varianz in den Novizenratings. Die Autoren interpretieren dies dahingehend, dass die deutsche TEAM-Version – analog zur Originalfassung – ein gemeinsames latentes Konstrukt abbildet und sich die Einzelitems daher sinnvoll zu einem allgemeinen Teamarbeitsmaß zusammenfassen lassen.

Ein zentrales Anliegen der Studie war die Untersuchung der Übereinstimmung zwischen Novizen- und Expertenratings. Diese wurde zunächst über einen Intraklassenkorrelationskoeffizienten berechnet, der sich auf die TEAM-Summenscores bezog. Der resultierende ICC von 0,66 wurde als moderat bis gut interpretiert. Diese Befunde werden ergänzt durch die Beobachtung, dass Experten und Novizen in 75 % der Fälle darin übereinstimmten, welche Teams innerhalb eines Szenarios zu den zwei besten beziehungsweise zu den zwei schwächsten gehörten. Damit zeigte sich, dass beide Gruppen die relativen Leistungsunterschiede zwischen den Teams weitgehend ähnlich einschätzten.

Auf Ebene der Einzelitems wurden die Urteile beider Gruppen mithilfe von Mann-Whitney-U-Tests verglichen. Für sieben der elf TEAM-Items ergaben sich keine signifikanten Unterschiede zwischen Experten- und Novizenbewertungen. Dies betraf die Items 1, 4, 5, 6, 7, 9 und 11. Für vier Items bewerteten Novizen die Teamarbeit hingegen signifikant höher als Experten. Dies galt für die Items 2, 3, 8 und 10. Auf Ebene des Summenscores ergab sich über alle Fälle hinweg kein statistisch signifikanter Unterschied zwischen beiden Gruppen. Die No-

vizen vergaben im Mittel 30,4 Punkte bei einer Standardabweichung von 8,6, die Experten 27,0 Punkte bei einer Standardabweichung von 8,4. Der Unterschied war mit einem t-Wert von 1,8 und einem p-Wert von 0,08 nicht signifikant. Anders verhielt es sich bei der globalen Ratingskala, auf der Novizen mit einem Mittelwert von 7,1 und einer Standardabweichung von 1,6 höhere Werte vergaben als Experten mit einem Mittelwert von 6,1 und einer Standardabweichung von 1,9. Dieser Unterschied war mit einem p-Wert von 0,02 statistisch signifikant. Die Autoren deuten dies als Hinweis auf eine tendenziell mildere oder großzügigere Bewertungsstrategie der Novizen.

Um die Verteilung der Bewertungen genauer zu betrachten, wurden sowohl die TEAM-Summenscores als auch die Werte der globalen Ratingskala z-standardisiert und graphisch dargestellt. Die Abbildung auf Seite 5 zeigt, dass die Verteilungen von Experten- und Novizenratings in standardisierter Form sehr ähnlich ausfielen. Die Interquartilsbereiche, Minima und Maxima überlappten stark. Dies stützt die Annahme, dass beide Gruppen zwar unterschiedliche Ausgangsniveaus oder Bewertungsbaselines haben könnten, jedoch in ihrer Einschätzungsmusterstruktur weitgehend ähnlich urteilten. Die Autoren führen die etwas höheren Bewertungen der Novizen darauf zurück, dass diese möglicherweise mit einem niedrigeren internen Referenzstandard urteilten als Experten, die aufgrund ihrer klinischen Erfahrung stärker für die potenziellen Folgen mangelhafter Teamarbeit sensibilisiert seien und deshalb strenger bewerteten.

Zur weiteren Differenzierung der Bewertungsunterschiede wurde ein Mixed-Effects-Modell berechnet, um die Quellen der Varianz in den globalen TEAM-Bewertungen zu identifizieren. Das Modell umfasste Zufallseffekte für einzelne Rater, Fälle, Raterstatus, Teams sowie Interaktionen zwischen Fall und Team sowie zwischen Fall und Raterstatus. Insgesamt erklärte das Modell 71,8 % der beobachteten Varianz. Der Raterstatus, also die Zugehörigkeit zur Novizen- oder Expertengruppe, erklärte 11,1 % der Varianz. Die Fälle selbst machten 10,2 % der Varianz aus. Demgegenüber trugen die Teams nur zu 2,6 % der beobachteten Varianz bei. Die mit Abstand größte Varianzquelle stellte die Interaktion zwischen Fall und Team mit 43,2 % dar. Dies bedeutet, dass die Unterschiede in den Bewertungen in erheblichem Maße davon abhingen, wie die jeweiligen Teams in den unterschiedlichen Szenarien performten, und nicht primär darauf zurückzuführen waren, dass bestimmte Teams grundsätzlich besser oder schlechter waren als andere. Eine kleinere, aber dennoch relevante Varianzquelle war die Interaktion zwischen Raterstatus und Fall mit 3,42 %. Die Autoren interpretieren diese Ergebnisse dahingehend, dass Teamleistung stark vom situativen Kontext abhängt und nicht als stabile, fallübergreifende Eigenschaft eines Teams verstanden werden sollte.

Diese Befunde führen zu mehreren inhaltlichen Schlussfolgerungen. Erstens sprechen sie dafür, dass TEAM grundsätzlich auch durch Personen mit begrenzter klinischer Erfahrung angewendet werden kann, sofern diese geschult werden und standardisierte Beobachtungsrahmen nutzen. Zweitens legen die Ergebnisse nahe, dass die Leistungsfähigkeit eines Teams nicht losgelöst vom konkreten Fall beurteilt werden darf. Die Autoren weisen darauf hin, dass in der vorliegenden Studie die Teamleitung zwischen den Szenarien wechselte, sodass die Effekte von Führungsverhalten und Fallspezifik statistisch nicht vollständig voneinander getrennt werden konnten. Dennoch machen die Daten deutlich, dass Teamleistung stark situationsgebunden ist. Daraus folgt nach Ansicht der Autoren, dass generische Aussagen über „gute Teamarbeit“ nur begrenzten Wert haben und dass zukünftig genauer untersucht werden sollte, welches Teamverhalten in welcher Situation und von welcher Person besonders hilfreich ist. In diesem Zusammenhang wird auch betont, dass TEAM-Scores nicht ohne Weiteres über verschiedene Fälle hinweg verglichen werden sollten, da noch keine verlässlichen Benchmarks und keine eindeutige Verknüpfung mit objektiven Leistungsmaßen vorliegen.

Hinsichtlich der Anwendungsbereiche ergibt sich aus der Studie, dass TEAM nicht nur durch Experten, sondern potenziell auch durch weniger erfahrene externe Beobachter eingesetzt werden kann. Dies eröffnet insbesondere für klinische Studien, Simulationstrainings und möglicherweise auch für die strukturierte Leistungsrückmeldung in realen Teams praktische Perspektiven, weil dadurch der Mangel an verfügbaren Experten teilweise kompensiert werden könnte. Gleichzeitig schließt die validierte deutsche TEAM-Version eine wichtige Lücke für den deutschsprachigen Raum und ermöglicht die standardisierte Erfassung von Teamarbeit in deutschen Ausbildungs- und Simulationskontexten.

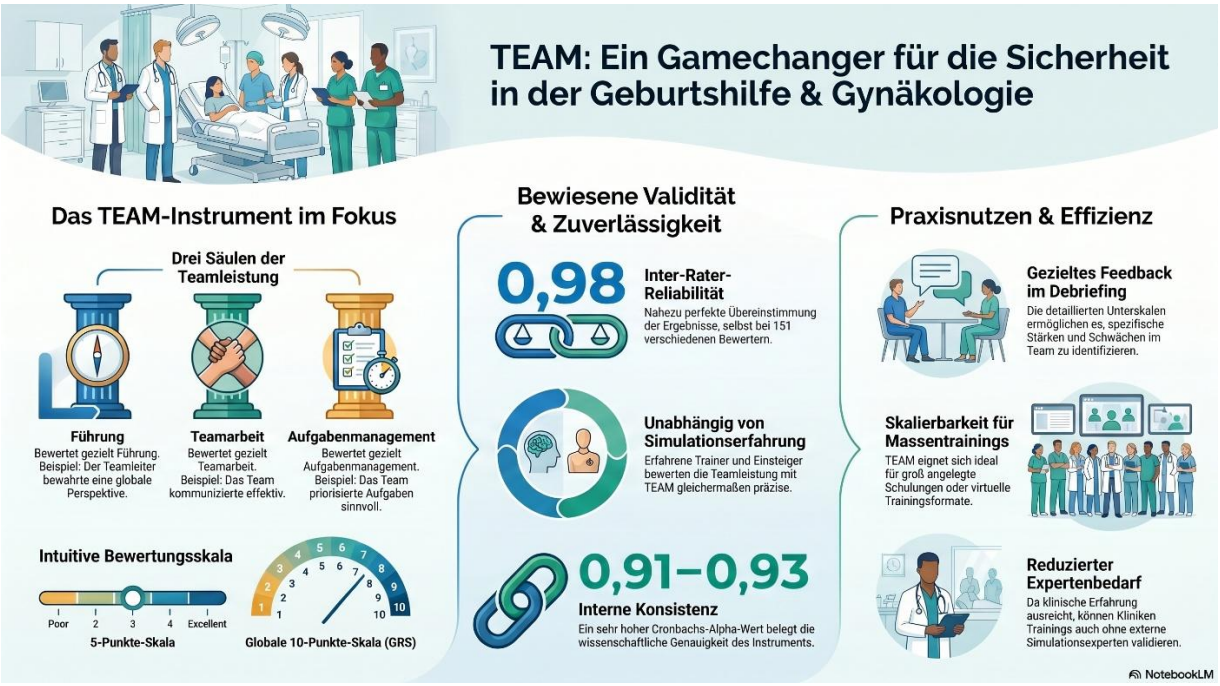
Die Autoren benennen jedoch auch mehrere Limitationen der Untersuchung. Erstens handelt es sich um eine Single-Center-Studie mit begrenzter Stichprobengröße. Zwar war die Zahl der Beobachtungen mit 84 im Vergleich zu anderen TEAM-Studien durchaus beachtlich, die Ergebnisse beruhen jedoch letztlich auf den Einschätzungen von lediglich sechs Novizen und sechs Experten, wobei jedes Szenario nur von einem Experten-Novizen-Paar bewertet wurde. Zweitens wurde die Untersuchung ausschließlich in einem Simulationssetting durchgeführt. Drittens handelte es sich bei den bewerteten Teams um monoprofessionelle Gruppen aus Medizinstudenten des letzten Studienjahres und nicht um reale multiprofessionelle Notfallteams. Viertens wurde TEAM in dieser Studie auch auf Notfallszenarien außerhalb typischer Reanimationskontexte angewandt, sodass weitere Forschung notwendig bleibt, um seine Eignung für andere Notfallsituationen sowie seine Fähigkeit zur Festlegung von Leistungsbenchmarks genauer zu prüfen.

Zusammenfassend zeigt die Studie, dass TEAM auch in seiner deutschen Fassung über gute psychometrische Eigenschaften verfügt und dass Novizen mit begrenzter klinischer Erfahrung zu weitgehend vergleichbaren Urteilen über Teamarbeit gelangen wie Experten. Die interne Konsistenz war hoch, die Hauptkomponentenanalyse bestätigte eine dominante allgemeine Teamarbeitskomponente, und die Übereinstimmung zwischen Experten- und Novizenratings lag im moderaten bis guten Bereich. Unterschiede zeigten sich vor allem in einer tendenziell großzügigeren Bewertung durch Novizen, insbesondere im globalen Gesamturteil und bei einzelnen Items. Insgesamt sprechen die Befunde dafür, dass TEAM ein geeignetes Instrument zur Beurteilung von Teamarbeit in simulierten Notfallsituationen ist und dass unter geeigneten Bedingungen auch weniger erfahrene Beobachter in die Leistungsbeurteilung einbezogen werden können. Zugleich machen die deutliche Kontextabhängigkeit der Teamleistung und die begrenzte Generalisierbarkeit der Ergebnisse deutlich, dass weitere Untersuchungen in multiprofessionellen und realklinischen Settings erforderlich sind.

5.25 Team Emergency Assessment Measure (TEAM) in geburtshilflich-gynäkologischen Reanimationsteams

Quelle: Carpini JA, Calvert K, Carter S, Epee-Bekima M, Leung Y. Validating the team emergency assessment measure (TEAM) in obstetric and gynaecologic resuscitation teams. Aust N Z J Obstet Gynaecol. (2021) 61:855–61.

Abbildung 28: Team Emergency Assessment Measure (TEAM) in geburtshilflich-gynäkologischen Reanimationsteams



Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Carpini et al. (2021)

Im vorliegenden Beitrag wird das Team Emergency Assessment Measure (TEAM) nicht neu entwickelt, sondern in einem spezifischen Anwendungskontext psychometrisch geprüft, nämlich in simulierten geburtshilflich-gynäkologischen Notfallsituationen. Die Studie geht von der Annahme aus, dass simulationsbasiertes Training im Bereich geburtshilflicher und gynäkologischer Notfälle ein wirksames Mittel zur Förderung fachlicher, technischer und nichttechnischer Kompetenzen darstellt. Zugleich wird darauf hingewiesen, dass die Implementierung solcher Trainings durch zwei zentrale Probleme erschwert wird, nämlich durch den Mangel an Simulationsexperten sowie durch das Fehlen flexibel einsetzbarer und belastbarer Outcome-Instrumente zur Erfassung von Teamleistung. Vor diesem Hintergrund verfolgte die Studie das Ziel, die psychometrischen Eigenschaften des TEAM für den Bereich der geburtshilflich-gynäkologischen Reanimations- und Notfallsimulationen zu untersuchen und damit zu klären, ob das Instrument in diesem multidisziplinären Setting als praktikables und aussagekräftiges Verfahren eingesetzt werden kann.

Das TEAM wird im Artikel als ein Instrument zur Bewertung nichttechnischer Teamleistung in Notfallsituationen beschrieben. Es umfasst elf Items, die drei inhaltlichen Bereichen zugeord-

net sind, nämlich Leadership, Teamwork und Task Management. Im Artikel werden beispielhaft einzelne Items aus diesen Domänen genannt, etwa das Aufrechterhalten einer globalen Übersicht durch die Teamleitung, die effektive Kommunikation im Team sowie die Priorisierung von Aufgaben. Die Beurteilung erfolgt über eine fünfstufige Skala mit den verbalen Ankern von „never/hardly ever“ bis „always/nearly always“. Aus den elf Einzelitems kann ein Gesamtscore mit einem Maximum von 44 Punkten gebildet werden. Zusätzlich enthält das Instrument ein zwölftes Item, das als globales Gesamturteil die Leistung des Teams auf einer Skala von eins bis zehn erfasst. Das TEAM verbindet damit analytische Einzelurteile mit einer globalen Einschätzung und erlaubt sowohl eine differenzierte als auch eine zusammenfassende Betrachtung nichttechnischer Teamleistung.

Die Untersuchung wurde im Rahmen einer SimWars-Veranstaltung durchgeführt, die Bestandteil einer internationalen Konferenz im Bereich Geburtshilfe und Gynäkologie in Australien war. SimWars wird im Artikel als etabliertes Simulationsformat beschrieben, das Teamarbeit und Kommunikation sichtbar machen soll und gleichzeitig einer größeren Zahl von Beobachter eine Beteiligung ermöglicht. Drei voneinander unabhängige multidisziplinäre Teams traten in drei hochrealistischen Simulationen gegeneinander an. Die Teams setzten sich aus erfahrenen und weniger erfahrenen Fachpersonen aus Geburtshilfe und Gynäkologie, Anästhesie, Pflege, Hebammenwesen und Rettungsdienst zusammen. Die Simulationen wurden auf einer Bühne vor einem Live-Publikum durchgeführt und mithilfe von hochauflösender audiovisueller Technik unterstützt. Als Simulationsmedium wurde ein hochrealistisches SimMom™-Modell verwendet, das auf die klinischen Maßnahmen des Teams mit situationsangemessenen Vitalparameterveränderungen und Rückmeldungen reagierte. Die drei Notfallszenarien umfassten einen Herzstillstand infolge Hypovolämie bei massiver antepartaler Blutung, einen Herzstillstand infolge Magnesiumtoxizität bei postpartaler Eklampsie sowie einen respiratorischen Kollaps bei schwerem ovariellen Hyperstimulationssyndrom. Damit deckten die Szenarien unterschiedliche, aber für den geburtshilflich-gynäkologischen Notfallbereich hochrelevante Situationen ab.

Die Datenerhebung erfolgte durch die Zuschauer der SimWars-Simulationen. Alle Simulations- und Survey-Teilnehmer unterzeichneten eine Einwilligungserklärung. Anschließend wurde an das Publikum ein Fragebogen verteilt, der sowohl das TEAM als auch demografische Angaben enthielt. Nach jeder Simulation wurden die Beobachter aufgefordert, die Skala unmittelbar auszufüllen, sodass die Einschätzungen zeitnah zur beobachteten Teamleistung erfolgten. Von insgesamt 412 potenziellen Teilnehmern gingen 151 Fragebögen zurück, was einer Rücklaufquote von 37 % entspricht. Diese 151 Personen erzeugten insgesamt 452 verwertbare Bewer-

tungen über die drei Simulationen hinweg. Die Stichprobe war überwiegend weiblich und bestand größtenteils aus Consultants, daneben waren aber auch Weiterbildungsassistenten der Geburtshilfe und Gynäkologie, Allgemeinmediziner, Medizinstudenten, Registrars und weitere medizinische Berufsgruppen vertreten. Die Teilnehmer verfügten im Mittel über 10,71 Jahre Berufserfahrung in ihrer aktuellen Rolle, wobei die Streuung relativ hoch war. Zugleich zeigte sich, dass 69 % der Teilnehmer nur über geringe Simulationserfahrung verfügten.

Ein besonderer Fokus der Studie lag auf der Frage, ob sich die Bewertungen des TEAM in Abhängigkeit vom Ausmaß der Simulationserfahrung der Beobachter unterscheiden. Hierzu wurde die Vorerfahrung mit Simulation auf einer fünfstufigen deskriptiven Skala erfasst. Personen ohne Simulationserfahrung oder mit Erfahrung ausschließlich als Teilnehmer wurden der Gruppe mit niedriger Simulationserfahrung zugeordnet. Personen, die bereits Simulationen moderiert, mitentwickelt oder evaluiert hatten, wurden der Gruppe mit hoher Simulationserfahrung zugerechnet. Diese Differenzierung diente dazu, die Robustheit des Instruments gegenüber unterschiedlichen Erfahrungsniveaus der Bewerter zu prüfen und damit eine der zentralen praktischen Fragen zu adressieren, ob der Einsatz des TEAM zwingend auf ausgebildete Simulationsexperten angewiesen ist.

Zur psychometrischen Analyse wurden zunächst deskriptive Statistiken für die Einzelitems, die Subskalen und die Gesamtscores des TEAM berechnet. Die Ergebnisse zeigen, dass die Teamleistungen über die drei Simulationsszenarien hinweg differenziert erfasst werden konnten. Die Mittelwerte einzelner Items reichten von relativ niedrigen Werten um 2,67 für das Item zur klaren Vermittlung von Erwartungen durch die Teamleitung bis hin zu höheren Werten von 3,68 für das Item zur Aufrechterhaltung einer globalen Perspektive durch die Teamleitung. Auch auf Ebene der Subskalen zeigten sich Unterschiede zwischen den Szenarien. Im Szenario X lag der Leadership-Wert bei 69 % des maximal erreichbaren Skalenwertes, der Teamwork-Wert bei 79 %, der Task-Management-Wert bei 75 % und der TEAM-Gesamtscore bei 76 %. Im Szenario Y lagen die Werte etwas niedriger, während Szenario Z die höchsten Werte aufwies und für den Gesamtscore 89 % des Maximums erreichte. Die globalen Teamleistungsurteile lagen je nach Simulation zwischen 7,35 und 8,96 Punkten. Die Autoren weisen darauf hin, dass diese differenzierten Ergebnisse auf Item-, Subskalen- und Gesamtebene Hinweise auf spezifische Stärken und Schwächen einzelner Teams liefern und damit direkt für strukturierte Debriefings nutzbar sind.

Die Konstruktvalidität des TEAM wurde mithilfe konfirmatorischer Faktorenanalysen geprüft. Anders als explorative Verfahren prüfen konfirmatorische Analysen direkt, ob ein theoretisch postuliertes Messmodell mit den beobachteten Daten übereinstimmt. Im vorliegenden Fall

wurde ein theoretisches Drei-Faktoren-Modell getestet, in dem die übergeordnete Teamleistung durch die drei latenten Dimensionen Leadership, Teamwork und Team Management repräsentiert wird. Die Ergebnisse zeigten für alle drei Szenarien, dass das Drei-Faktoren-Modell dem jeweiligen Basismodell überlegen war. Für Szenario X verbesserten sich die Modellfit-Indikatoren vom Basismodell mit einem Comparative Fit Index von 0,87 und einem Tucker-Lewis-Index von 0,84 auf ein Drei-Faktoren-Modell mit einem Comparative Fit Index von 0,95 und einem Tucker-Lewis-Index von 0,93. Gleichzeitig verbesserte sich der Root Mean Square Error of Approximation von 0,13 auf 0,08 und der Standardized Root Mean Square Residual von 0,06 auf 0,05. Auch in den Szenarien Y und Z ergaben sich vergleichbare Verbesserungen, wobei insbesondere für Szenario Z ein sehr guter Modellfit mit einem Comparative Fit Index von 0,97, einem Tucker-Lewis-Index von 0,96, einem Root Mean Square Error of Approximation von 0,07 und einem Standardized Root Mean Square Residual von 0,04 erreicht wurde. Alle Modellvergleiche fielen signifikant zugunsten der Drei-Faktoren-Lösung aus. Die Autoren interpretieren diese Befunde als robuste Evidenz dafür, dass die Struktur des TEAM in diesem Kontext durch die drei Domänen Leadership, Teamwork und Team Management angemessen beschrieben wird.

Zur Prüfung der konvergenten Validität wurden zunächst die Korrelationen zwischen den einzelnen Items und dem TEAM-Gesamtscore berechnet. Die durchschnittliche Item-zu-Skalen-Korrelation betrug 0,75 und war hochsignifikant. Auf Einzelitemebene zeigten sich durchgehend signifikante positive Korrelationen mit dem Gesamtscore. So korrelierte im Szenario X das Item zur globalen Perspektive der Teamleitung mit 0,83 mit dem Gesamtscore, während das Item zum ruhigen und kontrollierten Handeln des Teams mit 0,53 den niedrigsten, aber immer noch signifikanten Wert aufwies. Im Szenario Y lagen die Korrelationen durchweg zwischen 0,68 und 0,82, im Szenario Z zwischen 0,68 und 0,82. Zusätzlich berechneten die Autoren die Average Variance Extracted, also jenen Varianzanteil, den die Items im jeweiligen Faktor gemeinsam erklären. Für die aggregierte TEAM-Gesamtskala ergab sich über die drei Szenarien hinweg eine durchschnittliche AVE von 0,88. Auch die Subfaktoren zeigten hohe Werte, nämlich 0,91 für Leadership, 0,86 für Teamwork und 0,90 für Team Management. Da diese Werte deutlich über dem im Artikel genannten Schwellenwert von 0,50 liegen, sprechen sie klar für eine gute konvergente Validität sowohl der Gesamt- als auch der Subskalenwerte. Die Autoren leiten daraus ab, dass TEAM nicht nur als Gesamtscore, sondern auch in Form seiner drei Subskalen sinnvoll interpretiert werden kann.

Die Kriteriumsvalidität wurde über den Zusammenhang zwischen dem TEAM und dem globalen Einzelurteil über die Gesamtleistung des Teams geprüft. Hier wurde der TEAM-Gesamt-

score als Prädiktor und das globale Leistungsrating als Kriterium betrachtet. Die Ergebnisse zeigten moderate bis hohe positive Korrelationen in allen drei Szenarien, nämlich 0,62, 0,78 und 0,81, jeweils hochsignifikant. Diese Befunde deuten darauf hin, dass die differenzierten Bewertungen der elf TEAM-Items in enger Beziehung zu der globalen Einschätzung der Gesamtleistung stehen und damit das Konstrukt Teamleistung in einem klinisch sinnvollen Sinne erfassen.

Auch hinsichtlich der Reliabilität erwies sich TEAM in diesem Anwendungskontext als günstig. Die interne Konsistenz der Gesamtskala war über die drei Szenarien hinweg mit Cronbach-Alpha-Werten von 0,91, 0,92 und 0,93 sehr hoch. Auch die Subskalen zeigten gute Werte. Für Leadership lagen die Alpha-Koeffizienten zwischen 0,72 und 0,87, für Teamwork zwischen 0,87 und 0,91 und für Team Management zwischen 0,76 und 0,80. Diese Ergebnisse sprechen dafür, dass sowohl das Gesamtinstrument als auch seine Teilskalen intern konsistent sind und die jeweiligen Konstrukte zuverlässig erfassen. Besonders hervorzuheben ist die Interrater-Reliabilität. Diese wurde über den ICC1 als Maß absoluter Übereinstimmung bestimmt und ergab einen Wert von 0,98 bei einem 95 %-Konfidenzintervall von 0,96 bis 1,00. Der zugehörige F-Wert lag bei 183,10 und war hochsignifikant. Die Autoren bezeichnen diese Übereinstimmung als exzellent. Dieser Befund ist besonders bemerkenswert, da im vorliegenden Design eine große Zahl an Beobachtern beteiligt war und die Einschätzungen live im Rahmen einer Simulationsveranstaltung erfolgten.

Ein weiterer wichtiger Aspekt der Untersuchung betrifft die Robustheit des Instruments gegenüber Unterschieden in der Simulationserfahrung der Bewerter. Zur Prüfung dieser Frage wurden die TEAM- und globalen Gesamtleistungswerte zwischen Beobachtern mit niedriger und hoher Simulationserfahrung verglichen. In keinem der drei Szenarien ergaben sich signifikante Unterschiede. Für den TEAM-Gesamtscore lag der Mittelwert im Szenario X bei 29,76 Punkten in der Gruppe mit niedriger Erfahrung und bei 30,79 Punkten in der Gruppe mit hoher Erfahrung. Im Szenario Y lagen die Mittelwerte bei 28,15 beziehungsweise 29,08 und im Szenario Z bei 35,71 beziehungsweise 35,27. Auch die globalen Gesamturteile unterschieden sich nicht signifikant. Daraus schließen die Autoren, dass umfangreiche Simulationserfahrung offenbar keine Voraussetzung dafür ist, das TEAM in diesem Kontext konsistent und sinnvoll anzuwenden, sofern die Bewerter selbst über hinreichende klinische Erfahrung verfügen.

Die Diskussion hebt mehrere inhaltliche Implikationen dieser Ergebnisse hervor. Zunächst betonen die Autoren, dass diese Studie die erste sei, die die psychometrischen Eigenschaften des TEAM im spezifischen Kontext geburtshilflich-gynäkologischer Notfallsimulationen untersucht. Darüber hinaus wurde das Instrument nicht anhand von Videoaufzeichnungen mit der

Möglichkeit des Anhaltens und Wiederholens geprüft, sondern in einem Live-Simulationssetting, was seine praktische Anwendbarkeit unter realistischen Bedingungen unterstreicht. Die Identifikation einer stabilen Drei-Faktoren-Struktur wird zudem als didaktisch relevant hervorgehoben. Die Autoren argumentieren, dass Teams trotz identischer Gesamtscores sehr unterschiedliche Profile auf den drei Subskalen Leadership, Teamwork und Team Management aufweisen könnten. Solche differenzierten Profilinformationen seien für Debriefings besonders wertvoll, da sie gezieltere Rückmeldungen und damit potenziell bessere Lernprozesse ermöglichen. Zudem wird darauf hingewiesen, dass die sehr hohe Interrater-Reliabilität bei einer großen Zahl von Beobachtern den Einsatz des TEAM auch in groß angelegten Simulationsformaten oder möglicherweise in virtuellen Settings unterstützt und dadurch zur Verbesserung der Feasibility simulationsbasierter Trainings beitragen kann.

Gleichzeitig benennt die Studie mehrere Limitationen. Erstens wurden die Teilnehmer zwar nach ihrer allgemeinen Simulationserfahrung, nicht jedoch nach einer möglichen Vorerfahrung mit dem TEAM selbst befragt. Da TEAM typischerweise in Simulationskontexten verwendet wird, könnte eine gewisse Kovariation zwischen Simulationserfahrung und Instrumentenerfahrung bestehen. Zweitens war es aufgrund der Stichprobenzusammensetzung nicht möglich, andere potenzielle Quellen systematischer Bewertungsunterschiede, etwa unterschiedliche Fachrichtungen oder Berufsgruppen, näher zu analysieren. Die Stichprobe war überwiegend klinisch geprägt, weshalb die Übertragbarkeit der Ergebnisse auf andere Beobachtergruppen vorsichtig zu beurteilen ist. Drittens wäre eine strengere Prüfung der diskriminanten Validität sinnvoll gewesen, beispielsweise durch den Vergleich mit einem anderen Instrument zur Erfassung nichttechnischer Teamleistungen. Viertens könnten die drei gewählten Simulationen nicht die gesamte Bandbreite und Komplexität geburtshilflich-gynäkologischer Notfälle abbilden. Weitere Studien sollten daher zusätzliche Szenarien einbeziehen. Schließlich wurden keine objektiven Teamleistungsdaten erhoben, die die Kriteriumsvalidität zusätzlich hätten absichern können. Obwohl frühere Forschung Zusammenhänge zwischen nichttechnischen und technischen Leistungen nahelegt, bleibt eine direkte empirische Verknüpfung im hier untersuchten Kontext noch offen.

Zusammenfassend lässt sich festhalten, dass das TEAM im Kontext geburtshilflich-gynäkologischer Notfallsimulationen ein psychometrisch überzeugendes Instrument zur Erfassung nichttechnischer Teamleistung darstellt. Die Ergebnisse stützen eine hierarchische Drei-Faktoren-Struktur mit den Dimensionen Leadership, Teamwork und Team Management. Darüber hinaus zeigen sich deutliche Hinweise auf konvergente und kriteriumsbezogene Validität, eine hohe interne Konsistenz sowie eine exzellente Interrater-Reliabilität. Von besonderer prakti-

scher Bedeutung ist, dass sich die Bewertungen nicht systematisch nach dem Ausmaß der Simulationserfahrung der Beobachter unterschieden. Damit unterstützt die Studie die Annahme, dass TEAM in diesem Anwendungsfeld nicht nur valide und reliabel, sondern auch praktikabel einsetzbar ist und damit einen wichtigen Beitrag zur verbesserten Umsetzbarkeit simulationsbasierter Teamtrainings in der Geburtshilfe und Gynäkologie leisten kann.

5.26 TeamSTEPPS® 2.0 Team Performance Observation Tool (TPOT)

Quelle I: Zhang C, Miller C, Volkman K, Meza J, Jones K. Evaluation of the team performance observation tool with targeted behavioral markers in simulationbased interprofessional education. J Interprof Care. (2015) 29:202–8.

Quelle I: Maguire, Mary Beth R., "Psychometric Testing of the TeamSTEPPS® 2.0 Team Performance Observation Tool" (2016). Doctorate of Nursing Science Dissertations. Paper 2., *Frühjahr 03.01.2016*. Verfügbar unter: https://digitalcommons.kennesaw.edu/dns_etd/2/#:~:text=Data%20analysis%20provided%20baseline%20psychometric%20properties%20of%20the,included%20internal%20consistency%2C%20test-retest%2C%20and%20inter%20rater%20analysis.

Die Originalpublikation von Zhang et al. (Zhang C, Miller C, Volkman K, Meza J, Jones K. Evaluation of the team performance observation tool with targeted behavioral markers in simulation-based interprofessional education. J Interprof Care. 2015;29:202–208) ist für die vorliegende Arbeit nicht als frei zugänglicher Volltext verfügbar. Aus diesem Grund wurde zur vergleichenden Heranziehung und zur fundierten Diskussion psychometrischer Eigenschaften eines ähnlichen, verhaltensbasierten Beobachtungsinstruments die Dissertation von Maguire (Maguire MBR. Psychometric Testing of the TeamSTEPPS® 2.0 Team Performance Observation Tool. Doctorate of Nursing Science Dissertations. Paper 2. 2016) verwendet. Die Arbeit von Maguire liegt vollständig vor und beinhaltet eine ausführliche, nachvollziehbare psychometrische Prüfung eines Team-Performance-Beobachtungstools, das dem Ansatz von Zhang et al. (Zielrichtung: Beobachtung von Teamperformance anhand gezielter Verhaltensmarker in simulationsbasierten Settings) in wesentlichen methodischen Merkmalen entspricht. Vor diesem Hintergrund erschien die Dissertation als geeignete Vergleichsquelle, um fehlende Volltextdaten von Zhang et al. sachgerecht zu ergänzen und methodische Argumente (z. B. Reliabilitäts- und Validitätsprüfungen, Operationalisierung durch behaviorale Marker, Einsatz in

simulationsbasierten Lehrformaten) kritisch zu kontextualisieren. Es sei allerdings angemerkt, dass eine Dissertation nicht die Original-Begutachtung einer peer-reviewten Fachpublikation ersetzt; wo immer möglich sollte daher in späteren Arbeiten der vollständige Zhang-Artikel herangezogen und die hier getroffene Desiderata-Begründung überprüft werden.

Abbildung 29: TeamSTEPPS® 2.0 Team Performance Observation Tool (TPOT)

Quelle: erstellt mit KI-Tool NotebookLM, anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Maguire, Mary Beth R., "Psychometric Testing of the TeamSTEPPS® 2.0 Team Performance Observation Tool" (2016). Doctorate of Nursing Science Dissertations. Paper 2

Das TeamSTEPPS® 2.0 Team Performance Observation Tool (TPOT) wurde als Beobachtungsinstrument innerhalb des TeamSTEPPS®-Curriculums konzipiert, um Teamleistung in der Gesundheitsversorgung systematisch erfassen zu können. Im Zentrum der vorliegenden Dissertation von Maguire steht nicht die erstmalige Entwicklung des TeamSTEPPS®-Programms selbst, sondern die psychometrische Prüfung des dazugehörigen Team Performance Observation Tool in seiner überarbeiteten Version 2.0. Ausgangspunkt der Arbeit ist die Annahme, dass eine große Zahl vermeidbarer unerwünschter Ereignisse in Krankenhäusern mit Defiziten in der Teamleistung zusammenhängt und dass Verbesserungen in der Zusammenarbeit medizinischer Teams einen wesentlichen Beitrag zur Patientensicherheit leisten können.

Das TeamSTEPPS®-Programm wurde von der Agency for Healthcare Research and Quality in Zusammenarbeit mit dem Department of Defense als evidenzbasiertes Teamtrainingssystem entwickelt und adressiert die Kernkompetenzen Leadership, Situation Monitoring, Mutual Support und Communication. Mit der Einführung von TeamSTEPPS® 2.0 im Jahr 2014 wurde auch eine aktualisierte Version des TPOT bereitgestellt. Während für andere TeamSTEPPS®-Instrumente, insbesondere zur Erfassung von Einstellungen und Wahrnehmungen, bereits psychometrische Daten vorlagen, waren jene des TPOT bislang nur unzureichend dokumentiert. Der Autor verfolgte daher das Ziel, Baseline-Psychometrika dieses Instruments zu bestimmen und damit seine Einsetzbarkeit in Ausbildung und Praxis fundierter zu belegen.

Die Studie war als deskriptive quantitative Untersuchung angelegt und wurde durch das Dickinson and McIntyre Teamwork Model theoretisch gerahmt. Dieses Modell diente des Autors als konzeptuelle Grundlage, weil es zentrale Prozesse der Teamarbeit über die Komponenten Communication, Team Orientation, Team Leadership, Monitoring, Feedback, Backup und Coordination beschreibt. Die Dissertation legt ausführlich dar, dass die Inhalte des TPOT eng mit diesen theoretischen Elementen verknüpft sind. Damit erhielt die instrumentelle Prüfung nicht nur eine methodische, sondern auch eine konzeptionelle Fundierung.

In seiner ursprünglichen Fassung umfasst das TeamSTEPPS® 2.0 TPOT 23 Items, die auf fünf Domänen verteilt sind. Diese Domänen lauten Team Structure, Communication, Leadership, Situation Monitoring und Mutual Support. Das Instrument ist als beobachtungsbasierte Skala aufgebaut und verwendet eine fünfstufige Antwortskala von 1 für sehr schlechte bis 5 für exzellente Teamleistung. Die Abbildung des Instruments zeigt darüber hinaus Kommentarfelder sowie bereichsspezifische Gesamturteile für die einzelnen Domänen. Die Domäne Team Structure beinhaltet die Items „assembles a team“, „assigns or identifies team members’ roles and responsibilities“, „holds team members accountable“ und „includes patients and families as part of the team“. Im Bereich Communication werden die Fähigkeiten erfasst, kurze, klare, spezifische und rechtzeitige Informationen bereitzustellen, Informationen aus allen verfügbaren Quellen einzuholen, mittels Check-backs die Verständlichkeit von Mitteilungen zu sichern sowie SBAR, Call-outs und Handoff-Techniken effektiv einzusetzen. Die Domäne Leadership umfasst die Identifikation von Teamzielen und Vision, die effiziente Nutzung von Ressourcen zur Maximierung der Teamleistung, die Ausbalancierung der Arbeitslast, die angemessene Delegation von Aufgaben, das Durchführen von Briefs, Huddles und Debriefs sowie das Vorbildverhalten in Bezug auf Teamarbeit. Unter Situation Monitoring werden die Beobachtung des Patientenstatus, die Überwachung anderer Teammitglieder zur Sicherstellung von Sicherheit und Fehlervermeidung, die Beobachtung der Umgebung und der Res-

sourcenverfügbarkeit, die Überwachung des Fortschritts in Richtung Teamziel sowie die Förderung eines geteilten mentalen Modells erfasst. Die fünfte Domäne Mutual Support beinhaltet aufgabenbezogene Unterstützung, rechtzeitiges und konstruktives Feedback, das Eintreten für Patientensicherheit mithilfe von Assertive Statement, Two-Challenge Rule oder CUS sowie die Anwendung von Two-Challenge Rule oder DESC-Skript zur Konfliktbewältigung. Diese Struktur verdeutlicht, dass das TPOT die Teamleistung auf einer verhaltensnahen Ebene erfassen soll, ohne sich auf nur eine einzelne Form teambezogener Kompetenz zu beschränken.

Der Autor analysiert die TPOT-Domänen im Lichte des zugrunde gelegten Teamwork-Modells sehr detailliert. So werden die Kommunikationsitems unmittelbar dem theoretischen Konstrukt Communication zugeordnet. Team Orientation wird insbesondere über das Zusammenstellen eines Teams, die Einbindung von Patienten und Familien sowie das aktive Eintreten für Sicherheit repräsentiert. Team Leadership spiegelt sich in Rollenzuweisung, Verantwortungsübernahme, Zielklärung, Ressourcennutzung, Delegation und Vorbildfunktion wider. Monitoring wird durch die Beobachtung des Patienten, des Teams und der Umwelt operationalisiert. Feedback und Backup sind zwar nicht als eigenständige Domänen im TPOT benannt, finden sich jedoch in verschiedenen Items, etwa zur Förderung eines Shared Mental Model, zur Rückmeldung an Teammitglieder und zur Unterstützung anderer. Coordination wiederum ist in mehreren Domänen des TPOT eingebettet, insbesondere in Briefs, Huddles und Debriefs sowie in der laufenden Abstimmung zwischen Teammitgliedern. Diese enge theoretische Passung zwischen Instrument und Modell stellt ein zentrales Merkmal der Arbeit dar und wird vom Autor als wichtiger Schritt zur inhaltlichen Absicherung des TPOT verstanden.

Die psychometrische Prüfung erfolgte mit mehreren Stichproben. Für die Bestimmung der Inhaltsvalidität wurden sieben Teamwissenschafts-Experten einbezogen, die sich aus Autoren einschlägiger Publikationen, Mitgliedern hochfunktionaler medizinischer Teams sowie einem international tätigen Teamtrainer zusammensetzten. Diese Gruppe verfügte im Durchschnitt über mehr als 22 Jahre Erfahrung im Bereich Teamwissenschaften. Für die übrigen Analysen wurden 51 TeamSTEPPS®-geschulte Gesundheitsfachpersonen rekrutiert, die insgesamt 247 Beobachtungen an fünf vorab aufgezeichneten, jeweils etwa zehnminütigen Teamsimulationen bewerteten. Die Stichprobe war überwiegend weiblich, weiß und im akademischen Bereich tätig. Fast die Hälfte verfügte über einen Doktorgrad, und ein großer Teil hatte weniger als ein Jahr Erfahrung mit TeamSTEPPS®. Zudem hatte die Mehrheit keine oder keine aktuelle Erfahrung mit performance-basierten Assessments. Diese Zusammensetzung ist bedeutsam, weil sie die spätere Interpretation von Interrater-Reliabilität und Test-Retest-Befunden beeinflusst.

Die Datengrundlage bildeten fünf videografierte Teamperformanzen von Pflege-Studententeams, die im Rahmen eines Kurses zur Erkennung und Reaktion auf akute Patientenzustandsverschlechterung entstanden waren. Der Autor begründet die Wahl von Videoaufzeichnungen mit dem Vorteil der Standardisierung, Reproduzierbarkeit und Effizienz. Alle Bewerter sahen identische Performanzen in gleicher Reihenfolge. Diese bestand aus Teamleistungen aus verschiedenen Zeitpunkten eines Kurses und war bewusst nicht rein chronologisch angeordnet, um Vorannahmen über eine lineare Leistungssteigerung zu vermeiden. Neben dem TPOT bewerteten die Teilnehmer dieselben Performanzen zusätzlich mit dem Team Emergency Assessment Measure, um einen Vergleich mit einem bereits validierten Instrument zu ermöglichen.

Die Inhaltsvalidität wurde mithilfe des Content Validity Index überprüft. Die Experten bewerteten jedes der 23 TPOT-Items hinsichtlich seiner Relevanz auf einer vierstufigen Skala von „not relevant“ bis „highly relevant“. Für einzelne Items galt ein Item-CVI von 0,78 oder höher als akzeptabel, für die Gesamtskala ein Mittelwert von 0,90 oder höher als Hinweis auf starke Inhaltsvalidität. Die Ergebnisse zeigten, dass 22 der 23 Items diese Kriterien erfüllten. Lediglich das Item „includes patients and families as part of the team“ erreichte mit einem I-CVI von 0,71 nicht den geforderten Grenzwert und wurde daher aus den weiteren Analysen ausgeschlossen. Der Scale-Level Content Validity Index lag bei 0,95 und damit deutlich über dem geforderten Schwellenwert. Der Autor interpretiert dies als klaren Hinweis auf eine starke Inhaltsvalidität des Instruments, allerdings in einer reduzierten 22-Item-Version. Die Diskussion nennt mehrere mögliche Gründe für die vergleichsweise geringe Relevanz des ausgeschlossenen Items. Einerseits werde die Beteiligung von Patienten und Familien in der Praxis vielfach noch nicht als integraler Bestandteil des eigentlichen Gesundheitsteams verstanden. Andererseits werde diese Dimension auch im TeamSTEPPS®-Curriculum selbst nur am Rande behandelt.

Die Konstruktvalidität wurde zunächst über eine Itemanalyse geprüft. Dabei zeigten die Mittelwerte der 22 verbleibenden Items eine relativ stabile Verteilung, wobei die niedrigsten Mittelwerte bei jenen Items auftraten, die sich auf Konfliktmanagement bezogen, und die höchsten Mittelwerte im Bereich der Teambildung und Aufgabenverteilung. Die Inter-Item-Korrelationen lagen zwischen 0,39 und 0,88 und damit innerhalb des vom Autor als akzeptabel definierten Bereichs. Dies spricht dafür, dass die Items zwar miteinander zusammenhängen, aber keine problematische Redundanz oder Singularität aufweisen.

Die weiterführende Prüfung der Konstruktvalidität erfolgte mittels explorativer Faktorenanalyse mit Maximum-Likelihood-Extraktion und orthogonaler Varimax-Rotation. Die Eignung der Da-

ten für diese Analyse wurde zunächst statistisch abgesichert. Der Kaiser-Meyer-Olkin-Wert lag mit 0,973 im exzellenten Bereich, und auch der Bartlett-Test auf Sphärizität fiel hochsignifikant aus, sodass die Korrelationsmatrix als faktorierbar betrachtet werden konnte. Die Kommunalitäten der einzelnen Items lagen zwischen 0,62 und 0,88, was darauf hinweist, dass die extrahierten Faktoren einen hohen Anteil der Varianz der einzelnen Variablen erklären. Die Faktorenextraktion ergab zwei Faktoren mit Eigenwerten größer als eins. Der erste Faktor besaß einen Eigenwert von 16,32 und erklärte 74 % der Varianz, der zweite Faktor einen Eigenwert von 1,25 und erklärte weitere 6 %. Zusammen erklärten beide Faktoren somit 80 % der Gesamtvarianz. Der Scree-Plot bestätigte diese Zwei-Faktoren-Lösung, da nach dem zweiten Faktor ein deutlicher Knick in der Kurve auftrat.

Die anschließende Varimax-Rotation zeigte, dass 20 der 22 Items hoch auf einem ersten Faktor luden, während zwei Items auf einem zweiten Faktor konzentriert waren. Die meisten Ladungen lagen über 0,71 und wurden daher als exzellent interpretiert. Ein Item, nämlich „provides timely and constructive feedback to team members“, lud mit 0,65 auf den ersten Faktor und wurde damit ebenfalls klar zugeordnet. Ein weiteres Item, „conducts briefs, huddles, and debriefs“, zeigte eine Kreuzladung auf beiden Faktoren, wurde jedoch nach inhaltlicher Prüfung dem ersten Faktor zugeordnet. Die inhaltliche Interpretation der Faktoren stellt einen wesentlichen Befund der Arbeit dar, weil die empirisch gefundene Struktur nicht mit der ursprünglich theoretisch vorgesehenen Fünf-Domänen-Struktur des Instruments übereinstimmt. Der erste Faktor wurde vom Autor als Participative Leadership bezeichnet. Begründet wird dies damit, dass die zugehörigen Items nicht lediglich klassisches Führungsverhalten, sondern ein breiteres Verständnis geteilter Verantwortung, gemeinsamer Entscheidungsfindung, Ressourcennutzung, Rollenklarheit, Delegation, Monitoring, Unterstützung und Feedback repräsentieren. Diese Deutung passt nach Auffassung des Autors besonders gut zum Anliegen des TeamSTEPPS®-Programms, die traditionelle Hierarchie im Gesundheitswesen aufzubrechen und Verantwortung stärker im Team zu verteilen. Der zweite Faktor wurde als Conflict Management benannt, da er ausschließlich aus zwei Items bestand, die sich auf Patientensicherheitsadvokatur, Konfliktansprache und Konfliktlösung mittels strukturierter Kommunikationsstrategien beziehen. Der Autor weist ausdrücklich darauf hin, dass dieser Faktor in künftigen Versionen durch zusätzliche Items gestärkt werden sollte, da zwei Items nur eine sehr schmale Grundlage für die differenzierte Abbildung von Konfliktmanagement darstellen.

Die interne Konsistenzreliabilität des Instruments wurde mit Cronbachs Alpha untersucht. Für die 22-Item-Gesamtskala ergab sich ein Alpha von 0,98, was als exzellent bewertet wurde. Die Subskala Participative Leadership erreichte ein Alpha von 0,99, während für die Subskala Con-

flict Management ein Alpha von 0,90 berichtet wurde. Auch die korrigierten Item-Gesamt-Korrelationen lagen mit Werten zwischen 0,57 und 0,91 deutlich über dem geforderten Mindestwert. Zudem zeigte die Analyse „Cronbach's alpha if item deleted", dass durch das Entfernen keines Items die interne Konsistenz des Gesamtscores verbessert worden wäre. Diese Befunde sprechen dafür, dass die Items eng zusammenhängen und gemeinsam ein konsistentes Konstrukt abbilden. Der Autor weist allerdings darauf hin, dass die hohe interne Konsistenz der Conflict-Management-Subskala auch damit zusammenhängen könnte, dass sie nur zwei konzeptionell sehr ähnliche Items enthält, weshalb eine Erweiterung dieser Subskala aus inhaltlichen Gründen dennoch sinnvoll sei.

Zur Prüfung der konkurrenten Validität wurde das TPOT mit dem Team Emergency Assessment Measure verglichen. Der Einsatz des TEAM wurde damit begründet, dass dieses Instrument ähnliche Konstrukte der Teamleistung erfasst und bereits validiert worden war. Aufgrund der nicht normalverteilten Daten wurde Spearman's Rho als Korrelationsmaß verwendet. Zwischen TPOT- und TEAM-Gesamtscores ergab sich ein sehr starker positiver Zusammenhang von 0,930 bei einem Signifikanzniveau von $p < .001$. Daraus schließt der Autor, dass das TPOT in hohem Maße dasselbe übergeordnete Konstrukt erfasst, wie das etablierte Vergleichsinstrument und somit über eine gute konkurrente Validität verfügt.

Die Test-Retest-Reliabilität wurde mit elf zufällig ausgewählten Teilnehmern geprüft, die dieselben fünf Videos etwa zwei Wochen nach der ersten Bewertung erneut mit dem TPOT beurteilten. Für drei der fünf Videos zeigten sich signifikante Spearman-Korrelationen, nämlich für Video 1, Video 3 und Video 5. Für Video 2 und Video 4 waren die Korrelationen dagegen nicht signifikant. Der Autor interpretiert dies dahingehend, dass die Mehrheit der Bewertungen über die Zeit stabil blieb, also 60 % der Re-Test-Vergleiche zufriedenstellend ausfielen. Die fehlende Stabilität zweier Videos wird mit möglichen Reihenfolgeeffekten und mit der bewusst variierenden Leistungsqualität der Videosequenzen erklärt. Da die Reihenfolge der Videos konstant blieb und die Rater beim zweiten Durchgang möglicherweise bereits wussten, welche Teams schwächer oder stärker performten, könnten sich ihre Bezugsstandards verändert haben.

Die Interrater-Reliabilität wurde anhand von sechs zufällig ausgewählten Ratern untersucht, jeweils drei mit weniger als drei Jahren und drei mit mehr als drei Jahren TeamSTEPPS®-Erfahrung. Verwendet wurde ein zweifaktorielles Zufallseffekt-ICC-Modell mit absoluter Übereinstimmung. Für die Gesamtgruppe ergab sich ein ICC von 0,80, was als exzellente Interrater-Übereinstimmung gewertet wurde. Für die Gruppe mit weniger als drei Jahren Erfahrung lag der ICC bei 0,45, entsprechend einer fairen Übereinstimmung. Bei den erfahreneren Ra-

tern mit mehr als drei Jahren Erfahrung lag der Wert bei 0,68 und damit im guten Bereich. Der Autor schließt daraus, dass die Konsistenz der Bewertungen mit zunehmender Erfahrung in TeamSTEPPS® zunimmt und dass zusätzliche Schulung der Rater die Güte der Bewertungen weiter verbessern könnte.

Die Anwendungsbereiche des TPOT werden in der Dissertation breit angelegt. Das Instrument soll Teamleistung im Rahmen von TeamSTEPPS®-Trainings in Praxis und Ausbildung sichtbar und bewertbar machen. Es wird als Werkzeug beschrieben, das sich für unterschiedliche Disziplinen und Einsatzorte eignet, weil es nicht, wie etwa ANTS oder OTAS, auf einen eng umrissenen Fachbereich beschränkt ist. Für die Pflegepraxis wird insbesondere der Nutzen betont, unmittelbar nach Teamtrainings oder in wiederholten Abständen die Entwicklung von Teamstrategien zu messen. In der Pflegebildung soll das TPOT sowohl formative als auch summative Beurteilungen ermöglichen und damit eine bislang unterrepräsentierte Dimension der Kompetenzmessung abdecken, nämlich die Fähigkeit, als Mitglied eines Gesundheitsteams zu handeln. Auch für die Pflegeforschung und für das Gesundheitswesen insgesamt sieht der Autor im TPOT ein wichtiges Instrument, da es die Evaluation von Teamtrainings und damit auch die wissenschaftliche Untersuchung der Wirksamkeit des TeamSTEPPS®-Programms unterstützen könne.

Gleichzeitig benennt die Dissertation zahlreiche Limitationen. Eine erste Einschränkung betrifft die Stichprobengewinnung. Die Teilnehmer wurden über Convenience Sampling, Snowball Sampling und im Anschluss an TeamSTEPPS®-Schulungen rekrutiert, sodass die Repräsentativität der Stichprobe nicht gesichert werden kann. Hinzu kommt, dass ein Großteil der Teilnehmer nur über geringe TeamSTEPPS®-Erfahrung verfügte. Eine weitere Limitation besteht darin, dass das Vorwissen zu TeamSTEPPS® nicht formal getestet wurde, sondern allein auf Selbstbericht beruhte. Drittens basierten die Bewertungen auf vorab aufgezeichneten Simulationen von Pflege-Studententeams und nicht auf realen multidisziplinären Versorgungsteams. Dies begrenzt die Übertragbarkeit der Ergebnisse auf klinische Praxissettings. Auch die Heterogenität der Datenerhebung wird als Problem benannt, da ein Teil der Bewertungen online unter potenziell unkontrollierten Bedingungen erfolgte, während andere Teilnehmer in Gruppensettings bewerteten, in denen gegenseitige Beeinflussung nicht vollständig ausgeschlossen werden konnte. Der Autor diskutiert außerdem eine potenzielle Ermüdung der Teilnehmer, da jede Person fünf Videos mit zwei Instrumenten bewerten musste, sowie mögliche Effekte der festen Videosequenz auf die Stabilität der Test-Retest-Werte.

Eine konzeptionelle Limitation ergibt sich schließlich aus der Diskrepanz zwischen der ursprünglich theoretisch erwarteten Fünf-Domänen-Struktur des TPOT und der empirisch gefun-

denen Zwei-Faktoren-Struktur. Dies deutet darauf hin, dass die aktuelle Instrumentenstruktur die TeamSTEPPS®-Kernkonzepte nicht in der ursprünglich intendierten Form differenziert abbildet. Besonders der Bereich des Konfliktmanagements erscheint unterrepräsentiert. Aus diesem Grund empfiehlt der Autor eine Überarbeitung des Instruments zu einer neuen Version, die sie als TPOT 3.0 bezeichnet. Diese überarbeitete Fassung soll aus 22 Items bestehen und die empirisch gestützten Faktoren Participative Leadership und Conflict Management besser abbilden. Darüber hinaus fordert sie die Entwicklung eines standardisierten Scoring-Handbuchs, da ein solches für das aktuelle TPOT nicht verfügbar war und die Standardisierung der Anwendung dadurch erschwert wurde.

Zusammenfassend zeigt die Dissertation, dass das TeamSTEPPS® 2.0 Team Performance Observation Tool nach Ausschluss eines inhaltlich wenig relevanten Items in einer 22-Item-Version über günstige psychometrische Eigenschaften verfügt. Die Ergebnisse sprechen für eine starke Inhaltsvalidität, eine tragfähige Konstruktvalidität in Form einer Zwei-Faktoren-Struktur, eine sehr hohe interne Konsistenz, eine starke konkurrente Validität zum TEAM sowie insgesamt gute bis exzellente Interrater- und überwiegend zufriedenstellende Test-Retest-Reliabilität. Zugleich verdeutlicht die Arbeit, dass die empirische Struktur des Instruments von seiner ursprünglichen konzeptuellen Domänenlogik abweicht und deshalb eine Weiterentwicklung des TPOT angezeigt ist. Insgesamt kann das Instrument nach den vorliegenden Befunden als valides und reliables Verfahren zur Beobachtung von Teamleistung im Sinne des TeamSTEPPS®-Ansatzes verstanden werden, dessen Potenzial insbesondere für simulationsbasierte Ausbildung, Teamtraining und Qualitätsentwicklung im Gesundheitswesen hervorgehoben wird.

6. Diskussion

Die vorliegende Bachelorarbeit hatte die Intention, eine systematische Übersicht und Analyse von Instrumenten zur Erfassung von Non-Technical Skills (NTS) in High-Fidelity-Simulationen (HFS) im Gesundheitswesen zu erstellen. Im Rahmen der vorliegenden Untersuchung wurden folgende Forschungsfragen als zentral erachtet:

Welche Charakteristika weisen publizierte Instrumente zur Messung von Team-NTS in HFS auf?

Es stellt sich die Frage, welche Dimensionen von NTS (z. B. Situationsbewusstsein, Entscheidungsfindung, Kommunikation) abgedeckt werden.

Es soll der strukturelle Aufbau der Instrumente erörtert werden, wobei insbesondere die Kategorien, Elemente und Verhaltensanker zu berücksichtigen sind.

Es ist zu eruieren, für welche Zielgruppen (beispielsweise Ärzte, Pflegekräfte, multiprofessionelle Teams) und Settings (beispielsweise Anästhesie, Chirurgie, Notfallmedizin) die Maßnahmen konzipiert sind.

Es ist von Relevanz, die psychometrischen Eigenschaften der Instrumente zu ermitteln. Zu den relevanten Eigenschaften zählen dabei die Validität, die Reliabilität sowie die Praktikabilität.

Es soll eruiert werden, inwiefern die eingesetzten Tools als valide zu erachten sind, wobei insbesondere die Inhalts-, Konstrukt- und Kriteriumsvalidität zu berücksichtigen sind.

Es stellt sich die Frage, inwiefern die eingesetzten Instrumente als reliabel einzustufen sind. Als Kriterien können beispielsweise die Interrater-Reliabilität und die interne Konsistenz herangezogen werden.

Es soll eruiert werden, inwiefern eine praktische Anwendbarkeit für den Einsatz in Simulationen und klinischen Settings gegeben ist. Zu den zu berücksichtigenden Faktoren zählen der Zeitaufwand, der Bedarf an Schulungen sowie die Anwendbarkeit in Echtzeit.

Es stellt sich die Frage, ob sich die Instrumente für den praktischen Einsatz in klinischen und edukativen Kontexten eignen.

Es stellt sich die Frage, welche Tools sich für formative Assessments eignen, beispielsweise in Form von Feedback in Trainings.

Es stellt sich die Frage, welche sich für summative Assessments eignen (z. B. Prüfungen, Zertifizierungen).

Es stellt sich die Frage, auf welche Art und Weise eine Anpassung der Instrumente an spezifische Kontexte, wie beispielsweise Trauma-Teams oder die Geburtshilfe, zu bewerkstelligen ist.

Die Ergebnisse der Arbeit zeigen, dass eine Vielzahl von Instrumenten zur Erfassung von NTS existiert, die sich in Struktur, Zielgruppe, psychometrischer Qualität und Anwendungsbereich deutlich unterscheiden. Im Folgenden werden die zentralen Erkenntnisse diskutiert, Limitationen der Instrumente und der vorliegenden Arbeit aufgezeigt sowie Empfehlungen für die Praxis und weitere Forschung abgeleitet.

6.1 Charakteristika der Instrumente: Struktur, Dimensionen und Zielgruppen

Die analysierten 20 Instrumente (vgl. Anhang, Tabelle 8) zur Erfassung nicht-technischer Fähigkeiten (NTS) decken ein breites Spektrum ab, das sich in drei Hauptkategorien unterteilen lässt (vgl. Tab. 1): kognitive Fähigkeiten, soziale Fähigkeiten und persönliche Ressourcen. Jede dieser Kategorien umfasst spezifische Unterdimensionen, die von verschiedenen Assessment-Tools unterschiedlich gewichtet werden.

Tabelle 1: ***NTS - Struktur***

Hauptkategorie	Unterkategorien (Beispiele)	Instrumente (Beispiele)
Kognitive Fähigkeiten	Situationsbewusstsein, Entscheidungsfindung, Aufgabenmanagement, Antizipation, Metakognition	NOTSS, TEAM, OSCAR, T-NOTECHS, OSANTS
Soziale Fähigkeiten	Kommunikation, Teamarbeit, Führung, Followership, Konfliktmanagement, Rollenklarheit	MHPTS, CALM, AOTP, TPOT, TEAM
Persönliche Ressourcen	Stressmanagement, Selbstreflexion, Coping-Strategien, professionelles Verhalten	ANTSdk, NTS-NAS, HFRS

Quelle: eigene Darstellung

Kognitive Fähigkeiten umfassen beispielsweise Situationsbewusstsein, Entscheidungsfindung, Aufgabenmanagement, Antizipation und Metakognition und werden unter anderem von Instrumenten wie ANTS, NOTSS, TEAM, OSCAR, T-NOTECHS und OSANTS erfasst. Soziale Fähigkeiten beinhalten Kommunikation, Teamarbeit, Führung, Followership, Konfliktmanagement und Rollenklarheit und finden sich in Tools wie MHPTS, CALM, AOTP, TPOT und TEAM wieder. Persönliche Ressourcen wie Stressmanagement, Selbstreflexion, Coping-Strategien und professionelles Verhalten werden dagegen von Instrumenten wie ANTSdk, NTS-NAS und HFRS abgedeckt.

Besonders häufig werden in den 20 ausgewählten Instrumenten folgende NTS-Dimensionen berücksichtigt: Situationsbewusstsein gilt als zentrale Komponente und wird in fast allen Tools (z. B. ANTS, NOTSS, TEAM, OSCAR) erfasst. Es beschreibt die Fähigkeit, relevante Informationen zu sammeln, zu interpretieren und zukünftige Entwicklungen vorherzusehen. Die Entscheidungsfindung wird in Instrumenten wie ANTS, NOTSS und TEAM operationalisiert und bezieht sich auf das Abwägen von Handlungsoptionen, die Auswahl einer Strategie und deren kontinuierliche Überprüfung. Kommunikation ist ein Querschnittsthema, das in nahezu allen Instrumenten (z. B. TEAM, MHPTS, CALM) eine Rolle spielt, wobei besonders geschlossene Kommunikationsschleifen (*Closed-Loop Communication*) und strukturierte Übergaben (z. B. SBAR) betont werden. Führung und Teamarbeit werden in Tools wie CALM, TPOT und AOTP differenziert erfasst, wobei Führung nicht nur die Übernahme von Verantwortung, sondern auch die Einbindung von Teammitgliedern und die Verteilung der Arbeitslast umfasst. Das Aufgabenmanagement wird in ANTS, NOTSS und TEAM bewertet und bezieht sich auf Priorisierung, Ressourcennutzung und die Einhaltung von Standards.

Beispielsweise setzen die folgenden acht Instrumente jeweils unterschiedliche Schwerpunkte: ANTS (Anaesthetists' Non-Technical Skills) und NOTSS (Non-Technical Skills for Surgeons) sind professionsspezifisch und legen besonderen Wert auf kognitive Prozesse wie Situationsbewusstsein und Entscheidungsfindung. TEAM (Team Emergency Assessment Measure) und T-NOTECHS (Trauma Non-Technical Skills Scale) sind teamorientiert und betonen Kommunikation, Führung und Stressbewältigung in Notfallsituationen. CALM (Concise Assessment of Leader Management) und TPOT (TeamSTEPPS® Performance Observation Tool) fokussieren auf Führungsverhalten und Teamkoordination, während AOTP (Assessment of Obstetrical Team Performance) und GAOTP (Global Assessment of Obstetrical Team Performance) kontextspezifisch für die Geburtshilfe sind und auch patienten- und familienzentrierte Aspekte einbeziehen.

Die meisten NTS-Instrumente folgen einem hierarchischen Aufbau mit drei Ebenen: Auf der obersten Ebene stehen Kategorien (z. B. Leadership, Communication, Situation Awareness), gefolgt von Elementen (z. B. „Informationssammlung", „Geschlossene Kommunikation") und schließlich Verhaltensankern, die konkrete Beispiele für gutes oder schlechtes Verhalten liefern. So besteht ANTS aus 4 Kategorien, 15 Elementen und über 60 Verhaltensankern, während TEAM 3 Kategorien (Leadership, Teamwork, Task Management) mit 11 Items und einem Globalrating umfasst. NOTSS setzt sich aus 5 Kategorien, 14 Elementen und über 50 Verhaltensankern zusammen, und TPOT (TeamSTEPPS®) enthält 23 Items in 5 Domänen, wobei

empirische Prüfungen eine Zwei-Faktoren-Struktur (*Participative Leadership*, *Conflict Management*) ergaben.

Für die Bewertung kommen meist 4- bis 9-stufige Likert-Skalen (z. B. 1 = „schlecht“ bis 5 = „exzellent“) zum Einsatz. Einige Instrumente wie TEAM und ANTS bieten zusätzlich eine globale Gesamteinschätzung an. Ein zentrales Merkmal zur Erhöhung der Objektivität sind Verhaltensanker, die konkrete Verhaltensbeispiele liefern (z. B. *„Der Teamleiter delegiert Aufgaben klar und eindeutig“* vs. *„Die Aufgabenverteilung bleibt unklar“*).

In der praktischen Umsetzung lassen sich die Instrumente in drei Ansätze unterteilen: Checklistenbasierte Tools (z. B. OSCAR, STAT) eignen sich für detaillierte Analysen, sind jedoch zeitaufwendig. Globale Ratingskalen (z. B. Ottawa GRS, GAOTP) sind schneller anwendbar, bieten aber weniger differenziertes Feedback. Hybride Ansätze (z. B. TEAM, NOTSS) kombinieren Einzelitems mit einem Globalrating und ermöglichen sowohl formative als auch summative Bewertungen.

Die Instrumente sind für unterschiedliche Zielgruppen und klinische Settings konzipiert. Für Ärzte (Einzelpersonen) eignen sich beispielsweise ANTS, NOTSS, ANTSdk, AS-NTS und OSANTS in Bereichen wie Anästhesie, Chirurgie, Notfallmedizin und Geburtshilfe. Pflegekräfte werden mit Instrumenten wie NTS-NAS, MHPTS und TPOT in Notfallpflege, Intensivpflege und Geburtshilfe erfasst. Für multiprofessionelle Teams stehen Tools wie TEAM, T-NOTECHS, AOTP, STAT, OSCAR und CALM zur Verfügung, die in Reanimationsteams, Traumateams, OP-Teams und Geburtshilfe-Teams eingesetzt werden. Studenten können mit AS-NTS und MHPTS in Simulationstrainings für Anästhesie, Notfallmedizin und Pflege geschult werden (vgl. Tab. 2).

Tabelle 2: ***Zielgruppen und Settings der Instrumente***

Zielgruppe	Instrumente (Beispiele)	Einsatzbereich
Ärzte (Einzelpersonen)	ANTS, NOTSS, ANTSdk, AS-NTS und OSANTS	Anästhesie, Chirurgie, Notfallmedizin, Geburtshilfe
Pflegekräfte	NTS-NAS, MHPTS und TPOT	Notfallpflege, Intensivpflege, Geburtshilfe
Multiprofessionelle Teams	TEAM, T-NOTECHS, AOTP, STAT, OSCAR, CALM	Reanimationsteams, Traumateams, OP-Teams, Geburtshilfe-Teams
Studenten	AS-NTS, MHPTS	Simulationstrainings in Anästhesie, Notfallmedizin, Pflege

Quelle: eigene Darstellung

Dabei sind viele Instrumente kontextspezifisch angepasst: ANTS und NOTSS berücksichtigen die Rollenverteilung im OP und sind professionsspezifisch für Anästhesie und Chirurgie. TEAM und T-NOTECHS sind notfallmedizinisch ausgerichtet und erfassen Teamleistung unter Zeitdruck. AOTP und GAOTP sind geburtshilflich konzipiert und beziehen Patienten- und Angehörigenkommunikation ein, während CALM und TPOT auf Führungsverhalten in Reanimationsteams fokussieren.

Fazit: Die meisten NTS-Instrumente sind nicht universell einsetzbar, sondern müssen an Zielgruppe, Setting und Lernziele angepasst werden. Während ANTS und NOTSS sich besonders für individuelle Bewertungen eignen, sind TEAM und T-NOTECHS besser für Teamassessments geeignet. Die Wahl des passenden Instruments sollte daher stets kontextbezogen erfolgen, um eine valide und praxisnahe Anwendung zu gewährleisten.

6.2 Psychometrische Eigenschaften: Validität, Reliabilität und Praktikabilität

Die Qualität von Instrumenten zur Erfassung nicht-technischer Fähigkeiten (NTS) wird anhand zentraler Gütekriterien bewertet, wobei Validität, Reliabilität und Praktikabilität die entscheidenden Maßstäbe darstellen. Die Validität gibt an, ob ein Instrument tatsächlich das misst, was es messen soll, und wurde in dieser Arbeit hinsichtlich Inhaltsvalidität, Konstruktvalidität und Kriteriumsvalidität untersucht.

Die Inhaltsvalidität wird durch Expertenurteile und eine literaturbasierte Entwicklung sichergestellt. Einige Instrumente weisen eine besonders hohe Inhaltsvalidität auf, da sie systematisch entwickelt wurden: ANTS (Fletcher et al., 2003) entstand durch kognitive Aufgabenanalysen mit 29 Anästhesisten, während NOTSS (Yule et al., 2006) aus Critical-Incident-Interviews mit 27 Chirurgen abgeleitet wurde. TEAM (Cooper et al., 2010) wurde von einem Expertenpanel mit 148 Notfallmedizinern validiert, die die Relevanz der Items bestätigten. Allerdings zeigen sich auch Probleme: Einige Instrumente, wie das HFRS, wurden aus Luftfahrt-Checklisten adaptiert und sind nicht ausreichend an den medizinischen Kontext angepasst. Zudem wurde beim TPOT (Maguire, 2016) das Item *„Includes patients and families as part of the team“* von Experten als weniger relevant eingestuft.

Die Konstruktvalidität wird durch Faktorenanalysen und Gruppenvergleiche (z. B. Experten vs. Novizen) überprüft. Einige Instrumente zeigen hier eine gute Validität: TEAM (Freytag et al., 2019) bestätigte in einer Hauptkomponentenanalyse eine eindimensionale Struktur, während T-NOTECHS (Repo et al., 2019) durch eine konfirmatorische Faktorenanalyse eine stabile 5-Domänen-Struktur aufwies. OSANTS (Dedy et al., 2015) konnte zudem zwischen Weiterbildungsjahren (PGY-1 vs. PGY-3) diskriminieren. Allerdings gibt es auch kritische Befunde: Das HFRS (Morgan et al., 2007) zeigte keine klare Faktorenstruktur und wurde als nicht valide eingestuft. Beim TPOT (Maguire, 2016) wich die empirisch gefundene Zwei-Faktoren-Lösung von der ursprünglichen 5-Domänen-Struktur ab.

Die Kriteriumsvalidität wird durch Korrelationen mit externen Kriterien wie klinischer Leistung oder technischen Skills geprüft. Hier zeigen sich bei einigen Instrumenten gute Ergebnisse: ANTS (Fletcher et al., 2003) korrelierte in 81 % der Fälle mit hoher klinischer Performance, während TEAM (Carpini et al., 2021) eine starke Korrelation mit globalen Leistungsratings ($r = 0{,}62–0{,}81$) aufwies. T-NOTECHS (Steinemann et al., 2012) zeigte höhere Scores bei schnellerer Aufgabenbearbeitung ($r = 0{,}50$). Allerdings wurden viele Instrumente nur in Simulationen und nicht in realen klinischen Settings validiert, und es fehlen Studien, die den Zusammenhang zwischen NTS-Bewertungen und Patientenoutcomes (z. B. Mortalität, Komplikationsraten) untersuchen.

Die Reliabilität gibt an, wie zuverlässig ein Instrument misst, und wird anhand verschiedener Kennwerte bewertet (vgl. Tab. 3). Ein zentrales Maß ist die interne Konsistenz (Cronbachs α), die bei den meisten Instrumenten akzeptable bis hohe Werte aufweist: TEAM ($\alpha = 0{,}91–0{,}93$, Freytag et al., 2019), ANTS ($\alpha = 0{,}79–0{,}86$, Fletcher et al., 2003) und TPOT ($\alpha = 0{,}98$, Maguire, 2016). Die Interrater-Reliabilität (ICC) variiert dagegen stärker: Während OSANTS (CCI = 0,95, Dedy et al., 2015) und TEAM (CCI = 0,66, Freytag et al., 2019) gute Werte zeigen, liegen

T-NOTECHS (CCI = 0,54, Repo et al., 2019) und ANTS (CCI = 0,34–0,81, Jirativanont et al., 2017) im moderaten bis niedrigen Bereich. Die Test-Retest-Reliabilität ist oft schlecht dokumentiert, wie etwa bei TEAM (r = 0,53, Cooper et al., 2010), während beim TPOT (Maguire, 2016) immerhin 60 % der Videos stabile Bewertungen zeigten.

Tabelle 3: ***Reliabilität***

Reliabilitätsmaß	**Akzeptabler Wert**	**Beispiele aus den Instrumenten**
Interne Konsistenz (Cronbachs α)	α ≥ 0.70	**TEAM:** α = 0,91–0,93 (Freytag et al., 2019) **ANTS:** α = 0,79–0,86 (Fletcher et al., 2003) **TPOT:** α = 0,98 (Maguire, 2016)
Interrater-Reliabilität (ICC)	ICC ≥ 0,70	**OSANTS:** CCI = 0,95 (Dedy et al., 2015) **TEAM:** CCI = 0,66 (Freytag et al., 2019) **T-NOTECHS:** CCI = 0,54 (Repo et al., 2019) **ANTS:** CCI = 0,34–0,81 (Jirativanont et al., 2017)
Test-Retest-Reliabilität	r ≥ 0,70	**TEAM**: r = 0,53 (Cooper et al., 2010) **TPOT**: 60 % der Videos zeigten stabile Bewertungen (Maguire, 2016)

Quelle: eigene Darstellung

Die Reliabilität wird von verschiedenen Faktoren beeinflusst: Ratertraining spielt eine entscheidende Rolle – Instrumente mit strukturiertem Training (z. B. OSANTS, TEAM) zeigen höhere ICC-Werte. Auch die Komplexität des Instruments wirkt sich aus: Einfache Skalen wie das Ottawa GRS sind reliabler als detaillierte Checklisten wie OSCAR. Zudem ist das Beobachtungssetting relevant: Videobasierte Bewertungen sind reliabler als Echtzeitbeobachtungen (z. B. T-NOTECHS: ICC = 0,71 vs. 0,48), während Live-Bewertungen zwar praktikabler, aber anfälliger für Verzerrungen sind.

Die Praktikabilität entscheidet schließlich darüber, ob ein Instrument im klinischen Alltag oder in Simulationen einsetzbar ist. Hier spielen Zeitaufwand, Schulungsbedarf, Anwendbarkeit in Echtzeit und Zielgruppenfreundlichkeit eine zentrale Rolle. Einige Instrumente sind besonders praktikabel: TEAM, Ottawa GRS und CALM lassen sich schnell (< 5 Min.) ausfüllen und erfordern nur geringen Schulungsaufwand, während ANTS, NOTSS und OSCAR detaillierter sind, aber mehr Training (4–20 Stunden) und Zeit (5–20 Min.) benötigen. STAT und AOTP sind sehr umfangreich und eher für Forschungszwecke geeignet. Die Anwendbarkeit in Echtzeit variiert ebenfalls: TEAM, T-NOTECHS und CALM eignen sich gut für Live-Bewertungen, während ANTS, NOTSS und OSCAR eher für Videobewertungen konzipiert sind. Die Zielgruppenfreundlichkeit ist bei einfachen Instrumenten wie TEAM und MHPTS hoch, während komplexere Tools wie ANTS für Novizen schwer verständlich sein können.

Die meisten NTS-Instrumente weisen eine akzeptable bis gute Validität und Reliabilität auf, allerdings mit deutlichen Unterschieden zwischen den Tools. Während TEAM, Ottawa GRS und CALM besonders praktikabel und für Echtzeitbewertungen geeignet sind, erfordern ANTS, NOTSS und OSCAR mehr Aufwand, bieten dafür aber detailliertere Analysen. Die Wahl des Instruments sollte daher stets an den konkreten Einsatzkontext, die Zielgruppe und die verfügbaren Ressourcen angepasst werden. Zudem besteht weiterer Forschungsbedarf, insbesondere zur Kriteriumsvalidität in realen klinischen Settings und zum Zusammenhang zwischen NTS und Patientenoutcomes.

6.3 Eignung für klinische und edukative Kontexte

Die Eignung von Instrumenten zur Erfassung nicht-technischer Fähigkeiten (NTS) hängt maßgeblich davon ab, ob sie für formative (Feedback, Training) oder summative Zwecke (Prüfungen, Zertifizierungen) eingesetzt werden sollen (vgl. Tab. 4). Während einige Tools beide Funktionen erfüllen können, sind andere aufgrund ihrer Struktur, Reliabilität oder Praktikabilität nur für bestimmte Anwendungsbereiche geeignet.

Tabelle 4: ***Formative vs. summative Bewertungen***

Instrument	Formativ (Feedback, Training)	Summativ (Prüfungen, Zertifizierungen)	Begründung
TEAM	Sehr gut	Eingeschränkt	**Schnell anwendbar**, gute **Interrater-Reliabilität**, aber **keine Cut-off-Werte** definiert.
AMEISEN	Sehr gut	Nicht geeignet	**Detailliertes Feedback**, aber **komplex und zeitaufwendig**.
NOTSS	Sehr gut	Eingeschränkt	**Chirurgiespezifisch**, aber **keine ausreichende Reliabilität für High-Stakes**.
Ottawa GRS	Gut	Gut	**Einfache Skala**, **hohe Reliabilität**, aber **weniger differenziert**.
RUHIG	Sehr gut	Nicht geeignet	**Fokus auf Führung**, aber **keine Validierung für summative Zwecke**.
T-NO-TECHS	Gut	Eingeschränkt	**Teamorientiert**, aber **moderate Reliabilität**.
TPOT	Gut	Eingeschränkt	**TeamSTEPPS®-basiert**, aber **empirische Struktur weicht von Theorie ab**.
OSCAR	Gut	Nicht geeignet	**Sehr detailliert**, aber **zu komplex für Routineeinsatz**.
STAT	Gut	Nicht geeignet	**Umfangreich**, aber **nur für Videobewertungen**.

Quelle: eigene Darstellung

TEAM eignet sich besonders gut für formative Assessments, da es schnell anwendbar ist und eine gute Interrater-Reliabilität aufweist. Allerdings ist sein Einsatz für summative Bewertungen nur eingeschränkt möglich, da keine klaren Cut-off-Werte definiert sind. ANTS und NOTSS sind ebenfalls hervorragend für formative Zwecke geeignet, da sie detailliertes Feedback er-

möglichen – allerdings sind sie aufgrund ihrer Komplexität und des hohen Zeitaufwands weniger für summative Prüfungen geeignet. Ottawa GRS stellt hier eine Ausnahme dar, da es sowohl für formative als auch summative Bewertungen gut einsetzbar ist: Die einfache Skala und hohe Reliabilität machen es zu einem praktikablen Tool für beide Kontexte, auch wenn es weniger differenzierte Analysen bietet. CALM und RUHIG sind primär für formative Assessments konzipiert, wobei RUHIG aufgrund fehlender Validierung für summative Zwecke ungeeignet ist. T-NOTECHS und TPOT können sowohl formativ als auch eingeschränkt summativ eingesetzt werden, wobei ihre moderate Reliabilität und die Abweichung der empirischen Struktur von der theoretischen Grundlage bei TPOT Grenzen setzen. OSCAR und STAT sind aufgrund ihrer hohen Komplexität und des großen Zeitaufwands vor allem für Forschungszwecke geeignet, während sie für den Routineeinsatz oder summative Bewertungen weniger praktikabel sind.

Für formative Assessments (Feedback, Debriefing) empfehlen sich insbesondere TEAM, ANTS, NOTSS, CALM und T-NOTECHS, da sie eine einfache Anwendung und gute Feedbackqualität bieten. Für summative Assessments (Prüfungen, Zertifizierungen) sind dagegen Ottawa GRS und – mit definierten Cut-off-Werten – TEAM die besten Optionen, da sie eine hohe Reliabilität und einfache Handhabung gewährleisten. Für Forschungszwecke eignen sich vor allem OSCAR, STAT und AOTP, da sie detaillierte Analysen ermöglichen, auch wenn sie mit einem höheren Aufwand verbunden sind.

Neben der Unterscheidung zwischen formativen und summativen Anwendungen ist die kontextspezifische Anpassung der Instrumente entscheidend für ihre Wirksamkeit. Die meisten NTS-Tools wurden für bestimmte klinische Settings entwickelt und lassen sich mit unterschiedlichen Herausforderungen an andere Kontexte anpassen (vgl. Tab. 5).

Tabelle 5: ***Kontextspezifische Anpassungen***

Schauplatz	Geeignete Instrumente	Anpassungsbedarf	Beispiele
Anästhesie	ANTS, ANTSdk, AS-NTS	**Keine großen Anpassungen nötig**, da bereits anästhesie-spezifisch.	**ANTS** wurde erfolgreich in **Dänemark (ANTSdk)** adaptiert (Jepsen et al., 2015).
Chirurgie	NOTSS, DISSANCE	**Fokus auf intraoperative Führung und Entscheidungsfindung**.	**NOTSS** wurde in **Schottland und weltweit** eingesetzt (Yule et al., 2008).
Notfallmedizin	TEAM, T-NOTECHS, OSCAR, RUHIG	**Teamorientierung**, **Stressbewältigung**, **Schnelligkeit**.	**TEAM** wurde in **geburtshilflichen Notfällen** validiert (Carpini et al., 2021).
Traumaversorgung	T-NOTECHS, SOFORT	**Multiprofessionelle Teams**, **Priorisierung unter Zeitdruck**.	**T-NOTECHS** wurde in **Finnland** übersetzt und validiert (Repo et al., 2019).
Geburtshilfe	AOTP, GAOTP, TEAM	**Patienten- und Angehörigenkommunikation**, **interdisziplinäre Zusammenarbeit**.	**AOTP** wurde in **Kanada** entwickelt und getestet (Tregunno et al., 2009).
Pädiatrie	SOFORT, RUHIG	**Fokus auf Führung in Reanimationen**, **Teamkoordination**.	**STAT** wurde für **pädiatrische Reanimationen** entwickelt (Reid et al., 2012).
Pflege	NTS-NAS, MHPTS, TPOT	**Teamarbeit in Pflegekontexten**, **Kommunikation mit Ärzten**.	**MHPTS** wurde in **Frankreich** für Pflegestudenten validiert (Gosselin et al., 2019).

Quelle: eigene Darstellung

In der Anästhesie sind ANTS, ANTSdk und AS-NTS besonders geeignet, da sie bereits anästhesiespezifisch entwickelt wurden und kaum Anpassungen benötigen. So wurde ANTS erfolgreich in Dänemark adaptiert (ANTSdk, Jepsen et al., 2015). Für die Chirurgie eignen sich NOTSS und DISSANCE, da sie den Fokus auf intraoperative Führung und Entscheidungsfindung legen und bereits international eingesetzt wurden (Yule et al., 2008). In der Notfallmedizin kommen vor allem TEAM, T-NOTECHS, OSCAR und RUHIG zum Einsatz, da sie Teamorientierung, Stressbewältigung und Schnelligkeit betonen. TEAM wurde beispielsweise in geburtshilflichen Notfällen validiert (Carpini et al., 2021). Für die Traumaversorgung sind T-NOTECHS und SOFORT besonders geeignet, da sie auf multiprofessionelle Teams und Priorisierung unter Zeitdruck ausgelegt sind. T-NOTECHS wurde in Finnland übersetzt und validiert, wobei einzelne Begriffe angepasst werden mussten (Repo et al., 2019). In der Geburtshilfe kommen AOTP, GAOTP und TEAM zum Einsatz, da sie Patienten- und Angehörigenkommunikation sowie interdisziplinäre Zusammenarbeit berücksichtigen. AOTP wurde in Kanada entwickelt und getestet (Tregunno et al., 2009). Für die Pädiatrie eignen sich SOFORT und RUHIG, da sie auf Führung in Reanimationen und Teamkoordination fokussieren. STAT wurde speziell für pädiatrische Reanimationen entwickelt (Reid et al., 2012). In der Pflege kommen NTS-NAS, MHPTS und TPOT zum Einsatz, da sie Teamarbeit in Pflegekontexten und die Kommunikation mit Ärzten abbilden. MHPTS wurde in Frankreich für Pflegestudenten validiert (Gosselin et al., 2019).

Die Anpassung der Instrumente an verschiedene Kontexte ist jedoch mit Herausforderungen verbunden. Kulturelle Unterschiede spielen eine wichtige Rolle: So mussten bei der Übersetzung von ANTSdk (Dänemark) und der italienischen Version des Ottawa GRS kulturelle Anpassungen vorgenommen werden. Auch T-NOTECHS funktionierte nach der Übersetzung ins Finnische gut, allerdings waren Anpassungen einzelner Begriffe notwendig. Ein weiteres Problem stellen professionsspezifische Anforderungen dar: ANTS und NOTSS sind ärztezentriert und für Pflegekräfte weniger geeignet, während MHPTS und NTS-NAS pflegespezifisch sind und für Ärzte nicht optimal passen. Zudem muss die Teamgröße und -zusammensetzung berücksichtigt werden: TEAM und T-NOTECHS eignen sich für große, multiprofessionelle Teams, während CALM und ANTS besser für kleinere Teams oder Einzelpersonen geeignet sind.

Zusammenfassend lässt sich sagen, dass die Wahl des passenden NTS-Instruments nicht nur von der gewünschten Bewertungsform (formativ vs. summativ), sondern auch vom klinischen Kontext abhängt. Während einige Tools wie TEAM und Ottawa GRS vielseitig einsetzbar sind, erfordern andere wie ANTS, NOTSS oder AOTP spezifische Anpassungen an Setting, Ziel-

gruppe und kulturellen Hintergrund. Eine sorgfältige Auswahl und gegebenenfalls Modifikation der Instrumente ist daher essenziell, um valide und reliable Ergebnisse zu erzielen.

6.4 Stärken und Limitationen der vorliegenden Arbeit

Die vorliegende Bachelorarbeit leistet einen wissenschaftlichen Beitrag zur systematischen Analyse von Non-Technical Skills (NTS)-Instrumenten im Kontext von High-Fidelity-Simulationen (HFS) und deren Anwendung in interprofessionellen Gesundheitsteams. Durch eine strukturierte Aufarbeitung der bestehenden Literatur wird eine evidenzbasierte Grundlage für die Bewertung und Auswahl geeigneter Assessment-Tools geschaffen, die sowohl für die klinische Praxis als auch für die medizinische Ausbildung von Relevanz sind. Die Arbeit verbindet theoretische Grundlagen mit praxisorientierten Empfehlungen und trägt damit zur Förderung der Patientensicherheit sowie zur Optimierung interprofessioneller Zusammenarbeit bei.

Ein zentraler Mehrwert der Untersuchung besteht in der detaillierten Analyse struktureller und psychometrischer Eigenschaften verschiedener NTS-Instrumente. Die vergleichende Betrachtung von ausgewählten zwanzig etablierten Bewertungstools (vgl. Anhang, Tabelle 8) – darunter ANTS, NOTSS, TEAM und die Ottawa GRS – ermöglicht eine differenzierte Evaluation ihrer Anwendungsmöglichkeiten in unterschiedlichen Settings. Die Ergebnisse zeigen, dass kein universell einsetzbares Instrument existiert, sondern die Auswahl von spezifischen Kontextfaktoren abhängt, insbesondere der Zielsetzung (formativ vs. summativ), der Zielgruppe (mono- vs. multiprofessionell) sowie dem Anwendungssetting (Notaufnahme, Operationssaal, Intensivstation). Diese Erkenntnis unterstreicht die Notwendigkeit einer kontextspezifischen Instrumentenauswahl und bietet eine fundierte Orientierungshilfe für Praktiker:innen und Forschende.

Besonders hervorzuheben ist die praxisnahe Aufbereitung der Ergebnisse durch die Entwicklung einer Instrumentenauswahlmatrix (Tabelle 6), die eine unmittelbare Anwendbarkeit der Erkenntnisse ermöglicht. Die Matrix dient als Entscheidungshilfe für die Implementierung von NTS-Instrumenten in verschiedenen klinischen und edukativen Kontexten. So wird beispielsweise das TEAM-Instrument für multiprofessionelle Debriefings in der Notaufnahme empfohlen, während ANTS in anästhesiologischen Simulationen und NOTSS in chirurgischen Trainings eingesetzt werden sollte. Zudem wird das DESC-Tool (TeamSTEPPS) als strukturierte Methode zur Konfliktlösung in interprofessionellen Teams vorgestellt. Die theoretische Fundierung der Arbeit verdeutlicht die Bedeutung von NTS für die Patientensicherheit und zeigt auf, dass nicht-technische Fähigkeiten wie Situation Awareness, Decision Making und Communi-

cation entscheidend für die Fehlervermeidung und die Verbesserung der Teamperformance sind.

Ein weiterer Beitrag der Arbeit liegt in der systematischen Darstellung der NTS-Domänen in den Tabellen 7 bis 9. Tabelle 7 bietet eine abstrakte Gesamtübersicht über zentrale NTS-Domänen im Vergleich zu zwanzig Bewertungsinstrumenten und ermöglicht eine schnelle Orientierung über inhaltliche Schwerpunkte und Überschneidungen. Die Obstetric Global Rating Scale (GRS) verwendet kein NTS-Kategoriensystem im klassischen Sinne. Das Instrument besteht ausschließlich aus einer fünfstufigen globalen Bewertungsskala. Diese Zuordnung ist interpretativ und basiert auf inhaltlicher Analyse der Items. Sie weicht von der Originalstruktur des Instruments ab und ist daher mit Vorsicht zu interpretieren. Da diese Bewertungsstufen keine inhaltlichen NTS-Dimensionen darstellen, ist eine direkte Zuordnung zu den übrigen Kategorien (Situation awareness, Communication, Leadership etc.) nicht möglich. Die GRS erhält daher in Tabelle 7 eine eigene Kategorienspalte mit den fünf Leistungsstufen als Zeilen. Alle anderen Instrumente lassen diese Zeilen leer. Diese Darstellung dient der Vollständigkeit des Vergleichs, ist jedoch strukturell nicht mit den übrigen 19 Instrumenten vergleichbar. Tabelle 8 vertieft diese Übersicht durch die Darstellung der Top-Level-Kategorien und Elemente, während Tabelle 9 die vollständige Hierarchie der Instrumente abbildet und eine detaillierte Analyse einzelner Behavioral Marker ermöglicht. Diese strukturierte Aufbereitung erleichtert nicht nur die Auswahl geeigneter Instrumente, sondern schafft auch eine solide Grundlage für weiterführende Forschungsarbeiten. Bei der Erstellung der Tabellen wurde eine Vereinheitlichung der Terminologie vorgenommen, um die Vergleichbarkeit der Instrumente zu gewährleisten. So wurden beispielsweise „Team Management“ (CALM) unter „Leadership“ subsumiert, da es sich um einen Führungsstil handelt, während „Human Factors“ (STAT) als Oberbegriff behandelt wurde, da es sich nicht um eine spezifische NTS-Kategorie handelt. „Environment in the Room“ (AOTP) wurde der „Situation Awareness“ zugeordnet, da es sich um ein klassisches SA-Element handelt. Technische Skills wie „Basic Assessment Skills“ (STAT) wurden separat als „Clinical Assessment“ klassifiziert, da sie nicht zu den NTS im engeren Sinne zählen.

Trotz dieser Stärken weist die Untersuchung methodische und inhaltliche Limitationen auf. Eine zentrale Einschränkung besteht in der selektiven, nicht-systematischen Literaturrecherche. Da die Auswahl der analysierten Studien nicht nach den strengen Kriterien einer systematischen Übersichtsarbeit erfolgte, besteht die Gefahr, dass relevante Studien unberücksichtigt blieben und die Ergebnisse durch subjektive Präferenzen beeinflusst wurden. Eine systematische Literaturrecherche mit klar definierten Suchstrategien und transparenten Auswahlkriterien hätte die wissenschaftliche Aussagekraft und Nachvollziehbarkeit der Arbeit erhöht. Wei-

tere methodische Herausforderungen ergeben sich aus der Sprachbarriere, da alle analysierten Quellen in englischer Sprache vorlagen. Die Übersetzung beschränkte sich auf einzelne Fachbegriffe, was mögliche Verzerrungen durch Übersetzungsungenauigkeiten nicht ausschließt. Zudem konnten aufgrund des begrenzten zeitlichen Rahmens nicht alle potenziell relevanten Datenbanken berücksichtigt werden, und bei vier Arbeiten waren die Volltexte nicht zugänglich. Dies könnte dazu geführt haben, dass wichtige Studien und Inhalte nicht in die Analyse einflossen.

Ein weiterer Kritikpunkt betrifft die Validierungslücken der analysierten NTS-Instrumente. Viele Tools wurden primär in simulierten Settings validiert, während Daten aus klinischen Kontexten fehlen. So zeigt beispielsweise das TEAM-Instrument in Simulationen eine Interrater-Reliabilität (ICC) von 0,66, in der realen Notaufnahme jedoch nur eine ICC von 0,42 (Freytag et al., 2019). Diese Diskrepanz wirft Fragen zur Übertragbarkeit der Ergebnisse auf die klinische Praxis auf. Zudem variiert die psychometrische Robustheit der Instrumente erheblich: Während TPOT eine hohe interne Konsistenz ($\alpha = 0,98$) aufweist, zeigt NOTSS eine deutlich geringere Interrater-Reliabilität (ICC = 0,29–0,66). Diese Unterschiede unterstreichen den Bedarf an weiteren Validierungsstudien in realen klinischen Settings. Auch die kulturelle Adaptierbarkeit der Instrumente stellt eine Herausforderung dar. Viele Tools wurden für englischsprachige Kontexte entwickelt und sind nicht ausreichend an andere kulturelle und sprachliche Gegebenheiten angepasst. Zwar existieren erfolgreiche Adaptationen wie ANTSdk (Dänemark) oder die Italian Ottawa GRS, jedoch fehlt es an systematischen Validierungsstudien für den deutschen Sprachraum. Dies schränkt die Übertragbarkeit der Ergebnisse ein und verdeutlicht den Bedarf an kulturspezifischen Anpassungen.

Trotz dieser Limitationen liefert die Arbeit wertvolle Impulse für Praxis und Forschung. NTS-Instrumente sind unverzichtbar für die Förderung von Teamkompetenz und Patientensicherheit. Ihre Weiterentwicklung sollte sich auf technologische Innovationen wie KI-gestützte Videoanalysen, Virtual Reality (VR)-Simulationen und Wearables konzentrieren, die vielversprechende Ansätze für die automatisierte Erfassung und Bewertung von NTS bieten. Diese Technologien könnten die Objektivität und Effizienz der Assessments erhöhen und den Rater-Bias reduzieren. Zudem sollten modulare und adaptive Instrumente entwickelt werden, die eine flexible Anpassung an verschiedene Kontexte ermöglichen. Kurzversionen bestehender Tools oder hybride Instrumente könnten die Praktikabilität und Akzeptanz verbessern. Ein weiterer wichtiger Forschungsbereich ist die evidenzbasierte Validierung der Instrumente in klinischen Kontexten. Längsschnittstudien zur Übertragbarkeit von Simulationsdaten auf die Praxis sowie die Korrelation von NTS-Bewertungen mit klinischen Outcomes sind notwendig, um die Aus-

sagekraft der Instrumente zu stärken. Zudem sollten kulturelle und interprofessionelle Anpassungen durch systematische Validierungsstudien vorangetrieben werden

Zusammenfassend bietet die Arbeit eine fundierte Grundlage für die Auswahl und Implementierung von NTS-Instrumenten in der medizinischen Ausbildung und klinischen Praxis. Die strukturierte Analyse der bestehenden Tools und die praxisnahen Empfehlungen tragen dazu bei, die Teamperformance in kritischen Situationen zu optimieren und die Patientensicherheit nachhaltig zu verbessern. Durch die Weiterentwicklung der Instrumente im Hinblick auf Technologie, Modularität und Validierung können diese einen noch größeren Beitrag zur Förderung interprofessioneller Kompetenzen und zur Fehlervermeidung im Gesundheitswesen leisten.

6.5 Empfehlungen für Praxis und Forschung

Die Auswahl und Anwendung von Instrumenten zur Erfassung nicht-technischer Fähigkeiten (NTS) sollte stets am konkreten Anwendungszweck, Setting und Zielgruppe ausgerichtet sein (vgl. Tab. 6). Für formatives Feedback eignen sich besonders TEAM, ANTS, NOTSS, CALM und T-NOTECHS, da sie eine einfache Anwendung, gute Feedbackqualität und hohe Akzeptanz bei den Anwendern bieten. Soll hingegen eine summative Bewertung – etwa im Rahmen von Prüfungen oder Zertifizierungen – erfolgen, sind Ottawa GRS und TEAM (mit definierten Cut-off-Werten) die besten Optionen, da sie eine hohe Reliabilität und einfache Handhabung gewährleisten. Für Teamtrainings empfehlen sich TEAM, T-NOTECHS, TPOT und MHPTS, da sie teamorientiert, praktikabel und für multiprofessionelle Teams geeignet sind. Bei der individuellen Bewertung von Fachkräften kommen dagegen ANTS, NOTSS, AS-NTS und OSANTS infrage, da sie professionsspezifisch sind und eine detaillierte Analyse ermöglichen.

Tabelle 6: ***Auswahl geeigneter Instrumente***

Anwendungszweck	**Empfohlene Instrumente**	**Begründung**
Formatives Feedback	TEAM, AMEISEN, NOTSS, CALM, T-NOTECHS	**Einfache Anwendung**, **gute Feedbackqualität**, **hohe Akzeptanz**.
Summative Bewertungen	Ottawa GRS, TEAM (mit Cut-off-Werten)	**Hohe Reliabilität**, **einfache Handhabung**, **globale Bewertung möglich**.
Teamtraining	TEAM, T-NOTECHS, TPOT, MHPTS	**Teamorientiert**, **praktikabel**, **für multiprofessionelle Teams geeignet**.
Individuelle Bewertung	AMEISEN, NOTSS, AS-NTS, OSANTS	**Professionsspezifisch**, **detaillierte Analyse**.
Notfallmedizin	TEAM, T-NOTECHS, OSCAR, RUHIG	**Schnelle Anwendung**, **Fokus auf Teamkoordination**.
Geburtshilfe	AOTP, GAOTP, TEAM	**Patienten- und familienzentriert**, **interdisziplinär**.
Chirurgie	NOTSS, DISSANCE	**Intraoperative Führung**, **Entscheidungsfindung**.
Anästhesie	ANTS, ANTSdk, AS-NTS	**Anästhesiespezifisch**, **kognitive und soziale Fähigkeiten**.
Pflegeausbildung	NTS-NAS, MHPTS, TPOT	**Pflegespezifisch**, **einfache Anwendung**.

Quelle: eigene Darstellung

In der Notfallmedizin haben sich TEAM, T-NOTECHS, OSCAR und RUHIG bewährt, da sie eine schnelle Anwendung ermöglichen und den Fokus auf Teamkoordination legen. Für die Geburtshilfe eignen sich AOTP, GAOTP und TEAM, da sie patienten- und familienzentrierte Aspekte sowie interdisziplinäre Zusammenarbeit berücksichtigen. In der Chirurgie sind NOTSS und DISSANCE die Instrumente der Wahl, da sie intraoperative Führung und Entscheidungsfindung abbilden. Für die Anästhesie empfehlen sich ANTS, ANTSdk und AS-NTS,

da sie anästhesiespezifisch sind und sowohl kognitive als auch soziale Fähigkeiten erfassen. In der Pflegeausbildung kommen NTS-NAS, MHPTS und TPOT zum Einsatz, da sie pflegespezifisch und einfach anwendbar sind.

Damit die Instrumente zuverlässige Ergebnisse liefern, ist eine strukturierte Schulung der Anwender unerlässlich. Ein Ratertraining – etwa ein 4-stündiges Training für ANTS oder ein 1-stündiges Training für TEAM – erhöht die Reliabilität der Bewertungen deutlich. Besonders effektiv ist das Frame-of-Reference-Training, bei dem die Rater gemeinsam Beispielvideos bewerten und ihre Einschätzungen diskutieren, um eine einheitliche Bewertungsgrundlage zu schaffen. Zudem sollte eine regelmäßige Kalibrierung – beispielsweise durch monatliche Besprechungen von Bewertungsdifferenzen – erfolgen, um die Interrater-Übereinstimmung langfristig zu sichern.

Die Integration der Instrumente in Simulationstrainings bietet eine ideale Möglichkeit, NTS gezielt zu schulen und zu bewerten. Eine sinnvolle Kombination mit technischen Checklisten – etwa TEAM mit dem PALS-Algorithmus – ermöglicht umfassende Bewertungen. Das Debriefing sollte mit konkreten Verhaltensankern strukturiert werden, um den Lernern klare Handlungsempfehlungen zu geben (z. B. *„Im nächsten Szenario sollten Sie die Aufgabenverteilung klarer kommunizieren“*). Zudem kann eine Kombination aus Selbst- und Fremdbewertung – etwa TEAM für die Teamleistung und individuelle Reflexion – den Lerneffekt verstärken.

Für den Einsatz in klinischen Settings sind einfache Instrumente wie TEAM oder Ottawa GRS praktikabler als komplexe Checklisten, da sie schnell anwendbar sind und weniger Ressourcen erfordern. Nach kritischen Ereignissen – etwa Reanimationen oder Traumaversorgungen – können kurze Feedbackrunden (5–10 Minuten) wertvolle Erkenntnisse liefern. Zudem empfiehlt sich die Dokumentation der Bewertungen in Logbüchern, beispielsweise ANTS für die Anästhesie-Weiterbildung, um den Lernfortschritt zu verfolgen und gezielt zu fördern.

Neben diesen praktischen Empfehlungen besteht auch weiterer Forschungsbedarf, um die Instrumente kontinuierlich zu verbessern und ihre Aussagekraft zu erhöhen. Eine wichtige Aufgabe ist die Weiterentwicklung der Tools, etwa durch die Vereinfachung komplexer Instrumente wie OSCAR für den Echtzeiteinsatz oder die Erweiterung von Subskalen – beispielsweise mehr Items zu Konfliktmanagement im TPOT. Zudem sollten Cut-off-Werte für summative Bewertungen definiert werden, um klare Kompetenzstandards zu setzen (z. B. *„Ab 30 Punkten gilt ein Team als kompetent“*).

Ein zentraler Forschungsbereich ist die Validierung der Instrumente in realen klinischen Settings. Bisher wurden die meisten Tools nur in Simulationen getestet, sodass Längsschnittstu-

dien notwendig sind, um ihren Einfluss auf Patientenoutcomes zu untersuchen (z. B. *„Führen höhere TEAM-Scores zu weniger Komplikationen?“*). Zudem sollten multizentrische Studien die Generalisierbarkeit der Ergebnisse prüfen, etwa durch den Einsatz von TEAM in verschiedenen Ländern. Vergleichende Studien – etwa TEAM vs. T-NOTECHS in Traumateams – könnten Aufschluss darüber geben, welches Instrument in welchem Kontext am besten geeignet ist. Auch Kosten-Nutzen-Analysen wären sinnvoll, um zu bewerten, ob sich der Aufwand für komplexe Instrumente wie OSCAR im Vergleich zu einfacheren Tools wie TEAM lohnt.

Ein weiterer wichtiger Aspekt ist die kulturelle und kontextuelle Anpassung der Instrumente. Systematische Übersetzungen und Validierungen – etwa von TEAM in weitere Sprachen – sind notwendig, um ihre internationale Einsetzbarkeit zu gewährleisten. Zudem sollten die Tools an lokale Gegebenheiten angepasst werden, beispielsweise T-NOTECHS für deutsche Traumateams. Die technologische Unterstützung bietet hier neue Möglichkeiten, etwa durch App-basierte Bewertungen oder KI-gestützte Auswertungen von Videodaten, die automatische Analysen – beispielsweise zur Erkennung von Closed-Loop Communication – ermöglichen.

Schließlich sollten interprofessionelle Perspektiven stärker berücksichtigt werden. Die Entwicklung von Instrumenten für multiprofessionelle Teams – etwa Ärzte, Pflegekräfte und Rettungsdienst – könnte die Zusammenarbeit verbessern. Zudem wäre es sinnvoll, Patienten und Angehörige in die Bewertung einzubeziehen, wie es bereits bei AOTP in der Geburtshilfe der Fall ist. Durch diese Maßnahmen könnten die NTS-Instrumente weiter optimiert und ihr Nutzen für die Patientensicherheit und die Qualität der Versorgung nachhaltig gesteigert werden.

6.6 Fazit und Ausblick

Die vorliegende Arbeit verdeutlicht, dass nicht-technische Fähigkeiten (NTS) einen entscheidenden Einfluss auf Patientensicherheit und Teamleistung haben und dass es eine Vielzahl von Instrumenten zu ihrer Erfassung gibt. Diese Tools unterscheiden sich jedoch deutlich in ihrer Struktur, Zielgruppe, psychometrischen Qualität und Praktikabilität, was eine differenzierte Betrachtung erfordert.

Ein zentrales Ergebnis ist, dass es kein universell einsetzbares "One-Size-Fits-All"-Instrument gibt. Die Wahl des passenden Tools hängt vielmehr von der spezifischen Zielgruppe, dem klinischen Setting und dem Verwendungszweck ab. Während TEAM und Ottawa GRS durch ihre Praktikabilität und Vielseitigkeit überzeugen und sich für verschiedene Anwendungsbereiche eignen, sind ANTS und NOTSS zwar professionsspezifisch und detailliert, erfordern jedoch einen höheren Aufwand in der Anwendung. Die psychometrische Qualität der Instrumente va-

riiert ebenfalls stark: Zwar ist die interne Konsistenz bei den meisten Tools hoch (Cronbachs α > 0,80), doch die Interrater-Reliabilität liegt häufig nur im moderaten Bereich (ICC = 0,40–0,70) und ist stark vom Training der Anwender abhängig. Dies unterstreicht die Notwendigkeit strukturierter Schulungen, um zuverlässige Bewertungen zu gewährleisten.

Ein weiterer entscheidender Faktor ist die Praktikabilität der Instrumente. Einfache Tools wie TEAM oder Ottawa GRS eignen sich besonders für Echtzeitbewertungen im klinischen Alltag, während komplexe Instrumente wie OSCAR oder STAT aufgrund ihres hohen Zeitaufwands vorrangig für Forschungszwecke geeignet sind. Zudem zeigt sich, dass die Instrumente an verschiedene kulturelle und kontextuelle Gegebenheiten angepasst werden müssen. So erfordern Übersetzungen – etwa von T-NOTECHS ins Finnische – nicht nur sprachliche, sondern auch inhaltliche Anpassungen, um valide Ergebnisse zu liefern. Auch professionsspezifische Instrumente wie ANTS für die Anästhesie sind nicht ohne Weiteres auf andere Berufsgruppen übertragbar, was ihre universelle Einsetzbarkeit einschränkt.

Für die Zukunft ergeben sich mehrere vielversprechende Ansätze zur Weiterentwicklung der NTS-Instrumente. Eine wichtige Aufgabe besteht darin, die Tools einfacher, reliabler und valider zu gestalten, um ihre Anwendung im klinischen Alltag zu erleichtern. Zudem sollte die Integration in Aus- und Weiterbildung vorangetrieben werden, beispielsweise durch den Einsatz von ANTS in der Anästhesie-Weiterbildung oder TEAM in Notfallkursen. Ein besonders spannendes Forschungsfeld ist die Untersuchung des Zusammenhangs zwischen NTS und Patientenoutcomes – etwa die Frage, ob bessere nicht-technische Fähigkeiten tatsächlich zu weniger Komplikationen führen. Auch die technologische Unterstützung bietet neue Möglichkeiten, etwa durch KI-gestützte Bewertungen oder digitale Feedback-Tools, die die Erfassung und Auswertung von NTS effizienter gestalten könnten.

Abschließend lässt sich festhalten, dass die systematische Erfassung von Non-Technical Skills einen wichtigen Beitrag zur Verbesserung der Patientensicherheit leisten *kann*. Die vorliegende Arbeit bietet eine umfassende Übersicht über die bestehenden Instrumente und zeigt deren Stärken, Schwächen sowie Anwendungsmöglichkeiten auf. Für die Praxis empfiehlt sich der Einsatz von TEAM, Ottawa GRS oder CALM für formative Bewertungen, während Ottawa GRS oder TEAM (mit definierten Cut-off-Werten) für summative Assessments geeignet sind. Um die Qualität und Praktikabilität der Instrumente weiter zu verbessern, sind jedoch weiterführende Validierungsstudien, kulturelle Anpassungen und technologische Innovationen notwendig. Nur so kann sichergestellt werden, dass NTS-Instrumente ihr volles Potenzial entfalten und langfristig zu einer höheren Sicherheit und Qualität in der Patientenversorgung *beitragen können*.

7. Empfehlungen

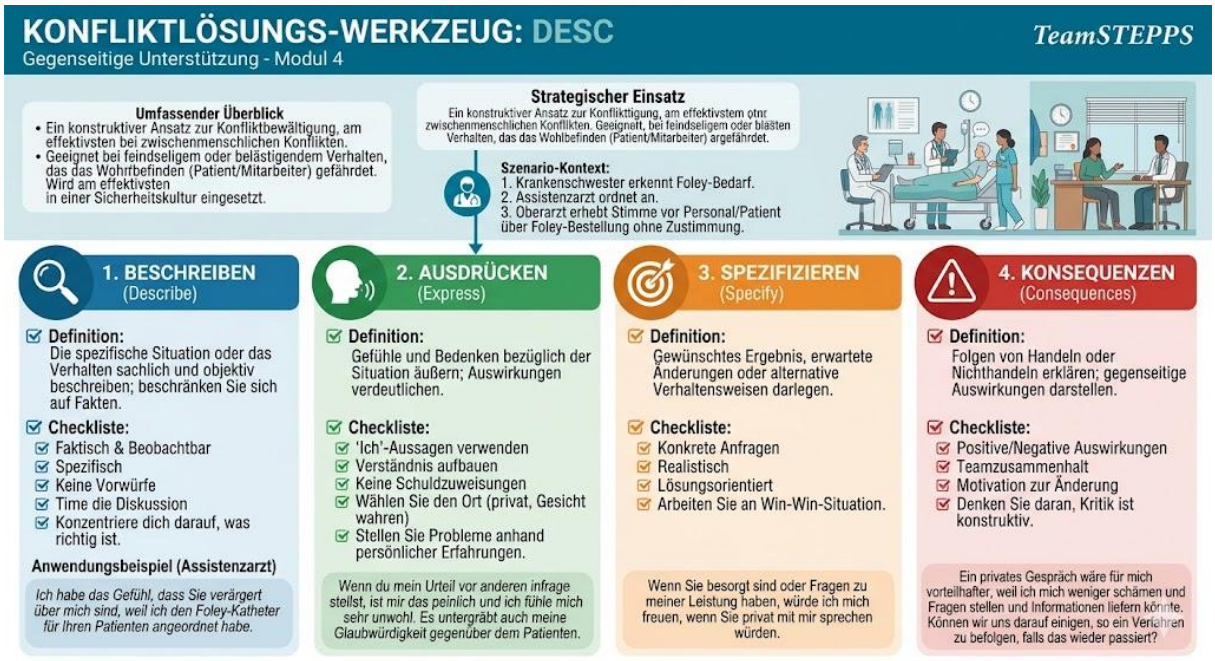
Non-Technical Skills (NTS) sind in der klinischen Praxis längst kein Randthema mehr, sondern ein zentraler Baustein für Patientensicherheit und Teamperformance. Doch wie lassen sich die theoretischen Erkenntnisse über NTS-Instrumente konkret in die Praxis übertragen? Und welche innovativen Ansätze könnten die Erfassung dieser Fähigkeiten in Zukunft revolutionieren? Um diese Fragen zu beantworten, lohnt ein Blick auf konkrete Anwendungsbeispiele, die sowohl die Stärken als auch die Grenzen bestehender Instrumente aufzeigen – und gleichzeitig den Weg für zukünftige Entwicklungen ebnen.

Ein besonders anschauliches Beispiel ist der Einsatz des **TeamSTEPPS® Team Performance Observation Tool (TPOT)** in interprofessionellen Schockraum-Teams. Man kann sich das so vorstellen: Ein Notfallteam – bestehend aus Notarzt, Pflegekraft und Rettungssanitäter – behandelt einen polytraumatisierten Patienten in einer High-Fidelity-Simulation. Während des Szenarios wird das Team mit TPOT bewertet, einem Instrument, das speziell für die Erfassung von Teamwork, Führung und Kommunikation entwickelt wurde. Die Stärken von TPOT liegen auf der Hand: Es deckt interprofessionelle Dynamiken ab, indem es Items wie *"Delegiert Aufgaben angemessen"* oder *"Verwendet das DESC-Skript zur Lösung von Konflikten"* integriert. Die Bewertung wird durch konkrete Verhaltensanker erleichtert – etwa die Frage, ob alle Teammitglieder nach ihrer Einschätzung gefragt werden. Die 5-stufige Likert-Skala ermöglicht zudem ein differenziertes Feedback, das nicht nur Stärken, sondern auch konkrete Verbesserungspotenziale aufzeigt, wie etwa fehlende Closed-Loop Communication.

Doch trotz dieser Vorzüge offenbart TPOT auch Herausforderungen, die typisch für viele NTS-Instrumente sind. So zeigt die empirische Forschung, dass die ursprünglich postulierte 5-Domänen-Struktur des Tools in der Praxis auf zwei zentrale Faktoren – *Partizipative Führung* und *Konfliktmanagement* – reduziert werden kann. Diese Diskrepanz zwischen Theorie und Empirie stellt Rater vor ein Dilemma: Sollen sie sich an der theoretischen Struktur orientieren oder an der empirisch validierten? Eine mögliche Lösung wäre ein Hybridmodell, das beide Perspektiven vereint – etwa indem die Bewertung nach den fünf Domänen erfolgt, die Auswertung jedoch auf die beiden empirisch bestätigten Faktoren fokussiert. Eine weitere Hürde ist der Schulungsaufwand: Während erfahrene Rater eine gute Interrater-Reliabilität (ICC = 0,68) erreichen, liegt diese bei Novizen deutlich niedriger (ICC = 0,45). Hier könnten standardisierte Schulungsvideos helfen, die anhand von Beispielen zeigen, wie NTS wie das DESC-Skript korrekt angewendet werden (vgl. Abb. 30). Zudem wäre eine Erweiterung der Conflict Manage-

ment-Subskala sinnvoll, die derzeit mit nur zwei Items unterrepräsentiert ist. Regelmäßige Kalibrierungssitzungen, in denen Rater gemeinsam Videos analysieren, könnten die Reliabilität zusätzlich steigern.

Abbildung 30: Werkzeug: DESC



Quelle: *Erstellt mit dem KI-Tool NotebookLM und anschließend manuell überarbeitet; inhaltliche Grundlage vgl. Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (2023). Verfügbar unter: https://www.ahrq.gov/teamstepps-program/curriculum/mutual/tools/desc.html*

Ein anderes Instrument, das sich in der Praxis bewährt hat, ist die **Team Notfallbewertungsmaßnahme (TEAM)**, die besonders in geburtshilflichen Simulationen zum Einsatz kommt. In einem typischen Szenario trainiert ein interdisziplinäres Team – bestehend aus Hebamme, Gynäkologe, Anästhesisten und Kinderarzt – die Bewältigung einer postpartalen Hämorrhagie im Rahmen einer SimWars-Simulation. Dabei bewertet das Publikum, darunter Studenten und Fachkräfte, das Team live mit TEAM. Die Stärken dieses Instruments liegen in seiner Einfachheit: Mit nur 11 Items und einem globalen Rating lässt sich eine Bewertung in weniger als fünf Minuten durchführen. Die hohe Interrater-Reliabilität (ICC = 0,98) zeigt, dass auch Laien-Rater konsistente Ergebnisse liefern können. Zudem ermöglicht das Live-Feedback eine unmittel-

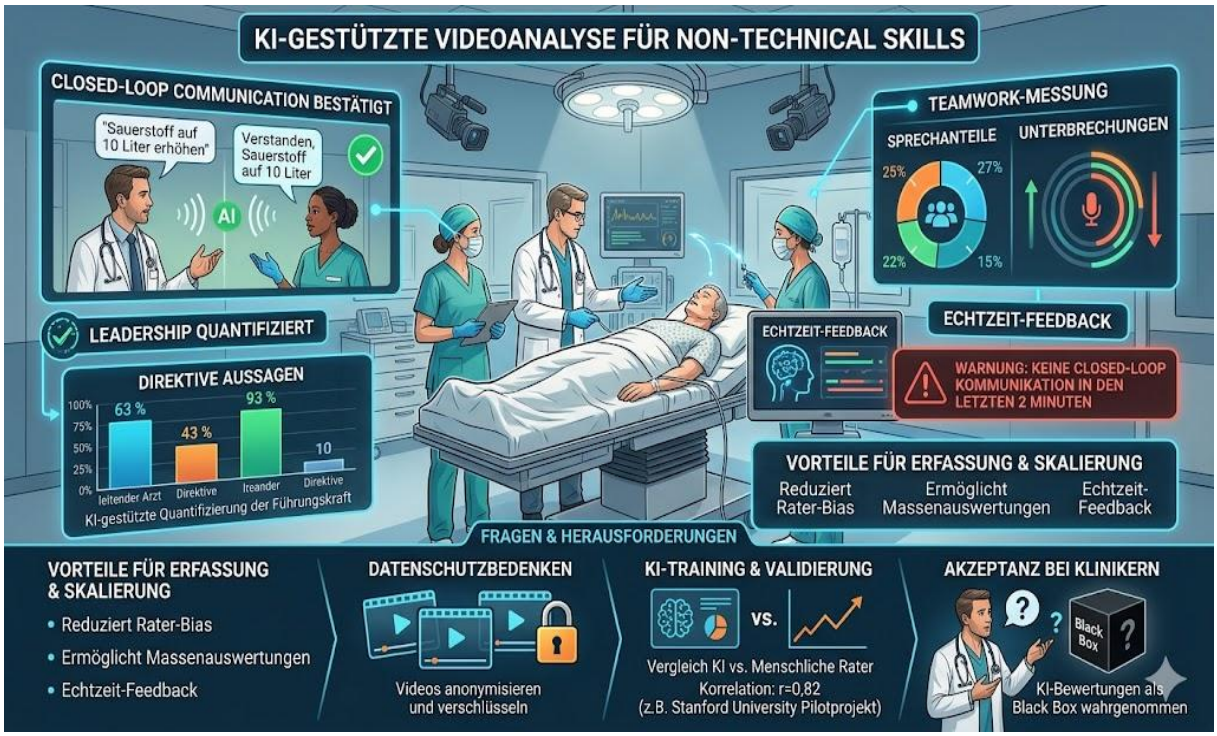
bare Rückmeldung, etwa wenn die Leadership zwar klar, die Aufgabenverteilung jedoch unstrukturiert erscheint.

Doch auch TEAM ist nicht frei von Limitationen. Ein zentrales Problem ist das Fehlen von Cut-off-Werten, die eine summative Bewertung – etwa im Sinne von "bestanden" oder "nicht bestanden" – ermöglichen würden. Hier könnte eine Delphi-Studie mit Experten helfen, Schwellenwerte zu definieren, etwa dass Teams mit sieben oder weniger von zehn Punkten nachschulen müssen. Noch gravierender ist jedoch die fehlende Validierung in realen klinischen Settings. Bisher wurde TEAM ausschließlich in Simulationen geprüft, sodass unklar bleibt, ob die Ergebnisse auf echte Notfälle übertragbar sind. Eine prospektive Studie in Geburtskliniken, die TEAM-Scores mit Outcome-Parametern wie Blutverlust oder Sectio-Rate vergleicht, wäre ein wichtiger nächster Schritt. Zudem könnte das globale Rating durch konkrete Verhaltensanker objektiviert werden, etwa indem definiert wird, dass zehn Punkte nur vergeben werden, wenn alle Items perfekt erfüllt sind.

Während diese Beispiele zeigen, wie NTS-Instrumente heute eingesetzt werden, deuten sich bereits die Konturen einer Zukunft an, in der Technologie eine zentrale Rolle spielen könnte. Ein besonders vielversprechender Ansatz ist die **KI-gestützte Videoanalyse** (vgl. Abb. 31)., die das Potenzial hat, die Erfassung von NTS zu objektivieren und zu skalieren. Stellen wir uns vor, eine KI analysiert Simulationsvideos und erkennt automatisch, ob Closed-Loop Communication angewendet wird – etwa indem sie bestätigt, dass ein Befehl wie *"Sauerstoff auf 10 Liter erhöhen"* mit *"Verstanden, Sauerstoff auf 10 Liter"* quittiert wird. Auch Leadership könnte die KI quantifizieren, indem sie erfasst, wer die meisten direktiven Aussagen trifft. Selbst Teamwork ließe sich messen, etwa durch die Analyse von Sprechanteilen und Unterbrechungen. Die Vorteile liegen auf der Hand: KI reduziert Rater-Bias, ermöglicht Massenauswertungen und könnte sogar Echtzeit-Feedback geben – etwa eine Warnung, wenn in den letzten zwei Minuten keine Closed-Loop Communication stattgefunden hat.

Doch der Einsatz von KI wirft auch Fragen auf. Datenschutzbedenken sind ein zentrales Thema, da Videos anonymisiert und verschlüsselt werden müssen. Zudem muss die KI zunächst trainiert und validiert werden, etwa durch den Vergleich mit menschlichen Ratern. Nicht zuletzt könnte die Akzeptanz bei Klinikern leiden, die KI-Bewertungen als "Black Box" wahrnehmen. Pilotprojekte wie an der Stanford University (vgl. Stanford University. (2026). *Use of artificial intelligence to assess trainee communication compared to human assessment)*, wo KI die Teamwork-Qualität in OP-Simulationen mit einer Korrelation von r = 0,82 zu menschlichen Bewertungen erfasst, zeigen jedoch, dass dieser Ansatz vielversprechend ist.

Abbildung 31: KI-gestützte Videoanalyse



Quelle: Mit dem KI-Tool Gemini erstellt und anschließend manuell überarbeitet

Ein weiterer innovativer Ansatz ist der Einsatz von **Wearables**, die physiologische Daten wie Herzfrequenz oder Hautleitfähigkeit nutzen, um Stresslevel und Teamdynamik in Echtzeit zu erfassen. So könnte eine hohe Herzfrequenzvariabilität auf Überforderung hinweisen, während die Synchronisation von Bewegungen – etwa, wenn mehrere Teammitglieder gleichzeitig nach einem Instrument greifen – mit besserer Performance korreliert. Wearables hätten den Vorteil, dass sie auch unbewusste Verhaltensmuster erfassen, die menschlichen Ratern entgehen. Zudem könnten sie als Frühwarnsystem fungieren, etwa indem sie anzeigen, wenn ein Teammitglied Anzeichen von Überlastung zeigt.

Doch auch hier gibt es Herausforderungen. Wie lassen sich relevante Signale von Rauschen trennen? Und wie steht es um die ethischen Implikationen, wenn Teammitglieder kontinuierlich überwacht werden? Zudem bleibt fraglich, ob physiologische Daten tatsächlich mit NTS korrelieren. Pilotprojekte wie am MIT Media Lab, wo Wearables die Team-Synchronisation in OP-Teams messen, oder an der Universität Heidelberg, wo Herzfrequenzvariabilität als Prädiktor für Fehler in Simulationen untersucht wird, könnten hier wertvolle Erkenntnisse liefern (vgl. Franklin et al. 2023).

Nicht zuletzt könnte **Gamification** die NTS-Erfassung und -Schulung revolutionieren. Virtuelle Realität (VR) und spielerische Elemente wie Punkte oder Leaderboards machen das Training attraktiver und motivierender – besonders für Studenten. In VR-Szenarien könnten Teams Notfallsituationen in immersiven Umgebungen üben, etwa eine Reanimation auf einem virtuellen Schiff. Serious Games könnten Spieler dazu auffordern, NTS anzuwenden, um im Spiel voranzukommen – etwa indem sie ein Team koordinieren, um einen Patienten zu retten. VR-Systeme könnten sogar Echtzeit-Feedback geben, etwa indem sie darauf hinweisen, dass Anweisungen unklar waren und Closed-Loop Communication fehlt.

iVR bietet vor allem Standardisierung und Reproduzierbarkeit. In der Studie schnitt Team-iVR in der realen Simulation besser ab als Einzeltraining: höhere NTS-Werte, weniger technische Fehler und kürzere Operationszeit. Gleichzeitig ist iVR laut Edwards et al. (2023) leicht zugänglich und vergleichsweise ressourcenschonend. Grenzen bleiben aber: Das Training war zeitintensiv, fand nur in einer Simulation statt und der Nutzen für echte Patientensicherheit ist noch ungeklärt.

Trotz dieser Fortschritte bleiben zentrale Forschungsfragen offen. Wie lassen sich NTS-Instrumente in reale klinische Settings übertragen? Können KI und Wearables die Erfassung objektiver gestalten als menschliche Rater? Und wie wirken sich VR-Trainings auf die NTS in echten Notfällen aus? Um diese Fragen zu beantworten, sind prospektive Studien in Krankenhäusern ebenso notwendig wie der Vergleich von KI-Bewertungen mit menschlichen Ratern oder randomisierte kontrollierte Studien zu VR-Trainings. Zudem gilt es, durch Delphi-Studien zu klären, welche NTS in welchen Kontexten besonders relevant sind – etwa in der Notaufnahme oder im Schockraum. Nicht zuletzt muss die Integration von NTS-Instrumenten in bestehende Curricula vorangetrieben werden, etwa durch Implementierungsstudien in der Pflegeausbildung.

Die systematische Analyse zeigt: Non-Technical Skills sind kein "Nice-to-have", sondern ein unverzichtbarer Bestandteil sicherer Patientenversorgung. Die Wahl des richtigen Instruments hängt dabei von Zielgruppe, Setting und Zweck ab. Für interprofessionelle Teams eignen sich TPOT oder TEAM, während für formative Assessments Instrumente wie ANTS oder die Ottawa GRS zu empfehlen sind. Summative Bewertungen sollten auf Tools mit definierten Cut-off-Werten wie der Ottawa GRS basieren. Doch die Zukunft der NTS-Erfassung liegt nicht allein in bewährten Methoden, sondern auch in innovativen Ansätzen wie KI, Wearables und VR. Diese Technologien haben das Potenzial, die Erfassung zu objektivieren, zu skalieren und attraktiver zu gestalten – doch sie müssen validiert, ethisch vertretbar und praxistauglich sein.

Die Forschung sollte sich daher auf vier zentrale Schwerpunkte konzentrieren: die Validierung von Instrumenten in realen klinischen Settings, die Weiterentwicklung bestehender Tools, die Erschließung technologischer Unterstützung und die nachhaltige Integration in medizinische Curricula. Kliniken und Bildungseinrichtungen sind gefordert, NTS-Instrumente systematisch in Simulationstrainings und Assessments zu integrieren – und dabei sowohl bewährte Tools als auch innovative Ansätze zu nutzen. Die Zukunft liegt in einer kombinierten Strategie: bewährte Methoden validieren, neue Technologien erforschen und beides praxistauglich umsetzen. Nur so lässt sich das volle Potenzial von Non-Technical Skills für die Patientensicherheit ausschöpfen.

8. Zusammenfassung

8.1 Problemstellung

Die moderne Gesundheitsversorgung ist geprägt durch hohe Komplexität, Zeitdruck und interdisziplinäre Zusammenarbeit. In diesem Kontext gewinnen Non-Technical Skills (NTS) – definiert als kognitive, soziale und persönliche Fähigkeiten, die technische Fertigkeiten ergänzen – zunehmend an Bedeutung. NTS umfassen drei zentrale Dimensionen:

1. **Kognitive Fähigkeiten** (z. B. Situationsbewusstsein, Entscheidungsfindung),
2. **Soziale/interpersonelle Fähigkeiten** (z. B. Kommunikation, Teamarbeit, Führung),
3. **Persönliche Ressourcen** (z. B. Stressmanagement, Metakognition).

Studien zeigen, dass 70–80 % der medizinischen Fehler auf Defizite in der Teamarbeit und NTS zurückzuführen sind (Cooper et al., 2010; St. Pierre, 2018). Trotz dieser Evidenz fehlt es an standardisierten, validen und praktikablen Instrumenten zur Beurteilung der Teamperformance – insbesondere in klinischen und edukativen Kontexten.

Vor diesem Hintergrund widmete sich die vorliegende Bachelorarbeit der **systematischen Evaluation von Instrumenten zur Erfassung von NTS in Healthcare-Teams**, mit folgenden Forschungsfragen:

1. Welche Charakteristika (Dimensionen, Struktur, Zielgruppen, Settings) weisen publizierte Instrumente zur Messung von Team-NTS in High-Fidelity-Simulationen (HFS) und klinischen Settings auf?
2. Welche psychometrischen Eigenschaften (Validität, Reliabilität, Praktikabilität) kennzeichnen diese Instrumente?
3. Eignen sich die Instrumente für den praktischen Einsatz in klinischen und edukativen Kontexten, und wie lassen sie sich an spezifische Anforderungen anpassen?

8.2 Methodik

Zur Beantwortung der Forschungsfragen wurde eine selektive, nicht-systematische Literaturrecherche durchgeführt. Die Suche erfolgte in den Datenbanken PubMed, Google Scholar sowie der Bibliothek der Katholischen Hochschule Köln mit Schlüsselbegriffen wie „Crew Resource Management (CRM)“, „Non-Technical Skills (NTS)“, „Simulationstraining“, „Teamper-

formance“ und „Behavioral Marker Systems (BMS)“. Ergänzt wurde die Recherche durch Handrecherche und Schneeballverfahren, um relevante Quellen zu identifizieren.

Analysiert wurden über 20 etablierte Instrumente, darunter:

- **ANTS** (*Anaesthetists' Non-Technical Skills*),
- **NOTSS** (*Non-Technical Skills for Surgeons*),
- **TEAM** (*Team Emergency Assessment Measure*),
- **Ottawa GRS** (*Global Rating Scale*),
- **OSCAR** (*Observational Skill-Based Clinical Assessment Tool for Resuscitation*),
- **T-NOTECHS** (*Trauma Non-Technical Skills Scale*),
- **TPOT** (*Team Performance Observation Tool*).

Die Analyse umfasste:

- **Struktur**: Hierarchischer Aufbau (Kategorien → Elemente → Verhaltensanker), Skalierung (Likert-Skalen, Option „nicht beobachtbar“),
- **Zielgruppen**: Ärztliche Teams, Pflegekräfte, multiprofessionelle Teams,
- **Settings**: Anästhesie, Chirurgie, Notfallmedizin, Geburtshilfe, Trauma,
- **Psychometrische Eigenschaften**: Inhaltsvalidität, Konstruktvalidität, Interrater-Reliabilität, interne Konsistenz,
- **Praktikabilität**: Zeitaufwand, Schulungsbedarf, Echtzeitanwendbarkeit.

8.3 Ergebnisse: Struktur, psychometrische Qualität und Einsatzszenarien

I. Hierarchische Struktur und NTS-Dimensionen

Die analysierten Instrumente folgen einem dreistufigen hierarchischen Aufbau:

a) **Kategorien (Domänen)** (z. B. „Führung“, „Teamarbeit“),

b) **Elemente** (z. B. „Priorisierung“, „Kommunikation“),

c) **Verhaltensanker** (konkrete Beobachtungskriterien).

Die meisten Instrumente decken die drei zentralen NTS-Dimensionen ab, wobei die Gewichtung setting- und zielgruppenspezifisch variiert (vgl. Tabelle 7):

- **Kognitive Fähigkeiten**: Situationsbewusstsein, Entscheidungsfindung, Aufgabenmanagement,
- **Soziale/interpersonelle Fähigkeiten**: Kommunikation, Teamarbeit, Leadership,
- **Persönliche Ressourcen**: Stressmanagement, Metakognition.

Beispiele für zielgruppenspezifische Instrumente:

Ärztliche Teams:

- **ANTS**: 4 Kategorien, 15 Elemente (Fokus: Anästhesie),
- **NOTSS**: 5 Kategorien, 14 Elemente (Fokus: Chirurgie).

Pflegekräfte:

- **NTS-NAS** (*Non-Technical Skills for Nursing*),
- **MHPTS** (*Mayo High Performance Teamwork Scale*).

Multiprofessionelle Teams:

- **TEAM**: 3 Kategorien, 11 Items (Notfallmedizin),
- **T-NOTECHS**: 5 Domänen (Trauma-Teams),
- **OSCAR**: 6 Domänen (Reanimation).

Settings:

- Anästhesie (ANTS, ANTSdk),
- Chirurgie (NOTSS),
- Notfallmedizin (TEAM, OSCAR),
- Geburtshilfe (AOTP, GAOTP, TEAM),

- Trauma (T-NOTECHS),
- Allgemeine klinische Praxis (Ottawa GRS, TPOT).

II. Psychometrische Qualität

Die **Validität** der Instrumente ist überwiegend gut, weist jedoch Lücken auf:

- **Inhaltsvalidität**: Basierend auf Expertenkonsens (z. B. TEAM: *Content Validity Index [CVI] = 0,96*),
- **Konstruktvalidität**: Häufig durch Faktorenanalysen bestätigt (z. B. TEAM: eindimensionale Struktur; T-NOTECHS: 5-Domänen-Struktur),
- **Kriteriumsvalidität**: **Selten geprüft** – Ausnahmen wie ANTS (Korrelation mit klinischer Performance in **81 % der Fälle** (Brogaard et al., 2024).

Die **Reliabilität** zeigt folgende Muster:

- **Interne Konsistenz**: Meist hoch (*Cronbachs α > 0,80*), z. B.:
  - TEAM: *α = 0,91–0,93* (Freytag et al., 2019),
  - OSCAR: *α = 0,84–0,96*.
- **Interrater-Reliabilität**: **Variiert stark** (*ICC 0,30–0,95*), abhängig von Training und Komplexität (vgl. Koo & Li, 2016):
  - **Hohe Reliabilität**: Ottawa GRS (*ICC = 0,80–0,87*; Jirativanont et al., 2017), TEAM (*ICC = 0,66*),
  - **Moderate Reliabilität**: ANTS (*ICC = 0,55–0,67*; Fletcher et al., 2003), NOTSS (*ICC = 0,29–0,66*),
  - **Niedrige Reliabilität**: T-NOTECHS (*ICC = 0,54*).

Die **Praktikabilität** hängt von der Komplexität ab:

- **Kompakte Instrumente** (z. B. TEAM, Ottawa GRS):
  - Zeitaufwand: **< 5 Minuten**,

  - Schulungsbedarf: **gering**,
  - Einsatz: **Echtzeit-Assessments** in klinischen Settings.
- **Detaillierte Instrumente** (z. B. ANTS, OSCAR):
  - Zeitaufwand: **5–20 Minuten**,
  - Schulungsbedarf: **hoch**,
  - Einsatz: **Simulationstrainings mit Debriefing**.

III. Einsatzszenarien: Formativ vs. Summativ

Non-Technical Skills (NTS) lassen sich je nach Zielsetzung entweder **formativ** (für Feedback und Lernprozesse) oder **summativ** (für Prüfungen und Zertifizierungen) erfassen. Beide Ansätze haben spezifische Vor- und Nachteile, die bei der Auswahl des passenden Instruments berücksichtigt werden müssen.

Formative Assessments: Detailliertes Feedback für Trainingszwecke

Für Lern- und Entwicklungsprozesse eignen sich besonders Instrumente wie ANTS, NOTSS, TEAM und CALM. Diese Tools bieten eine hohe Differenziertheit und klare Verhaltensanker, die präzises Feedback ermöglichen – etwa in Simulationstrainings oder strukturierten Debriefings. Durch die detaillierte Bewertung können Stärken und Schwächen in der Teamperformance gezielt identifiziert und verbessert werden.

Allerdings sind diese Instrumente zeitaufwendig in der Anwendung und erfordern einen hohen Schulungsbedarf für die Beobachter. Die komplexe Struktur und die Vielzahl an Bewertungskriterien machen sie weniger geeignet für schnelle Echtzeit-Assessments, dafür aber ideal für pädagogische Settings, in denen Lernfortschritte im Vordergrund stehen.

Summative Assessments: Effiziente Bewertung für Prüfungen und Zertifizierungen

Für Prüfungen, Zertifizierungen oder hochstake Bewertungen kommen dagegen kompaktere Instrumente wie die Ottawa GRS oder das TEAM-Tool mit definierten Cut-off-Werten zum Einsatz. Diese ermöglichen eine schnelle Anwendbarkeit und liefern eine globale Bewertung, die für summative Zwecke ausreichend ist.

Der Vorteil dieser Tools liegt in ihrer Praktikabilität – sie sind weniger zeitintensiv und erfordern weniger Schulungsaufwand. Allerdings geht dies zu Lasten der Differenziertheit: Feinere Nu-

ancen in der Teamperformance lassen sich mit diesen Instrumenten nicht abbilden. Dennoch sind sie aufgrund ihrer hohen Reliabilität (z. B. Ottawa GRS mit ICC = 0,80–0,87) besonders für hochstake Assessments geeignet, bei denen eine zuverlässige, aber weniger detaillierte Bewertung ausreicht.

Flexible Instrumente: TEAM als Hybridlösung

Ein besonderer Fall ist das TEAM-Tool, das sich sowohl für formative als auch summative Zwecke eignet. In Simulationstrainings kann es zunächst für detailliertes Feedback genutzt werden, während ein globaler Rating-Score von ≥4/5 als Kompetenznachweis für Zertifizierungen dient. Diese duale Nutzbarkeit macht TEAM zu einem vielseitigen Instrument, das sowohl Lernprozesse als auch Prüfungssituationen abdeckt.

Es lässt sich feststellen, dass die Wahl des passenden NTS-Assessment-Tools stets am konkreten Einsatzzweck ausgerichtet werden sollte:

- **Formative Instrumente** (ANTS, NOTSS, TEAM, CALM) eignen sich für **Trainings und Feedback**, da sie detaillierte Analysen ermöglichen.
- **Summative Instrumente** (Ottawa GRS, TEAM mit Cut-offs) sind ideal für **Prüfungen und Zertifizierungen**, da sie schnell und zuverlässig bewerten.
- **Hybridlösungen** wie TEAM bieten die Möglichkeit, **beide Ansätze zu kombinieren** und so die Stärken beider Methoden zu nutzen.

TEAM eignet sich aufgrund seiner dualen Nutzbarkeit (formativ/summativ) besonders für Notfallteams (Carpini et al., 2021): Ein globaler Rating-Score von ≥ 4/5 gilt als „kompetent“.

Ottawa GRS ist aufgrund seiner hohen Reliabilität (ICC = 0,80–0,87) für hochstake Assessments prädestiniert.

IV. Kontextspezifische und kulturelle Anpassungen

Die Instrumente wurden für **spezifische Settings und Zielgruppen** entwickelt und teilweise **kulturell adaptiert**:

- **Anästhesie**: ANTS, ANTSdk (dänische Version),
- **Chirurgie**: NOTSS,

- **Notfallmedizin**: TEAM, T-NOTECHS, OSCAR,
- **Geburtshilfe**: AOTP (*Anaesthetists' Non-Technical Skills for Obstetrics*), GAOTP (*Global Assessment of Obstetric Team Performance*),
- **Pflege**: NTS-NAS, MHPTS,
- **Kulturelle Anpassungen**: Italienische Ottawa GRS, finnische T-NOTECHS, dänische ANTSdk.

8.4 Diskussion: Schlussfolgerungen und Handlungsempfehlungen

I. Beantwortung der Forschungsfragen

Charakteristika der Instrumente:

- Die Instrumente sind hierarchisch aufgebaut und decken kognitive, soziale und persönliche NTS-Dimensionen ab.
- Sie sind zielgruppenspezifisch (Ärzte, Pflegekräfte, multiprofessionelle Teams) und settingabhängig (Anästhesie, Chirurgie, Notfallmedizin etc.).
- Unterschiede bestehen in Komplexität, Gewichtung der NTS-Dimensionen und kulturellen Anpassungen.

Psychometrische Eigenschaften:

- **Validität**: Meist gut (Inhalts- und Konstruktvalidität), Kriteriumsvalidität selten geprüft.
- **Reliabilität**: Interne Konsistenz meist hoch, Interrater-Reliabilität variiert stark (ICC 0,30–0,95).
- **Praktikabilität**: **Kompakte Instrumente sind schnell anwendbar, detaillierte Instrumente benötigen mehr Training.**

Eignung für klinische und edukative Kontexte:

- Formative Instrumente (ANTS, TEAM) eignen sich für Feedback in Trainings.
- Summative Instrumente (Ottawa GRS) sind für Prüfungen geeignet.

- Kontextspezifische Anpassungen sind notwendig (z. B. ANTSdk für Anästhesie, AOTP für Geburtshilfe).

II. Zentrale Schlussfolgerungen

- „One size fits all“ gibt es nicht: Die Wahl des Instruments muss sich an Zweck (formativ vs. summativ), Setting, Zielgruppe und Ressourcen orientieren.
- Formative Assessments: Detaillierte Behavioral Marker Systeme (z. B. ANTS, NOTSS) liefern tiefgreifendes Feedback für Trainings und Debriefings.
- Summative Assessments: Kompakte Skalen (z. B. TEAM, Ottawa GRS) sind praktikabel für Echtzeit-Assessments und Prüfungen.
- Rater Training ist entscheidend: Regelmäßige Schulungen und Kalibrierung verbessern die Interrater-Reliabilität deutlich.
- Kulturelle und kontextuelle Validierung: Instrumente müssen sprachlich und kulturell adaptiert werden (z. B. ANTSdk, italienische Ottawa GRS).

III. Forschungslücken und Limitationen

Limitationen der Arbeit:

- Selektive Literaturrecherche ohne systematische Suche,
- Fehlende empirische Datenerhebung in realen klinischen Settings,
- Kulturelle Lücken (z. B. Übertragbarkeit auf das deutsche Gesundheitssystem),
- Psychometrische Lücken (z. B. fehlende Kriteriumsvalidität).

Forschungsbedarf:

- Validierung in realen Settings: Untersuchung des Zusammenhangs zwischen NTS-Scores und Patientenoutcomes,
- Kulturelle Anpassungen: Entwicklung deutschsprachiger Versionen (z. B. ANTS, TEAM),
- Technologische Innovationen: Integration von KI-gestützter Videoanalyse, Wearables (Eye-Tracking, Stresssensoren) und Virtual Reality (VR),

- Definition von Cut-off-Werten: Klare Schwellenwerte für summative Assessments,
- Interprofessionelle Instrumente: Entwicklung von Tools für multiprofessionelle Teams.

8.5 Praktische Empfehlungen

Für die **Integration von NTS-Assessment-Tools in die Praxis** ergeben sich folgende Handlungsempfehlungen:

Kontextspezifische Auswahl:

- Kein Instrument ist universell einsetzbar – die Auswahl sollte sich an Zielgruppe, Setting und Zweck orientieren (vgl. Abbildung 12: Entscheidungsbaum für die Tool-Auswahl).

Strukturiertes Ratertraining:

- Regelmäßige Schulungen und Kalibrierung der Rater (z. B. 4-stündiges ANTS-Training).

Integration in Simulationstrainings:

- Nutzung von ANTS, TEAM oder OSCAR für formative Assessments in HFS.

Einbindung in klinische Routine:

- Kompakte Instrumente wie TEAM oder Ottawa GRS für Echtzeit-Feedback in Notaufnahmen oder OP-Sälen.

Weiterentwicklung von Curricula:

- NTS-Trainings als fester Bestandteil von Aus- und Weiterbildungsprogrammen (z. B. CRM-Kurse für Ärzte und Pflegekräfte).

8.6 NTS-Assessment als Baustein für Patientensicherheit

Die vorliegende Arbeit unterstreicht die zentrale Bedeutung von NTS für die Patientensicherheit und zeigt, dass eine Vielzahl von Instrumenten existiert, die sich in Struktur, Zielgruppe und psychometrischer Qualität unterscheiden. Kein Instrument ist universell einsetzbar – die

Auswahl muss kontextspezifisch erfolgen. Die psychometrischen Eigenschaften sind überwiegend gut, wobei die Interrater-Reliabilität eine zentrale Herausforderung darstellt.

Für die Praxis empfiehlt sich die Integration der Instrumente in Simulationstrainings und klinische Routine, kombiniert mit strukturiertem Ratertraining. Die Forschung sollte sich auf die Validierung in realen Settings, kulturelle Anpassungen und technologische Innovationen konzentrieren, um die Messung von NTS weiter zu verbessern. Nachhaltige Patientensicherheit erfordert die systematische Entwicklung und Evaluation von Teamkompetenzen.

9. Abstract

Zielsetzung

Ziel der Arbeit war die systematische Evaluation von Instrumenten zur Erfassung non-technischer Fähigkeiten (Non-Technical Skills, NTS) in Healthcare-Teams während High-Fidelity-Simulationen (HFS). Untersucht wurde, welche Tools verfügbar sind, welche Strukturen und Zielgruppen sie haben sowie welche psychometrischen Eigenschaften (Validität, Reliabilität, Praktikabilität) sie aufweisen und inwieweit sie sich für formative bzw. summative Einsätze eignen.

Hintergrund

Non-technical Skills (kognitive, soziale und personale Ressourcen) sind entscheidend für Teamleistung und Patientensicherheit in Hochrisikobereichen (z. B. OP, Notaufnahme, Geburtshilfe). Behavioral Marker Systems (z. B. ANTS, NOTSS, TEAM, Ottawa GRS, OSCAR, T-NOTECHS, TPOT u. a.) übersetzen diese latenten Konstrukte in beobachtbares Verhalten und bilden die Grundlage für Training, Feedback und Evaluation in Simulation und Praxis.

Methoden

Durchgeführt wurde eine selektive, nicht-systematische Literaturanalyse (PubMed, Google Scholar, manuelle Suche, Hochschulbibliothek; Publikationssprachen Deutsch/Englisch). Es wurden etablierte Instrumente identifiziert und vergleichend hinsichtlich Aufbaus (Kategorien, Elemente, Verhaltensanker), Zielgruppe, Settings sowie Gütekriterien (Inhalts-, Konstrukt- und Kriteriumsvalidität; Interrater-Reliabilität; interne Konsistenz; Praktikabilität) ausgewertet.

Ergebnisse

Die Instrumente folgen meist einer hierarchischen Struktur (Kategorien → Elemente → Verhaltensanker) und erfassen drei Hauptdimensionen: kognitive Fähigkeiten (Situationsbewusstsein, Decision Making), soziale/interpersonelle Fähigkeiten (Kommunikation, Teamwork, Leadership) und personale Ressourcen (Stress-/Müdigkeitsmanagement, Metakognition). Psychometrisch zeigen sich heterogene Befunde: interne Konsistenzen sind häufig gut bis sehr gut, die Interrater-Reliabilität variiert jedoch stark zwischen Tools, Szenarien und Rater-Training. Praktikabilität ist ein Kompromiss zwischen Detailliertheit und Einsatzaufwand: kompakte Globalratings (z. B. Ottawa GRS, TEAM) sind schnell und alltagstauglich, detaillierte Behavioral-Marker-Systeme (z. B. ANTS, OSCAR, STAT) liefern differenziertes Feedback, erfordern aber mehr Schulung und Zeit. Zentrale Empfehlungen sind kontextspezifische Instru-

mentenauswahl, systematische Rater-Training und kulturelle/sprachliche Adaptation; weiterhin fehlen verbindliche Feldvalidierungen und Evidenz zu Zusammenhängen zwischen NTS-Scores und Patienten-Outcomes.

10. Quellenverzeichnis

Ausbildung und Training für komplexe Situationen: 26. Jahrestagung der Plattform e. V. - Menschen in komplexen Arbeitswelten (Plattform Menschen in komplexen Arbeitswelten e. V.- Tagungsdokumentation, Band 5): Heimann, R., Hörnberger, C. (2026, 25. März). Verfügbar unter **https://www.amazon.de/gp/product/B0DFVC7S8N?ref_=dbs_m_mng_rwt_calw_tpbk_4&storeType=ebooks**

Agentur für Gesundheitsforschung und -qualität. (2023a). Abschnitt 1: Überblick über Schlüsselkonzepte und Werkzeuge. **https://www.ahrq.gov/teamstepps-program/curriculum/intro/overview.html**

Agentur für Gesundheitsforschung und -qualität. (2023b). TeamSTEPPS Updates. **https://www.ahrq.gov/teamstepps-program/updated/index.html**

Brogaard, L.; Rosvig, L.; Hjorth-Hansen, K. R.; Hvidman, L.; Hinshaw, K.; Kierkegaard, O.; Uldbjerg, N.; Manser, T. (2024). Team performance during vacuum-assisted vaginal delivery: video review of obstetric multidisciplinary teams. Frontiers in Medicine, 11. **https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2024.1330457**

Carpini, J. A., Calvert, K., Carter, S., Epee-Bekima, M., & Leung, Y. (2021). Validating the Team Emergency Assessment Measure (TEAM) in obstetric and gynaecologic resuscitation teams. Australian and New Zealand Journal of Obstetrics and Gynaecology, 61(6), 855–861. **https://doi.org/10.1111/ajo.13362**

Cooper, S., Cant, R., Porter, J., Sellick, K., Somers, G., Kinsman, L., & Nestel, D. (2010). Rating medical emergency teamwork performance: Development of the Team Emergency Assessment Measure (TEAM). *Resuscitation, 81*(4), 446–452. **https://doi.org/10.1016/j.resuscitation.2009.11.027**

Dedy, N. J., Szasz, P., Louridas, M., Bonrath, E. M., Husslein, H., & Grantcharov, T. P. (2015). Objective structured assessment of nontechnical skills: reliability of a global rating scale for the in-training assessment in the operating room. Surgery, 157(6), 1002–1013. **https://doi.org/10.1016/j.surg.2014.12.023**

Edwards, T. C., Soussi, D., Gupta, S., Khan, S., Patel, A., Patil, A., Liddle, A. D., Cobb, J. P., & Logishetty, K. (2023). Collaborative team training in virtual reality is superior to individual lear-

ning for performing complex open surgery: A randomized controlled trial. *Annals of Surgery, 278*. **https://doi.org/10.1097/SLA.0000000000006079**

Gosselin, É., Marceau, M., Vincelette, C., Daneau, C.-O., Lavoie, S., & Ledoux, I. (2019). French translation and validation of the Mayo High Performance Teamwork Scale for nursing students in a high-fidelity simulation context. Clinical Simulation in Nursing, 30, 25–33. **https://doi.org/10.1016/j.ecns.2019.03.002**

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., & Patey, R. (2003). Anaesthetists' non-technical skills (ANTS): evaluation of a behavioural marker system. British Journal of Anaesthesia, 90(5), 580–588. **https://doi.org/10.1093/bja/aeg112**

Flowerdew, L., Brown, R., Vincent, C., & Woloshynowych, M. (2012). Development and validation of a tool to assess emergency physicians' nontechnical skills. Annals of Emergency Medicine, 59(5), 376–385.e4. **https://doi.org/10.1016/j.annemergmed.2011.11.022**

Franc, J. M., Verde, M., Gallardo, A. R., Carenzo, L., & Ingrassia, P. L. (2017). An Italian version of the Ottawa Crisis Resource Management Global Rating Scale: a reliable and valid tool for assessment of simulation performance. Internal and Emergency Medicine, 12(5), 651–656. **https://doi.org/10.1007/s11739-016-1486-7**

Franklin, D., Tzavelis, A., Lee, J. Y., Chung, H. U., Trueb, J., Arafa, H., Kwak, S. S., Huang, I., Liu, Y., Rathod, M., et al. (2023). Synchronized wearables for the detection of haemodynamic states via electrocardiography and multispectral photoplethysmography. Nature Biomedical Engineering, 7, 1229–1241. **https://doi.org/10.1038/s41551-023-01098-y**

Freytag, J., Stroben, F., Hautz, W. E., Schauber, S. K., & Kämmer, J. E. (2019). Rating the quality of teamwork — a comparison of novice and expert ratings using the Team Emergency Assessment Measure (TEAM) in simulated emergencies. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 27(1), 12. **https://doi.org/10.1186/s13049-019-0591-9**

Gawronski, O., Thekkan, K. R., Genna, C., Egman, S., Sansone, V., Erba, I., Vittori, A., Varano, C., Dall'Oglio, I., Tiozzo, E., & Chiusolo, F. (2022). Instruments to evaluate non-technical skills during high-fidelity simulation: A systematic review. *Frontiers in Medicine, 9*. **https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2022.986296**

Jepsen, R. M. H. G., Spanager, L., Lyk-Jensen, H. T., Dieckmann, P., & Østergaard, D. (2015). Customisation of an instrument to assess anaesthesiologists' non-technical skills. International Journal of Medical Education, 6, 17–25. **https://doi.org/10.5116/ijme.54be.8f08**

Jepsen, R. M. H. G.; Dieckmann, P.; Spanager, L.; Lyk-Jensen, H. T.; Konge, L.; Ringsted, C.; Østergaard, D. (2016). Evaluating structured assessment of anaesthesiologists' non-technical skills. Acta Anaesthesiologica Scandinavica, 60(6), 756–766. **https://doi.org/10.1111/aas.12709**

Jirativanont, T., Raksamani, K., Aroonpruksakul, N., Apidechakul, P., & Suraseranivongse, S. (2017). Validity evidence of non-technical skills assessment instruments in simulated anaesthesia crisis management. Anaesthesia and Intensive Care, 45(4), 469–475. **https://doi.org/10.1177/0310057X1704500410**

Kelly, F. E.; Frerk, C.; Bailey, C. R.; Cook, T. M.; Ferguson, K.; Flin, R.; Fong, K.; Groom, P.; John, C.; Lang, A. R.; Meek, T.; Miller, K. L.; Richmond, L.; Sevdalis, N.; Stacey, M. R. (2023). Implementing human factors in anaesthesia: guidance for clinicians, departments and hospitals. Anaesthesia, 78(4), 458–478. **https://doi.org/10.1111/anae.15941**

Kim, J., Neilipovitz, D., Cardinal, P., Chiu, M., & Clinch, J. (2006). A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: the University of Ottawa Critical Care Medicine, high-fidelity simulation, and crisis resource management I study. Critical Care Medicine, 34(8), 2167–2174. **https://doi.org/10.1097/01.CCM.0000229877.45125.CC**

Kranz, K., & Regener, H. (2023). Trainieren für den Notfall. vsao, (6), 23–27. **https://www.paraplegie.ch/sites/default/files/2023-12/kranz_k._regener_h._2023._trainieren_fuer_den_notfall_vsao-journal.pdf**

Maguire, M. B. R. (2016). Psychometric testing of the TeamSTEPPS® 2.0 Team Performance Observation Tool (Doctoral dissertation). Kennesaw State University. **https://digitalcommons.kennesaw.edu/dns_etd/2/#:~:text=Data%20analysis%20provided%20baseline%20psychometric%20properties%20of%20the,included%20internal%20consistency%2C%20test-retest%2C%20and%20inter%20rater%20analysis**

Malec, J. F., Torsher, L. C., Dunn, W. F., Wiegmann, D. A., Arnold, J. J., Brown, D. A., & Phatak, V. (2007). The Mayo High Performance Teamwork Scale: Reliability and validity for evaluating key crew resource management skills. *Simulation in Healthcare, 2*(1), 4–10. **https://doi.org/10.1097/SIH.0b013e31802b68ee**

Moll-Khosrawi, P., Kamphausen, A., Hampe, W., Schulte-Uentrop, L., Zimmermann, S., & Kubitz, J. C. (2019). Anaesthesiology students' non-technical skills: development and evaluation

of a behavioural marker system for students (AS-NTS). BMC Medical Education, 19(1), 205. **https://doi.org/10.1186/s12909-019-1609-8**

Moorthy, K., Munz, Y., Adams, S., Pandey, V., & Darzi, A. (2005). A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. Annals of Surgery, 242(5), 631–639. **https://doi.org/10.1097/01.sla.0000186298.79308.a8**

Morgan, P. J., Pittini, R., Regehr, G., Marrs, C., & Haley, M. F. (2007). Evaluating teamwork in a simulated obstetric environment. Anesthesiology, 106(5), 907–915. **https://doi.org/10.1097/01.anes.0000265149.94190.04**

Nadkarni, L. D., Roskind, C. G., Auerbach, M. A., Calhoun, A. W., Adler, M. D., & Kessler, D. O. (2018). The development and validation of a concise instrument for formative assessment of team leader performance during simulated pediatric resuscitations. Simulation in Healthcare, 13(2), 77–82. **https://doi.org/10.1097/SIH.0000000000000267**

Pires, S. M. P., Monteiro, S. O. M., Pereira, A. M. S., Stocker, J. N. M., Chaló, D. d. M., & Melo, E. M. O. P. de. (2018). Non-technical skills assessment scale in nursing: construction, development and validation. Revista Latino-Americana de Enfermagem, 26, e3042. **https://doi.org/10.1590/1518-8345.2383.3042**

Reid, J., Stone, K., Brown, J., Caglar, D., Kobayashi, A., Lewis-Newby, M., Partridge, R., Seidel, K., & Quan, L. (2012). The Simulation Team Assessment Tool (STAT): Development, reliability and validation. *Resuscitation, 83*(7), 879–886. **https://doi.org/10.1016/j.resuscitation.2011.12.012**

Repo, J. P., Rosqvist, E., Lauritsalo, S., & Paloneva, J. (2019). Translatability and validation of non-technical skills scale for trauma (T-NOTECHS) for assessing simulated multi-professional trauma team resuscitations. BMC Medical Education, 19(1), 40. **https://doi.org/10.1186/s12909-019-1474-5**

St. Pierre, M. (2018). Simulation in der Medizin: grundlegende Konzepte — klinische Anwendung (2. Aufl.). Springer. **https://livivo.idm.oclc.org/login?url=https://ebookcentral.proquest.com/lib/zbmed-ebooks/detail.action?docID=5451757**

St. Pierre, M. (2020). Human factors und Patientensicherheit in der Akutmedizin (4. Aufl.). Springer. **https://livivo.idm.oclc.org/login?url=https://ebookcentral.proquest.com/lib/zbmed-ebooks/detail.action?docID=6167593**

Steinemann, S., Berg, B., DiTullio, A., Skinner, A., Terada, K., Anzelon, K., & Ho, H. C. (2012). Assessing teamwork in the trauma bay: Introduction of a modified “NOTECHS” scale for trauma. *American Journal of Surgery, 203*(1), 69–75. **https://doi.org/10.1016/j.amjsurg.2011.08.004**

Stanford University. (2026). Use of artificial intelligence to assess trainee communication compared to human assessment (ClinicalTrials.gov ID: NCT07107880). **https://clinicaltrials.gov/study/NCT07107880**

Tregunno, D., Pittini, R., Haley, M., & Morgan, P. J. (2009). Development and usability of a behavioural marking system for performance assessment of obstetrical teams. Quality & Safety in Health Care, 18(5), 393–396. **https://doi.org/10.1136/qshc.2007.026146**

Walker, S., Brett, S., McKay, A., Lambden, S., Vincent, C., & Sevdalis, N. (2011). Observational skill-based clinical assessment tool for resuscitation (OSCAR): development and validation. Resuscitation, 82(7), 835–844. **https://doi.org/10.1016/j.resuscitation.2011.03.009**

Wauben, L. S. G. L.; Dekker-van Doorn, C. M.; van Wijngaarden, J. D. H.; Goossens, R. H. M.; Huijsman, R.; Klein, J.; Lange, J. F. (2011). Discrepant perceptions of communication, teamwork and situation awareness among surgical team members. International Journal for Quality in Health Care, 23(2), 159–166. **https://doi.org/10.1093/intqhc/mzq079**

Watkins, S. C., Roberts, D. A., Boulet, J. R., McEvoy, M. D., & Weinger, M. B. (2017). Evaluation of a simpler tool to assess nontechnical skills during simulated critical events. Simulation in Healthcare, 12(2), 69–75. **https://doi.org/10.1097/SIH.0000000000000199**

Yule, S., Flin, R., Paterson-Brown, S., Maran, N., & Rowley, D. (2006). Development of a rating system for surgeons’ non-technical skills. Medical Education, 40(11), 1098–1104. **https://doi.org/10.1111/j.1365-2929.2006.02610.x**

Yule, S., Flin, R., Maran, N., Rowley, D., Youngson, G., & Paterson-Brown, S. (2008). Surgeons’ non-technical skills in the operating room: reliability testing of the NOTSS behaviour rating system. World Journal of Surgery, 32(4), 548–556. **https://doi.org/10.1007/s00268-007-9320-z**

Zhang, C., Miller, C., Volkman, K., Meza, J., & Jones, K. (2015). Evaluation of the Team Performance Observation Tool with targeted behavioral markers in simulation-based interprofessional education. Journal of Interprofessional Care, 29(3), 202–208. **https://doi.org/10.3109/13561820.2014.982789**

Werkzeug: DESC. (2023, Juli). Agentur für Gesundheitsforschung und -qualität (AHRQ). **https://www.ahrq.gov/teamstepps-program/curriculum/mutual/tools/desc.html**

11. Anhang

Tabelle 7: Domänen-Überblick

BMS Instrument Comparison – Hauptkategorien										
Bezeichnung	**ANTS**	**ANTSdk**	**Ottawa GRS**	**TEAM**	**NOTSS**	**BARS**	**MHPTS**	**CALM**	**ENTS**	**OSCAR**
	Fletcher et al. (2003)	Jepsen et al. (2015)	Kim et al. (2006)	Cooper et al. (2010)	Yule et al. (2006)	Watkins et al. (2015)	Malec et al. (2007)	Nadkarni et al. (2018)	Flowerdew et al. (2012)	Walker et al. (2011)
Task management	x			x	x				x	
Team working	x	x		x		x	x		x	x
Situation awareness	x	x	x	x	x	x	x		x	x
Decision making	x	x	x		x	x		x	x	x
Leadership		x	x	x	x		x	x		x
Overall			x							
Resource management			x							
Communication			x	x	x	x	x	x		x
Error management							x			

BMS Instrument Comparison – Hauptkategorien										
Bezeichnung	STAT	T NOTECHS	AOTP	AS NTS	LOSA	NTS NAS	HFRS	TPOT	OSANTS	GRS obs
	Reid et al. (2012)	Repo et al. (2019)	Tregunno et al. (2009)	Moll-Khosrawi et al. (2019)	Moorthy et al. (2005)	Pires et al. (2018)	Morgan et al. (2007)	AHRQ / DoD (2014)	Dedy et al. (2015)	Morgan et al. (2007)
Task management			x	x	x	x	x		x	
Team working		x	x	x		x	x		x	
Situation awareness		x	x		x	x		x	x	
Decision making	x	x							x	
Leadership		x			x	x	x	x	x	
Overall										
Resource management						x				
Communication		x	x		x	x	x	x	x	
Error management						x	x			
Clinical assessment	x									
Communication with patient and partner			x							
Mutual support						x		x		
Team Structure								x		
Professionalism									x	
1 – Unacceptable Performance										x
2 – Borderline Performance										x
3 – Acceptable Performance										x
4 – Good Performance										x
5 – Superior Performance										x

Quelle: Eigene Darstellung

Tabelle 8: Top-Level (Kategorien und Elemente)

BMS Instrument Comparison – Kategorien und Elemente											
Typ	Bezeichnung	ANTS	ANTSdk	Ottawa GRS	TEAM	NOTSS	BARS	MHPTS	CALM	ENTS	OSCAR
		Fletcher et al. (2003)	Jepsen et al. (2015)	Kim et al. (2006)	Cooper et al. (2010)	Yule et al. (2006)	Watkins et al. (2015)	Malec et al. (2007)	Nadkarni et al. (2018)	Flowerdew et al. (2012)	Walker et al. (2011)
Kategorie	**Task management**	x			x	x				x	
Element	• Planning and preparing	x				x					
Element	• Prioritizing	x									
Element	• Providing and maintaining standards	x									
Element	• Identifying and utilizing resources	x									
Element	• Flexibility / responding to change					x					
Element	• Maintaining Standards									x	
Element	• Managing Workload									x	
Element	• Supervising and Providing Feedback									x	
Kategorie	**Team working**	x	x		x		x			x	x
Element	• Co-ordinating activities with team members	x	x								
Element	• Exchanging information	x	x								
Element	• Using authority and assertiveness	x									
Element	• Assessing capabilities	x	x								
Element	• Supporting others	x	x								
Element	• Team Building									x	
Element	• Communicating Effectively									x	
Element	• Authority & Assertiveness									x	
Kategorie	**Situation awareness**	x	x	x	x	x	x			x	x
Element	• Gathering information	x	x			x				x	
Element	• Recognizing and understanding	x	x								
Element	• Anticipating	x	x							x	
Element	• Demonstrating self-awareness		x								

BMS Instrument Comparison – Kategorien und Elemente											
Typ	Bezeichnung	ANTS	ANTSdk	Ottawa GRS	TEAM	NOTSS	BARS	MHPTS	CALM	ENTS	OSCAR
		Fletcher et al. (2003)	Jepsen et al. (2015)	Kim et al. (2006)	Cooper et al. (2010)	Yule et al. (2006)	Watkins et al. (2015)	Malec et al. (2007)	Nadkarni et al. (2018)	Flowerdew et al. (2012)	Walker et al. (2011)
Element	• Understanding information					x					
Element	• Projecting and anticipating future state					x					
Element	• Updating the Team									x	
Kategorie	**Decision making**	x	x	x		x	x		x	x	x
Element	• Identifying options	x	x								
Element	• Balancing risks and selecting options	x									
Element	• Re-evaluating	x									
Element	• Choosing, communicating and implementing decisions		x								
Element	• Reassessing decisions		x								
Element	• Considering options					x					
Element	• Selecting and communicating options					x					
Element	• Implementing and reviewing decisions					x					
Element	• Generating Options									x	
Element	• Selecting & Communicating Options									x	
Element	• Reviewing Outcomes									x	
Kategorie	**Leadership**		x	x	x	x			x		x
Element	• Planning and preparing		x								
Element	• Prioritizing		x								
Element	• Identifying and utilizing resources		x								
Element	• Using authority and assertiveness		x								
Element	• Providing and maintaining standards		x								
Element	• Setting and maintaining standards					x					

BMS Instrument Comparison – Kategorien und Elemente											
Typ	Bezeichnung	ANTS	ANTSdk	Ottawa GRS	TEAM	NOTSS	BARS	MHPTS	CALM	ENTS	OSCAR
		Fletcher et al. (2003)	Jepsen et al. (2015)	Kim et al. (2006)	Cooper et al. (2010)	Yule et al. (2006)	Watkins et al. (2015)	Malec et al. (2007)	Nadkarni et al. (2018)	Flowerdew et al. (2012)	Walker et al. (2011)
Element	• Supporting others					x					
Element	• Coping with pressure					x					
Kategorie	**Overall**			x							
Kategorie	**Resource management**			x							
Kategorie	**Communication**			x	x	x	x		x		x
Element	• Exchanging information					x					
Element	• Establishing a shared understanding					x					
Element	• Co-ordinating team activities					x					

BMS Instrument Comparison – Kategorien und Elemente											
Typ	Bezeichnung	STAT	T NOTECHS	AOTP	AS NTS	LOSA	NTS NAS	HFRS	GRS obs	TPOT	OSANTS
		Reid et al. (2012)	Repo et al. (2019)	Tregunno et al. (2009)	Moll-Khosrawi et al. (2019)	Moorthy et al. (2005)	Pires et al. (2018)	Morgan et al. (2007)	Morgan et al. (2007)	AHRQ / DoD (2014)	Dedy et al. (2015)
Kategorie	**Task management**			x		x	x	x			x
Element	• Plan of action			x							
Element	• Problem solving			x							
Element	• Resource utilisation			x							
Element	• Introduction to team members					x					
Element	• Preoperative instrument and equipment check					x					
Element	• Briefing					x					
Kategorie	**Team working**		x	x	x		x	x			x
Element	• Leadership			x							
Element	• Role assignment			x							
Element	• Team interaction			x							
Kategorie	**Situation awareness**		x	x		x	x			x	x
Element	• Anticipating			x							
Element	• Realising limitations			x							
Element	• Fixation			x							
Element	• Responsiveness			x							
Element	• Vigilance			x							
Element	• Management of disruptive behaviour			x							
Element	• Atmosphere of the room			x							
Element	• Monitored patient's parameters throughout the procedure					x					
Element	• Awareness of anesthetist					x					

BMS Instrument Comparison – Kategorien und Elemente											
Typ	Bezeichnung	STAT	T NOTECHS	AOTP	AS NTS	LOSA	NTS NAS	HFRS	GRS obs	TPOT	OSANTS
		Reid et al. (2012)	Repo et al. (2019)	Tregunno et al. (2009)	Moll-Khosrawi et al. (2019)	Moorthy et al. (2005)	Pires et al. (2018)	Morgan et al. (2007)	Morgan et al. (2007)	AHRQ / DoD (2014)	Dedy et al. (2015)
Element	• Actively initiates communication with anesthetist					x					
Kategorie	**Decision making**	x	x								x
Element	• All team members exhibit professional attitude and interactions (Team)	x									
Element	• There is a clearly identified team leader (Leadership)	x									
Element	• Assigns roles to team members (Leadership)	x									
Element	• Maximizes skill sets of personnel in assigned roles (Leadership)	x									
Element	• Directs/Redirects team members effectively (Leadership)	x									
Element	• Monitors actions of team members (Leadership)	x									
Kategorie	**Leadership**		x			x	x	x		x	x
Element	• Adherence to best practice during the procedure					x					
Element	• Resource utilization, ie, appropriate task-load distribution and delegation of responsibilities					x					
Element	• Authority/assertiveness					x					
Kategorie	**Overall**										
Kategorie	**Resource management**						x				
Kategorie	**Communication**		x	x		x	x	x		x	x
Element	• Focussed communication			x							
Element	• Closing the loop			x							
Element	• Instructions to assistant/scrub nurse: clear and polite					x					

BMS Instrument Comparison – Kategorien und Elemente											
Typ	Bezeichnung	ANTS	ANTSdk	Ottawa GRS	TEAM	NOTSS	BARS	MHPTS	CALM	ENTS	OSCAR
		Fletcher et al. (2003)	Jepsen et al. (2015)	Kim et al. (2006)	Cooper et al. (2010)	Yule et al. (2006)	Watkins et al. (2015)	Malec et al. (2007)	Nadkarni et al. (2018)	Flowerdew et al. (2012)	Walker et al. (2011)
Element	• Generating Options									x	
Element	• Selecting & Communicating Options									x	
Element	• Reviewing Outcomes									x	
Kategorie	**Leadership**		x	x	x	x			x		x
Element	• Planning and preparing		x								
Element	• Prioritizing		x								
Element	• Identifying and utilizing resources		x								
Element	• Using authority and assertiveness		x								
Element	• Providing and maintaining standards		x								
Element	• Setting and maintaining standards					x					
Element	• Supporting others					x					
Element	• Coping with pressure					x					
Kategorie	**Overall**			x							
Kategorie	**Resource management**			x							
Kategorie	**Communication**			x	x	x	x		x		x
Element	• Exchanging information					x					
Element	• Establishing a shared understanding					x					
Element	• Co-ordinating team activities					x					

BMS Instrument Comparison – Kategorien und Elemente											
Typ	Bezeichnung	STAT	T NOTECHS	AOTP	AS NTS	LOSA	NTS NAS	HFRS	GRS obs	TPOT	OSANTS
		Reid et al. (2012)	Repo et al. (2019)	Tregunno et al. (2009)	Moll-Khosrawi et al. (2019)	Moorthy et al. (2005)	Pires et al. (2018)	Morgan et al. (2007)	Morgan et al. (2007)	AHRQ / DoD (2014)	Dedy et al. (2015)
Element	• Awaits acknowledgment from the assistant/scrub nurse					x					
Element	• Assistance sought from team members					x					
Element	• Acknowledges help/advice from team members					x					
Kategorie	**Clinical assessment**	x									
Kategorie	**Communication with patient and partner**			x							
Element	• Information sharing			x							
Element	• Reassuring attitude			x							
Element	• Partner management			x							
Kategorie	**Mutual support**						x			x	
Kategorie	**Error management**						x	x			
Kategorie	**Team Structure**									x	
Kategorie	**Professionalism**										x

Quelle: Eigene Darstellung

Tabelle 9: vollständige Hierarchie (BMS Taxonomie – Vergleich aller 28 Studien)

BMS Taxonomie – Vergleich aller 28 Studien															
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Kategorie	**Task management**	1	x	x				x	x		x					
Element	• Planning and preparing	2	x	x							x					
Element	• Prioritizing	2	x								x					
Element	• Providing and maintaining standards	2	x	x							x					
Element	• Identifying and utilizing resources	2	x								x					
Kategorie	**Team working**	1	x	x		x	x	x	x	x	x			x		
Element	• Co-ordinating activities with team members	2	x								x					
Element	• Exchanging information	2	x	x		x	x				x					
Element	• Using authority and assertiveness	2	x	x							x					
Element	• Assessing capabilities	2	x	x							x					
Element	• Supporting others	2	x	x		x	x				x					
Kategorie	**Situation awareness**	1	x	x	x	x	x		x	x	x		x			x
Element	• Gathering information	2	x	x		x	x		x		x					
Element	• Recognizing and understanding	2	x								x					
Element	• Anticipating	2	x	x					x		x					
Kategorie	**Decision making**	1	x	x	x	x	x		x		x		x			x
Element	• Identifying options	2	x	x		x	x				x					
Element	• Balancing risks and selecting options	2	x								x					
Element	• Re-evaluating	2	x								x					

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Marker	◦ Does not alter physical layout of workspace to improve data visibility	3	x													
Marker	◦ Does not ask questions to orient self to situation during hand-over	3	x													
Element	• Prioritising	2		x												
Element	• Identifying and utilising resources	2		x												
Element	• Coordinating with team	2		x												
Element	• Recognising and understanding	2		x												
Element	• Balancing risk and selecting options	2		x												
Element	• Re-evaluation	2		x												
Kategorie	**Overall**	1			x								x			x
Kategorie	**Leadership**	1			x	x	x					x	x	x		x
Kategorie	**Resource management**	1			x								x			x
Kategorie	**Communication**	1			x					x	x	x	x	x		x
Element	• Recognising and understanding contexts	2				x	x									
Element	• Anticipating and thinking ahead	2				x	x									
Element	• Demonstrating self-awareness	2				x	x									
Element	• Choosing, communicating and implementing decisions	2				x	x									
Element	• Reassessing decisions	2				x	x									

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Element	• Assessing competencies	2				x	x									
Element	• Coordinating activities	2				x	x									
Element	• Planning and preparing	2				x	x									
Element	• Prioritising	2				x	x									
Element	• Identifying and utilising resources	2				x	x									
Element	• Using authority and assertiveness	2				x	x									
Element	• Providing and maintaining standards	2				x	x									
Marker	◦ Describes relevant changes in the patient's status to the team and ensures that appropriate action is taken when needed	3				x										
Marker	◦ Informs team members when a situation could develop critically	3				x										
Marker	◦ Summarises the situation for the team when needed; for example, using ABCDE systematics	3				x										
Marker	◦ Introduces her/himself to new team members and states competencies	3				x										
Marker	◦ Reacts to signals from team members when they are losing focus and no longer can manage the task	3				x										
Marker	◦ Includes knowledge about team members' competences when tasks are distributed	3				x										
Marker	◦ Does not use systematics when gathering information about the situation	3				x										

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Marker	◦ Does not point out relevant changes in a patient's condition to the team	3				x										
Marker	◦ Exhibits inappropriate behaviour in relation to the situation	3				x										
Marker	◦ Stays passive without participating in the coordination of activities	3				x										
Marker	◦ Fixates on using a single guideline although it does not fit the situation	3				x										
Kategorie	**Teamwork and leadership (exchanging information and leading the team)**	1						x								
Marker	◦ Systematically assesses the situation and formulates a plan	3						x								
Marker	◦ Prioritises tasks correctly according to clinical urgency	3						x								
Marker	◦ Adapts the plan to new information or changing conditions	3						x								
Marker	◦ Introduces him/herself and states competencies to team members	3						x								
Marker	◦ Communicates clearly and concisely with team members	3						x								
Marker	◦ Coordinates and directs team members effectively	3						x								
Marker	◦ Actively supports team members in their tasks	3						x								

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Element	• Team Building	2							x							
Element	• Communicating Effectively	2							x							
Element	• Authority & Assertiveness	2							x							
Element	• Generating Options	2							x							
Element	• Selecting & Communicating Options	2							x							
Element	• Reviewing Outcomes	2							x							
Element	• Updating the Team	2							x							
Marker	◦ Notices doctor's illegible notes and explains the value of good note keeping	3							x							
Marker	◦ Explains importance of ensuring sick patient is stable prior to transfer	3							x							
Marker	◦ Ensures clinical guidelines are followed and appropriate pro forma is complete	3							x							
Marker	◦ Sees a doctor has spent a long time with a patient and ascertains the reason	3							x							
Marker	◦ Ensures both themselves and other team members take appropriate breaks	3							x							
Marker	◦ Takes the opportunity to teach whilst reviewing patient with junior doctor	3							x							
Marker	◦ Gives positive feedback to junior doctor who has made a difficult diagnosis	3							x							
Marker	◦ Even when busy, reacts positively to a junior doctor asking for help	3							x							

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Marker	◦ Demonstrates awareness of team members' workload and offers help	3						x								
Marker	◦ Maintains a positive team atmosphere	3						x								
Marker	◦ Does not formulate a plan or acts without a clear strategy	3						x								
Marker	◦ Fails to prioritise tasks appropriately	3						x								
Marker	◦ Does not adapt plan despite clear changes in the situation	3						x								
Marker	◦ Fails to communicate relevant information to the team	3						x								
Marker	◦ Does not coordinate team members' activities	3						x								
Marker	◦ Gives unclear or contradictory instructions	3						x								
Marker	◦ Ignores team members' difficulties or needs	3						x								
Marker	◦ Does not support team members when they are overloaded	3						x								
Marker	◦ Creates a tense or negative team atmosphere	3						x								
Marker	◦ REFERENZIERT (nicht für Datenerhebung angewandt): ANTS (Fletcher et al. 2003) – als Vergleichsinstrument zitiert; als zu komplex für Studierende befunden.	3						x								
Element	• Maintaining Standards	2							x							
Element	• Managing Workload	2							x							
Element	• Supervising and Providing Feedback	2							x							

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Marker	◦ Gives clear referral to specialty doctor with reason for admission (e.g. SBAR)	3							x							
Marker	◦ Uses appropriate degree of assertiveness when inpatient doctor refuses referral	3							x							
Marker	◦ Goes to see patient to get more information when junior is unclear about history	3							x							
Marker	◦ Discusses the contribution of false positive and false negative test results	3							x							
Marker	◦ Follows up with doctor to see if provisional plan needs revising	3							x							
Marker	◦ Uses Patient Tracking System appropriately to monitor state of the department	3							x							
Marker	◦ Eyeballs' patients during long wait times to identify anyone who looks unwell	3							x							
Marker	◦ Discusses contingencies with nurse-in-charge during periods of overcrowding	3							x							
Marker	◦ Updates team about new issues such as bed availability or staff shortages	3							x							
Marker	◦ Communicates a change in patient status to relevant inpatient team	3							x							
Marker	◦ Does not wash hands (or use alcohol gel) after reviewing patient	3							x							
Marker	◦ Fails to act when a junior is overloaded and patient care is compromised	3							x							

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Marker	◦ Focuses on one particular patient and loses control of the department	3							x							
Marker	◦ Does not adequately supervise junior doctor with a sick patient	3							x							
Marker	◦ Fails to ask if junior doctor is confident doing a practical procedure unsupervised	3							x							
Marker	◦ Harasses team members rather than giving assistance or advice	3							x							
Marker	◦ Repeatedly interrupts doctor who is presenting a patient's history	3							x							
Marker	◦ Fails to persevere when inpatient doctor refuses appropriate referral	3							x							
Marker	◦ Fails to ensure all relevant information is available when advising referral	3							x							
Marker	◦ Sticks rigidly to plan despite availability of new information	3							x							
Marker	◦ Fails to notice that patient is about to breach and no plan has been made	3							x							
Marker	◦ Fails to notice that CDU is full when arranging new transfers	3							x							
Marker	◦ Fails to anticipate and prepare for difficulties or complications during a practical procedure	3							x							
Marker	◦ Fails to ensure that breaks are planned to maintain safe staffing levels	3							x							

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Marker	◦ Fails to anticipate and plan for clinical deterioration during patient transfer	3							x							
Marker	◦ Notices the long wait but fails to check the rest of the team is aware	3							x							
Kategorie	**Communication with patient and partner**	1								x						
Element	• Information sharing	2								x						
Element	• Reassuring attitude	2								x						
Element	• Partner management	2								x						
Kategorie	**Task/case management**	1								x						
Element	• Plan of action	2								x						
Element	• Problem solving	2								x						
Element	• Resource utilisation	2								x						
Element	• Leadership	2								x						
Element	• Role assignment	2								x						
Element	• Team interaction	2								x						
Element	• Anticipation	2								x						
Element	• Realising limitations	2								x						
Element	• Fixation	2								x						
Element	• Responsiveness	2								x						
Element	• Vigilance	2								x						
Element	• Management of disruptive behaviour	2								x						

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Element	• Atmosphere of the room	2								x						
Marker	◦ ⚠ Table 2 enthält nur Auszug (Theme 1, Subthemes 1.1–1.3) mit Poor/Excellent-Deskriptoren — keine vollständige Marker-Liste publiziert	3								x						
Marker	◦ [1.1 poor] No introductions offered by any of the team members; team members 'talk to the room'; patient/partner has to ask what is happening and about the baby's condition	3								x						
Marker	◦ [1.1 excellent] Introduction and role identification by team members as they enter the room; addresses patient and partner directly; team communicates throughout the entire episode	3								x						
Marker	◦ [1.2 poor] No one assumes responsibility for the patient and/or partner to the point where they become increasingly anxious as the situation unfolds	3								x						
Marker	◦ [1.2 excellent] At least one team member assumes responsibility for the patient and partner and spontaneously provides ongoing information and reassurance	3								x						
Marker	◦ [1.3 poor] No one assumes responsibility for monitoring change in partner behaviours; team members do not intervene during early signs of potentially disruptive behaviour	3								x						

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Marker	◦ [1.3 excellent] Team anticipates disruptive behaviour; at least one team member intervenes early to prevent escalation of disruptive behaviour	3								x						
Marker	◦ REFERENZIERT (nicht für Datenerhebung angewandt): MHPTS – Mayo High Performance Teamwork Scale (Malec et al. 2007) – als Vergleich zitiert; unterschiedliche Struktur zu AOTP.	3								x						
Element	• Focussed communication	2								x						
Element	• Closing the loop	2								x						
Kategorie	**Vigilance/awareness**	1									x					
Marker	◦ AUCH ANGEWANDT: ANTS – Anaesthetists' Non-Technical Skills (Fletcher et al. 2003)	3									x					
Kategorie	**Medical management**	1										x				
Marker	◦ The obstetrician encouraged questions from the obstetric resident.	3												x		
Marker	◦ The anesthesiologist encouraged questions from the anesthesia resident.	3												x		
Marker	◦ The successful management of the scenario was mainly a function of the obstetrician's expertise.	3												x		
Marker	◦ The successful management of the scenario was mainly a function of the anesthesiologist's expertise.	3												x		
Marker	◦ The obstetric resident should have been more involved in the patient's care.	3												x		

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Marker	◦ The anesthesia resident should have been more involved in the patient's care.	3												x		
Marker	◦ The successful management of the case was mainly due to the technical proficiency of the physicians.	3												x		
Marker	◦ During the critical event management, the nurses were appropriately consulted by the physicians.	3												x		
Kategorie	**The nurses assumed a leadership role during the scenario.**	1												x		
Kategorie	**Error management**	1												x		
Marker	◦ During the critical event management, the most senior physician was in charge.	3												x		
Marker	◦ The residents questioned the actions of the consultant physicians.	3												x		
Marker	◦ When a critical event occurred, the nurse(s) questioned the actions of the physician.	3												x		
Marker	◦ When a critical event occurred, the obstetrician questioned the actions of the anesthesiologist.	3												x		
Marker	◦ When a critical event occurred, the anesthesiologist questioned the actions of the obstetrician.	3												x		
Marker	◦ If there was uncertainty on the part of any team member, the team openly questioned that team member on his or her actions.	3												x		

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Marker	◦ The attending anesthesiologist clearly verbalized his/her plan of action when a critical event occurred.	3												x		
Marker	◦ The attending obstetrician clearly verbalized his or her plan of action when a critical event occurred.	3												x		
Marker	◦ The anesthesiologist ensured that requests for action were acknowledged by the receiver.	3												x		
Marker	◦ The obstetrician ensured that requests for action were acknowledged by the receiver.	3												x		
Marker	◦ The nurse(s) ensured that requests for action were acknowledged by the receiver.	3												x		
Marker	◦ The obstetric resident ensured that requests for action were acknowledged by the receiver.	3												x		
Marker	◦ The anesthesia resident ensured that requests for action were acknowledged by the receiver.	3												x		
Kategorie	**The team shared information efficiently.**	1												x		
Kategorie	**Communication between physicians was effective.**	1												x		
Kategorie	**Communication between physicians and nurses was effective.**	1												x		
Kategorie	**Communication between nurses was effective.**	1												x		
Kategorie	**Obstetricians gave feedback to the anesthesiologist.**	1												x		
Kategorie	**Anesthesiologists gave feedback to the obstetricians.**	1												x		

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Kategorie	**Physicians gave feedback to the nurses.**	1												x		
Kategorie	**Nurses gave feedback to the physicians.**	1												x		
Marker	◦ The anesthesiologist took charge of coordinating the team effort.	3												x		
Marker	◦ The obstetrician took charge of coordinating the team effort.	3												x		
Kategorie	**The nurses took charge of coordinating the team effort.**	1												x		
Kategorie	**The team effectively prioritized activities.**	1												x		
Kategorie	**Conflicts were openly resolved.**	1												x		
Kategorie	**The team worked well together.**	1												x		
Kategorie	**Basic rules were broken during the management of the case.**	1												x		
Kategorie	**Mistakes were made and not voiced by team members.**	1												x		
Kategorie	**Errors that were committed resulted from lack of knowledge.**	1												x		
Marker	◦ Errors that were committed resulted from lack of communication.	3												x		
Kategorie	**Errors that were committed resulted from lack of equipment.**	1												x		
Marker	◦ Errors that were committed resulted from lack of technical skills.	3												x		

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	ANTS	ANTS	Ottawa GRS	ANTSdk	ANTSdk	AS-NTS	ENTS	AOTP	BARS	CALM	Ottawa GRS	HFRS	GRS	Ottawa GRS
			Fletcher et al. (2003)	Jirativanont et al. (2017)	Kim et al. (2006/2009) zit. in Jirativanont et al. (2017)	Jepsen et al. (2015)	Jepsen et al. (2016)	Moll-Khosrawi et al. (2019)	Flowerdew et al. (2012)	Tregunno et al. (2009)	Watkins et al. (2015)	Nadkarni et al. (2018)	Kim et al. (2006)	Morgan et al. (2007)	Morgan et al. (2007)	Franc et al. (2016)
Marker	◦ Errors that were committed resulted from lack of following guidelines.	3												x		
Kategorie	**Errors that were committed resulted from lack of experience.**	1												x		
Kategorie	**Errors that were committed resulted from lack of resources.**	1												x		
Kategorie	**1 – Unacceptable Performance**	1													x	
Kategorie	**2 – Borderline Performance**	1													x	
Kategorie	**3 – Acceptable Performance**	1													x	
Kategorie	**4 – Good Performance**	1													x	
Kategorie	**5 – Superior Performance**	1													x	

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	NTS-Assessment	MHPTS	MHPTS	NOTSS v1.1	NOTSS	NTS-NAS	OSANTS	OSCAR	STAT	T-NOTECHS	TEAM	TEAM	TEAM	TPOT
			Moorthy et al. (2005)	Malec et al. (2007)	Gosselin et al. (2019)	Yule et al. (2006)	Yule et al. (2008)	Pires et al. (2018)	Dedy et al. (2015)	Walker et al. (2011)	Reid et al. (2012)	Repo et al. (2019)	Cooper et al. (2010)	Freytag et al. (2019)	Carpini et al. (2021)	AHRQ / DoD (2014)
Kategorie	**Task management**	1				x		x	x					x	x	
Kategorie	**Team working**	1						x	x	x				x	x	
Kategorie	**Situation awareness**	1				x		x	x	x						x
Element	• Gathering information	2				x			x							
Kategorie	**Decision making**	1				x			x	x		x				
Kategorie	**Overall**	1									x			x		
Kategorie	**Leadership**	1	x			x		x	x	x		x			x	x
Kategorie	**Resource management**	1										x				
Kategorie	**Communication**	1	x			x		x	x	x		x				x
Kategorie	**Error management**	1						x								
Kategorie	**Preoperative preparation**	1	x													
Element	• Introduction to team members	2	x													
Element	• Preoperative instrument and equipment check	2	x													
Element	• Briefing	2	x													
Element	• Instructions to assistant/scrub nurse: clear and polite	2	x													
Element	• Awaits acknowledgment from the assistant/scrub nurse	2	x													
Element	• Assistance sought from team members	2	x													
Element	• Acknowledges help/advice from team members	2	x													
Kategorie	**Vigilance/situation awareness**	1	x													

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	NTS-Assessment	MHPTS	MHPTS	NOTSS v1.1	NOTSS	NTS-NAS	OSANTS	OSCAR	STAT	T-NOTECHS	TEAM	TEAM	TEAM	TPOT
			Moorthy et al. (2005)	Malec et al. (2007)	Gosselin et al. (2019)	Yule et al. (2006)	Yule et al. (2008)	Pires et al. (2018)	Dedy et al. (2015)	Walker et al. (2011)	Reid et al. (2012)	Repo et al. (2019)	Cooper et al. (2010)	Freytag et al. (2019)	Carpini et al. (2021)	AHRQ / DoD (2014)
Element	• Monitored patient's parameters throughout the procedure	2	x													
Element	• Awareness of anesthetist	2	x													
Element	• Actively initiates communication with anesthetist	2	x													
Element	• Adherence to best practice during the procedure	2	x													
Element	• Resource utilization, ie, appropriate task-load distribution and delegation of responsibilities	2	x													
Element	• Authority/assertiveness	2	x													
Element	• Understanding information	2				x			x							
Element	• Projecting and anticipating future state	2				x			x							
Element	• Considering options	2				x			x							
Element	• Selecting and communicating option	2				x			x							
Element	• Implementing and reviewing decisions	2				x			x							
Element	• Planning and preparation	2				x			x							
Element	• Flexibility / responding to change	2				x			x							
Element	• Setting and maintaining standards	2				x			x							
Element	• Supporting others	2				x			x							
Element	• Coping with pressure	2				x			x							
Element	• Exchanging information	2				x			x							
Element	• Establishing a shared understanding	2				x			x							
Element	• Co-ordinating team activities	2				x			x							
Kategorie	**Mutual support**	1						x								x
Kategorie	**Use cognitive aids**	1						x								

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	NTS-Assessment	MHPTS	MHPTS	NOTSS v1.1	NOTSS	NTS-NAS	OSANTS	OSCAR	STAT	T-NOTECHS	TEAM	TEAM	TEAM	TPOT
			Moorthy et al. (2005)	Malec et al. (2007)	Gosselin et al. (2019)	Yule et al. (2006)	Yule et al. (2008)	Pires et al. (2018)	Dedy et al. (2015)	Walker et al. (2011)	Reid et al. (2012)	Repo et al. (2019)	Cooper et al. (2010)	Freytag et al. (2019)	Carpini et al. (2021)	AHRQ / DoD (2014)
Kategorie	**Professionalism (PRO)**	1							x							
Marker	◦ AUCH ANGEWANDT: NOTSS – Non-Technical Skills for Surgeons (Yule et al. 2006)	3							x							
Marker	◦ REFERENZIERT (nicht für Datenerhebung angewandt): OTAS (Healey et al. 2004) – diente als Vorlage für OSCAR-Entwicklung; Exemplar-Behaviors adaptiert. NOTECHS – als Referenz zitiert.	3								x						
Kategorie	**Basic assessment skills**	1									x					
Kategorie	**Airway/breathing**	1									x					
Kategorie	**Circulation**	1									x					
Element	• All team members exhibit professional attitude and interactions (Team)	2									x					
Element	• There is a clearly identified team leader (Leadership)	2									x					
Element	• Assigns roles to team members (Leadership)	2									x					
Element	• Maximizes skill sets of personnel in assigned roles (Leadership)	2									x					
Element	• Directs/Redirects team members effectively (Leadership)	2									x					
Element	• Monitors actions of team members (Leadership)	2									x					
Kategorie	**Situation awareness/Coping with stress**	1										x				
Marker	◦ REFERENZIERT: T-NOTECHS (Steinemann et al. 2012, englische Original-Version) – Basis der finnischen Übersetzung.	3										x				

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	NTS-Assessment	MHPTS	MHPTS	NOTSS v1.1	NOTSS	NTS-NAS	OSANTS	OSCAR	STAT	T-NOTECHS	TEAM	TEAM	TEAM	TPOT
			Moorthy et al. (2005)	Malec et al. (2007)	Gosselin et al. (2019)	Yule et al. (2006)	Yule et al. (2008)	Pires et al. (2018)	Dedy et al. (2015)	Walker et al. (2011)	Reid et al. (2012)	Repo et al. (2019)	Cooper et al. (2010)	Freytag et al. (2019)	Carpini et al. (2021)	AHRQ / DoD (2014)
Kategorie	**The team leader let the team know what was expected of them through direction and command**	1											x			
Kategorie	**The team leader maintained a global perspective**	1											x			
Kategorie	**The team communicated effectively**	1											x			
Kategorie	**The team worked together to complete tasks in a timely manner**	1											x			
Kategorie	**The team acted with composure and control**	1											x			
Kategorie	**The team morale was positive**	1											x			
Kategorie	**The team adapted to changing situations**	1											x			
Kategorie	**The team monitored and reassessed the situation**	1											x			
Kategorie	**The team anticipated potential actions**	1											x			
Kategorie	**The team prioritized tasks**	1											x			
Kategorie	**The team followed approved standards/guidelines**	1											x			
Marker	◦ REFERENZIERT (nicht für Datenerhebung angewandt): ANTS, NOTSS, NOTECHS, MHPTS – in Literaturübersicht als verwandte Instrumente aufgeführt.	3											x			
Marker	◦ The team leader let the team know what was expected of them through direction and command	3												x		
Marker	◦ The team leader maintained a global perspective	3												x		
Marker	◦ The team communicated effectively	3												x		
Marker	◦ The team worked together to complete tasks in a timely manner	3												x		

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	NTS-Assessment	MHPTS	MHPTS	NOTSS v1.1	NOTSS	NTS-NAS	OSANTS	OSCAR	STAT	T-NOTECHS	TEAM	TEAM	TEAM	TPOT
			Moorthy et al. (2005)	Malec et al. (2007)	Gosselin et al. (2019)	Yule et al. (2006)	Yule et al. (2008)	Pires et al. (2018)	Dedy et al. (2015)	Walker et al. (2011)	Reid et al. (2012)	Repo et al. (2019)	Cooper et al. (2010)	Freytag et al. (2019)	Carpini et al. (2021)	AHRQ / DoD (2014)
Marker	◦ The team acted with composure and control	3												x		
Marker	◦ The team morale was positive	3												x		
Marker	◦ The team adapted to changing situations	3												x		
Marker	◦ The team monitored and reassessed the situation	3												x		
Marker	◦ The team anticipated potential actions	3												x		
Marker	◦ The team leader let the team know what was expected of them through direction and command	3												x		
Marker	◦ The team leader maintained a global perspective	3												x		
Marker	◦ The team prioritized tasks	3												x		
Marker	◦ The team followed approved standards/guidelines	3												x		
Marker	◦ The team leader let the team know what was expected of them through direction and command	3												x		
Marker	◦ The team leader maintained a global perspective	3												x		
Marker	◦ Q1. The team leader let the team know what was expected of them through direction and command	3													x	
Marker	◦ Q2. The team leader maintained a global perspective	3													x	
Marker	◦ Q4. The team worked together to complete tasks in a timely manner	3													x	
Marker	◦ Q5. The team acted with composure and control	3													x	
Marker	◦ Q7. The team adapted to changing situations	3													x	
Marker	◦ Q8. The team monitored and reassessed the situation	3													x	
Marker	◦ Q9. The team anticipated potential situations	3													x	

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	NTS-Assessment	MHPTS	MHPTS	NOTSS v1.1	NOTSS	NTS-NAS	OSANTS	OSCAR	STAT	T-NOTECHS	TEAM	TEAM	TEAM	TPOT
			Moorthy et al. (2005)	Malec et al. (2007)	Gosselin et al. (2019)	Yule et al. (2006)	Yule et al. (2008)	Pires et al. (2018)	Dedy et al. (2015)	Walker et al. (2011)	Reid et al. (2012)	Repo et al. (2019)	Cooper et al. (2010)	Freytag et al. (2019)	Carpini et al. (2021)	AHRQ / DoD (2014)
Marker	◦ Q11. The team followed approved standards/guidelines	3													x	
Kategorie	**Team Structure**	1														x
Marker	◦ b. Assigns or identifies team members' roles/responsibilities	3														x
Marker	◦ d. Includes patients and families as part of the team	3														x
Marker	◦ a. Provides brief, clear, specific, and timely information to team members	3														x
Marker	◦ b. Seeks information from all available sources	3														x
Marker	◦ c. Uses check-backs to verify information that is communicated	3														x
Marker	◦ d. Uses SBAR, call-outs, and handoff techniques to communicate effectively with team members	3														x
Marker	◦ b. Uses resources efficiently to maximize team performance	3														x
Marker	◦ d. Delegates tasks or assignments, as appropriate	3														x
Marker	◦ e. Conducts briefs, huddles, and debriefs	3														x
Marker	◦ b. Monitors fellow team members to ensure safety and prevent errors	3														x
Marker	◦ c. Monitors the environment for safety and availability of resources	3														x
Marker	◦ d. Monitors progress toward the goal and identifies changes that could alter the plan of care	3														x

BMS Taxonomie – Vergleich aller 28 Studien																
Typ	Bezeichnung	Ebene	NTS-Assessment	MHPTS	MHPTS	NOTSS v1.1	NOTSS	NTS-NAS	OSANTS	OSCAR	STAT	T-NOTECHS	TEAM	TEAM	TEAM	TPOT
			Moorthy et al. (2005)	Malec et al. (2007)	Gosselin et al. (2019)	Yule et al. (2006)	Yule et al. (2008)	Pires et al. (2018)	Dedy et al. (2015)	Walker et al. (2011)	Reid et al. (2012)	Repo et al. (2019)	Cooper et al. (2010)	Freytag et al. (2019)	Carpini et al. (2021)	AHRQ / DoD (2014)
Marker	◦ e. Fosters communication to ensure that team members have a shared mental model	3														x
Marker	◦ a. Provides task-related support and assistance	3														x
Marker	◦ b. Provides timely and constructive feedback to team members	3														x
Marker	◦ c. Effectively advocates for patient safety using the Assertive Statement, Two-Challenge Rule or CUS	3														x
Marker	◦ d. Uses the Two-Challenge Rule or DESC Script to resolve conflict	3														x

Quelle: Eigene Darstellung

12. Abbildungsverzeichnis

13. Tabellenverzeichnis

14. Abkürzungen / Glossar

Abkürzung	Bedeutung	Hinweise / Kontext
ABCDE	Airway, Breathing, Circulation, Disability, Exposure	Strukturiertes Schema zur systematischen Patientenbeurteilung in Notfallsituationen.
ACRM	Anesthesia Crisis Resource Management	Adaption des CRM-Konzepts für die Anästhesiologie.
ANTS	Anaesthetists' Non-Technical Skills	Verhaltensmarkierungssystem zur Bewertung nicht-technischer Fähigkeiten in der Anästhesie.
ANTSdk	Anaesthesiologists' Non-Technical Skills in Denmark	Dänische Adaption des ANTS-Instruments.
AS-NTS	Anaesthesiology Students' Non-Technical Skills	Instrument zur Erfassung nicht-technischer Fähigkeiten von Medizinstudenten in der Anästhesiologie.
AOTP	Assessment of Obstetrical Team Performance	Instrument zur Bewertung der Teamleistung in der Geburtshilfe.
BARS	Behaviorally Anchored Rating Scale Tool	Verhaltensverankerte Ratingskala zur Bewertung nicht-technischer Fähigkeiten.
BMS	Behavioral Marker Systems	Systeme zur Erfassung beobachtbarer, nicht-technischer Verhaltensweisen in Hochrisikoumgebungen.
CALM	Concise Assessment of Leader Management	Instrument zur Bewertung von Führungsverhalten in pädiatrischen Reanimationsteams.
CRM	Crew Resource Management	Konzept zur Optimierung der Teamleistung durch Nutzung aller verfügbaren Ressourcen (Mensch, Technik, Information).

CUS	Concerned, Uncomfortable, Safety (I am Concerned, Uncomfortable, this is a Safety issue)	Kommunikationswerkzeug zur Eskalation von Sicherheitsbedenken.
CV	Content Validity	Inhaltsvalidität; Maß dafür, ob ein Instrument alle relevanten Aspekte eines Konstrukts abdeckt.
CVI	Content Validity Index	Index zur quantitativen Bewertung der Inhaltsvalidität eines Instruments.
DESC	Describe, Express, Suggest, Consequences	Strukturiertes Skript zur Konfliktlösung in Teams.
FOR-DEC	Facts, Options, Risks – Decision, Execution, Check	Entscheidungsmodell zur strukturierten Entscheidungsfindung unter Stress.
GAOTP	Global Assessment of Obstetrical Team Performance	Globale Bewertungsskala für die Teamleistung in der Geburtshilfe.
GRS	Global Rating Scale	Globale Bewertungsskala zur summarischen Einschätzung von Leistungen.
HFRS	Human Factors Rating Scale	Instrument zur Bewertung von Human-Factors-Aspekten in der Teamarbeit.
HFS	High-Fidelity Simulation	Hochrealistische Simulationen zur Schulung und Bewertung klinischer Fertigkeiten.
ICC	Intraclass Correlation Coefficient	Statistisches Maß für die Übereinstimmung zwischen mehreren Beurteilern (Interrater-Reliabilität).
LOSA	Line Operations Safety Audit	Instrument zur Sicherheitsbewertung in der Luftfahrt, adaptiert für medizinische Kontexte.
MHPTS	Mayo High Performance Teamwork Scale	Instrument zur Bewertung von Teamarbeit in simulationsbasierten Kontexten.
NTS	Non-Technical Skills	Nicht-technische Fähigkeiten wie Kommunikation, Führung, Teamarbeit, Situati-

		onsbewusstsein und Entscheidungsfindung.
NTS-NAS	Non-Technical Skills – Nursing Assessment Scale	Instrument zur Erfassung nicht-technischer Fähigkeiten in der Pflegeausbildung.
NOTSS	Non-Technical Skills for Surgeons	Verhaltensmarkierungssystem zur Bewertung nicht-technischer Fähigkeiten von Chirurgen.
NOTECHS	Non-Technical Skills (ursprünglich aus der Luftfahrt)	Instrument zur Bewertung nicht-technischer Fähigkeiten, adaptiert für medizinische Kontexte.
OSANTS	Objective Structured Assessment of Nontechnical Skills	Instrument zur strukturierten Bewertung nicht-technischer Fähigkeiten im Operationssaal.
OSCAR	Observational Skill-Based Clinical Assessment Tool for Resuscitation	Instrument zur Bewertung nicht-technischer Fähigkeiten in Reanimationsteams.
OSCE	Objective Structured Clinical Examination	Strukturierte Prüfungsform zur Bewertung klinischer Fertigkeiten.
OTAS	Observational Teamwork Assessment for Surgery	Instrument zur Bewertung von Teamarbeit im Operationssaal.
PALS	Pediatric Advanced Life Support	Fortgeschrittenes Reanimationskonzept für pädiatrische Notfälle.
SBAR	Situation, Background, Assessment, Recommendation	Strukturiertes Kommunikationswerkzeug zur klaren Informationsweitergabe.
SPLINTS	Scrub Practitioners' List of Intraoperative Non-Technical Skills	Instrument zur Bewertung nicht-technischer Fähigkeiten von OP-Pflegekräften.
STAT	Simulation Team Assessment Tool	Instrument zur Bewertung der Teamleistung in simulierten pädiatrischen Reanimationssituationen.

T-NO-TECHS	Trauma Non-Technical Skills Scale	Instrument zur Bewertung nicht-technischer Fähigkeiten in Traumateams.
TEAM	Team Emergency Assessment Measure	Instrument zur Bewertung der Teamleistung in medizinischen Notfallsituationen.
TPOT	TeamSTEPPS® Team Performance Observation Tool	Instrument zur Bewertung von Teamleistung im Rahmen des TeamSTEPPS®-Curriculums.
TRACS	Tool for Resuscitation Assessment using Computerized Simulation	Instrument zur Bewertung von Reanimationsleistungen in computergestützten Simulationen.
Two-Challenge Rule	Regel zur Eskalation von Sicherheitsbedenken: Zweimalige Wiederholung einer Warnung, falls keine Reaktion erfolgt.	Kommunikationswerkzeug zur Durchsetzung von Sicherheitsmaßnahmen.