Aus der Klinik und Poliklinik für Psychosomatische Medizin und Psychotherapie
der Universitätsmedizin der Johannes Gutenberg-Universität Mainz

Einsatz eines digitalen Rückmeldesystem (EvaSys) zur Beurteilung der kommunikativen Kompetenz bei Studierenden der Medizin im Vergleich mit einem papierbasierten Verfahren

Inauguraldissertation
zur Erlangung des Doktorgrades der
Medizin
der Universitätsmedizin
der Johannes-Gutenberg-Universität Mainz

Vorgelegt von

Jisung Hong
aus Masan (Südkorea)

Mainz, 2025

Wissenschaftlicher Vorstand: Univ.-Prof. Dr. med. Philipp Drees

Tag der Promotion: 05.05.2026

Inhaltsverzeichnis

Abkürzungen

A

AS ..... Aufgaben-Schwierigkeit

D

DiFS ..... digitalbasiertes Feedback-System
DOZ ..... Checklisten-Bewertung der Dozierenden

F

FG ..... Fragengruppe
FG1 ..... Fragengruppe 1
FG2 ..... Fragengruppe 2
FG3 ..... Fragengruppe 3
FG4 ..... Fragengruppe 4
FG5 ..... Fragengruppe 5
FG6 ..... Fragengruppe 6

G

GED ..... Gesamt-Note in Checkliste
GLT ..... Gesamt-Punkte in Checkliste
GNT ..... Gesamt-Note in Fragebogen
GPT ..... Gesamt-Punkte in Fragebogen

O

OSCE ..... Objective Structured Clinical Examination

P

PaFS ..... papierbasiertes Feedback-System

R

RP ..... Role Play (Simulationsgespräch)
RP1 ..... Simulationsgespräch 1 Thema Anamnese
RP2 ..... Simulationsgespräch 2 Thema Gesprächförderung
RP3 ..... Simulationsgespräch 3 Thema Informationsvermittlung
RP4 ..... Simulationsgespräch 4 Thema Partizipative Entscheidung
RP5 ..... Simulationsgespräch 5 Thema Compliance
RP6 ..... Simulationsgespräch 6 Thema Motivationales Interview
RP7 ..... Simulationsgespräch 7 Thema Stressreaktion
RP8 ..... Simulationsgespräch 8 Thema Krebsaufklärung

S

SD ..... Standardabweichung
SoSe21 ..... Sommersemester2021
SP ..... Fragebogen-Bewertung der Simulationspatienten und -innen
STU ..... Checklisten-Bewertung der Studierenden

W

WiSe21/22 ..... Wintersemester 2021/2022

Tabellen

Abbildung

Abstrakt

Einleitung Durch Simulationsgespräche mit standardisierten Simulationspatienten und -innen können die Studierenden im Studiengang Medizin ihre ärztlichen Kommunikationstechniken praktisch ausbilden. Zur Evaluation des Simulationsgesprächs sind Checklisten für Studierenden-, Dozierenden und Simulationspatienten-Feedback entwickelt worden. Das Ziel dieser Studie ist, die testtheoretische Qualität der Checklisten als Feedback-System zu evaluieren und die Einsetzbarkeit des Online-Systems EvaSys als ein digitalbasiertes Feedback-System zu überprüfen.

Methode Im Sommersemester 2021 und Wintersemester 2021/2022 führten n = 473 Studierende jeweils zwei Simulationsgespräche durch. Die Themen des Gespräches waren Anamnese, Gesprächsförderung, Informationsvermittlung, partizipative Entscheidung, Compliance, Motivationales Interview, Stressreaktion und Krebsaufklärung. Diese Gespräche wurden anhand von lernzielbasierten Checklisten bewertet und statistisch analysiert.

Ergebnis Eine gute Internale-Konsistenz, eine leichte Objektivität, eine akzeptable Validität, eine leichte Aufgabenschwierigkeit, eine schwache Trennschärfe und eine homogene Dimensionalität wurden zur testtheoretischen Qualität der Checkliste-Bewertung nachgewiesen. Die mit EvaSys erhobenen Daten unterscheiden sich bei allen acht Themen nicht signifikant von den Ergebnissen der Papier-Checklisten. Bei den EvaSys-Chekclisten wurden signifikant weniger Fehler bei der Datenauswertung festgestellt als bei den Papier-Checklisten.

Konklusion Die Checkliste-Bewertung ist ein zuverlässiges Messverfahren des Simulationsgesprächs für Arzt-Patient-Kommunikation bei Medizinstudierenden. Das digitalbasierte Rückmeldesystem EvaSys kann für die untersuchten Checklisten zur Beurteilung der kommunikativen Kompetenz bei Simulationsgespräch das papierbasierte Rückmeldesystem ohne Qualitätsverlust ersetzen.

Schlüsselwörter Checkliste, digital, Evaluation, EvaSys, Gespräch, Kommunikation

1. Einleitung

1.1 Hintergrund

Kommunikation ist eine der wichtigsten Säule der ärztlichen Kompetenz. Neben fachlichem Wissen sowie technischer Fertigkeit ist die Kommunikation mit Patienten und -innen im ärztlichen Berufsleben in nahezu allen Fachrichtungen unabdingbar. Bei fast allen Tätigkeiten im klinischen Einsatz sollten Ärzte und -innen kontinuierlich mit Patienten und -innen sprechen. Es treten verschiedene Situationen auf, wie z.B. Patientenaufnahme, Anamnese, Untersuchung, Diagnose- bzw. Informationenmitteilung, Beratung, Entscheidung für weitere Maßnahmen und Förderung der Compliance. In jeder Situation müssen sie das Gespräch angemessen und effektiv führen, um die Patienten und -innen erfolgreich zu behandeln. Aus diesem Grund nimmt die Bedeutung der Lehre von Kommunikation in der ärztlichen Ausbildung stetig zu.

Im Nationaler Kompetenzbasierter Lernzielkatalog Medizin (NKLM) und der neu ab 2025 geltenden Approbationsordnung für Ärzte wird der Kommunikationsfähigkeit von Ärzten und -innen noch mehr Bedeutung beigemessen [1]. Diese neuen Veränderungen weisen darauf hin, in welche Richtung die medizinische Ausbildung gehen sollte. Besonders in den mündlichen Prüfungen für die Approbation wird die Abfrage von Fachwissen durch OSCE (Objective Structured Clinical Examination) ersetzt, um neben Fachwissen auch dessen Anwendung in der Praxis zu prüfen. Das bedeutet, dass zukünftige Ärzte und -innen bereits während ihres Studiums noch praxisnäher ausgebildet werden sollen. Daher sind die Medizinstudierenden aufgefordert, sich intensiv und praktisch mit der Arzt-Patient-Kommunikation auseinanderzusetzen und sie zu üben.

1.2. Simulationsgespräch mit standardisierten Simulationspatienten und -innen (SP)

Für bessere ärztliche Kommunikationstechniken werden im Studiengang Medizin bereits verschiedene Methoden eingesetzt. Unter dem Ansatz von „Learning by Doing“ bietet das Simulationsgespräch mit standardisierten Simulationspatienten und -innen (SP) den Studierenden eine erweiterte Möglichkeit für praktische Übungen [2–11]. Indem die Studierenden Gespräche mit Patienten und -innen führen und Feedback erhalten, können sie die erlernten Kommunikationstechniken anwenden und ihre Leistung verbessern. Darüber hinaus haben sie auch die Gelegenheit, die Gespräche ihrer Mitstudierenden mit Patienten und -innen zu beobachten und sich metakommunikativ damit auseinanderzusetzen [12]. Dadurch gewinnen sie verschiedene gute Beispiele für Stärken und Verbesserungspotenzial in der Arzt-Patient-Kommunikation und können ihre Technik verbessern [10,11].

Die Studierenden, die diese Lernmethode anwenden, sind überwiegend damit zufrieden und zeigen bessere Leistungen in der Kommunikationstechnik [13–21]. Die Erfahrungen mit Patientinnen und Patienten stärken auch das Selbstvertrauen der Studierenden [22]. Die praktische Anwendung, wie die Verwendung der erlernten Kommunikationstechniken in einem Gespräch mit Simulationspatienten und -innen, kann einen besseren Lernerfolg bringen als die reine Theorie. Auf diese Weise können die Studierenden nicht nur Flexibilität im Umgang mit Patienten und -innen entwickeln, sondern auch die theoretischen Grundlagen der Arzt-Patient-Kommunikation verinnerlichen. Dies führt einerseits zu hoher Zufriedenheit bei den Studierenden und andererseits entspricht es den Veränderungen des NKLM und der neuen Approbationsordnung, die eine größeren Wert auf praktische ärztliche Kommunikationskompetenz legt.

Die Studierenden lernen nicht nur durch die aktive Gesprächsführung. Sie können auch als Zuschauer die Gespräche ihrer Mitstudierenden und deren unterschiedlichen Herangehensweisen an

Patienten und -innen betrachten. Jeder Studierende hat seine eigene Vorgehensweise, was zu unterschiedlichen Interaktionen mit den Patienten und -innen führt. Dieses breite Spektrum an Herangehensweisen verdeutlicht, dass es nicht nur eine einzige richtige Lösung für ein Gespräch gibt, sondern viele verschiedene Möglichkeiten bestehen. Dann tauschen die Studierenden ihre Eindrücke und Meinungen dazu aus. Durch diese Metakommunikation analysieren sie Stärke und Schwächen der Gespräche, was zur Verbesserung ihrer eigenen Kommunikationstechniken beiträgt. Dies vertieft bei den Studierenden auch ihr Verständnis für die Arzt-Patient-Kommunikation. So ermöglicht ihnen das Rollenspiel mit Simulationspatienten und -innen, bei dem die Studierenden sowohl das Gespräch üben als auch die Gespräche beobachten und bewerten können, ein besseres Lernen im Vergleich zu anderen Lernmethoden wie z.B. Beispiel-Videos nach Skript [16,18,20].

Jedoch hat diese Lernmethode Einschränkungen. Es ist deutlich aufwendiger als andere Methoden. Zunächst müssen die Simulationspatienten und -innen für das Rollenspiel organisiert werden [23]. Die rekrutierten Simulationspatienten und -innen benötigen Zeit zur Vorbereitung, damit sie die Fallvignette verstehen und sich damit vertraut machen, genauso wie die Studierenden. Im Kurs ist es auch zeitaufwendig, jedes Rollenspiel durchzuführen und darauf Feedback zu geben. Nicht jede medizinische Bildungseinrichtung kann die dafür benötigte Kapazität aufbringen. Ob das Rollenspiel mit Simulationspatienten und -innen trotz dieser Umstände eine hervorragende Lernmethode ist, bleibt eine offene Frage [24]. Als Alternative können die Studierenden das Rollenspiel innerhalb einer Peer-Gruppe ohne Simulationspatienten und -innen üben. Allerdings scheint es, dass die Studierenden mit dieser Methode weniger profitieren als mit dem Rollenspiel mit Simulationspatienten und -innen [16,25].

1.3. Checklisten als Feedback-System

Zudem müssen die lernzielbasierten Fallvignetten für das Gespräch und die Checkliste für das Feedback vorbereitet werden. Sie sind wesentliche Bestandteile, um das Gespräch in einer standardisierten Form ablaufen zu lassen. Das hilft den Studierenden, trotz ihrer noch begrenzten Erfahrungen, alle wichtigen Punkte des Gesprächs abzudecken. Ein improvisiertes Gespräch könnte sich häufig in unerwartete Richtungen führen und somit die geplanten Übungen, die den Lernzielen entsprechen, erschweren.

Mit Hilfe der Checkliste können die Studierenden das Gespräch ihrer Mitstudierenden strukturiert beobachten. Indem sie die Checkliste beachten, können sie das Gespräch detailliert bewerten, was zu einer noch vertiefenden Diskussion in der Metakommunikation führt. Daher ist die Checkliste von entscheidender Bedeutung für einen optimalen Lernerfolg durch Rollenspiel. Ebenfalls benötigen die Dozierenden eine gut strukturierte Checkliste, die zur expliziten Leistungsüberprüfung der Rollenspiele dienen kann. Die Bewertung anhand der Checkliste gibt eine direkte Übersicht über Stärke und Schwächen, die beobachtet wurden. Dies hilft den Dozierenden, ausführlich das Gespräch zu kommentieren und objektiv zu beurteilen.

In unserer Studie wurden die Checklisten verwendet, die speziell für die jeweiligen Gesprächskonzepte entwickelt wurden [26]. Anhand der Checkliste bzw. Fallvignette bereiten sich die Studierenden auf das Gespräch mit Simulationspatienten und -innen vor. Auch für die Bewertung des Gesprächs und die anschließende Diskussion wird die Checkliste von Dozierenden und Studierenden genutzt, um systematisch gemäß den Lernzielen den Verlauf des Gesprächs zu analysieren. Also, die Checkliste ist für diese Lernmethode wesentlich und von grundlegender Bedeutung [26–28]. Daher ist es für die Checkliste erforderlich, eine gute testtheoretische Qualität aufzuweisen. Zu diesem Zweck

wurden eine Reihe von Tests durchgeführt, bei denen die Checkliste-Bewertungen statistisch analysiert wurden. Dadurch konnten wir die testtheoretische Qualität der Checklisten als Feedback-System abschätzen. Diese Auseinandersetzung gab uns Hinweise, wie die Checkliste noch optimiert werden kann und in welche Richtung sich die Lernmethode für Arzt-Patient-Kommunikation weiterentwickeln sollte.

1.4. Das Rückmeldesystem

Neben einem zu schriftlich auszufüllenden Papierbogen zur Bewertung gibt es eine moderne Methode mit digitalem Endgerät. Im Vergleich zum papierbasierten Verfahren erleichtert uns das digitalbasierte Rückmeldesystem die Datenverwaltung und statistische Analyse. Zudem ist es umweltfreundlicher, da weniger Papier benötigt wird. Daher wird ein Umstieg auf das digitalbasierte System im medizinischen Bereich angestrebt. Voraussetzung für diese Umstellung ist, dass der Einfluss des Rückmeldesystems auf die Bewertung ausgeschlossen wird. Fällt das Ergebnis der Bewertung je nach dem Bewertungssystem unterschiedlich aus, kann das System nicht einfach umgestellt werden. Aus diesem Grund wurden bereits in einigen Studien versucht, diese Frage zu beantworten [29–32]. Auch bei den in unserer Studie verwendeten Checklisten soll die Ersatzbarkeit überprüft werden, um eine zuverlässige Umstellung auf ein digitalbasiertes System zu ermöglichen.

Für das Feedback wurde die Checkliste normalerweise in Papierform an die Bewertenden ausgehändigt und bearbeitet. Allerdings war dies unter den Bedingungen der COVID19-Pandemie in der Online-Lehre sehr aufwändig, vor allem hinsichtlich der Rückgabe. Unter solche Umstände wurde die Umstellung zum digitalen System beschleunigt, indem dies einen kontaktlosen Verlauf ermöglichte [33,34]. Die webbasierte Befragungssoftware EvaSys bietet ein Online Feedback-System an [35]. Mit dieser Software kann man eine Online-Umfrage erstellen und von den Teilnehmern die Daten einsammeln. Die gesammelten Daten werden digital ausgewertet und analysiert. So kann die Software zum digitalen Rückmeldesystem für die Checkliste angewendet werden. Aber es gibt aktuell nur begrenzte Daten, um zu entscheiden, ob das papierbasierte Verfahren schlicht durch die Onlineform ersetzt werden kann [5,6]. So überprüften wir die Einsatzbarkeit dieses Online Feedback-Systems, indem wir die Checkliste-Bewertungen aus beiden Rückmeldesystemen statistisch verglichen.

Aus diesem Hintergrund führten wir diese Studie durch. Das Ziel dieser Studie bestand darin, zunächst die testtheoretische Qualität der verwendeten Checklisten sowohl im papierbasierten Feedback-System (PaFS) als auch im digitalbasierten Feedback-System (DiFS) abzuschätzen und anschließend die Einsatzbarkeit des digitalbasierten Feedback-Systems (DiFS) für diese Checklisten zu überprüfen.

2. Literatur Diskussion

Bezüglich unseres Themas gibt es bereits einige publizierte Studien. Durch die Recherche dieser Studien können wir den aktuellen Forschungs- sowie Wissensstand erlangen. Das ist entscheidend wichtig, um unser Studiendesign effizient aufzubauen. Im Folgenden werden einige Studien diskutiert, die uns die wichtigen Hinweise gaben.

Tabelle 1 Übersicht der diskutierten Literaturen

Autor	Titel	Publikation	Fragenstellung	Probanden	Schlussfolgerung
Setyonugroho W. et al.	Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review	Patient Education and Counseling 2015:98(12)	Qualität der RP-Bewertung	Systemic Review: 3 Studien	Kein Goldstandard für Bewertung der OSCE
Natt N. et al.	High-Value, Cost-Conscious Communication Skills in Undergraduate Medical Education: Validity Evidence for Scores Derived from Two Standardized Patient Scenarios	Simulation in Healthcare 2018:13(5)	Qualität der RP-Bewertung	n = 97 Studierende der Medizin	Validitätsprüfung für Checklisten-Bewertungen der Simulationsgespräche nach Messick's Schema
Phillips A. et al.	A comparison of electronic and paper-based clinical skills assessment: Systematic review	Medical Teacher 2019:41(10)	DiFS und PaFS	Systemic Review: 5 Studien	Reliable Ergebnisse des DiFS zu PaFS, Geringer Fehlerrate des DiFS als PaFS
Felicitas L. et al.	Usability and preference of electronic vs. paper and pencil OSCE checklists by examiners and influence of checklist type on missed ratings in the Swiss Federal Licensing Exam	GMS Journal of Medical Education 2022:39(2)	DiFS und PaFS	N = 377 Studierende der Medizin	Präferenz der Benutzer zu DiFS
Schmitz F. et al.	Electronic Rating of Objective Structured Clinical Examinations: Mobile Digital Forms Beat Paper and Pencil Checklists in a Comparative Study	Lecture Notes in Computer Science 2011:7058(S 11)	DiFS und PaFS	N = 240 Studierenden der Medizin	Reliable Ergebnisse des DiFS zu PaFS, Geringer Fehlerrate des DiFS als PaFS
Daniels V. et al.	Impact of tablet-scoring and immediate score sheet review on validity and educational impact in an internal medicine residency Objective Structured Clinical Exam (OSCE)	Medical Teacher 2019:41(9)	DiFS und PaFS	N = 67 Assistenzärzte und -innen	Geringer Fehlerrate des DiFS als PaFS
Markland A. et al.	Psychometric evaluation of an online and paper accidental bowel leakage questionnaire: The ICIQ-B questionnaire	Neurourology and Urodynamics 2017:36(1)	DiFS und PaFS	N = 65 Patienten mit Fäkal-inkontinenz	Reliable Ergebnisse des DiFS zu PaFS

Li J. et al.	A Brief Online and Offline (Paper-and-Pencil) Screening Tool for Generalized Anxiety Disorder: The Final Phase in the Development and Validation of the Mental Health Screening Tool for Anxiety Disorders (MHS: A)	Frontiers in Psychology 2021:12(639366)	DiFS und PaFS	N = 527 Gesunde und Patienten mit generalisierter Angststörung	Reliable Ergebnisse des DiFS zu PaFS

2.1. Qualitätsprüfung der Checkliste-Bewertung

Das systemische Review von Setyonugroho W. et al. und die Studie von Natt N. et al. zeigen uns die wichtigen Aspekte für das Evaluationsverfahren der Kommunikationstechniken [27,36].

Es bestand noch kein standardisiertes Studiendesign für die Evaluierung der Kommunikationstechniken [27]. Trotz ausführlicher Recherche in den 23 eingeschlossenen Studien konnten Setyonugroho W. et al. keinen eindeutigen Goldstandard für Bewertung der Kommunikationstechnik feststellen. In den recherchierten Studien waren die Beschreibungen des gesamten Verlaufs der Studie nicht einheitlich, sodass die systematische Analyse zwischen den Studien schwierig war. Nur 30% der eingeschlossenen Studien gaben die Dauer der einzelnen Stationen an. Der Anteil der Studien, die keine klare Erläuterung zum Messverfahren lieferten, war sogar 83%, was zu einer Fehlinterpretation der Ergebnisse führen könnte. Daher empfahlen sie für die zukünftigen Studien eine detaillierte Beschreibung und Angabe des Studiendesigns.

Markland A. et al hatten auch ein ähnliches Problem in ihrer Studie über den klinischen Fragebogen [29]. Die Ergebnisse vom geprüften Fragebogen (ICIQ-B) wurden mit denen von den gängigen Fragebögen (BSFS, Vaizey Severity Score und MOS SF-12) verglichen, um die Validität von ICIQ-B zu überprüfen. Aber von den genannten Fragebögen gab es keinen anerkannten Goldstandard. So wurde es auch zur Einschränkung der Studie angegeben, ob die für die Validitätsprüfung angewendeten Teste selbst valid sind. Für die Prüfung der Validität ist auf einen Goldstandard aufgewiesen [37–41].

Trotz des fehlenden Goldstandards versuchten Natt N. et al. mit n = 97 Studierenden die Validität der zwei neu entwickelten RP-Szenarien und Checklisten für die Kommunikationstechnik zu prüfen. Diese Validität prüften sie nach dem Schema von Messick [40]. Es bestand aus fünf Komponenten: nämlich Testinhalt, Reaktionsprozess, internale Struktur, Relation zu anderen Variablen und Konsequenz von Test. Testinhalt und Reaktionsprozess konnten sie aus der Vorbereitung und dem Ablauf des RP ableiten. Die Entwicklung des RP und der damit verbundenen Checkliste, die Protokolle für das Training von Simulationspatienten und -innen bzw. Rater und die Interrater-Reliabilität der Checkliste-Bewertungen waren Evidenz für Testinhalt und Reaktionsprozess. Internale Struktur und Relation zu anderen variables wurden durch die weiteren statistischen Verfahren geprüft.

Die Rückmeldungen von Studierenden bewiesen letztlich die Konsequenz von Test. Über 90% der Studierenden stimmten zu, dass diese Übungen ihre kommunikativen Fähigkeiten verbesserten. Allerdings bleibt fraglich, ob die Zufriedenheit der Studierenden tatsächlich eng mit der Verbesserung ihrer Leistung zusammen korrelierte. Die Teilnehmer bekamen zwar als Konsequenz der OSCE-Prüfung eine Überzeugung, die damit Ihre Fähigkeiten verbessert wurden, aber diese Einschätzung war weder das zu messende Testergebnis noch das erwartete Ziel des Tests [40,41]. Die Konsequenzen in diesem Zusammenhang sind erst messbar, wenn die Teilnehmer als Ärzte in ihrer Berufstätigkeit die Kommunikationstechnik anwenden.

2.2. Vergleich zwischen DiFS und PaFS

Es gibt bereits mehrere Studien, die das DiFS und das PaFS verglichen, sowohl in der OSCE-Bewertung als auch in der Anwendung der klinischen Fragebogen.

Phillips A. et al. recherchierten die Studien zum Thema: A comparison of electronic and paper-based clinical skills assessment: Systematic review [30]. Dabei wurden schließlich fünf Studien eingeschlossen und analysiert. Davon wurde OSCE nur in einer Studie durchgeführt, in den anderen vier Studien wurde eine andere Form verwendet, z.B. Face-to-face-training. Die vier von fünf Studien gaben keine Informationen über Dauer und Ablauf des Moduls an, wie bereits in der Studie von Setyonugroho W. et al. vermerkt wurde. Die Teilnehmer befanden sich im zweiten oder dritten Jahr ihrer medizinischen Ausbildung, während die Studierenden unserer Studie im zweiten Semester des ersten Jahres waren. Allerdings hatten sie bereits erfolgreich den Kurs Medizinische Psychologie und Medizinische Soziologie I absolviert, was ein gewisses Vor- und Fachwissen nachweist. Die Ergebnisse waren reliabel zwischen DiFS und PaFS. Aber beim DiFS gab es niedrigere Fehlerrate, und die Benutzer gaben Präferenz zum DiFS gegenüber dem PaFS an.

Von Felicitas L. et al. wurde einen Vergleich zwischen PaFS und DiFS durchgeführt: Usability and preference of electronic vs. paper and pencil OSCE checklists by examiners and influence of checklist type on missed ratings in the Swiss Federal Licensing Exam [31]. In den beiden darauffolgenden Jahren wurden die Daten aus den fünf Universitäten gesammelt. Die gesammelten Daten umfassten die OSCE-Bewertungen jeweils mit PaFS im Jahr 2014 sowie mit DiFS im Jahr 2015, und die Rückmeldung der Prüfenden zur OSCE-Checkliste. Wie die OSCE-Bewertung fand diese Umfrage auch im Jahr 2014 mit Papierbogen, und im Jahr 2015 mit einem digitalen App statt. Die große Teilnehmerzahl (n=377) sowie die Teilnahme mehrerer Fakultäten boten die erweiterte Generalisierbarkeit des Ergebnisses an. Aber die OSCE-Checklisten gestalteten in den beiden Rückmeldesystemen unterschiedlich (PaFS mit 25-38 Items je nach OSCE-Station, DiFS mit 21-42 Items je nach OSCE-Station), sodass leider ein direkter Vergleich der OSCE-Bewertungen aus den beiden Rückmeldesystemen schwierig war. Stattdessen wurde die Anzahl der fehlenden Bewertungen analysiert, wobei das DiFS deutlich besser abschnitt als das PaFS.

Schmitz F. et al. untersuchten systematisch den Unterschied zwischen PaFS und DiFS beim OSCE-Bewertung: Electronic Rating of Objective Structured Clinical Examinations: Mobile Digital Forms Beat Paper and Pencil Checklists in a Comparative Study [42]. Aus fünf Aspekten wurden die beiden Systeme verglichen, zwar subjektive Nützlichkeit, mentale Anstrengung, fehlende Daten, Score Ergebnis und Präferenz. Die n = 10 OSCE-Prüfenden bewerteten die Leistung der insgesamt n = 240 Studierenden im OSCE entweder mit PaFS oder mit DiFS. Anschließend beantworteten sie eine Reihe der Umfragen, darunter PSSUQ für die subjektive Nützlichkeit, RSME für die mentale Anstrengung, und die Präferenz beider Systeme. Die Score Ergebnisse und die fehlenden Daten wurden aus den OSCE-Bewertungen erhoben. Die Einzelheiten der Untersuchung sowie der Ablauf des OSCE bzw. der anschließenden Umfrage wurden ausführlich beschrieben. Somit konnte man Einzelheiten aus der Studie entnehmen, was zu einer besseren Qualität der zukünftigen Metaanalyse führt. Die Analysen durch Vergleich der Mittelwerte sowie Mann-Whitney-U-Test wiesen einige Vorteile des DiFS und die zuverlässige Übereinstimmung der Ergebnisse aus den beiden Rückmeldesystemen auf.

Durch den Vergleich zwischen DiFS und PaFS beschäftigten sich Daniels V. et al. mit den Vorteilen von DiFS: Impact of tablet-scoring and immediate score sheet review on validity and educational

impact in an internal medicine residency Objective Structured Clinical Exam (OSCE) [43]. Neben den fehlenden Bewertungen wurden auch die fehlenden Kommentare und die Dauer der Ergebnisermittlung analysiert, um die Überlegenheit von DiFS gegenüber PaFS zu quantifizieren. Die Studienteilnehmer bestanden aus n = 67 Assistenzärzte in ersten bzw. zweiten Jahr in ihrer Facharztausbildung. Die n = 23 Teilnehmer erhielten die OSCE-Bewertung mit PaFS, bei den n = 44 anderen Teilnehmer wurde die OSCE-Bewertung hingegen mit DiFS durchgeführt. Bei der Bewertung wurden die Prüfenden gebeten, zusätzlich zur Bewertung auch deren Begründung gemäß konkreter Leitlinie abzugeben. Die Ergebnisse von DiFS zeigten eine signifikant geringere Anzahl fehlender Bewertungen. Dies wies die klaren Vorteile von DiFS bei der OSCE-Bewertung auf.

Es gibt noch Studien, in denen die Fragebogen für die klinische Anwendung in Papierform und Digitalsystem verglichen. Eine davon ist: Psychometric evaluation of an online and paper accidental bowel leakage questionnaire: The ICIQ-B questionnaire [29]. Markland A. et al beschäftigten sich mit einem Fragebogen um Fäkalinkontinenz. Es ging zwar nicht um OSCE für Kommunikationstechnik, jedoch um die psychometrische Evaluation, aus der wir einige wichtigen Aspekte für unsere Studie entnehmen konnten. Die rekrutierten n = 65 Patienten mit Fäkalinkontinenz wurden in zwei Gruppen randomisiert, und nahmen im Verlaufe der nichtoperativen Behandlung dreimal (Baseline, zwei Wochen, drei Monate) an der Umfrage teil. Bei jedem der drei Termine füllten die Patienten die als valid anerkannten Fragebogen aus, darunter BSFS (Bristol Stool Form Scale), Vaizey Severity Score und MOS SF-12, und den zu untersuchenden Test ICIQ-B entweder in der Papierform oder im Digitalsystem an. Das ICIQ-B bestand aus 17 Items in drei Fragengruppen, die größtenteils in der Likert-Skala angegeben werden sollten. Die beantworteten Items wurden ohne Gewichtung summiert und interpretiert. Die für die konvergente Validität eingesetzten Fragebögen BSFS und Vaizey Severity Score untersuchten die Symptome und den Schwergrad von Fäkalinkontinenz, und MOS SF-12 erfasste die Lebensqualität der Patienten. Die Homogenität beider durch Randomisierung geteilten Stichproben wurde in Bezug auf Alter, Geschlecht und Berufsstatus getestet. Dafür ließen sich Fischer's-exact-test und Welch-t-Test anwenden. Abgesehen vom Berufsstatus waren die beiden Stichproben homogen. Bei den Ergebnissen beider Rückmeldesystemen gab es keinen signifikanten Unterschied.

Die Studie von Li J. et al. hatte noch größere Stichprobe als die von Markland A. et al. Aber es handelt sich um eine psychometrische Evaluation für GAD (Generalized Anxiety Disorder): A Brief Online and Offline (Paper-and-Pencil) Screening Tool for Generalized Anxiety Disorder: The Final Phase in the Development and Validation of the Mental Health Screening Tool for Anxiety Disorders (MHS: A) [32]. Insgesamt wurden 527 Daten von Gesunden und Patienten mit GAD erhoben. Die Teilnehmer wurden gebeten, eine Reihe von Fragebogen auszufüllen. Zuerst bekamen sie den zu untersuchenden Fragebogen MHS: A in Papierform. Anschließend folgten vier weitere Teste, MINI Plus Version 5.0.0, BAI, GAD-7 und PSWQ, deren Ergebnisse zur Validitätsprüfung dienten. Danach bearbeiteten sie erneut das MHS: A, diesmal jedoch in Digitalform. Wie Markland A. et al., konnten Li J. et al. auch keinen signifikanten Unterschied von den beiden Rückmeldesystemen herausfinden. So konnte ihre Studie bestätigen, dass der Ersatz des PaFS durch DiFS kein Störfaktor bei ihrer Fragebogen-Umfrage war.

In den genannten Studien zeigten sich generell die reliablen Ergebnisse zwischen PaFS und DiFS. Dadurch kamen die Schlussfolgerungen, dass das PaFS für ihre Checkliste-Bewertungen bzw. Fragebögen durch das DiFS ersetzt werden kann. Bei statistischen Analysen fielen die Studien von Felicitas L. et al. und Schmitz F. et al. auf. Sie führten den Vergleich zwischen den beiden Rückmeldesystemen mit Mann-Whitney-U-Test durch, da die Daten auf Ordinal Skala vorlagen.

Um das Nutzen von den beiden Rückmeldesystemen zu vergleichen, wurden in einigen Studien die Fehlerrate untersucht. In den zwei von Phillips A. et al. auseinandergesetzten Studien wurde die Anzahl der Missing-Data aufgezählt. Auch in den Studien von Felicitas L. et al., Schmitz F. et al. und Daniels V. et al. wurden die Fehlerrate von den beiden Rückmeldesystem ermittelt. Hier zeigte DiFS eine deutliche Überlegenheit gegenüber dem PaFS.

Neben der deutlich besseren Fehlerrate ließen sich die weiteren Vorteile des DiFS nennen. Mit dem elektronischen System ist weniger Nacharbeit erforderlich. Die Bewertungen werden automatisch ins zentrale System übertragen und gesammelt, ohne dass Fehler auftreten. Beim PaFS muss man die Bewertungen entweder einscannen oder manuell eintippen, um die Daten in einen Datensatz zu integrieren. Dabei ist der Einsatz der menschlichen Arbeitskraft unvermeidbar, sodass viele Fehler auftreten könnten. Für die Aufbewahrung der originalen Daten hat DiFS noch einen Vorteil. Die digitalisierten Daten besetzen kaum physischen Platz, während die gesammelten Papierbogen auf viel Raum angewiesen sind.

Phillips A. et al., Felicitas L. et al. und Schmitz F. et al. fragten die Teilnehmer zur Zufriedenheit sowie Präferenz des Rückmeldesystems. Sie antworteten, dass sie mit dem DiFS weniger müde und insgesamt zufriedenstellend waren als mit PaFS. Zudem war das DiFS viel einfacher zu benutzen ist als das PaFS. Daher bevorzugten die meisten Befragten das DiFS. Allerdings waren diese Ergebnisse aus der Umfrage die subjektiven Präferenzen der Benutzer, wobei diese positiven Präferenzen zum DiFS nicht die gute testtheoretische Qualität der Checkliste bzw. des Fragebogens im DiFS gewährleisten.

3. Fragestellung

In dieser Studie wurden hauptsächlich zwei Fragen untersucht.

Zuerst wurde die testtheoretische Qualität der eingesetzten Checkliste zur RP-Bewertung überprüft. Diese Überprüfung wurde bereits bei der Einführung des RP und der Checkliste von Fischbeck S. et al. durchgeführt [26,45]. In unserer Studie wurde die testtheoretische Qualität dieses Messverfahrens erneut mit größeren Stichproben überprüft, sowohl mit PaFS als auch mit DiFS. Dies dient einerseits zur Überprüfung der Einsatzfähigkeit von DiFS für unsere RP, indem die beiden Feedback-Systems in verschiedenen Aspekten verglichen werden. Andererseits liefert es Evidenz für die Etablierung dieser Lernmethode in der Arzt-Patient-Kommunikation.

Die zweite Frage lautet, ob das papierbasierte Feedback-System (PaFS) beim RP zum Thema Arzt-Patient-Kommunikation durch das digitalbasierte Feedback-System (DiFS) ersetzbar ist. Sollte es trotz ihrer Unterschiede möglich sein, die zuverlässigen Ergebnisse mit den beiden Feedback-System zu erzielen, könnte man das digitalbasierte System zu Nutze bringen, da es praktisch eine bessere Effektivität bei Bearbeitung bieten kann [33,34,37,46,47].

3.1. Frage 1: Testtheoretische Qualität der Checkliste bei RP für Arzt-Patient-Kommunikation

Die Evaluation und das Feedback nach dem RP sind wichtige Verfahren, um die Kommunikationstechnik der Studierenden zu verbessern. Dabei unterstützt die Checkliste, indem sie sicherstellt, dass alle zu prüfende Punkte behandelt werden. Die Checklistenfragen ermöglichen es den Prüfenden, keine Einzelheiten zu übersehen. Eine genaue und ausführliche Evaluation ermöglicht ein detailliertes Feedback und trägt zur Verbesserung des Lernerfolgs der Studierenden bei. So spielen die Struktur und der Inhalt der Checkliste bei der Bewertung des RP eine wesentliche Rolle. Aus diesem Grund untersuchten wir in unserer Studie die testtheoretische Qualität der Feedback-Systeme, einschließlich der Gütekriterien (Objektivität, Reliabilität und Validität), der Aufgaben-Schwierigkeit und der Trennschärfe.

Die verwendete Checkliste umfasst je nach dem Fall Vignette 18-21 Fragen, die man entweder mit „Ja“ oder „Nein“ ankreuzen kann, eine Note für den Gesamteindruck, die man von 1 „Sehr gut“ bis 5 „Mangelhaft“ bewerten kann, und einen Freitextkommentar. (s. 11. Anhang) Aufgrund dessen Formular als Fragebogen sollte die Checkliste bestimmte Merkmale aufweisen, um eine gute testtheoretische Qualität zu gewährleisten. Dazu gehören hohe Reliabilität, hohe Testobjektivität, hohe Testvalidität, hohe Trennschärfe der einzelnen Items, ein breites Spektrum an Schwierigkeiten der Items und Homogenität der Items (Dimensionalität) [46].

Die Qualität des Feedback-Systems wurde separat für PaFS und für DiFS überprüft. Es diente einerseits zum wiederholten Qualitätstest, andererseits zur Untersuchung der Einsatzbarkeit von DiFS beim RP. Nur unter der guten und homogenen Qualität beider Feedback-Systeme lässt sich die Verwendung von DiFS empfehlen.

3.2. Frage 2: Ersetzbarkeit durch DiFS bei RP für Arzt-Patient-Kommunikation mit PaFS

Bisher wurden die Checklisten bzw. Fragebogen für ein psychometrisches Messverfahren in konventioneller Form auf Papier verwendet. Mit Fortschritt der Technik lässt sich das Verfahren in ein digitales System überführen, um Daten effektiv zu analysieren und sicher aufzubewahren [30,31,42]. Allerdings entsteht bei dieser Umwandlung die Frage, ob der Unterschied zwischen beiden Methoden als Störfaktor die Qualität des Messverfahrens beeinträchtigen kann. Zwar sind die Inhalte der Checkliste identisch, aber der Umgang mit einem digitalen Endgerät könnte im Vergleich zur Handhabung eines Papierbogens die Bewertenden bei der Evaluation beeinflussen [30,31,42]. So sollten die spezifischen Eigenschaften beider Methoden auseinandergesetzt und verglichen werden, um festzustellen, welche Methode eine bessere Effektivität des Messverfahrens bietet, und ob das DiFS trotz des Unterschieds von PaFS eine akzeptable testtheoretische Qualität des Messverfahrens gewährleistet. Erst nach dieser Prüfung kann das DiFS als zuverlässige Methode für das getestete Messverfahren etabliert werden.

3.2.1. Papier basiertes Feedback-System (PaFS)

Das Ausfüllen eines Fragebogens auf Papier erfordert keine besondere Schulung oder Vorbereitung seitens der Bewertenden. Mit einem Stift können sie ihn intuitiv bearbeiten. Das PaFS bietet den Bewertenden einen freien Raum für Bearbeitung. Auf einem Papierbogen können sie beispielsweise neben den vorgegebenen Fragen in einer Ecke kurze Notizen machen, die später als Hinweise oder Anmerkungen dienen können. Vor allem für Personen, die Schwierigkeiten im Umgang mit digitalen Geräten haben, ist das PaFS eine gute Methode zur Evaluation. Deshalb ist das PaFS noch bei der Verwendung der Checkliste insbesondere für ältere Benutzer wichtig [32].

Aber bei der Personengruppe, die sich täglich mit dem digitalen Umfeld beschäftigen, verliert das PaFS heutzutage seinen Stellenwert. Beim PaFS müssen alle einzelnen Daten abgelesen und ggf. umgerechnet werden, um das Ergebnis zu interpretieren. Diese Arbeit ist umso aufwendiger, wenn die Anzahl der zu analysierenden Fragebögen größer ist. Die Aufbewahrung der gesammelten Daten stellt ebenfalls ein Nachteil von PaFS dar, da physisch viel Platz für die Lagerung der Papierbögen benötigt wird. Im Gegensatz dazu benötigen die digitalisierten Daten nur einen minimalen Speicherumfang.

Ein weiteres Problem bei handschriftlich ausgefüllten Fragebögen besteht in der Interpretation, wenn die Handschrift nicht gut lesbar ist. Das kann zu Fehlinterpretationen der Angaben führen. Bei der Eingabe über eine Tastatur besteht zwar das Risiko von Tippfehlern bei freien Textfeldern, aber das Risiko für Missverständnis ist geringer als bei schlecht lesbarer Handschrift.

3.2.2. Digital basiertes Feedback-System (DiFS)

Die oben genannten Probleme von PaFS kann man mit DiFS lösen. Digitales System braucht weniger Aufwand für die Datenspeicherung, und ermöglicht eine schnellere statistische Datenanalyse. Die digitale Datenverarbeitung erfordert nur einen minimalen Einsatz menschlicher Arbeitskräfte, angefangen von der Integration der Datensätze bis hin zur komplexen Analyse. Dadurch kann man viel Zeit und Kosten ersparen, im Vergleich zur Arbeit mit PaFS. Zudem bietet das digitale System ökologische Vorteile, da der Druck von Papierbögen entfällt.

Moderne Mobilgeräte mit stabiler Netzwerkverbindung ermöglichen reibungslose Arbeit mit dem digitalen System [27,33,34]. Ein großer Bildschirm und eine benutzerfreundliche Schnittstelle entlasten die Benutzer bei der Bedienung und stören sie weniger bei der Evaluation. Deshalb gaben die Mehrheit der Benutzer die Präferenz zum DiFS an [27,30,31,42].

Auf diesem Hintergrund werden die Umstellung von PaFS zu DiFS vorgenommen [33,34]. Aber für ein psychometrisches Messverfahren muss es zuerst geklärt werden, ob die Qualität des Messverfahrens durch den Einsatz des neuen Systems nicht beeinträchtigt wird. Daher wurde bereits in mehreren Studien der Unterschied der Ergebnisse mit den beiden Systemen untersucht, um die Einsatzbarkeit von DiFS zu prüfen [29–31,42].

Damit das neue DiFS als zuverlässiger Standard in einem psychometrischen Messverfahren verwendet werden kann, sollten die Ergebnisse von dem Feedback-System unabhängig sein. Das bedeutet, dass die Ergebnisse mit PaFS und mit DiFS nicht signifikant unterschiedlich sein sollten. Ein signifikanter Unterschied zwischen den Ergebnissen mit beiden Feedback-Systemen deutet darauf hin, dass das Feedback-System einen bestimmten Einfluss auf das Messverfahren ausübt. Dann kann das DiFS nicht als zuverlässige Methode für dieses Messverfahren anerkannt werden. Wenn die Ergebnisse aber keinen signifikanten Unterschied aufweisen, kann man das DiFS als Ersatz für PaFS in der Messung einsetzen.

4. Stichprobe und Methode

Die für unsere Studie rekrutierten Stichproben und die erhobenen Daten werden dargestellt. Zudem sind die zur Analyse eingesetzten Methoden auf dem Analyseplan aufgelistet.

4.1. Stichprobe und Material

Im Sommersemester 2021 und im Wintersemester 2021/2022 im Studiengang Humanmedizin an der Johannes-Gutenberg-Universität Mainz wurden im Rahmen des Kursus (Teil II) der Medizinischen Psychologie und Medizinischen Soziologie die Arzt-Patient-Gespräche von den Studierenden und den Simulationspatienten und -innen durchgeführt. Den Studierenden wurde zu Beginn des Semesters die Zuweisung von zwei RP-Themen mitgeteilt. Zur Vorbereitung standen ihnen Skript, Videos und schriftliche praxisbezogene Übungen im Rahmen des Kurses zur Verfügung. 14 Tage vor dem RP-Termin erhielten sie dann die konkrete Fallvignette sowie Checkliste für das RP.

Die Dozierenden (SoSe 2021: 6 Kurse weibliche, 8 Kurse männliche; WiSe 21/22: 11 Kurse weibliche, 3 Kurse männliche) bekamen etwa 14 Tage vor Semesterbeginn den Zugang zu allen Materialien für die Studierenden.

Die Simulationspatienten und -innen hatten im Sommersemester 2020 einen vier-stündigen Termin für die Übung im Lernlabor Universität Mainz. Etwa eine Woche vor diesem Termin erhielten sie die Fallvignetten. Nach der Übung sollten sie weiterhin persönlich üben, jedoch sind die genaue Dauer sowie Intensität des Übens unbekannt. Die Fragebögen für die Patientenrolle erhielten sie etwa eine Woche vor Beginn der Gespräche.

Das Arzt-Patient-Gespräch dauerte zehn Minuten. Die Studierenden in der Kursgruppe und die oder der für diese Kursgruppe zuständige Dozierende bewerteten das Gespräch anhand der Checkliste. (s. 11. Anhang) Nach dem Gespräch gab die Patientenrolle zuerst der Arztrolle ihre Eindrücke und die Rückmeldungen zum Gespräch ab, dann konnten die zugeschauten Studierenden ihr Feedback geben. Zum Schluss gab die oder der Dozierende einen Kommentar ab. Die gesamte Rückmelderunde dauerte ca. fünf Minuten. Die Patientenrolle konnte den Fragebogen erst nach dem Gespräch ausfüllen. Die Studierenden bekamen ca. eine Woche nach Semesterende per E-Mail alle Checkliste-Ratings für ihr eigenes RP. Die gesammelten Checklisten- sowie Fragebogen-Bewertungen wurden statistisch analysiert. Die Aufzeichnung bzw. die Aufschreibung der Gespräche sowie der Feedback-Runde wurden aus Datenschutzgründen nicht stattgefunden.

Für die Bewertungen wurde den Dozierenden und Studierenden die lernzielbasiert entwickelte Checkliste ausgehändigt. Die Checkliste umfasst Fragen zu den einzelnen Leistungen, den Gesamteindruck (GED), und den freien Kommentar [26]. Für die Gesamtnote sind einzelne Leistungen zu bewerten, ob die gestellte Aufgabe während des Gesprächs erfüllt wurde. Jede Aufgabe ist je nach Schwergrad 0,5 oder 1 bepunktet, insgesamt konnten für die gesamte Leistung (GLT) maximal auf 14 Punkte mit Ausnahme von RP4 auf 15 Punkte erreicht werden. Für GED kann man auf einer Skala von 1 bis 5 bewerten, wobei 1 für „sehr gut“, 2 für „gut“, 3 für „befriedigend“, 4 für „ausreichend“ und 5 für „mangelhaft“ stand. Daher wurden für jedes Gespräch eine GLT und GED von Dozierenden sowie eine bis vier GLT und GED von STU erhoben. Im SoSe21 gaben die Dozierenden und Studierenden ihre Bewertungen mit PaFS ab, im WiSe21/22 arbeiteten sie mit DiFS (EvaSys).

Die Checklistefragen bestehen je nach RP aus 18 bis 21 Fragen, und sind nach ihrem Konzept in drei bis sechs Gruppe geteilt. Jede Fragengruppe (FG) formulieren die Fragen unter verschiedenen

Bewertungskriterien, damit man das Gespräch den Lernzielen entsprechend konkret bewerten kann. Die FG erfassen die Einführung sowie die Qualität des Gesprächs, Behandlung der gegebenen Aufgaben und die Anwendung der angewiesenen Kommunikationstechniken.

Das Thema für RP1 ist „Anamnese“. Die Checkliste für RP1 beinhaltet 19 Fragen in vier FG. Bei FG1 geht es um die Einleitung des Gesprächs, bei FG2 um die Anamnese der Hauptbeschwerden, bei FG3 um die ergänzende Anamnese für Erstgespräch, und bei FG4 um die Qualität des Gesprächs.

Das Thema für RP2 ist „Gesprächsförderer“. Die Checkliste für RP2 beinhaltet 20 Fragen in fünf FG. Bei FG1 geht es um die Einleitung des Gesprächs, bei FG2 um die Gesprächsförderer, bei FG3 um die Gesprächsstörer, bei FG4 um die Behandlung der Hintergrundproblematik und bei FG5 um die Qualität des Gesprächs.

Das Thema für RP3 ist „Informationsvermittlung“. Die Checkliste für RP3 beinhaltet 18 Fragen in fünf FG. Bei FG1 geht es um die Einleitung des Gesprächs, bei FG2 um die Vorbereitung der Informationsvermittlung, bei FG3 um die Informationsvermittlung, bei FG4 um die Rückversicherung nach Informationsvermittlung und bei FG5 um die Qualität des Gesprächs.

Das Thema für RP4 ist „Partizipative Entscheidungsfindung / SDM“. Die Checkliste für RP4 beinhaltet 18 Fragen in drei FG. Bei FG1 geht es um die Einleitung des Gesprächs, bei FG2 um die partizipative Entscheidungsfindung, und bei FG3 um die Qualität des Gesprächs.

Das Thema für RP5 ist „Compliance-Förderung“. Die Checkliste für RP5 beinhaltet 19 Fragen in sechs FG. Bei FG1 geht es um die Einleitung des Gesprächs, bei FG2 um die Informierung, bei FG3 um die Überzeugung, bei FG4 um die Erleichterung, bei FG5 um die Unterstützung und bei FG6 um die Qualität des Gesprächs.

Das Thema für RP6 ist „Motivationales Interview (MI)“. Die Checkliste für RP6 beinhaltet 21 Fragen in fünf FG. Bei FG1 geht es um die Einleitung des Gesprächs, bei FG2 um die Vermittlung von MI-Spirit, bei FG3 um das Chang-Talk, bei FG4 um den geschmeidigen Umgang mit Widerstand und bei FG5 um die Qualität des Gesprächs.

Das Thema für RP7 ist „Stress- und Stressbewältigung“. Die Checkliste für RP7 beinhaltet 21 Fragen in fünf FG. Bei FG1 geht es um die Einleitung des Gesprächs, bei FG2 um die Stressorenanalyse, bei FG3 um die Analyse der Stressreaktion, bei FG4 um die Ansätze der Stressbewältigung und bei FG5 um die Qualität des Gesprächs.

Das Thema für RP8 ist „Krebsaufklärungsgespräch“. Die Checkliste für RP8 beinhaltet 20 Fragen in sechs FG. Bei FG1 geht es um die Einleitung des Gesprächs (Setting up the Interview), bei FG2 um Perception und Invitation, bei FG3 um Knowledge, bei FG4 um Emotion, bei FG5 um Strategy und bei FG6 um die Qualität des Gesprächs.

Nach dem Gespräch füllten die Simulationspatienten und -innen einen Fragebogen aus, um die Leistungen sowie den Gesamteindruck der Arztrolle im Gespräch zu bewerten. Der Fragebogen beinhaltet vier Fragen zur Qualität bzw. Zufriedenheit mit dem Gespräch und eine gesamte Note für das Gespräch, die man von 1 bis 5 geben kann, wie bei der Checkliste. Auf die vier Fragen stehen jeweils fünf Antwortmöglichkeit in Adjektivform zur Verfügung: „sehr“, „ziemlich“, „mittelmäßig“, „wenig“ und „nicht“. Die Antworten wurden ins Notensystem von 1 bis 5 umgerechnet und die Summe ergab die Gesamtpunkte (GPT), wobei dann die beste Leistung GPT 4 hat und die schlechteste GPT 20. Die Bewertung der Leistungen ergab somit die Gesamtpunkte (GPT), und Gesamteindruck die Gesamtnote (GNT). Aus einem RP wurden daher eine GPT und GNT von SP erhoben. Die

Simulationspatienten und -innen bearbeiteten den Fragebogen auf dem Papier in den beiden Semestern (s. Tabelle 3).

Die Stichprobe für unsere Studie bestand aus den Teilnehmenden des Kursus (Teil II) der Medizinischen Psychologie und Medizinischen Soziologie im Sommersemester 2021 (SoSe21, n = 245) und im Wintersemester 2021/2022 (WiSe21/22, n = 228) im Studiengang Humanmedizin des zweiten Semesters (Vorklinik) der Universitätsmedizin, Johannes-Gutenberg-Universität (JGU) Mainz.

Aus n = 245 Studierenden im SoSe21 sind n = 152 Studierende weiblich und n = 93 Studierende männlich. Sie haben Alter von 18 bis 47 Jahre alt, wobei der Mittelwert 22,93 Jahre und die Standardabweichung (SA) 4,12 Jahre ist.

Aus n = 228 Studierenden im WiSe21/22 sind n = 144 Studierende weiblich und n = 84 Studierende männlich. Sie haben Alter von 18 bis 35 Jahre alt, wobei der Mittelwert 23,10 Jahre und die Standardabweichung (SA) 3,01 Jahre beträgt.

In den beiden Semestern waren alle sieben Dozierenden des Kurses beim RP als Prüfende beteiligt. Alle von denen waren bereits mindestens ein Semester zuvor in der Lehre.

Es wurden sechs Schauspielerinnen und Schauspieler aus der Lernklinik in Universitätsmedizin Mainz rekrutiert, um die Rolle der Patienten im RP zu übernehmen.

Im Rahmen des Kurses sollten die Studierenden in ihrer aus maximal 16 Studierenden bestehenden Kursgruppe jeweils zwei RPs mit unterschiedlichen Themen ausführen. Die Themen wurden den Studierenden nach dem Zufallsprinzip zugewiesen. Die Lernveranstaltungen mit RP fanden während des Semesters statt und erstreckten sich über einen Zeitraum von fünf Wochen ab der sechsten Unterrichtswoche. In den ersten beiden Wochen wurden RP1, RP2 und RP3 durchgeführt, in den folgenden dritten und vierten Wochen RP4, RP5 und RP6, und in der letzten Woche dann RP7 und RP8. Im SoSe21 wurden RP1 insgesamt 70-mal, RP2 59-mal, RP3 58-mal, RP4 63-mal, RP5 67-mal, RP6 62-mal, RP7 44-mal, RP8 42-mal im Verlauf des Semesters ausgeführt. Im WiSe21/22 wurden RP1 insgesamt 62-mal, RP2 61-mal, RP3 66-mal, RP4 53-mal, RP5 55-mal, RP6 57-mal, RP7 47-mal, RP8 46-mal im Verlauf des Semesters ausgeführt.

Tabelle 2 Anzahl der Teilnehmer und der Ausführung des RPs

Semester	Studierende*			Dozierende	Schauspieler	RP1	RP2	RP3	RP4	RP5	RP6	RP7	RP8
	m	w	Gesamt										
SoSe21	93	152	245	7	6	70	59	58	63	67	62	44	42
WiSe21/22	84	144	228	7	6	62	61	66	53	55	57	47	46

* Für Geschlecht gibt es keine Kategorie „divers“, weil diese Daten aus Verwaltungsakten entnommen wurden.

Somit wurden als Daten beim RP1 im SoSe21 256 GLT bzw. 249 GED von STU, 70 GLT bzw. 69 GED von DOZ und 68 GPT bzw. 64 GNT von SP erhoben. Im WiSe21/22 wurden 121 GLT bzw. GED von STU, 58 GLT bzw. GED von DOZ und 61 GPT bzw. GNT von SP erhoben.

Beim RP2 im SoSe21 wurden 216 GLT bzw. 212 GED von STU, 59 GLT bzw. 58 GED von DOZ und 57 GPT bzw. 50 GNT von SP erhoben. Im WiSe21/22 wurden 120 GLT bzw. 117 GED von STU, 57 GLT bzw. GED von DOZ und 59 GPT bzw. 58 GNT von SP erhoben.

Beim RP3 im SoSe21 wurden 212 GLT bzw. 205 GED von STU, 57 GLT bzw. 55 GED von DOZ und 56 GPT bzw. 52 GNT von SP erhoben. Im WiSe21/22 wurden 130 GLT bzw. 127 GED von STU, 63 GLT bzw. 62 GED von DOZ und 63 GPT bzw. 61 GNT von SP erhoben.

Beim RP4 im SoSe21 wurden 231 GLT bzw. 224 GED von STU, 57 GLT bzw. GED von DOZ und 58 GPT bzw. GNT von SP erhoben. Im WiSe21/22 wurden 67 GLT bzw. GED von STU, 45 GLT bzw. 44 GED von DOZ und 51 GPT bzw. 53 GNT von SP erhoben.

Beim RP5 im SoSe21 wurden 247 GLT bzw. 233 GED von STU, 58 GLT bzw. GED von DOZ und 63 GPT bzw. 61 GNT von SP erhoben. Im WiSe21/22 wurden 81 GLT bzw. GED von STU, 44 GLT bzw. 43 GED von DOZ und 55 GPT bzw. GNT von SP erhoben.

Beim RP6 im SoSe21 wurden 221 GLT bzw. 210 GED von STU, 55 GLT bzw. GED von DOZ und 58 GPT bzw. 57 GNT von SP erhoben. Im WiSe21/22 wurden 115 GLT bzw. GED von STU, 52 GLT bzw. 51 GED von DOZ und 53 GPT bzw. 53 GNT von SP erhoben.

Beim RP7 im SoSe21 wurden 163 GLT bzw. 153 GED von STU, 38 GLT bzw. GED von DOZ und 41 GPT bzw. 37 GNT von SP erhoben. Im WiSe21/22 wurden 85 GLT bzw. GED von STU, 42 GLT bzw. GED von DOZ und 42 GPT bzw. GNT von SP erhoben.

Beim RP8 im SoSe21 wurden 155 GLT bzw. 151 GED von STU, 36 GLT bzw. GED von DOZ und 40 GPT bzw. 37 GNT von SP erhoben. Im WiSe21/22 wurden 81 GLT bzw. GED von STU, 44 GLT bzw. GED von DOZ und 44 GPT bzw. GNT von SP erhoben.

Die Daten aus Papier wurden in digital eingetragen. Alle Daten wurden danach pseudonymisiert mit STATA17 analysiert.

Tabelle 3 Anzahl der erhobenen Daten

	SoSe21						WiSe21/22					
	STU		DOZ		SP		STU		DOZ		SP	
	GLT	GED	GLT	GED	GPT	GNT	GLT	GED	GLT	GED	GPT	GNT
RP1	256	249	70	69	68	64	121	121	58	58	61	61
RP2	216	212	59	58	57	50	120	117	57	57	59	58
RP3	212	204	57	55	56	52	130	127	63	62	63	61
RP4	231	224	57	57	58	58	67	67	45	44	51	53
RP5	247	233	58	58	63	61	81	81	44	43	55	55
RP6	221	210	55	55	58	57	115	115	52	51	53	53
RP7	163	153	38	38	41	37	85	85	42	42	42	42
RP8	155	151	36	36	40	37	81	81	44	44	44	44

STU: Checkliste-Bewertungen aus Studierenden, DOZ: Checkliste-Bewertungen aus Dozierenden, SP: Fragebogen-Bewertungen aus Simulationspatienten und -innen, GLT: Gesamt-Punkte in Checkliste, GED: Gesamt-Note in Checkliste, GPT: Gesamt-Punkte in Fragebogen, GNT: Gesamt-Note in Fragebogen, RP: Rollenspiel

4.2. Analyse und Methode

Zunächst wurden der Mittelwert sowie SD von GLT (STU bzw. DOZ), GED (STU bzw. DOZ), GPT (SP), GNT (SP) bei jedem RP ermittelt. Dies diente zur Beurteilung der Verteilung der Daten, was für die Methodenwahl zur Analyse entscheidend ist.

Die Homogenität zwischen den beiden Semestern wurde getestet. Es ist eine wesentliche Voraussetzung für die Analyse zur Frage 2, dass die Stichproben der beiden Semester homogen sind. Dafür wurden das Geschlecht und das Alter von den Stichproben jeweils mit einem χ^2-Test bzw. einem t-Test geprüft [46,48–50].

Die Frage 1 besteht aus fünf Unterfragen. Für die Intrarater-Reliabilität wurde Spearman-Korrelation zwischen GLT und GED aus DOZ bzw. STU angewendet [46,51,52]. Für die Interrater-Objektivität wurde Conger's Kappa eingesetzt, um die Checkliste-Bewertungen zwischen den Studierenden bzw. zwischen den Studierenden und Dozierenden zusammen zu analysieren [46,51,53–58]. Darüber hinaus wurde der Kappa-Koeffizient nach Fragengruppe (FG) unterteilt ermittelt. Wir brachten die Spearman-Korrelation erneut in Einsatz, um die Inhaltsvalidität zu prüfen [37–39,46,47,59–61]. GED aus DOZ und GNT aus SP wurden gegeneinander verglichen. Für die Aufgabenschwierigkeit wurde zunächst der Mittelwert jeder Fragengruppe in einem RP berechnet, und als Prozentsatz gegen den maximal erreichbaren Punkt angegeben [46]. Als Item wurden die Fragengruppen der Checkliste nochmal analysiert, um ihre Trennschärfe zu bestimmen. Dafür wurde Cronbach's Alpha ermittelt [46]

Für Frage 2 wurden die Checkliste-Bewertungen, also GLT und GED, zwischen den beiden Semestern verglichen. Da zwei unabhängige, nichtnormalverteilte Stichproben analysieret werden sollten, wurde der Mann-Whitney-U-Test (Mann-Whitney-U-Test) eingesetzt [46,62].

Tabelle 4 Analyseplan

Fragenstellung	Methode	Vergleich	Stichprobe	Daten
Statistische Beschreibung	Mittelwert und SA*	-	Alle	GLT*, GED*, GPT*, GNT*
Homogenität	χ^2-Test, t-Test	SoSe21/WiSe21/22	STU*	Geschlecht, Alter
Einsetzbarkeit des DiFS (Frage 1)	Mann-Whitney-U-Test	SoSe21/WiSe21/22	STU, DOZ*	GLT, GED
Reliabilität (Frage 2.1)	Spearman-Korrelation	GLT/GED	STU, DOZ	GLT, GED
Objektivität (Frage 2.2)	Conger's Kappa	STU, STU/DOZ	STU, DOZ	FG*, GLT, GED
Validität (Frage 2.3)	Spearman-Korrelation	GED/GNT	DOZ, SP*	GED, GNT
AS* (Frage 2.4)	Mittelwert	FG	DOZ	FG
Trennschärfe (Frage 2.5)	Cronbach's α	FG	DOZ	FG

* SA: Standardabweichung, GLT: Gesamt-Punkte in Checkliste, GED: Gesamt-Note in Checkliste, GPT: Gesamt-Punkte in Fragebogen, GNT: Gesamt-Note in Fragebogen, STU: Checkliste-Bewertung aus Studierenden, DOZ: Checkliste-Bewertung aus Dozierenden, SP: Fragebogen-Bewertung aus Simulationspatienten und -innen, FG: Fragengruppe, AS: Aufgabenschwierigkeit

4.2.1. Methode für Frage 1: Testtheoretische Qualität der Checkliste bei RP für Arzt-Patient-Kommunikation

Zur Prüfung der testtheoretischen Qualität sollten die Gütekriterien sowie noch AS und Trennschärfe getestet werden. In unserer Studie wurde daher eine Reihe von den Analysen durchgeführt, um die Reliabilität, Objektivität, Validität, AS und Trennschärfe zu überprüfen. Über die Details der einzelnen Prüfungen wird es im Folgenden erklärt.

4.2.1.1. Reliabilität

Alle RP wurden ohne Video-Aufnahme nur einmal während der Gespräche bewertet. Daher war in unserer Studie keine präzise Analyse für Test-Retest-Reliabilität möglich. Stattdessen wurde die interne Konsistenz Reliabilität für die Reliabilitätsschätzung ermittelt. Dafür ließen sich GLT mit GED

aus DOZ und STU vergleichen, sowie GPT mit GNT aus SP für die jeweilige Beobachtung im selben RP.

Häufig wird hierbei Cronbach's Alpha verwendet, um einen Koeffizienten zu ermitteln [49,63–67]. Allerdings besteht die Checkliste aus binären Checkliste-Fragen, einem GED in Likert-Skala und einem freien Kommentar. Die Checkliste-Fragen lassen sich durch gewichtete Bepunktung in eine Summe GLT umrechnen. Aufgrund der Struktur dieser Datensätzen ist es schwierig, einfach Cronbach's Alpha anzuwenden [68]. Die zu analysierenden Daten erweisen sich als nicht normalverteilt, so dass die Reliabilität mit Cronbach's Alpha eher relativ unterschätz wird [63,66].

Daher wurde in dieser Studie die Spearman-Korrelation als die Methode angewendet, da die zu analysierenden Daten aus zwei Variablen bestehen und deshalb mit Rangkorrelation noch genauer der Koeffizient zu berechnen sind.

Das ermittelte Spearman's Rho gibt an, wie stark die beiden Daten korrelieren. Ein Wert von -0,1 bis 0,1 gilt als keine Korrelation, bis (-)0,3 als eine schwache Korrelation, bis (-)0,6 als eine mittlere Korrelation, bis (-)0,8 als eine moderate Korrelation, ab (-)0,8 als eine sehr starke Korrelation und von (-)1 als eine perfekte Korrelation [64,66].

Neben dem Spearman's Rho wird auch der p-Wert angegeben, um zu beurteilen, ob das Ergebnis in einem bestimmten Signifikanzniveau anzunehmen ist.

4.2.1.2. Objektivität

Um die Interrater-Objektivität zu bestimmen, sollte die Kappa-Analyse verwendet werden. Zuerst wurden die Bewertungen aus den Studierenden (STU) analysiert. Anschließend wurden zusätzlich die Bewertungen aus Studierenden (STU) und Dozierenden (DOZ) gemeinsam analysiert. Als Datensätze ließen sich die GLT, GED sowie die Bewertungen aller FG auseinandersetzen.

Die Ergebnisse aus STU lieferten Hinweise darauf, wie die RP-Bewertungen aus verschiedenen STU-Ratern miteinander korrelierten. Da die Studierenden allerdings nicht qualifiziert für RP-Bewertung waren, wurden ihre Bewertungen erneut mit denen aus den qualifizierten Dozierenden verglichen. Bei dieser Analyse verbanden wir die Rohdaten von STU mit den Bewertungen aus Dozierenden, und nahmen nochmal Conger's Kappa, wie die erste Analyse allein mit STU. Dies berücksichtigte den Einfluss einzelner Bewertungen auf das Ergebnis. Der Interrater-Vergleich zwischen den qualifizierten Dozierenden war nicht möglich, weil bei jedem RP nur ein Dozierender oder eine Dozierende an der Bewertung teilnahm.

Da in dieser Studie mehr als zwei Rater an der Bewertung beteiligt waren, war das Cohen's Kappa nicht für die Analyse geeignet. Für die Analyse mit mehr als zwei Ratern stehen jedoch andere Kappa-Verfahren zur Verfügung, wie z.B. Fleiss' Kappa oder Conger's Kappa. Aufgrund der detailierten Ergebnisse mit Konfidenzintervall ließ sich das Conger's Kappa verwenden [51,53,55,56].

Es bestand auch die Möglichkeit, die gängigste Methode wie Cohen's Kappa oder t-Test einzusetzen, durch Matching der Bewertungen aus Dozierenden und dem Mittelwert von dementsprechenden Bewertungen aus Studierenden [57,58]. Allerdings ließen sich die Auswirkungen einzelner Ausreißer begrenzen, und diese könnten sogar völlig ignoriert werden, falls es bei einer Beobachtung gegenseitige Ausreißer gab. So könnte eine bessere Interrater-Übereinstimmung erzielt werden, jedoch könnte dies zu einer Überschätzung der Objektivität führen. Aus diesem Grund wurde auf die zwei Rater-Analyse durch das Matching verzichtet.

Die Interpretation des Kappa-Werts zwischen 0 bis 0,2 lautet eine geringe Übereinstimmung, zwischen 0,2 bis 0,4 eine ausreichende, zwischen 0,4 bis 0,6 eine mittelmäßige, zwischen 0,6 bis 0,8 eine beachtliche, und zwischen 0,8 bis 1 eine vollkommene. Falls der Wert kleiner als 0 ist, deutet dies auf eine geringere Übereinstimmung als Zufall hin, sodass eine Interpretation schwer möglich ist [51,53].

Das Conger's Kappa liefert neben den Kappa-Koeffizienten das Konfidenzintervall, was uns noch weitere Interpretation ermöglicht. Obwohl der Koeffizient auf eine gute Übereinstimmung zwischen den Ratern hindeutet, z.B. kann man das Ergebnis nicht als zuverlässig gelten lassen, wenn das Konfidenzintervall einen negativen Bereich umfasst.

4.2.1.3. Validität

Aufgrund mangelhafter bzw. fehlender Standard-Messverfahren steht die Messung der Validität im Bereich Psychologie sowie Medizin oft in Frage, ob eine präzise Abschätzung der Validität möglich ist [37,47,59,60]. Dieses Problem betrifft auch unsere Validitätsprüfung der Checkliste. Es gibt kein Gold-Standard für die Messung der Qualität von Arzt-Patient-Kommunikation, mit dem wir unsere Checkliste direkt vergleichen können. Als ein Alternativ gibt es das Schema nach Messick. Diese Methode bietet die Möglichkeit einer multidimensionalen Analyse für ein Messverfahren an [36,40,69]. So evaluierten wir die Validität unserer Checkliste anhand der Methode nach Messick.

Diese besteht aus fünf Komponenten zur Überprüfung der Validität: Testinhalt, Reaktionsprozess, internale Struktur, Relation zu anderen Variablen und Konsequenz von Test [41,44,70–72].

Die auf dem Testinhalt basierende Validität ließ sich durch den Entwicklungsprozess von RP-Szenarien und -Checkliste qualitativ begutachten.

Den Reaktionsprozess konnte man anhand der Vorbereitung und Durchführung des RP-Prüfung betrachten, wobei auch die Interrater Korrelation einen Hinweis geben konnte.

Für die Prüfung der internalen Struktur ist die internale Konsistenz der Checkliste zu ermitteln. Die Korrelationen zwischen GLT und GED, sowie mit einzelnen FG konnten durch das Cronbach's Alpha berechnet werden.

Das Ziel der RP-Prüfung besteht darin „den handlungsbezogenen Wissensstand der Studierenden der Medizin und ihre Fertigkeiten, mit den Patienten zu kommunizieren" zu testen [45]. Das Fachwissen konnte man eigentlich durch eine Klausur überprüfen, und die Kommunikationskompetenz mit Patientinnen und Patienten konnte von der Rückmeldung sowie Anmerkung der Patientinnen und Patienten ausgehend bewerten. Die Messung der Validität erfolgte dann durch die Ermittlung der Korrelationen zwischen diesen herkömmlichen Bewertungen und den RP-Bewertungen. Da in unserer Studie jedoch keine Daten zu den Klausurergebnissen der Studierenden erhoben wurden, konnte die Validität in Bezug auf den Wissensstand nicht gemessen werden. Die andere Komponente, die Validität für die kommunikativen Fertigkeiten wurden jedoch überprüft, indem wir die Übereinstimmung der RP-Bewertungen mit den Bewertungen aus Simulationspatienten und -innen ermittelten. Dieses Ergebnis diente als Nachweis für die Relation zu anderen Variablen.

Die letzte Komponente zur Evaluation der Validität ist die Konsequenz von Test. Dies ist aber ein langfristiger Prozess, um die Konsequenzen unserer RP-Prüfung zu messen. Es ist erst zu messen, wenn die teilgenommenen Studierenden nach Abschluss ihres Studiums als Ärzte mit Patienten kommunizieren. Da in unserer Studie nicht geplant war, dies bereits jetzt voranziehend zu messen, und

keiner der Studierenden aktuell als Ärzte tätig waren, konnten wir nichts bezüglich der Konsequenz von Test messen und evaluieren [73].

Die Interrater-Korrelation, Korrelationen zwischen GLT und GED und Cronbach's Alpha mit FG wurden bereits jeweils zur Überprüfung der Objektivität (s. 5.3.2), der Reliabilität (s. 5.3.1) und der Trennschärfe (s.5.3.5) durchgeführt, sodass wir nicht erneut berechnen mussten.

Die einzige noch zu durchführende Analyse war die Korrelation zwischen den Bewertungen aus Dozierenden und denen aus Simulationspatienten und -innen. Dabei wurden die GED aus DOZ und die GNT aus SP als Material analysiert. Die beiden waren in Schulnotensystem abgegeben. Mit Spearman-Korrelation ließ sich die Übereinstimmung prüfen. Die Interpretation lief gleich wie die von Reliabilität, weil dafür auch das Spearman-Korrelation eingesetzt wurde (s. 4.2.2.1.) [64,66].

4.2.1.4. Aufgabenschwierigkeit (AS)

Für AS der Checkliste wurden zunächst der Mittelwerte der von Dozierenden erhaltenen Punkte in jeder FG ermittelt, dann wurden die Ergebnisse in AS-Index umgestellt angegeben. Dadurch lässt sich beurteilen, wie gut die Checkliste die guten und schlechten Leistungen der Studierenden voneinander unterscheiden kann.

Ein Item mit einem AS-Index über 80% bzw. unter 20% wird in der Regel ausgeschlossen. Denn mit zu schweren bzw. zu leichten Aufgaben kann man die Leistungen der Teilnehmer nicht effektiv unterscheiden. Idealerweise sollte eine Prüfung aus den Aufgaben bestehen, die ein breites Spektrum an Schwierigkeit abdecken [46,74,75].

4.2.1.5. Trennschärfe

Cronbach's Alpha wurde verwendet, um die Trennschärfe einzelner FG zu berechnen. Dafür wurden die Bewertungen aus Dozierenden als Material analysiert. Bei der Interpretation wurden die zwei Werte in Betracht gezogen: Item-Rest-Korrelation und Alpha-Koeffizient. Da es sich bei dieser Analyse um die Interaktionen zwischen den Items handelt, sollte man die Ergebnisse im Zusammenhang mit anderen Werten verstehen.

Die Item-Rest-Korrelation, auch bekannt als Trennschärfe, gibt Aufschluss über den Zusammenhang zwischen dem getesteten Item und den restlichen Items. Es sollte zumindest größer als 0,3 sein, damit das Item nicht ausgeschlossen oder überarbeitet werden muss.

Neben der Item-Rest-Korrelation wurde auch der Alpha-Koeffizient ermittelt, um die Interaktionen zwischen den Items zu verstehen. Durch den Vergleich mit dem Test-Scale wurden die einzelnen Alpha-Koeffizienten interpretiert. Ein Alpha von kleiner als 0,5 ist inakzeptabel, während ein Wert ab 0,8 als gut interpretiert wird [46,49].

4.2.1.6. Dimensionalität

Unter Testtheoretischer Qualitätsprüfung spricht man von Dimensionalität die Erfassungsdimension des testtheoretischen Konstrukts [46]. Da es bei der Bewertung mit der Checkliste keine Untertests gibt, sollte es eindimensional sein.

Die Checkliste umfasst inhaltlich die Fragen zu verschiedenen Leistungen von Studierenden als Arztrolle aus verschiedenen Aspekten. Einige Fragen beziehen sich auf die Einführung ins Gespräch und den Einstieg ins Thema, sowie auf das Verhalten bzw. Non-Verbal Kommunikation. Darüber hinaus wird es gefragt, ob die gegebenen Aufgaben erfolgreich in die Leistung gebracht wurden. Diese Fragen sind in der Checkliste den jeweiligen Fragengruppen entsprechend zugeordnet, bei jedem RP besteht die Checkliste aus drei bis sechs Fragengruppen (FG). Nichtsdestotrotz wird die Checkliste nicht als multidimensionale Prüfung angesehen, weil die allen verschiedenen Items doch gemeinsam die Qualität der Arzt-Patient-Kommunikation beim RP abfragen.

In unserer Studie wurde keine weitere statistische Prüfung bzw. Analyse hinsichtlich der Dimensionalität durchgeführt. Dies wird im 6.1.6 weiter qualitativ diskutiert.

4.2.2. Methode für Frage 2: Ersetzbarkeit durch DiFS bei RP für Arzt-Patient-Kommunikation mit PaFS

Die erhobene GLT und GED erwiesen sich als nicht normalverteilt und lagen auf Ordinal-Skalenniveaus vor. Insofern war eine t-Test anzuwenden, um die zwei Datensätze zu vergleichen, nicht möglich [46,62]. Zudem handelte es sich bei den Studierenden in den beiden Semestern um unabhängige Stichproben, obwohl sie homogen waren. Aus diesem Grund wurde Mann-Whitney-U-Test als die geeignete Methode ausgewählt, um die Daten präzis zu analysieren [18,46,62].

Durch Mann-Whitney-U-Test wurden die Checkliste-Bewertungen aus Dozierenden und Studierenden im SoSe21 mit denen im WiSe21/22 hinsichtlich Ihrer Unterschiede zu verglichen. Die Bewertungen aus Simulationspatienten und -innen wurden ebenfalls mit dem gleichen Verfahren analysiert.

Als Ergebnis erhielten wir einen z-Wert und einen p-Wert. Ist z-Wert größer als 1,96 oder kleiner als -1,96, sind die verglichenen zwei Datensätze im 5% Signifikanzniveau signifikant unterschiedlich. Liegt z-Wert aber zwischen -1,96 und 1,96, sind sie im 5% Signifikanzniveau nicht signifikant unterschiedlich.

Darüber hinaus wurde die Anzahl der fehlenden Bewertungen aufgezählt. Es gab unvollständige Bewertungen, bei denen beispielsweise lediglich die Checkliste-Fragen angekreuzt, aber keine GED angegeben wurden. Indem wir die Anzahl des Fehlers bei beiden Rückmeldesystemen verglichen, konnten die Vor- und Nachteile beider Systeme statistisch festgehalten werden.

Die gesamte Anzahl der Fehler ließ sich als Prozentsatz der gesamten Anzahl der Beobachtungen angeben, und zwischen den beiden Semestern vergleichen.

5. Ergebnisse

Die Ergebnisse der durchgeführten Analysen werden präsentiert, um die Fragestellungen der vorliegenden Arbeit zu beantworten. In der deskriptiven Statistik (s. 5.1.) werden alle analysierten Daten statistisch beschrieben. Anschließend (s. 5.1. und 5.2.) werden die ermittelten Ergebnisse vorgestellt.

5.1. Deskriptive Statistik

Die Verteilung spielt eine wichtige Rolle bei der Wahl der Analysenmethode. Mit Shapiro-Wilk-Test wurde der Test auf Normalverteilung durchgeführt. Zudem wurde die Homogenität der Stichproben zwischen den beiden Semestern geprüft, was eine wichtige Voraussetzung für die zuverlässige Interpretation ist.

5.1.1. Statistische Beschreibungen

Bei RP1 (Anamnese) im SoSe21 (Tabelle 5) ergaben sich alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass die Daten nicht normalverteilt waren.

Im WiSe21/22 ergaben sich auch alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass sie nicht normalverteilt waren.

Tabelle 5 Statistische Beschreibung für RP1

		SoSe21				WiSe21/22			
		Obs	MW	SD	t	Obs	MW	SD	t
DOZ	GLT	70	12,95	1,41	0,00	58	13,03	1,33	0,00
	GED	69	1,36	0,54	0,00	58	1,5	0,66	0,00
STU	GLT	256	12,90	1,26	0,00	121	12,67	1,78	0,00
	GED	249	1,34	0,51	0,00	121	1,50	0,70	0,00
SP	GPT	68	5,74	2,30	0,00	61	6,02	2,71	0,00
	GNT	64	1,47	0,64	0,00	61	1,60	0,72	0,00

* Obs: Anzahl der Beobachtungen, MW: Mittelwert, SD: Standardabweichung, t: Ergebnis von Shapiro-Wilk-Test

Bei RP2 (Gesprächsförderung) im SoSe21 (Tabelle 6) ergaben sich alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass die Daten nicht normalverteilt waren.

Im WiSe21/22 ergaben sich auch alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass sie nicht normalverteilt waren.

Tabelle 6 Statistische Beschreibung für RP2

		SoSe21				WiSe21/22			
		Obs	MW	SD	t	Obs	MW	SD	t
DOZ	GLT	59	12,47	1,54	0,00	57	12,11	1,71	0,00
	GED	58	1,48	0,63	0,00	57	1,63	0,88	0,00
STU	GLT	216	12,39	1,70	0,00	120	12,12	1,60	0,00
	GED	212	1,46	0,56	0,00	117	1,62	0,68	0,00
SP	GPT	57	6,88	2,84	0,00	59	7,16	3,21	0,00

	GNT	50	1,76	0,77	0,00	58	1,90	0,85	0,00

Bei RP3 (Informationsvermittlung) im SoSe21 (Tabelle 7) ergaben sich alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass die Daten nicht normalverteilt waren.

Im WiSe21/22 ergaben sich, außer GLT aus DOZ und GNT aus SP, alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass sie nicht normalverteilt waren. Die GLT aus DOZ sowie GNT aus SP waren normalverteilt.

Tabelle 7 Statistische Beschreibung für RP3

		SoSe21				WiSe21/22			
		Obs	MW	SD	t	Obs	MW	SD	t
DOZ	GLT	57	12,01	1,61	0,00	63	11,57	1,71	0,38
	GED	55	1,62	0,67	0,01	62	1,76	0,69	0,00
STU	GLT	212	12,36	1,51	0,00	130	11,92	1,59	0,00
	GED	204	1,47	0,56	0,00	127	1,58	0,62	0,00
SP	GPT	56	6,18	2,43	0,00	63	6,75	2,67	0,00
	GNT	52	1,60	0,66	0,01	61	1,84	0,71	0,87

Bei RP4 (Partizipative Entscheidung) im SoSe21 (Tabelle 8) ergaben sich alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass die Daten nicht normalverteilt waren.

Im WiSe21/22 ergaben sich, außer GED aus STU, alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass sie nicht normalverteilt waren. Die GED aus STU war normalverteilt.

Tabelle 8 Statistische Beschreibung für RP4

		SoSe21				WiSe21/22			
		Obs	MW	SD	t	Obs	MW	SD	t
DOZ	GLT	57	13,02	2,10	0,00	45	13,42	2,01	0,00
	GED	57	1,63	0,67	0,00	44	1,84	0,93	0,01
STU	GLT	231	13,46	1,84	0,00	67	13,49	1,75	0,00
	GED	224	1,38	0,55	0,00	67	1,64	0,73	0,06
SP	GPT	58	6,39	3,07	0,00	51	7,06	3,11	0,00
	GNT	58	1,93	0,90	0,00	53	1,91	0,81	0,01

Bei RP5 (Compliance) im SoSe21 (Tabelle 9) ergaben sich alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass die Daten nicht normalverteilt waren.

Im WiSe21/22 ergaben sich alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass sie nicht normalverteilt waren.

Bei RP6 (Motivationales Interview) im SoSe21 (Tabelle 10) ergaben sich, außer GLT aus DOZ, alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass die Daten nicht normalverteilt waren. Die GLT aus DOZ war normalverteilt.

Im WiSe21/22 ergaben sich, außer GLT aus DOZ, alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass sie nicht normalverteilt waren. Die GLT aus DOZ war normalverteilt.

Tabelle 9 Statistische Beschreibung für RP5

		SoSe21				WiSe21/22			
		Obs	MW	SD	t	Obs	MW	SD	t
DOZ	GLT	58	11,90	1,76	0,00	43	11,88	1,81	0,02
	GED	58	1,7	0,71	0,00	43	1,84	0,84	0,00
STU	GLT	247	12,41	1,53	0,00	86	11,91	1,81	0,00
	GED	233	1,42	0,53	0,00	86	1,67	0,71	0,00
SP	GPT	63	6,24	2,47	0,00	55	7,51	3,40	0,00
	GNT	61	1,72	0,90	0,00	55	1,96	0,92	0,00

Tabelle 10 Statistische Beschreibung für RP6

		SoSe21				WiSe21/22			
		Obs	MW	SD	t	Obs	MW	SD	t
DOZ	GLT	55	10,96	1,88	0,67	51	11,36	1,69	0,99
	GED	55	2,20	0,90	0,02	51	1,80	0,78	0,01
STU	GLT	221	12,44	1,52	0,00	115	12,37	1,74	0,00
	GED	208	1,45	0,57	0,00	115	1,45	0,65	0,00
SP	GPT	58	6,83	3,22	0,00	61	6,53	2,89	0,00
	GNT	57	2,12	1,18	0,00	53	1,77	0,80	0,00

Bei RP7 (Stressreaktion) im SoSe21 (Tabelle 11) ergaben sich alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass die Daten nicht normalverteilt waren.

Im WiSe21/22 ergaben sich, außer GED aus DOZ, alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass sie nicht normalverteilt waren. Die GED aus DOZ war normalverteilt.

Tabelle 11 Statistische Beschreibung für RP7

		SoSe21				WiSe21/22			
		Obs	MW	SD	t	Obs	MW	SD	t
DOZ	GLT	38	12,15	1,38	0,03	42	12	2,03	0,00
	GED	38	1,63	0,85	0,00	42	1,78	0,77	0,40
STU	GLT	163	12,69	1,35	0,00	85	12,58	1,32	0,00
	GED	152	1,44	0,61	0,00	85	1,58	0,62	0,00
SP	GPT	41	6,73	3,54	0,00	42	6,05	2,50	0,00
	GNT	37	1,92	1,09	0,00	42	1,53	0,63	0,00

Bei RP8 (Krebsaufklärung) im SoSe21 (Tabelle 12) ergaben sich alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass die Daten nicht normalverteilt waren.

Im WiSe21/22 ergaben sich, außer GED aus DOZ, alle t-Werte für die Normalverteilungsprüfung unter 0,05, sodass es angenommen werden konnte, dass sie nicht normalverteilt waren. Die GED aus DOZ war normalverteilt.

Tabelle 12 Statistische Beschreibung für RP8

		SoSe21				WiSe21/22			
		Obs	MW	SD	t	Obs	MW	SD	t
DOZ	GLT	36	12,43	1,25	0,00	63	11,57	1,71	0,00
	GED	35	1,57	0,60	0,00	62	1,76	0,69	0,34
STU	GLT	155	12,82	1,42	0,00	130	11,92	1,59	0,00
	GED	151	1,43	0,61	0,00	127	1,58	0,62	0,01
SP	GPT	40	6,68	3,10	0,00	63	6,75	2,67	0,00
	GNT	37	1,81	0,85	0,00	61	1,84	0,71	0,00

Abgesehen von einzelnen Ausnahmen waren unsere Datensätze nicht normalverteilt. Dementsprechend wurden die Analysemethoden ausgewählt.

5.1.2. Homogenität beider Stichproben

Im SoSe21 und im WiSe21/22 nahmen dieselben Simulationspatienten und -innen am RP teil. Sie waren nach dem Zufallsprinzip, aber auch berücksichtigt in persönlicher Einsetzbarkeit, z.B. Austausch der Gruppe aufgrund von COVID-19 Infektion, den jeweiligen Gruppen zugeteilt. So waren die Simulationspatienten und -innen in beiden Veranstaltungen identisch, also die Stichproben sind homogen.

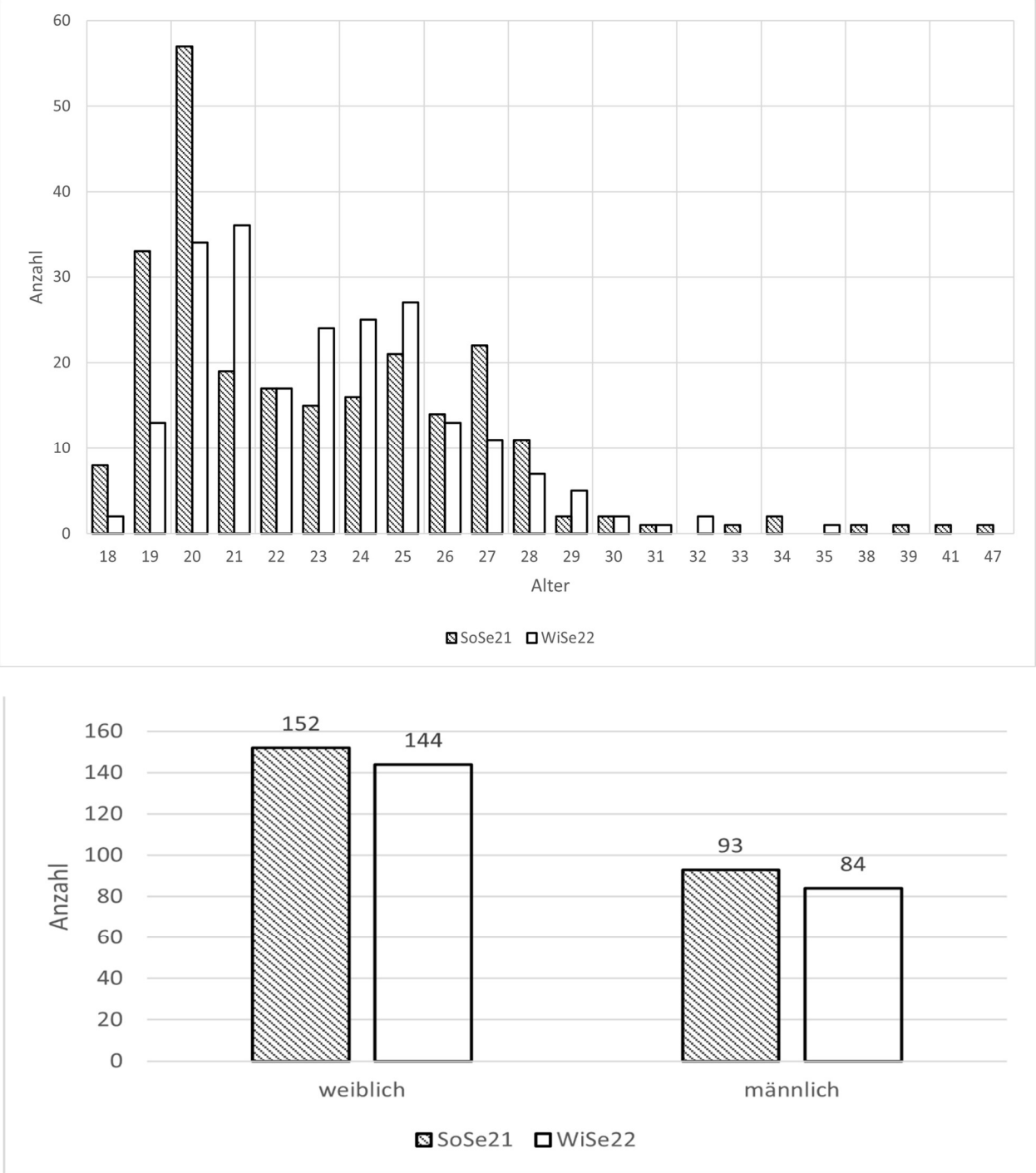


Abbildung 1 Homogenitätsprüfung der STU. Alter hat unter t-Test den p-Wert von 0,62 > 0,05, Geschlecht hat unter χ^2-Test den p-Wert von 0,63 > 0,05.

Insgesamt n = 10 Dozierende waren als Prüfende am RP beteiligt. Davon waren vier in beiden Semestern tätig, während die anderen sechs nur in einem Semester eingesetzt wurden. So hatten wir jeweils n = 7 Dozierende als Prüfende für das RP in den jeweiligen Semestern. Alle n = 10 Dozierenden hatten einen Diplom- oder Masterabschluss in Psychologie. Daher waren die beiden Stichproben hinsichtlich ihres Bildungsstatus homogen. Da die Größe der Stichproben so kleine war, wurden die statistischen Homogenitätsprüfungen nicht durchgeführt.

Anders als Dozierenden sowie Simulationspatienten und -innen, sind Studierenden in den beiden Semestern nicht identisch. Es gab keine Studierenden, die an beiden Veranstaltungen teilnahmen. Daher müssen die Studierendengruppen im SoSe21 und im WiSe21/22 geprüft werden, ob die beiden Stichproben statistisch homogen sind.

Die Studierenden im SoSe21 und WiSe21/22 wurden mit ihrem Geschlecht und Alter verglichen.

Zur Analyse für das Geschlecht wurde Chi-Quadrat-Test (χ^2-Test) angewendet. Der χ^2 Wert ergibt sich 0,235, ist mit Freiheitsgrad 1 sein p-Wert 0,63 > 0,05, sodass keinen signifikanten Unterschied beider Stichproben besteht. Der Freiheitsgrad ist 1, da das Geschlecht der Studierenden aus ihren Stammdaten erhoben wurde. Diese Daten enthielten lediglich die Angaben "männlich" oder "weiblich", ohne die Option "diverse".

Zweistichproben t-Test mit unterschiedlichen Varianzen gab für das Alter von beiden Stichproben einen zweiseitigen p-Wert 0,62 > 0,05 ab. Somit besteht im Alter auch keinen signifikanten Unterschied.

Der Bildungsstand der Studierenden war alle im zweiten Semester der Vorklinik in Medizinstudium, somit ist es identisch. Zur medizinischen Vorbildung wurden keine Daten erhoben.

Aus diesen Analysen kann davon ausgegangen werden, dass die Stichproben STU im SoSe21 und WiSe21/22 statistisch homogen sind.

5.2. Ergebnisse für Frage 1: Testtheoretische Qualität der Checkliste bei RP für Arzt-Patient-Kommunikation

Um die testtheoretische Qualität der in dieser Studie eingesetzten Checkliste zu prüfen, wurden verschiedene Analysen durchgeführt. Die Ergebnisse zur Reliabilität, Objektivität, Validität, AS und Trennschärfe werden nach RP sortiert vorgestellt.

5.2.1. Reliabilität

Mit Spearman-Korrelation wurden Rangkorrelationskoeffizienten zwischen GLT und GED ermittelt, um die Reliabilität der Checkliste zu prüfen. Während GLT mit guter Leistung hohe Punkte bis auf 14 gegeben wurde, wurde GED nach Schulnotensystem bewertet, bei dem eine niedrige Zahl die bessere Leistung darstellt. Deshalb ergab sich der Rangkorrelationskoeffizient negativ.

Die Simulationspatienten und -innen bearbeiteten einen anderen Fragebogen als Checkliste. Mit der testtheoretischen Qualität der Checkliste hat dieser Fragebogen im engeren Sinne keinen direkten Zusammenhang. Dennoch wurden hier GPT und GNT aus ihrer Bewertung analysiert, denn GNT sollte mit GED aus DOZ verglichen werden, um die Validität zu prüfen. Beide wurden nach dem Schulnotensystem bewertet, sodass sich der Rangkorrelationskoeffizient positiv ergab.

5.2.1.1. Reliabilität bei RP1 (Anamnese)

Beim RP1 (Tabelle 5) im SoSe21 ließen sich gute Korrelationen zwischen GLT und GED beobachten. Die 69 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,53, die 250 aus STU hatten das Ergebnis als -0,53. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “ausreichend” korrelieren.

Die GPT und GNT aus SP standen auch in einem starken Zusammenhang. Aus 64 Bewertungen war der Rangkorrelationskoeffizient 0,90, der p-Wert 0,00.

Im WiSe21/22 ließen sich auch gute Korrelationen zwischen GLT und GED beobachten. Die 58 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,43, die 121 aus STU hatten das Ergebnis als -0,40. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “ausreichend” korrelieren.

Die GPT und GNT aus SP standen auch in einem starken Zusammenhang. Aus 61 Bewertungen war der Rangkorrelationskoeffizient 0,82, der p-Wert 0,00.

Tabelle 13 Korrelationen der GLT mit GED sowie der GPT mit GNT (Spearman-Korrelationen) RP1

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	Rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ	GLT, GED	69	-0.5300	0.0000	58	-0.4304	0.0007
STU	GLT, GED	250	-0.5275	0.0000	121	-0.3945	0.0000
SP	GPT, GNT	64	0.8971	0.0000	61	0.8175	0.0000

5.2.1.2. Reliabilität bei RP2 (Gesprächsförderung)

Beim RP2 (Tabelle 6) im SoSe21 ließen sich gute Korrelationen zwischen GLT und GED beobachten. Die 58 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,62, die 212 aus STU hatten das Ergebnis als -0,59. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “mittelmäßig” korrelieren.

Die GPT und GNT aus SP standen auch in einem starken Zusammenhang. Aus 50 Bewertungen war der Rangkorrelationskoeffizient 0,87, der p-Wert 0,00.

Im WiSe21/22 ließen sich auch gute Korrelationen zwischen GLT und GED beobachten. Die 57 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,59, die 117 aus STU hatten das Ergebnis als -0,42. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau „mittelmäßig“ korrelieren.

Die GPT und GNT aus SP standen auch in einem starken Zusammenhang. Aus 58 Bewertungen war der Rangkorrelationskoeffizient 0,89, der p-Wert 0,00.

Tabelle 14 Korrelationen der GLT mit GED sowie der GPT mit GNT (Spearman-Korrelationen) RP2

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	Rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ	GLT, GED	58	-0.6165	0.0000	57	-0.5887	0.0000
STU	GLT, GED	212	-0.5918	0.0000	117	-0.4236	0.0000
SP	GPT, GNT	50	0.8694	0.0000	58	0.8866	0.0000

Tabelle 15 Korrelationen der GLT mit GED sowie der GPT mit GNT (Spearman-Korrelationen) RP3

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	Rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ	GLT, GED	55	-0.7404	0.0000	62	-0.6660	0.0000
STU	GLT, GED	205	-0.5347	0.0000	127	-0.4979	0.0000
SP	GPT, GNT	52	0.9261	0.0000	61	0.8536	0.0000

5.2.1.3. Reliabilität bei RP3 (Informationsvermittlung)

Beim RP3 (Tabelle 7) im SoSe21 ließen sich gute Korrelationen zwischen GLT und GED beobachten. Die 55 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,74, die 205 aus STU hatten das Ergebnis als -0,54. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “mittelmäßig” korrelieren.

Die GPT und GNT aus SP standen in einem fast perfekten Zusammenhang. Aus 52 Bewertungen war der Rangkorrelationskoeffizient 0,93, der p-Wert 0,00.

Im WiSe21/22 ließen sich auch gute Korrelationen zwischen GLT und GED beobachten. Die 62 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,67, die 127 aus STU hatten das

Ergebnis als -0,50. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “mittelmäßig” korrelieren.

Die GPT und GNT aus SP standen auch in einem starken Zusammenhang. Aus 61 Bewertungen war der Rangkorrelationskoeffizient 0,85, der p-Wert 0,00.

5.2.1.4. Reliabilität bei RP4 (Partizipative Entscheidung)

Beim RP4 (Tabelle 8) im SoSe21 ließen sich gute Korrelationen zwischen GLT und GED beobachten. Die 57 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,72, die 224 aus STU hatten das Ergebnis als -0,60. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “mittelmäßig” korrelieren.

Die GPT und GNT aus SP standen auch in einem starken Zusammenhang. Aus 58 Bewertungen war der Rangkorrelationskoeffizient 0,86, der p-Wert 0,00.

Im WiSe21/22 ließen sich auch gute Korrelationen zwischen GLT und GED beobachten. Die 44 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,65, die 67 aus STU hatten das Ergebnis als -0,64. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “mittelmäßig” korrelieren.

Die GPT und GNT aus SP standen auch in einem starken Zusammenhang. Aus 51 Bewertungen war der Rangkorrelationskoeffizient 0,87, der p-Wert 0,00.

Tabelle 16 Korrelationen der GLT mit GED sowie der GPT mit GNT (Spearman-Korrelationen) RP4

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	Rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ	GLT, GED	57	-0.7205	0.0000	44	-0.6456	0.0000
STU	GLT, GED	224	-0.6033	0.0000	67	-0.6403	0.0000
SP	GPT, GNT	58	0.8613	0.0000	51	0.8671	0.0000

5.2.1.5. Reliabilität bei RP5 (Compliance)

Beim RP5 (Tabelle 9) im SoSe21 ließen sich gute Korrelationen zwischen GLT und GED beobachten. Die 58 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,81, die 233 aus STU hatten das Ergebnis als -0,55. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “beachtlich“ bzw. “ausreichend” korrelieren.

Die GPT und GNT aus DOZ standen auch in einem starken Zusammenhang. Aus 61 Bewertungen war der Rangkorrelationskoeffizient 0,70, der p-Wert 0,00.

Im WiSe21/22 ließen sich auch gute Korrelationen zwischen GLT und GED beobachten. Die 43 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,59, die 81 aus STU hatten das Ergebnis als -0,48. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “ausreichend” korrelieren.

Die GPT und GNT aus DOZ standen auch in einem fast perfekten Zusammenhang. Aus 55 Bewertungen war der Rangkorrelationskoeffizient 0,92, der p-Wert 0,00.

Tabelle 17 Korrelationen der GLT mit GED sowie der GPT mit GNT (Spearman-Korrelationen) RP5

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	Rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ	GLT, GED	58	-0.8093	0.0000	43	-0.5851	0.0000
STU	GLT, GED	233	-0.5514	0.0000	81	-0.4825	0.0000
SP	GPT, GNT	61	0.6997	0.0000	55	0.9224	0.0000

5.2.1.6. Reliabilität bei RP6 (Motivationales Interview)

Beim RP6 (Tabelle 10) im SoSe21 ließen sich gute Korrelationen zwischen GLT und GED beobachten. Die 55 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,87, die 220 aus STU hatten das Ergebnis als -0,63. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “beachtlich“ bzw. “mittelmäßig” korrelieren.

Die GPT und GNT aus DOZ standen auch in einem moderaten Zusammenhang. Aus 57 Bewertungen war der Rangkorrelationskoeffizient 0,72, der p-Wert 0,00.

Im WiSe21/22 ließen sich auch gute Korrelationen zwischen GLT und GED beobachten. Die 51 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,56, die 115 aus STU hatten das Ergebnis als -0,52. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “ausreichend” korrelieren.

Die GPT und GNT aus DOZ standen auch in einem fast perfekten Zusammenhang. Aus 53 Bewertungen war der Rangkorrelationskoeffizient 0,94, der p-Wert 0,00.

Tabelle 18 Korrelationen der GLT mit GED sowie der GPT mit GNT (Spearman-Korrelationen) RP6

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	Rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ	GLT, GED	55	-0.8650	0.0000	51	-0.5644	0.0000
STU	GLT, GED	220	-0.6312	0.0000	115	-0.5154	0.0000
SP	GPT, GNT	57	0.7158	0.0000	53	0.9392	0.0000

5.2.1.7. Reliabilität bei RP7 (Stressreaktion)

Beim RP7 (Tabelle 11) im SoSe21 ließen sich gute Korrelationen zwischen GLT und GED beobachten. Die 38 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,68, die 153 aus STU hatten das Ergebnis als -0,54. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “mittelmäßig” korrelieren.

Die GPT und GNT aus DOZ standen auch in einem starken Zusammenhang. Aus 37 Bewertungen war der Rangkorrelationskoeffizient 0,87, der p-Wert 0,00.

Im WiSe21/22 ließen sich auch gute Korrelationen zwischen GLT und GED beobachten. Die 42 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,64, die 85 aus STU hatten das Ergebnis als -0,39. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “mittelmäßig” korrelieren.

Die GPT und GNT aus DOZ standen auch in einem starken Zusammenhang. Aus 42 Bewertungen war der Rangkorrelationskoeffizient 0,81, der p-Wert 0,00.

Tabelle 19 Korrelationen der GLT mit GED sowie der GPT mit GNT (Spearman-Korrelationen) RP7

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	Rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ	GLT, GED	38	-0.6766	0.0000	42	-0.6406	0.0000
STU	GLT, GED	153	-0.5405	0.0000	85	-0.3898	0.0002
SP	GPT, GNT	37	0.8710	0.0000	42	0.8046	0.0000

5.2.1.8. Reliabilität bei RP8 (Krebsaufklärung)

Beim RP8 (Tabelle 12) im SoSe21 ließen sich gute Korrelationen zwischen GLT und GED beobachten. Die 36 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,57, die 151 aus STU hatten das Ergebnis als -0,71. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “ausreichend” bzw. “mittelmäßig” korrelieren.

Die GPT und GNT aus DOZ standen auch in einem fast perfekten Zusammenhang. Aus 37 Bewertungen war der Rangkorrelationskoeffizient 0,93, der p-Wert 0,00.

Im WiSe21/22 ließen sich auch gute Korrelationen zwischen GLT und GED beobachten. Die 44 Bewertungen aus DOZ zeigten den Rangkorrelationskoeffizienten -0,64, die 81 aus STU hatten das Ergebnis als -0,61. Die beiden Werte hatten p-Wert von 0,00 < 0,05, sodass GLT und GED auf dem 5% Signifikanzniveau “mittelmäßig” korrelieren.

Die GPT und GNT aus DOZ standen auch in einem fast perfekten Zusammenhang. Aus 44 Bewertungen war der Rangkorrelationskoeffizient 0,98, der p-Wert 0,00.

Tabelle 20 Korrelationen der GLT mit GED sowie der GPT mit GNT (Spearman-Korrelationen) RP8

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	Rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ	GLT, GED	36	-0.5652	0.0003	44	-0.6421	0.0000
STU	GLT, GED	151	-0.7105	0.0000	81	-0.6097	0.0002
SP	GPT, GNT	37	0.9314	0.0000	44	0.9794	0.0000

Insgesamt zeigten bei allen Stichproben gute Korrelationen der Bewertungen und die Leistungen und den Gesamteindruck. Während die Fragen für GLT in der Checkliste im Laufe des RPs angekreuzt wurden, wurde GED nach Ende des Gesprächs abgegeben. Die beiden Bewertungen waren somit zeitlich versetzt, trotzdem korrelierten sie recht hoch miteinander.

5.2.2. Objektivität

Durch Conger´s Kappa wurden die Checkliste aus STU bzw. DOZ analysiert. Beim einzelnen RP waren maximal sechs Bewertende einschließlich einer oder eines Dozierenden an der Bewertung betätigt, die Kappa Analyse überprüfte ihre Übereinstimmung.

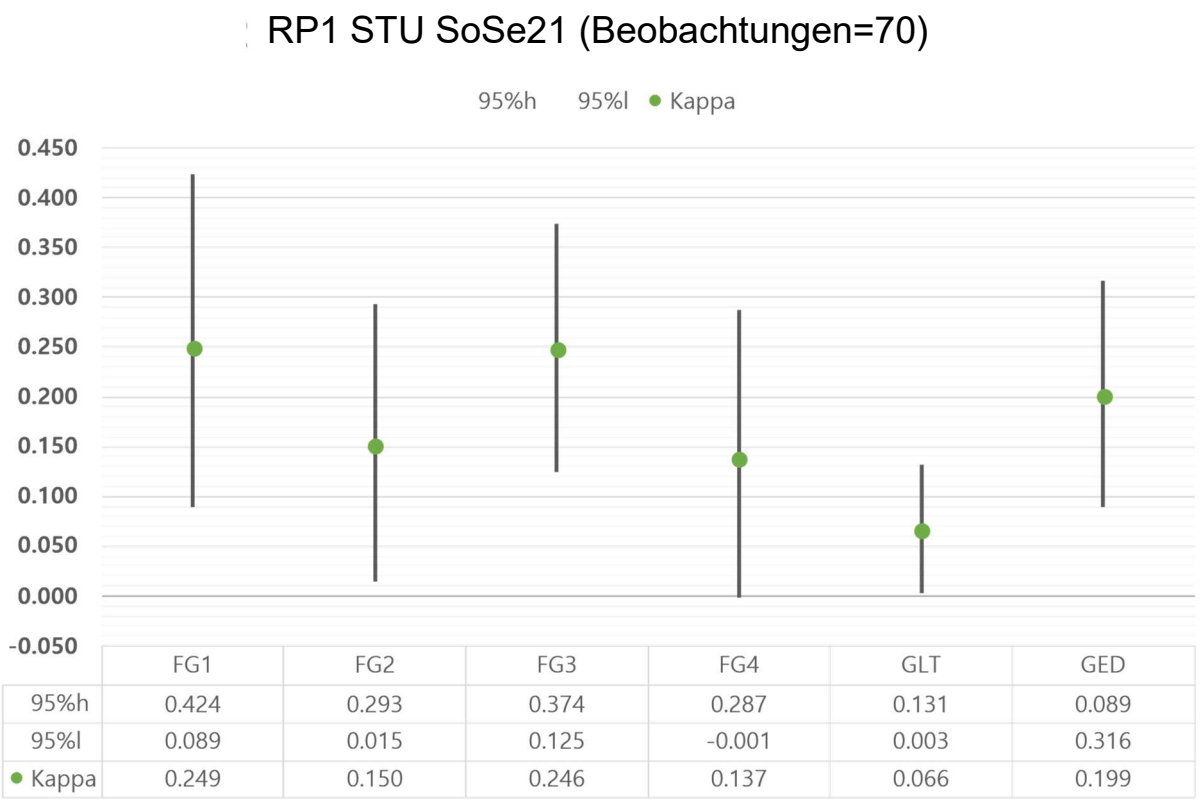
Zunächst wurden die Bewertungen von STU analysiert. Bei jedem Gespräch gaben zwei bis fünf Studierende ihre Checkliste-Bewertungen ab. Die GLT, GED und die Bewertungen jeder FG wurden zwischen den Ratern verglichen.

Anschließend wurde das gleiche Verfahren diesmal mit den Bewertungen von STU und DOZ zusammen durchgeführt.

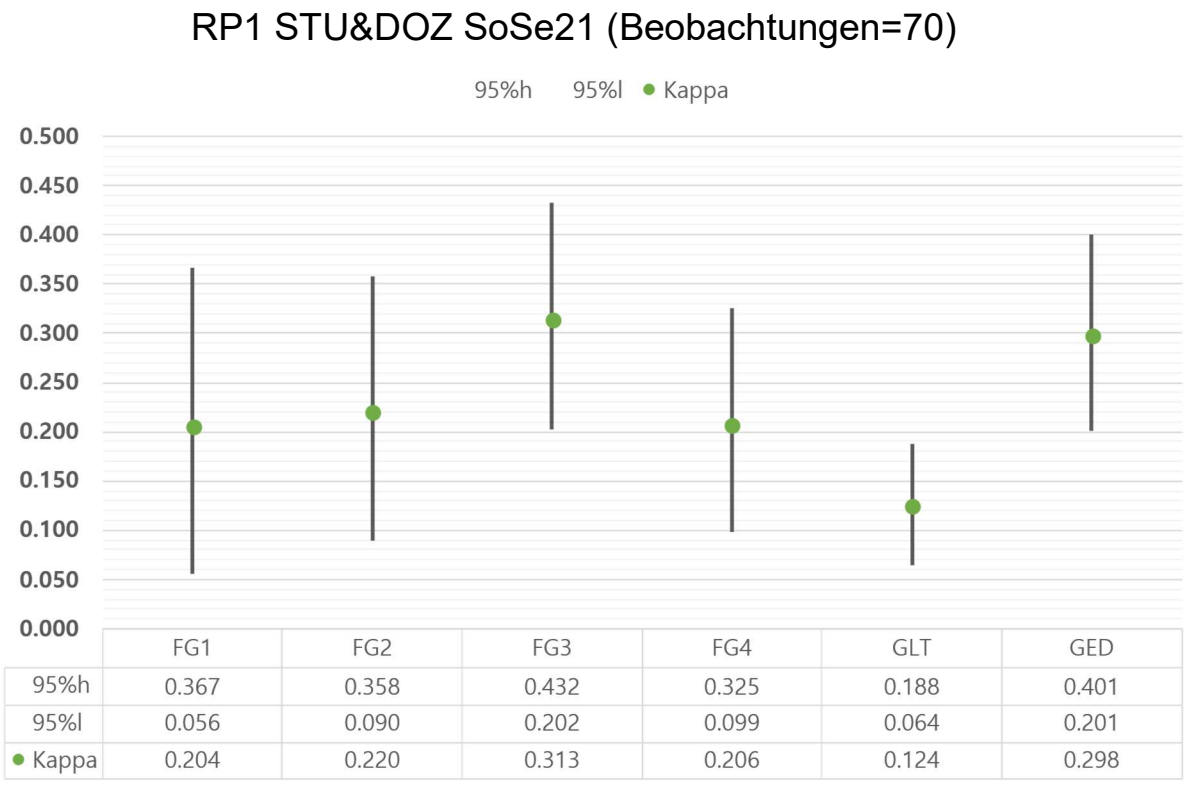
Die Ergebnisse von Conger´s Kappa lieferten den Kappa-Koeffizienten und dessen 95% Konfidenzintervall.

5.2.2.1. Objektivität bei RP1 (Anamnese)

Beim RP1 im SoSe21 (Abbildung 2) wiesen die Kappa-Koeffizienten geringe bis mittlere Übereinstimmungen zwischen den Ratern auf.

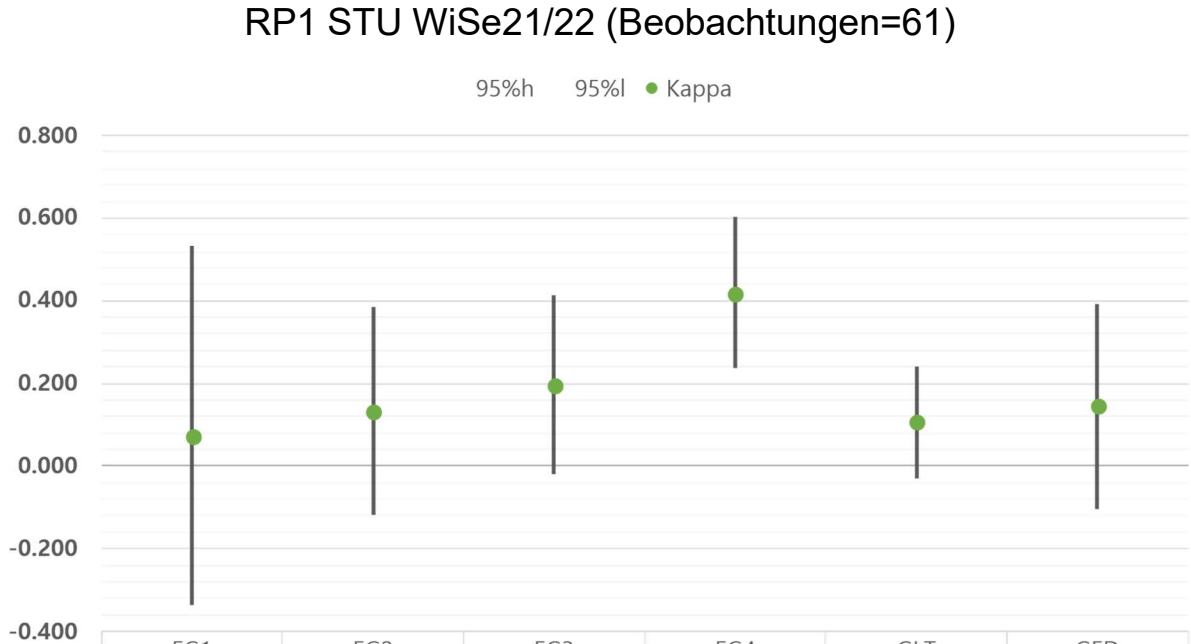


	FG1	FG2	FG3	FG4	GLT	GED
95%h	0.424	0.293	0.374	0.287	0.131	0.089
95%l	0.089	0.015	0.125	-0.001	0.003	0.316
Kappa	0.249	0.150	0.246	0.137	0.066	0.199

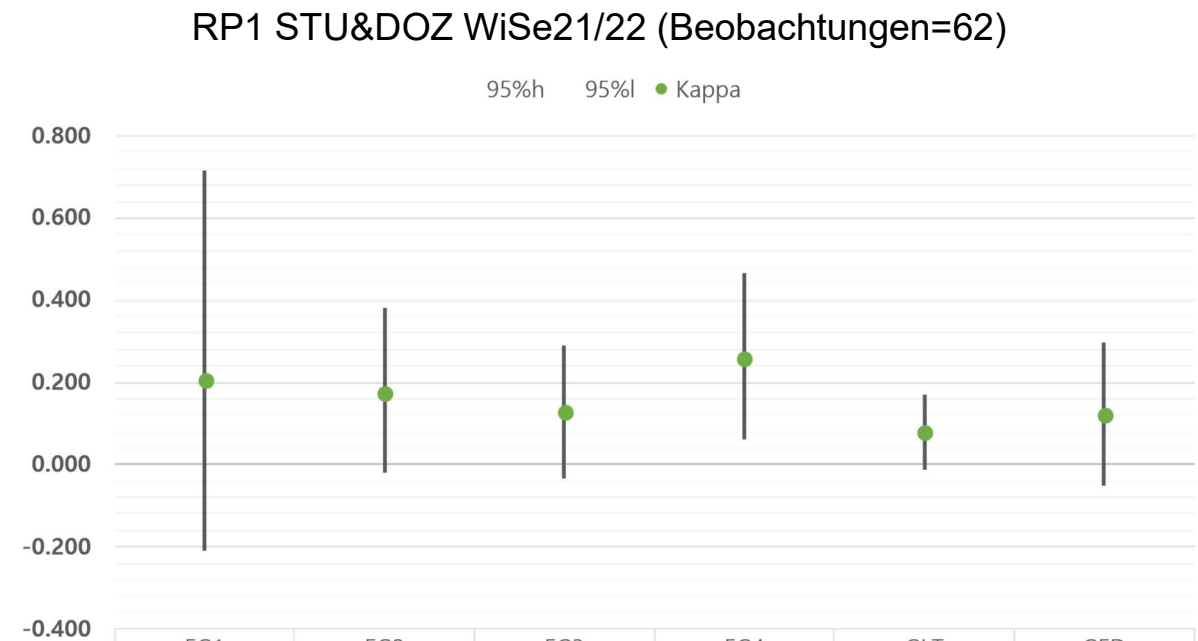


	FG1	FG2	FG3	FG4	GLT	GED
95%h	0.367	0.358	0.432	0.325	0.188	0.401
95%l	0.056	0.090	0.202	0.099	0.064	0.201
Kappa	0.204	0.220	0.313	0.206	0.124	0.298

Abbildung 2 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP1 SoSe21



	FG1	FG2	FG3	FG4	GLT	GED
95%h	0.533	0.385	0.413	0.602	0.242	0.392
95%l	-0.337	-0.119	-0.020	0.237	-0.031	-0.104
Kappa	0.070	0.129	0.193	0.416	0.105	0.142



	FG1	FG2	FG3	FG4	GLT	GED
95%h	0.717	0.381	0.290	0.467	0.169	0.295
95%l	-0.210	-0.021	-0.034	0.060	-0.014	-0.052
Kappa	0.205	0.173	0.125	0.257	0.076	0.119

Abbildung 3 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP1 WiSe21/22

Der Kappa-Koeffizient für GLT zwischen STU war 0,07, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,20, es besteht auch eine “geringe“ Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,25, für FG2 (Exploration der jetzigen Symptome) 0,15, für FG3 (Ergänzende Anamnesen) 0,25 und für FG4 (Qualität der Gesprächsführung) 0,14. Die FG1 und FG3 zeigten die “ausreichenden” Übereinstimmungen, und die FG2 und FG4 die „leichten“, wobei aber für FG4 das Konfidenzintervall in den negativen Bereich hineinragte, sodass man keine eindeutige Interpretation treffen kann.

Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab. Die Werte waren eher tendenziell leicht höher als die von STU, obwohl ein zusätzlicher Bewerter hinzukam.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,12, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,30, es besteht eine „mittelmäßige“ Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,20, für FG2 (Exploration der jetzigen Symptome) 0,22, für FG3 (Ergänzende Anamnesen) 0,31 und für FG4 (Qualität der Gesprächsführung) 0,21. Alle vier FG zeigten mittelmäßige Übereinstimmungen, die Konfidenzintervalle blieben alle im positiven Bereich.

Im WiSe21/22 (Abbildung 3) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “mittelhohen” Übereinstimmungen zwischen den Ratern auf.

Der Kappa-Koeffizient für GLT zwischen STU war 0,11, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,14, es besteht auch eine “geringe“ Übereinstimmung. Aber die Konfidenzintervalle für die beiden ragten in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,07, für FG2 (Exploration der jetzigen Symptome) 0,13, für FG3 (Ergänzende Anamnesen) 0,19 und für FG4 (Qualität der Gesprächsführung) 0,42. Die FG1, FG2 und FG3 zeigten “geringe“ Übereinstimmungen, und die FG4 eine „moderate“. Aber für FG1, FG2 und FG3 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

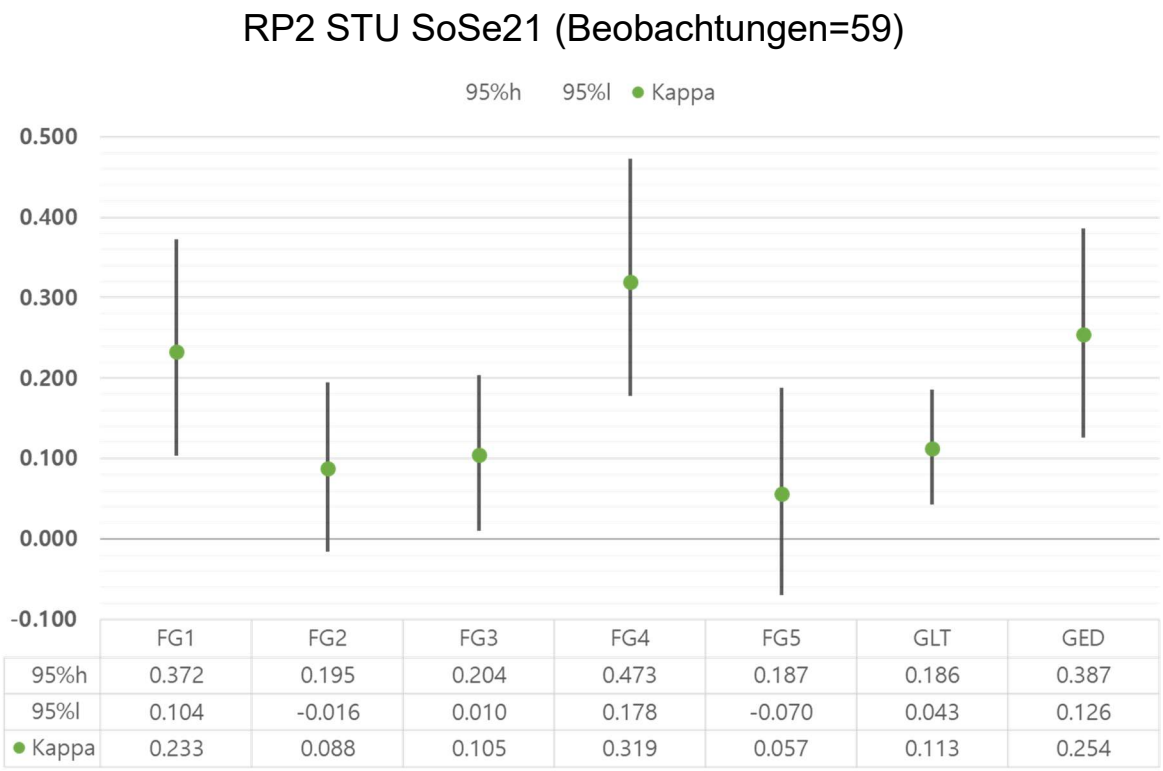
Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,08, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,12, es besteht auch eine “geringe“ Übereinstimmung. Aber das Konfidenzintervalle für die beiden ragten in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

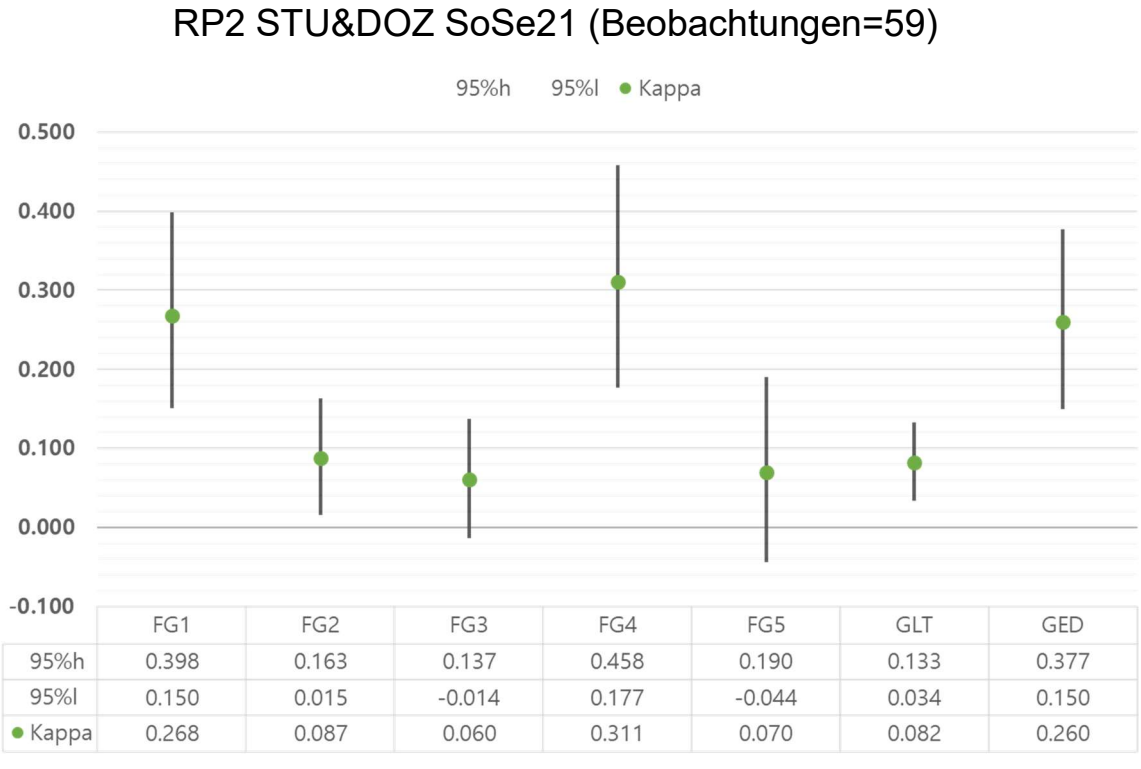
Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,21, für FG2 (Exploration der jetzigen Symptome) 0,17, für FG3 (Ergänzende Anamnesen) 0,13 und für FG4 (Qualität der Gesprächsführung) 0,26. Die FG1, FG4 zeigten die “ausreichenden” Übereinstimmungen, und die FG2 und FG3 die „leichten“. Aber für FG1, FG2 und FG3 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

5.2.2.2. Objektivität bei RP2 (Gesprächsförderung)

Beim RP2 im SoSe21 (Abbildung 4) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “ausreichenden” Übereinstimmungen zwischen den Ratern auf.



	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.372	0.195	0.204	0.473	0.187	0.186	0.387
95%l	0.104	-0.016	0.010	0.178	-0.070	0.043	0.126
Kappa	0.233	0.088	0.105	0.319	0.057	0.113	0.254



	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.398	0.163	0.137	0.458	0.190	0.133	0.377
95%l	0.150	0.015	-0.014	0.177	-0.044	0.034	0.150
Kappa	0.268	0.087	0.060	0.311	0.070	0.082	0.260

Abbildung 4 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP2 SoSe21

Der Kappa-Koeffizient für GLT zwischen STU war 0,11, es besteht einer „leichte“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,25, es besteht eine “ausreichenden” Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,23, für FG2 (Welche Gesprächsförderer haben Sie beobachtet?) 0,09, für FG3 (Waren Gesprächsstörer zu beobachten?) 0,11, für FG4 (Aufrollen der Hintergrundproblematik) 0,32 und für FG5 (Qualität der Gesprächsführung) 0,06. Die FG1 und FG4 zeigten die “ausreichenden” Übereinstimmungen, und die FG2, FG3 und FG5 die “geringen“. Aber für FG2 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,08, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,26, es besteht eine “ausreichende” Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,27, für FG2 (Welche Gesprächsförderer haben Sie beobachtet?) 0,09, für FG3 (Waren Gesprächsstörer zu beobachten?) 0,06, für FG4 (Aufrollen der Hintergrundproblematik) 0,31 und für FG5 (Qualität der Gesprächsführung) 0,07. Die FG1 und FG4 zeigten die “ausreichenden” Übereinstimmungen, und die FG2, FG3 und FG5 die “geringen“. Aber für FG2 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

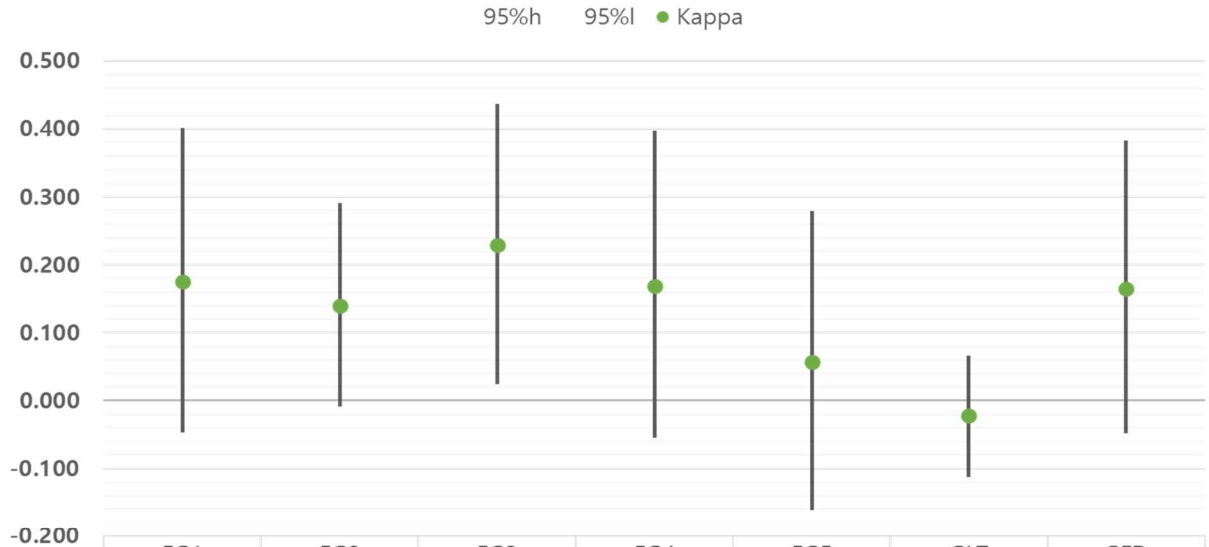
Im WiSe21/22 (Abbildung 5) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “ausreichenden” Übereinstimmungen zwischen den Ratern auf.

Der Kappa-Koeffizient für GLT zwischen STU war -0,02, es besteht eine „geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,17, es besteht ebenso eine “geringe“ Übereinstimmung. Aber die Konfidenzintervalle für die beiden ragten in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,18, für FG2 (Welche Gesprächsförderer haben Sie beobachtet?) 0,14, für FG3 (Waren Gesprächsstörer zu beobachten?) 0,23, für FG4 (Aufrollen der Hintergrundproblematik) 0,17 und für FG5 (Qualität der Gesprächsführung) 0,06. Die FG1, FG2, FG4 und FG5 zeigten die „leichten“ Übereinstimmungen, und die FG3 eine “ausreichende”. Aber für FG1, FG2, FG4 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

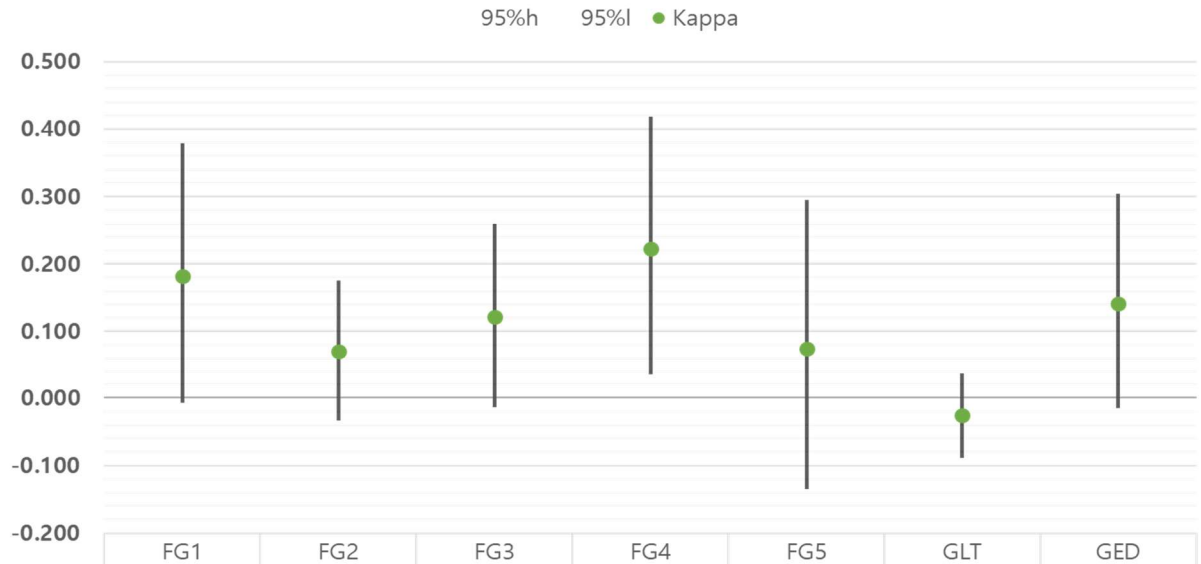
Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab.

RP2 STU WiSe21/22 (Beobachtungen=59)



	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.401	0.291	0.437	0.398	0.279	0.067	0.384
95%l	-0.047	-0.009	0.024	-0.055	-0.160	-0.112	-0.048
Kappa	0.176	0.140	0.230	0.169	0.057	-0.022	0.165

RP2 STU&DOZ WiSe21/22 (Beobachtungen=61)



	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.379	0.176	0.260	0.419	0.295	0.036	0.304
95%l	-0.008	-0.033	-0.014	0.034	-0.135	-0.089	-0.016
Kappa	0.182	0.071	0.122	0.223	0.075	-0.026	0.141

Abbildung 5 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP2 WiSe21/22

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war -0,03, es besteht eine „geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,14, es besteht eine “geringe“ Übereinstimmung.

Aber die Konfidenzintervalle für die beiden ragten in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann. Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,18, für FG2 (Welche Gesprächsförderer haben Sie beobachtet?) 0,07, für FG3 (Waren Gesprächsstörer zu beobachten?) 0,12, für FG4 (Aufrollen der Hintergrundproblematik) 0,23 und für FG5 (Qualität der Gesprächsführung) 0,08. Die FG1, FG2, FG3 und FG5 zeigten die „leichten“ Übereinstimmungen, und die FG4 eine “ausreichende”, Aber für FG1, FG2, FG3 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

5.2.2.3. Objektivität bei RP3 (Informationsvermittlung)

Beim RP3 im SoSe21 (Abbildung 6) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “geringen“ Übereinstimmungen zwischen den Ratern auf.

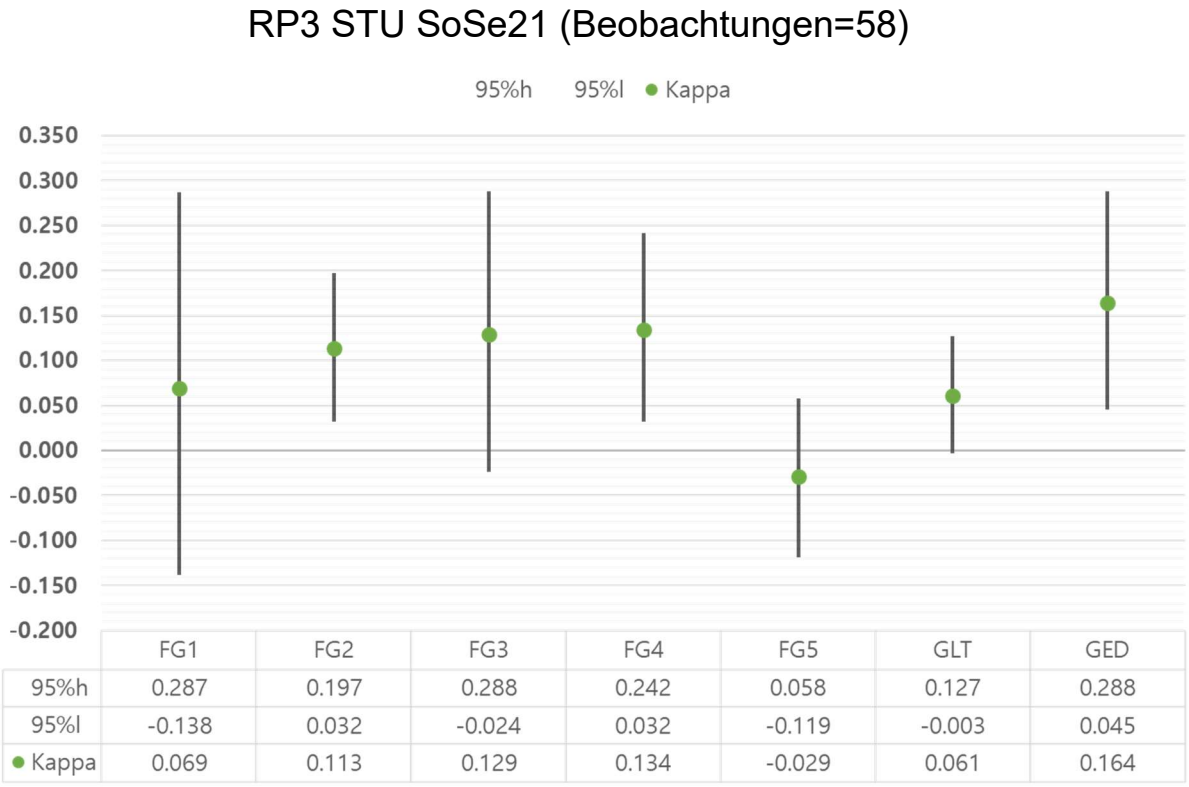
Der Kappa-Koeffizient für GLT zwischen STU war 0,06, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,16, es besteht auch eine “geringe“ Übereinstimmung. Das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,07, für FG2 (Vorbereitung) 0,11, für FG3 (Verständliche Vermittlung der Information, Gesagtes) 0,13, für FG4 (Rückversicherung) 0,13 und für FG5 (Qualität der Gesprächsführung) -0,03. Die FG1, FG2, FG3 und FG4 zeigten die “geringen“ Übereinstimmungen, und die FG5 eine „geringe“. Aber für FG1, FG3 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

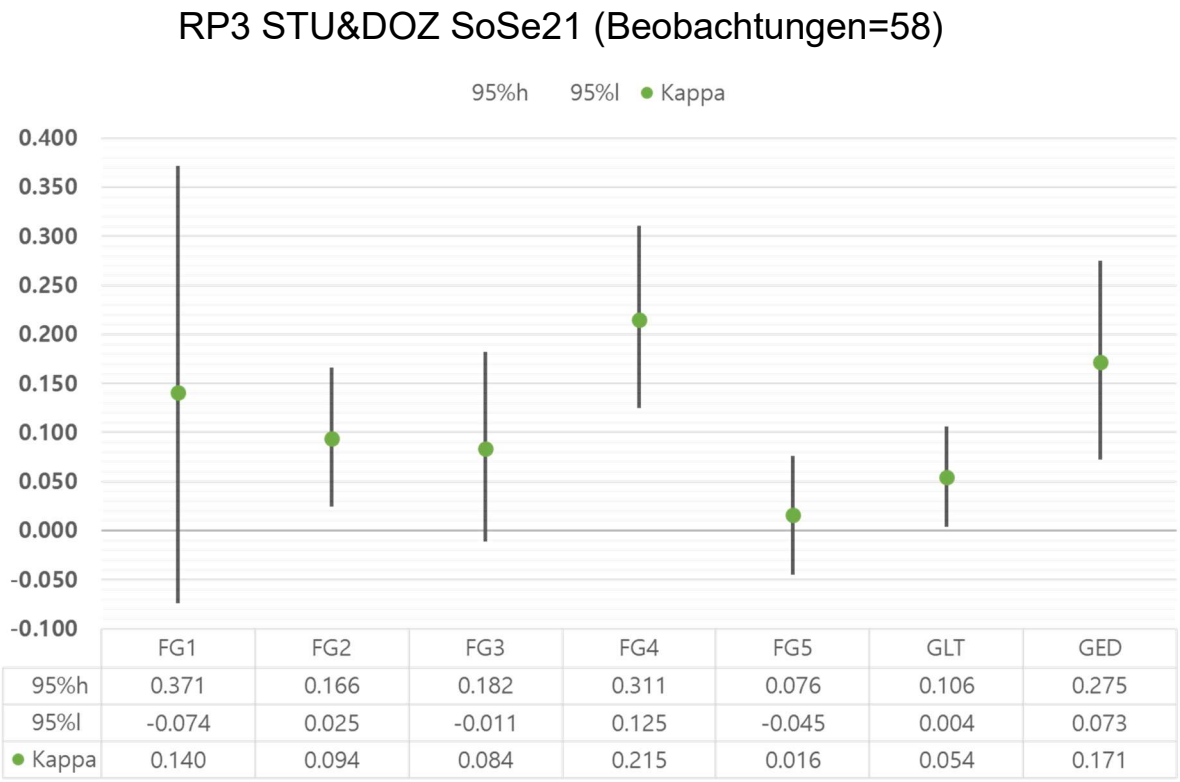
Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab. Die Werte waren eher tendenziell leicht höher als die von STU, obwohl ein zusätzlicher Bewerter hinzukam.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,05, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,17, es besteht auch eine “geringe“ Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,14, für FG2 (Vorbereitung) 0,09, für FG3 (Verständliche Vermittlung der Information, Gesagtes) 0,08, für FG4 (Rückversicherung) 0,22 und für FG5 (Qualität der Gesprächsführung) 0,02. Alle zeigten die “geringen“ Übereinstimmungen. Aber für FG1, FG3 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

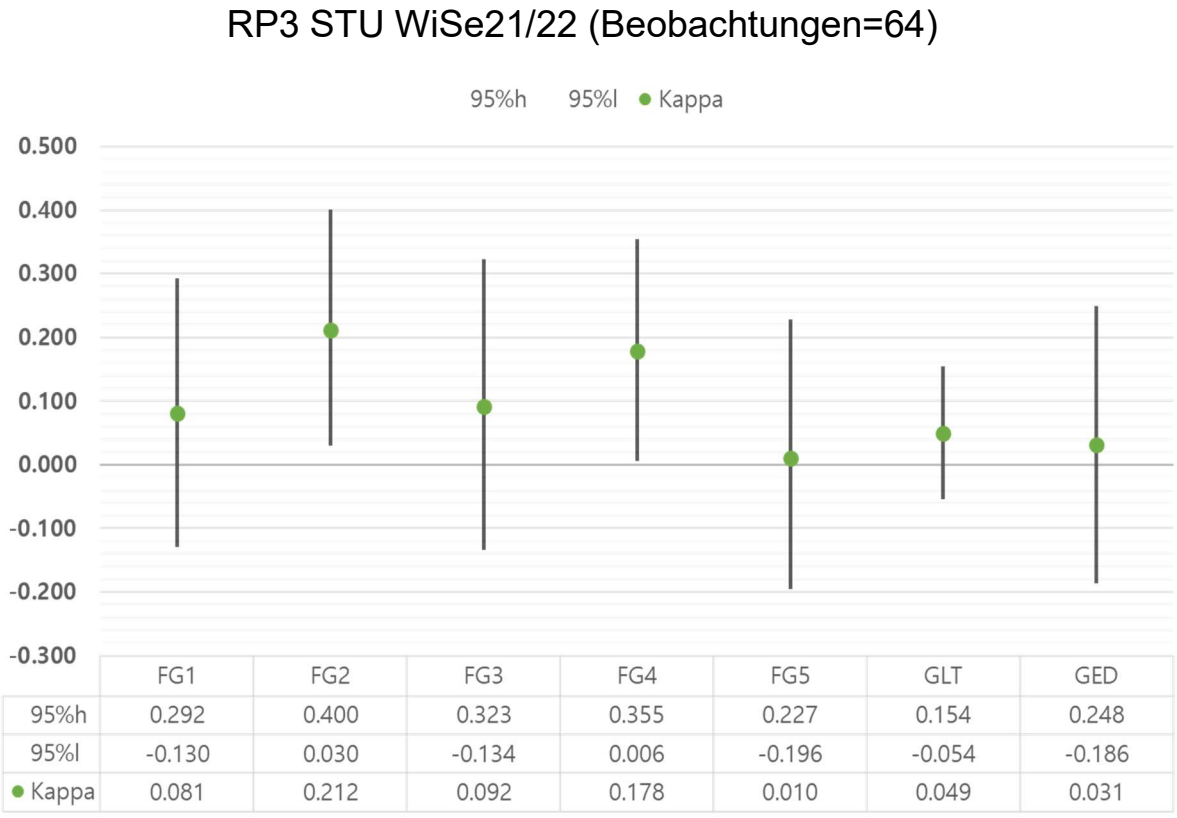


	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.287	0.197	0.288	0.242	0.058	0.127	0.288
95%l	-0.138	0.032	-0.024	0.032	-0.119	-0.003	0.045
Kappa	0.069	0.113	0.129	0.134	-0.029	0.061	0.164

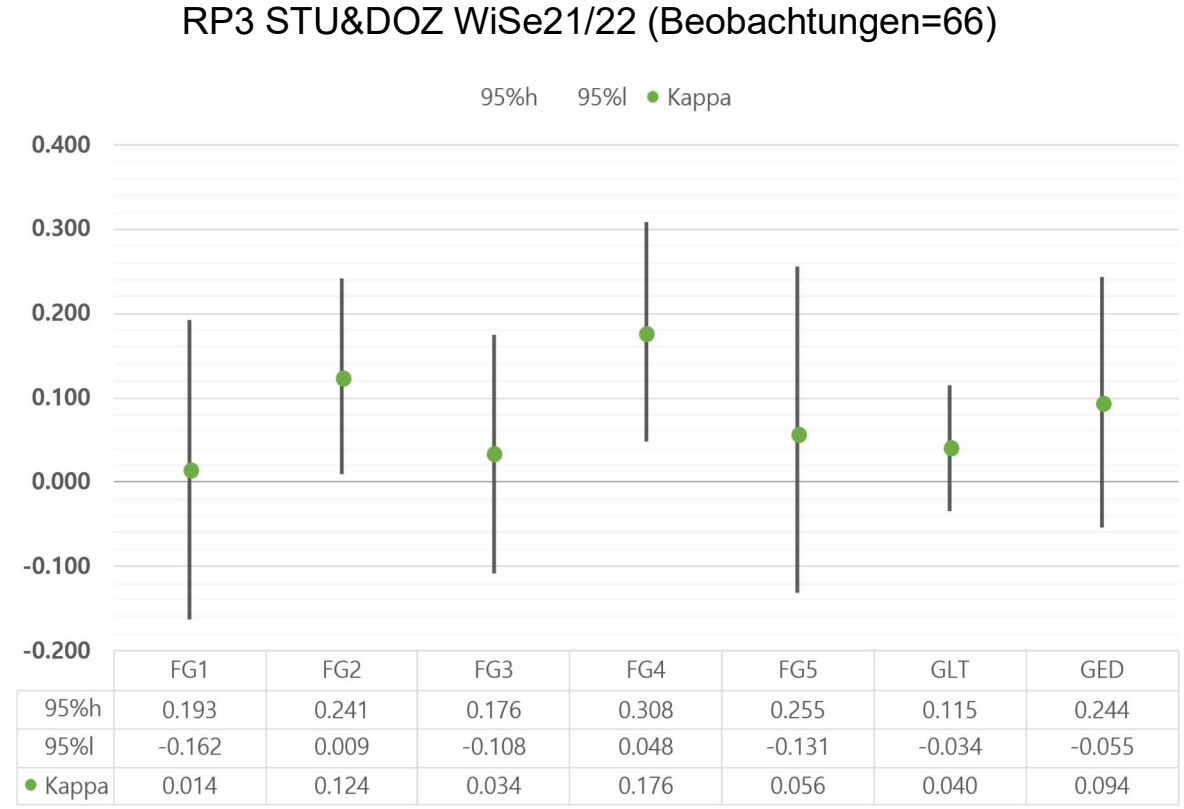


	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.371	0.166	0.182	0.311	0.076	0.106	0.275
95%l	-0.074	0.025	-0.011	0.125	-0.045	0.004	0.073
Kappa	0.140	0.094	0.084	0.215	0.016	0.054	0.171

Abbildung 6 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP3 SoSe21



	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.292	0.400	0.323	0.355	0.227	0.154	0.248
95%l	-0.130	0.030	-0.134	0.006	-0.196	-0.054	-0.186
Kappa	0.081	0.212	0.092	0.178	0.010	0.049	0.031



	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.193	0.241	0.176	0.308	0.255	0.115	0.244
95%l	-0.162	0.009	-0.108	0.048	-0.131	-0.034	-0.055
Kappa	0.014	0.124	0.034	0.176	0.056	0.040	0.094

Abbildung 7 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP3 WiSe21/22

Im WiSe21/22 (Abbildung 7) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “ausreichenden” Übereinstimmungen zwischen den Ratern auf.

Der Kappa-Koeffizient für GLT zwischen STU war 0,05, es besteht eine “geringen“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,03, es besteht auch eine “geringe“ Übereinstimmung. Aber die Konfidenzintervalle für die beiden ragten in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,08, für FG2 (Vorbereitung) 0,21, für FG3 (Verständliche Vermittlung der Information, Gesagtes) 0,09, für FG4 (Rückversicherung) 0,18 und für FG5 (Qualität der Gesprächsführung) 0,01. Die FG1, FG3, FG4 und FG5 zeigten die “geringen“ Übereinstimmungen, und die FG2 eine “ausreichende”. Aber für FG1, FG3 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,04, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,09, es besteht auch eine “geringe“ Übereinstimmung. Aber die Konfidenzintervalle für die beiden ragten in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,01, für FG2 (Vorbereitung) 0,12, für FG3 (Verständliche Vermittlung der Information, Gesagtes) 0,03, für FG4 (Rückversicherung) 0,18 und für FG5 (Qualität der Gesprächsführung) 0,06. Alle zeigten die “geringen“ Übereinstimmungen. Aber für FG1, FG3 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

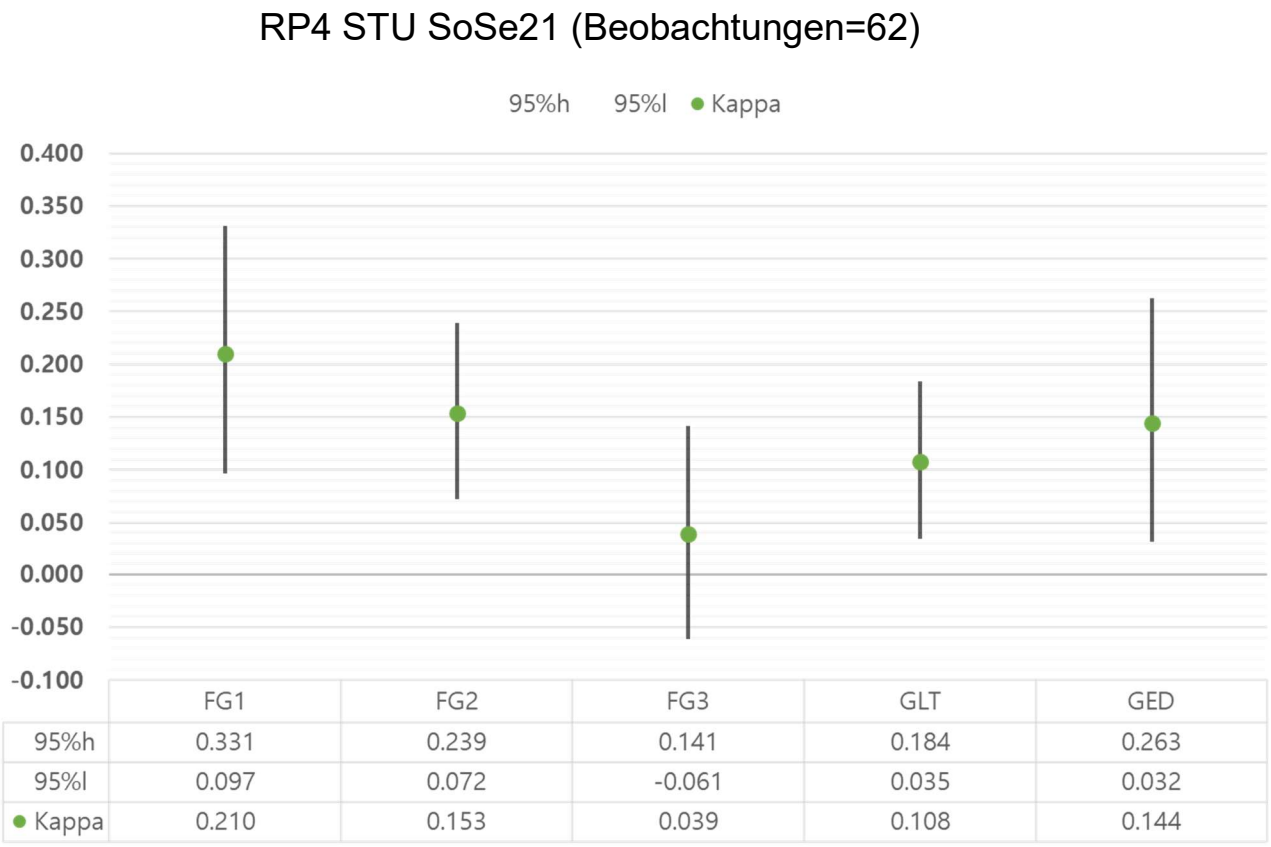
5.2.2.4. Objektivität bei RP4 (Partizipative Entscheidung)

Beim RP4 im SoSe21 (Abbildung 8) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “ausreichenden” Übereinstimmungen zwischen den Ratern auf.

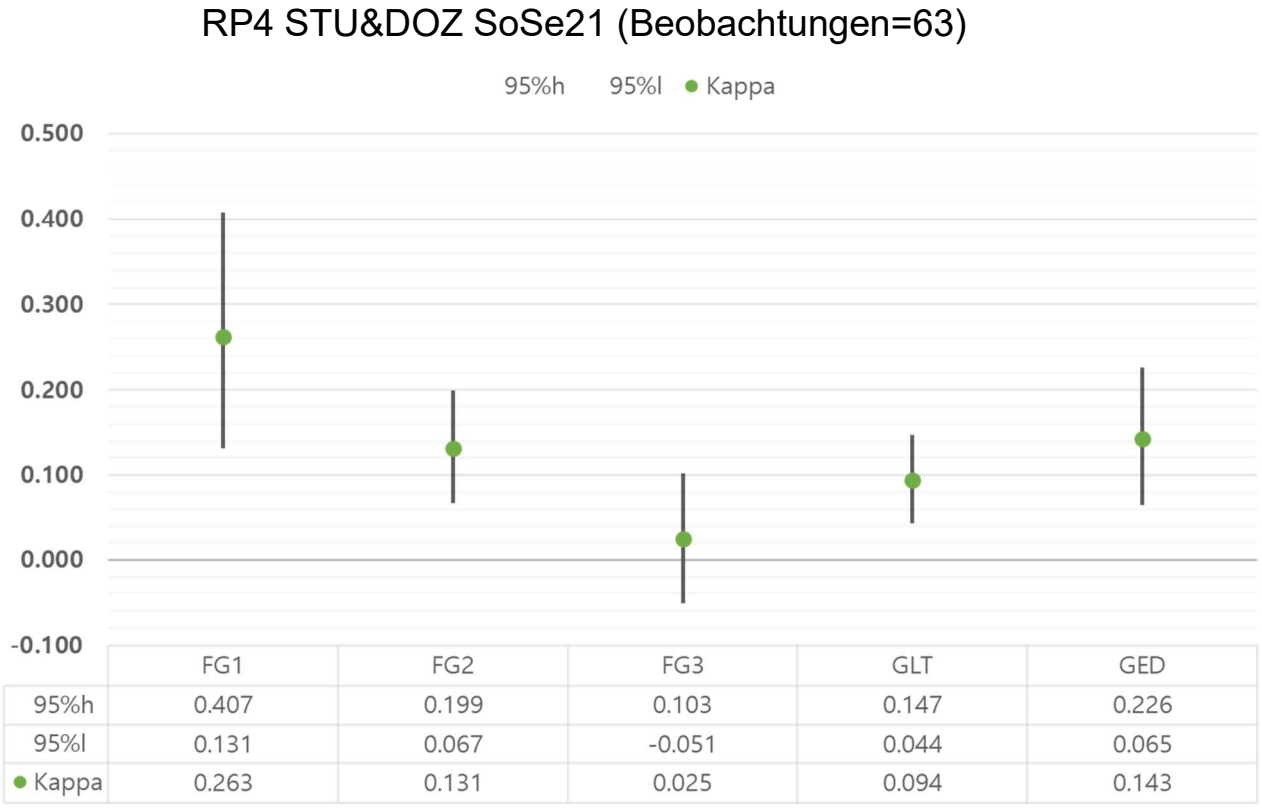
Der Kappa-Koeffizient für GLT zwischen STU war 0,11, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,14, es besteht auch eine “geringe“ Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,21, für FG2 (Schritte der partizipativen Entscheidungsfindung) 0,15 und für FG3 (Qualität der Gesprächsführung) 0,04. Die FG1 zeigte eine “ausreichende” Übereinstimmung, und die FG2 und FG3 die „leichten“. Aber für FG3 ragte das Konfidenzintervall in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab.

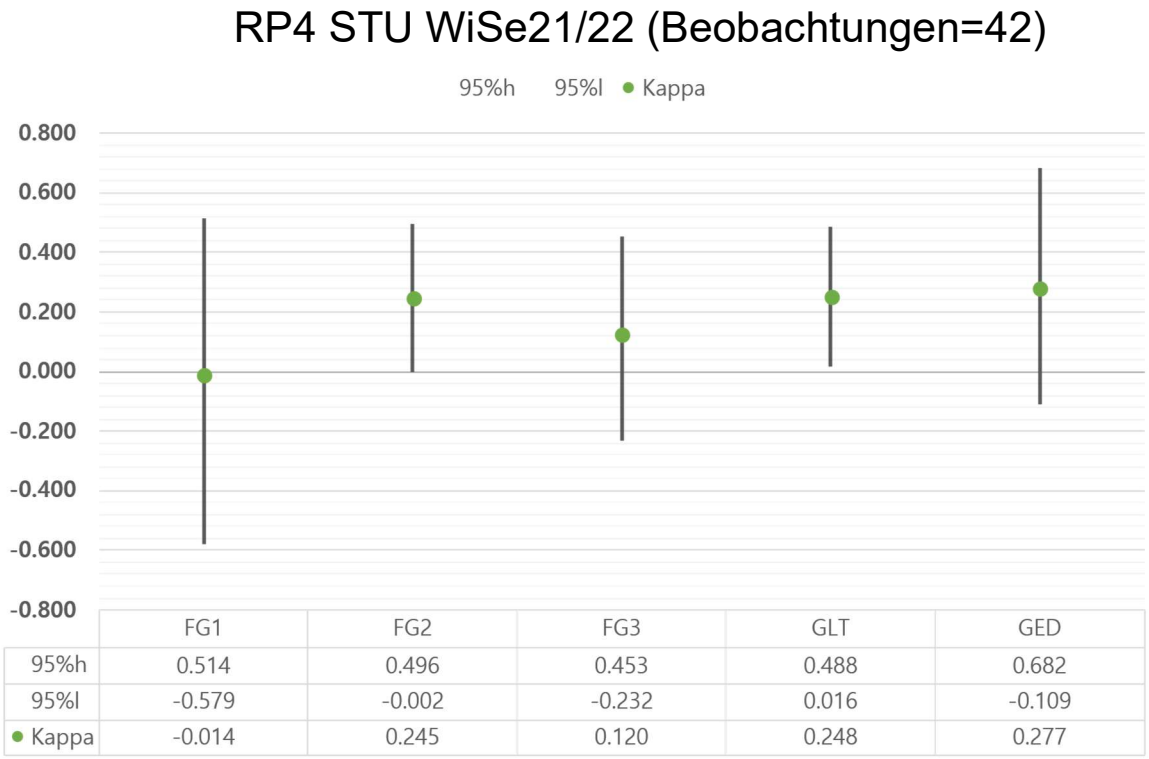


	FG1	FG2	FG3	GLT	GED
95%h	0.331	0.239	0.141	0.184	0.263
95%l	0.097	0.072	-0.061	0.035	0.032
Kappa	0.210	0.153	0.039	0.108	0.144

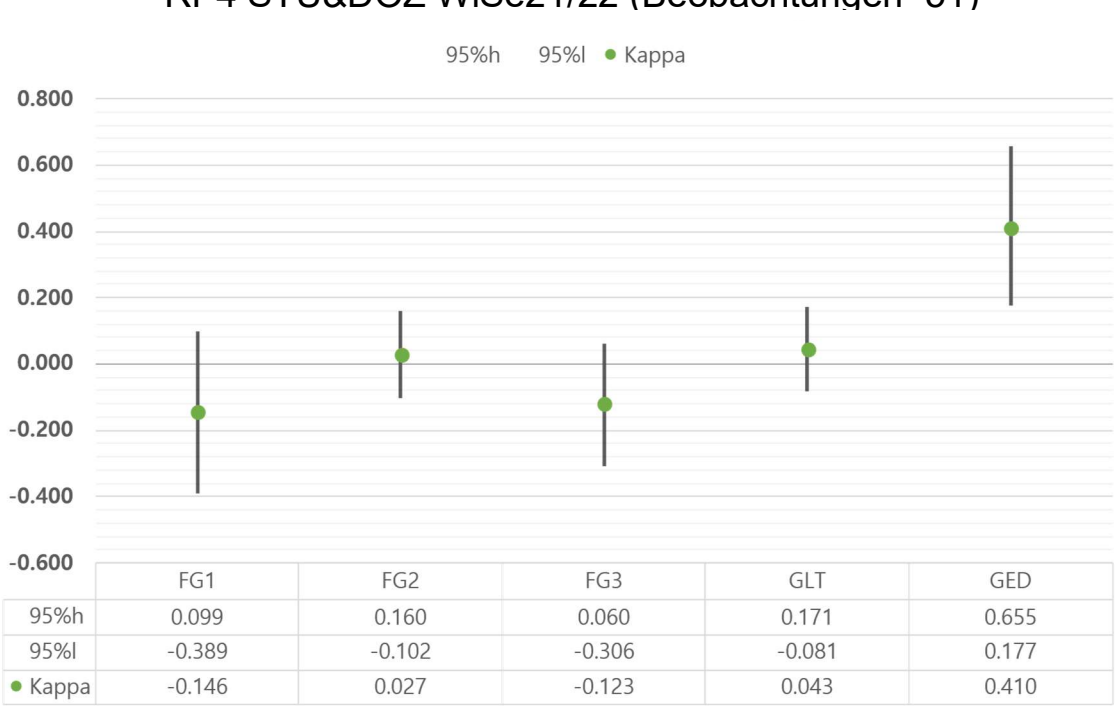


	FG1	FG2	FG3	GLT	GED
95%h	0.407	0.199	0.103	0.147	0.226
95%l	0.131	0.067	-0.051	0.044	0.065
Kappa	0.263	0.131	0.025	0.094	0.143

Abbildung 8 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP4 SoSe21



	FG1	FG2	FG3	GLT	GED
95%h	0.514	0.496	0.453	0.488	0.682
95%l	-0.579	-0.002	-0.232	0.016	-0.109
Kappa	-0.014	0.245	0.120	0.248	0.277



	FG1	FG2	FG3	GLT	GED
95%h	0.099	0.160	0.060	0.171	0.655
95%l	-0.389	-0.102	-0.306	-0.081	0.177
Kappa	-0.146	0.027	-0.123	0.043	0.410

Abbildung 9 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP4 WiSe21/22

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,09, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,14, es besteht auch eine “geringe“ Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,26, für FG2 (Schritte der partizipativen Entscheidungsfindung) 0,13 und für FG3 (Qualität der Gesprächsführung) 0,03. Die FG1 zeigte eine “ausreichende” Übereinstimmung, und die FG2 und FG3 die „leichten“. Aber für FG3 ragte das Konfidenzintervall in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Im WiSe21/22 (Abbildung 9) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “mittelmäßigen” Übereinstimmungen zwischen den Ratern auf.

Der Kappa-Koeffizient für GLT zwischen STU war 0,25, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,28, es besteht auch eine “geringe“ Übereinstimmung. Aber das Konfidenzintervall für GED ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) -0,01, für FG2 (Schritte der partizipativen Entscheidungsfindung) 0,25 und für FG3 (Qualität der Gesprächsführung) 0,12. Die FG1 zeigte eine „geringe“ Übereinstimmung, die FG2 eine “mittelmäßige” und FG3 eine “geringe“. Aber die Konfidenzintervalle für alle FG ragten in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

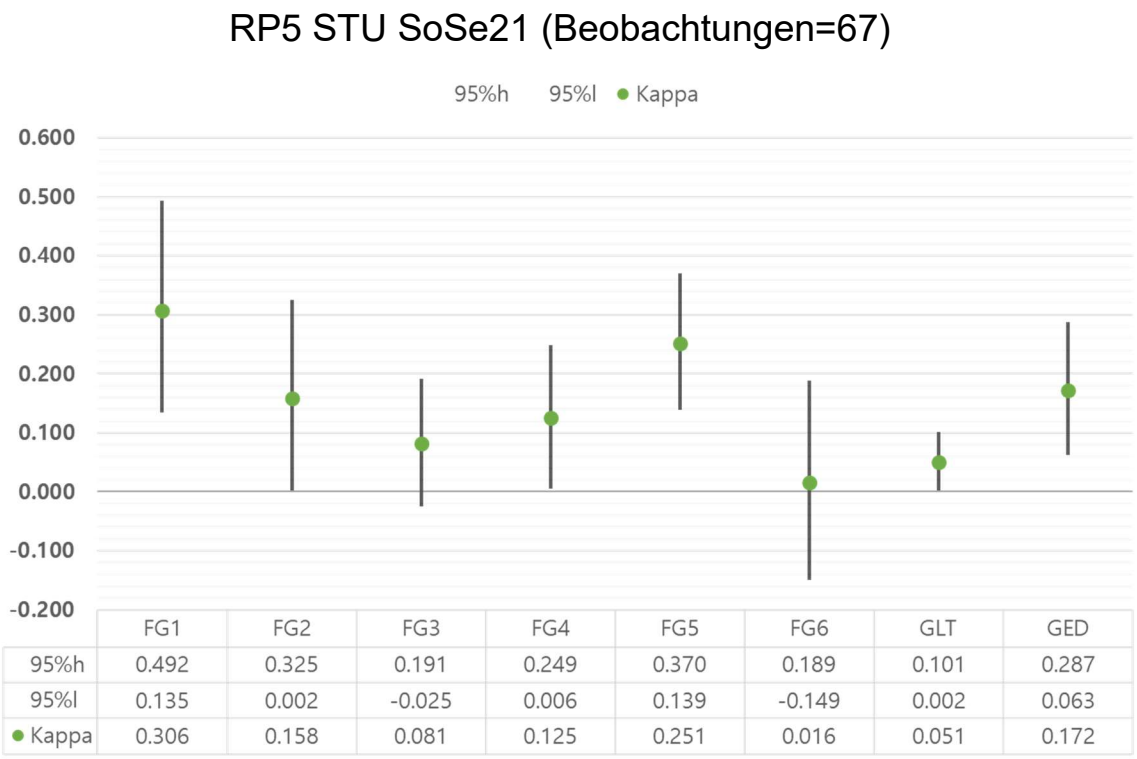
Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,04, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,41, es besteht eine “mittelmäßige” Übereinstimmung. Aber das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

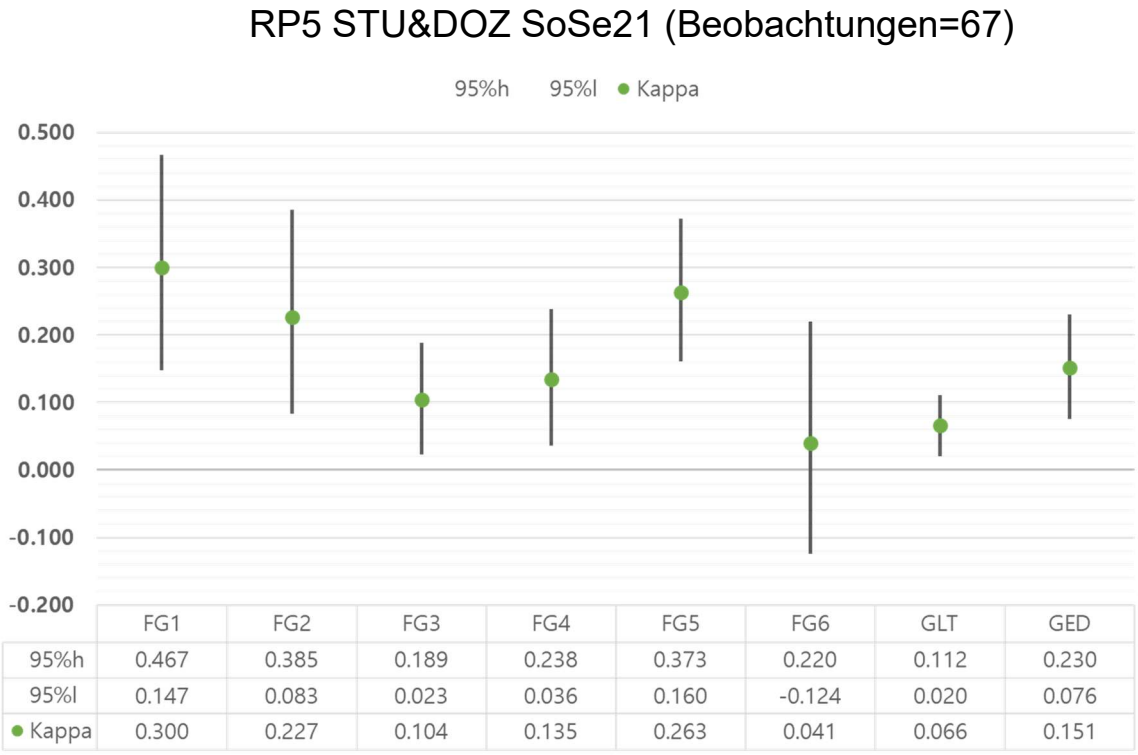
Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) -0,15, für FG2 (Schritte der partizipativen Entscheidungsfindung) 0,03 und für FG3 (Qualität der Gesprächsführung) -0,12. Die FG1 und FG3 zeigten die „geringen“ Übereinstimmungen, und die FG2 eine “geringen“. Aber die Konfidenzintervalle für alle FG ragten in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

5.2.2.5. Objektivität bei RP5 (Compliance)

Beim RP5 im SoSe21 (Abbildung 10) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “ausreichenden” Übereinstimmungen zwischen den Ratern auf.

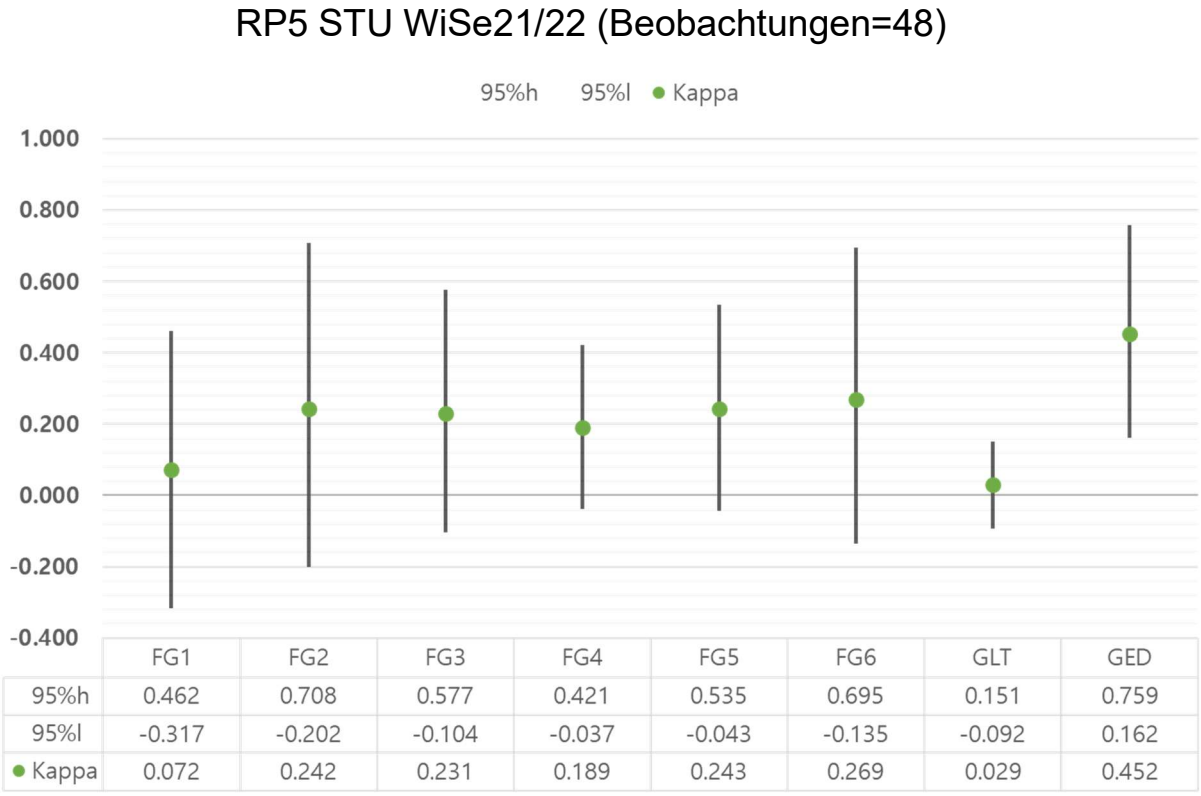


	FG1	FG2	FG3	FG4	FG5	FG6	GLT	GED
95%h	0.492	0.325	0.191	0.249	0.370	0.189	0.101	0.287
95%l	0.135	0.002	-0.025	0.006	0.139	-0.149	0.002	0.063
Kappa	0.306	0.158	0.081	0.125	0.251	0.016	0.051	0.172

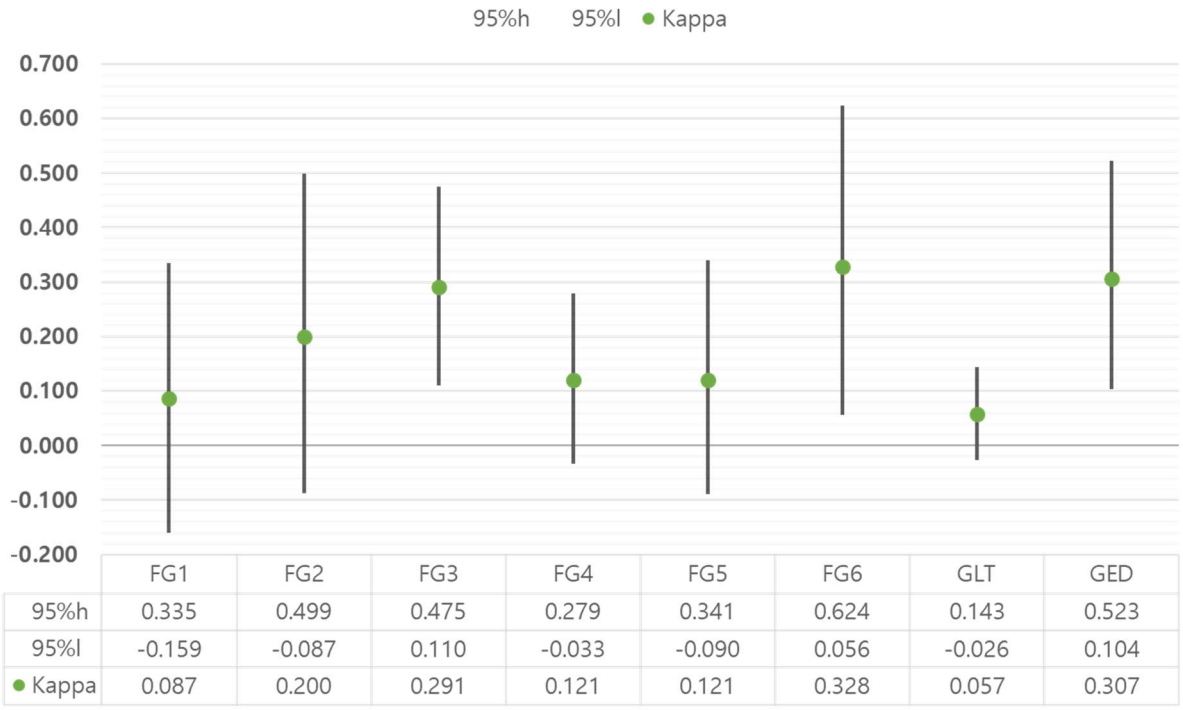


	FG1	FG2	FG3	FG4	FG5	FG6	GLT	GED
95%h	0.467	0.385	0.189	0.238	0.373	0.220	0.112	0.230
95%l	0.147	0.083	0.023	0.036	0.160	-0.124	0.020	0.076
Kappa	0.300	0.227	0.104	0.135	0.263	0.041	0.066	0.151

Abbildung 10 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP5 SoSe21



	FG1	FG2	FG3	FG4	FG5	FG6	GLT	GED
95%h	0.462	0.708	0.577	0.421	0.535	0.695	0.151	0.759
95%l	-0.317	-0.202	-0.104	-0.037	-0.043	-0.135	-0.092	0.162
Kappa	0.072	0.242	0.231	0.189	0.243	0.269	0.029	0.452



	FG1	FG2	FG3	FG4	FG5	FG6	GLT	GED
95%h	0.335	0.499	0.475	0.279	0.341	0.624	0.143	0.523
95%l	-0.159	-0.087	0.110	-0.033	-0.090	0.056	-0.026	0.104
Kappa	0.087	0.200	0.291	0.121	0.121	0.328	0.057	0.307

Abbildung 11 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP5 WiSe21/22

Der Kappa-Koeffizient für GLT zwischen STU war 0,05, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,17, es besteht auch eine “geringe“ Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,31, für FG2 (Informierung) 0,16, für FG3 (Gewinnen) 0,08, für FG4 (Erleichtern) 0,13, für FG5 (Unterstützen) 0,25 und für FG6 (Qualität der Gesprächsführung) 0,02. Die FG1 und FG5 zeigten die “ausreichenden” Übereinstimmungen, und die FG2, FG3, FG4 und FG6 die “geringen“. Aber für FG3 und FG6 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab. Die Werte waren eher tendenziell leicht höher als die von STU, obwohl ein zusätzlicher Bewerter hinzukam.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,07, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,15, es besteht auch eine “geringe“ Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,30, für FG2 (Informierung) 0,23, für FG3 (Gewinnen) 0,10, für FG4 (Erleichtern) 0,14, für FG5 (Unterstützen) 0,26 und für FG6 (Qualität der Gesprächsführung) 0,04. Die FG1, FG2 und FG5 zeigten die “ausreichenden” Übereinstimmungen, und die FG3, FG4 und FG6 die “geringen“. Aber für FG6 ragte das Konfidenzintervall in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Im WiSe21/22 (Abbildung 11) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “mittelmäßigen” Übereinstimmungen zwischen den Ratern auf.

Der Kappa-Koeffizient für GLT zwischen STU war 0,03, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,45, es besteht auch eine “geringe“ Übereinstimmung. Aber das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,07, für FG2 (Informierung) 0,24, für FG3 (Gewinnen) 0,23, für FG4 (Erleichtern) 0,19, für FG5 (Unterstützen) 0,24 und für FG6 (Qualität der Gesprächsführung) 0,27. Die FG1 und FG4 zeigten die „leichten“ Übereinstimmungen, und die FG2, FG3, FG5 und FG6 die “ausreichenden”. Aber für die alle FG ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab. Die Werte waren eher tendenziell leicht höher als die von STU, obwohl ein zusätzlicher Bewerter hinzukam.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,06, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,31, es besteht auch eine “geringe“ Übereinstimmung. Aber das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,09, für FG2 (Informierung) 0,20, für FG3 (Gewinnen) 0,29, für FG4 (Erleichtern) 0,12, für FG5 (Unterstützen) 0,12 und für FG6

(Qualität der Gesprächsführung) 0,33. Die FG1, FG2, FG4 und FG5 zeigten die „leichten“ Übereinstimmungen, und die FG3 und FG6 die „mittelmäßigen“. Aber für FG1, FG2, FG4 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

5.2.2.6. Objektivität bei RP6 (Motivationales Interview)

Beim RP6 im SoSe21 (Abbildung 12) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “geringen“ Übereinstimmungen zwischen den Ratern auf.

Der Kappa-Koeffizient für GLT zwischen STU war 0,04, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,17, es besteht eine “ausreichende” Übereinstimmung. Aber das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,13, für FG2 (Vermittlung von MI-Spirit) 0,13, für FG3 (Change Talk) 0,07, für FG4 (Geschmeidiger Umgang mit Widerstand) 0,09 und für FG5 (Qualität der Gesprächsführung) 0,17. Alle FG zeigten die „leichten“ Übereinstimmungen. Aber für die allen FG ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

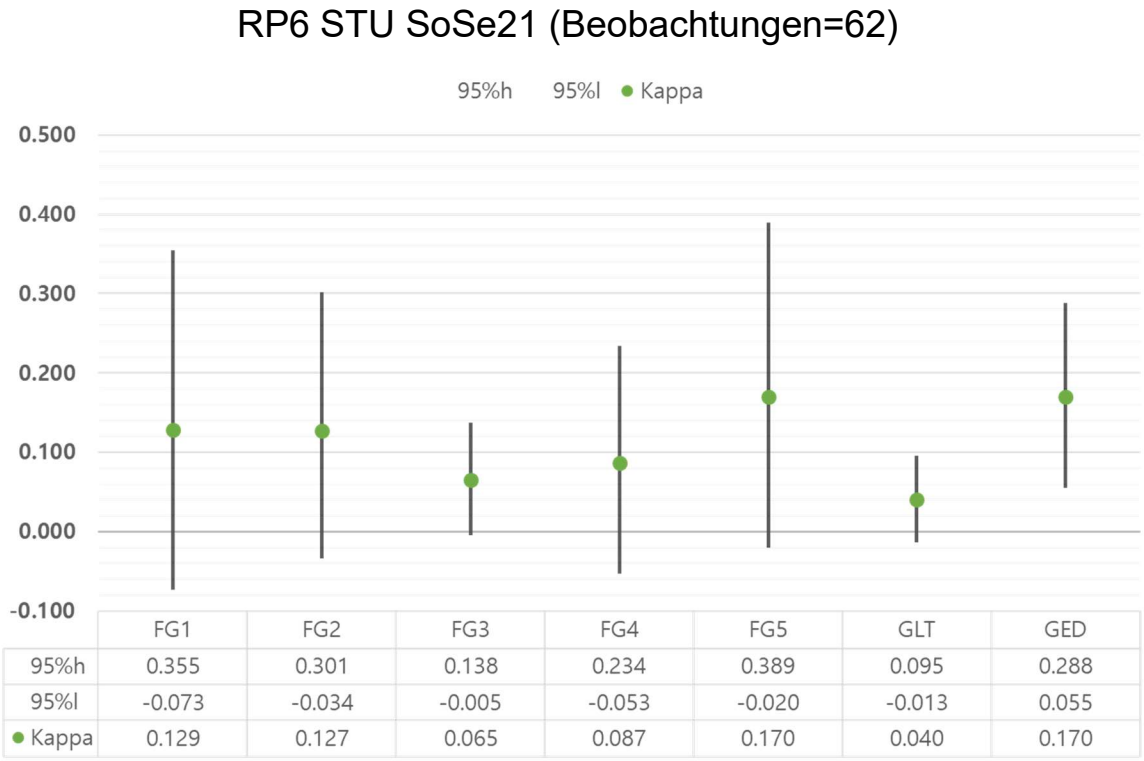
Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab. Die Werte waren eher tendenziell leicht höher als die von STU, obwohl ein zusätzlicher Bewerter hinzukam.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,04, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,14, es besteht auch eine “geringe“ Übereinstimmung. Aber das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

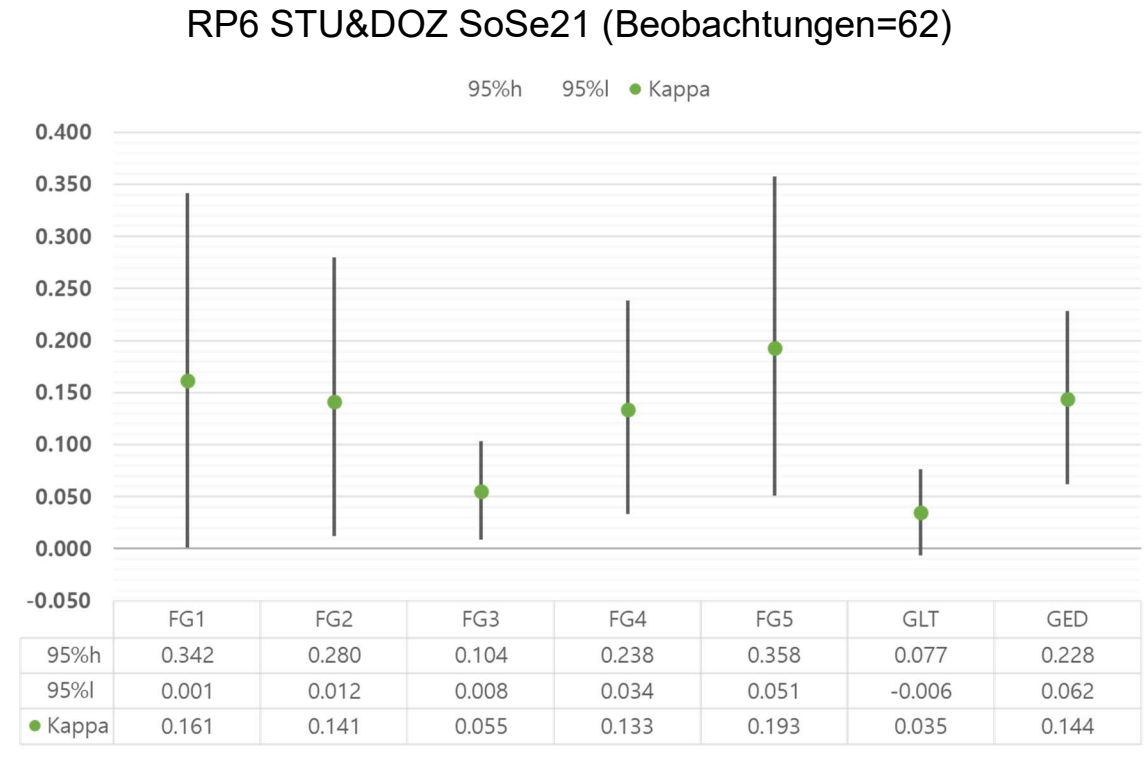
Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,16, für FG2 (Vermittlung von MI-Spirit) 0,14, für FG3 (Change Talk) 0,06, für FG4 (Geschmeidiger Umgang mit Widerstand) 0,13 und für FG5 (Qualität der Gesprächsführung) 0,19. Alle FG zeigten die „leichten“ Übereinstimmungen. Die Konfidenzintervalle blieben alle im positiven Bereich.

Im WiSe21/22 (Abbildung 13) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “ausreichenden” Übereinstimmungen zwischen den Ratern auf.

Der Kappa-Koeffizient für GLT zwischen STU war 0,04, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,67, es besteht auch eine “geringe“ Übereinstimmung. Aber für die beiden ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

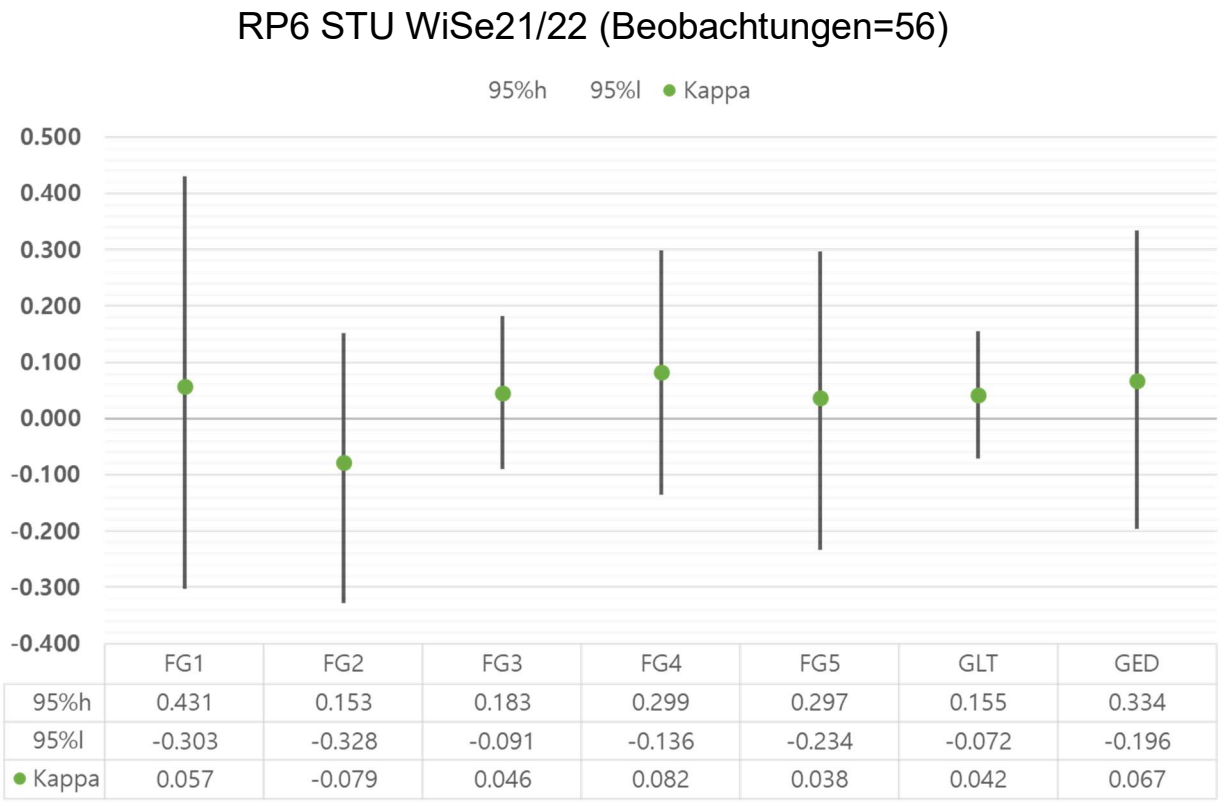


	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.355	0.301	0.138	0.234	0.389	0.095	0.288
95%l	-0.073	-0.034	-0.005	-0.053	-0.020	-0.013	0.055
Kappa	0.129	0.127	0.065	0.087	0.170	0.040	0.170

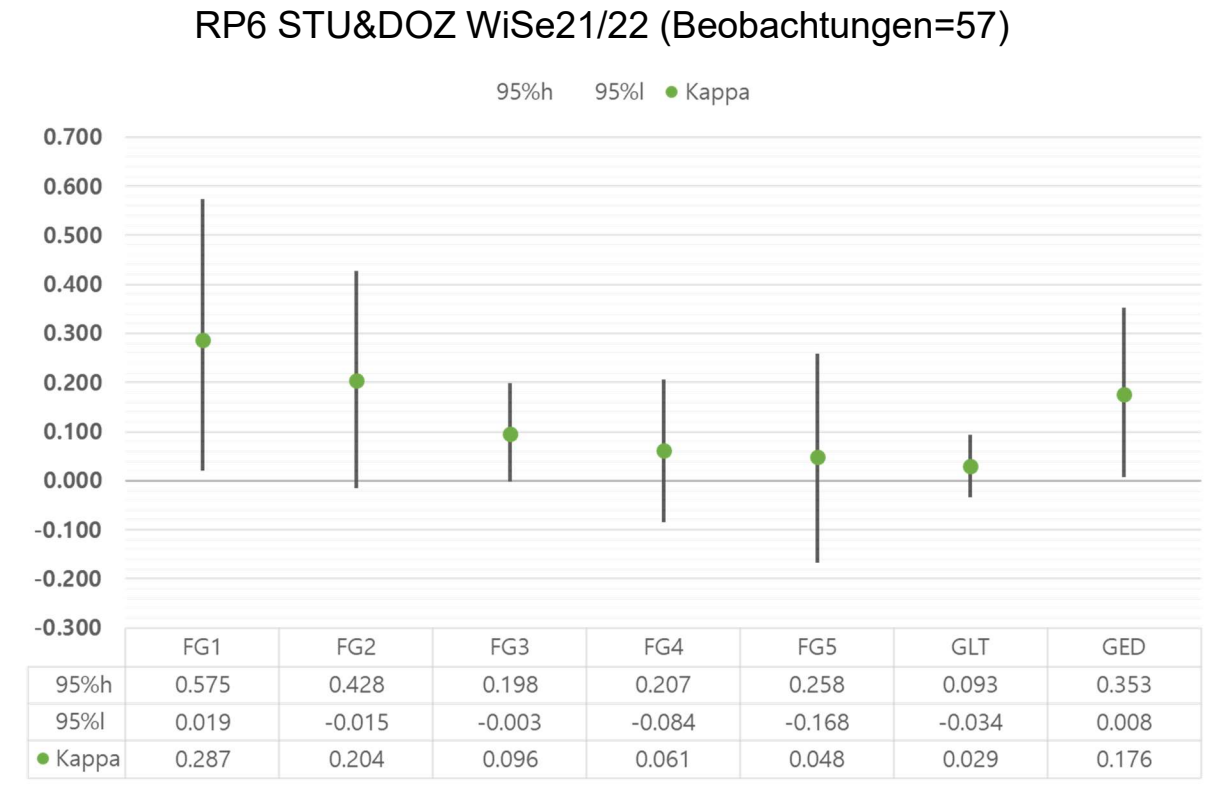


	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.342	0.280	0.104	0.238	0.358	0.077	0.228
95%l	0.001	0.012	0.008	0.034	0.051	-0.006	0.062
Kappa	0.161	0.141	0.055	0.133	0.193	0.035	0.144

Abbildung 12 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP6 SoSe21



	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.431	0.153	0.183	0.299	0.297	0.155	0.334
95%l	-0.303	-0.328	-0.091	-0.136	-0.234	-0.072	-0.196
Kappa	0.057	-0.079	0.046	0.082	0.038	0.042	0.067



	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.575	0.428	0.198	0.207	0.258	0.093	0.353
95%l	0.019	-0.015	-0.003	-0.084	-0.168	-0.034	0.008
Kappa	0.287	0.204	0.096	0.061	0.048	0.029	0.176

Abbildung 13 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP6 WiSe21/22

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,06, für FG2 (Vermittlung von MI-Spirit) -0,08, für FG3 (Change Talk) 0,05, für FG4 (Geschmeidiger Umgang mit Widerstand) 0,08 und für FG5 (Qualität der Gesprächsführung) 0,04. Die FG1, FG3, FG4 und FG5 zeigten die “geringen“ Übereinstimmungen, und die FG2 eine „geringe“. Aber für die allen FG ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

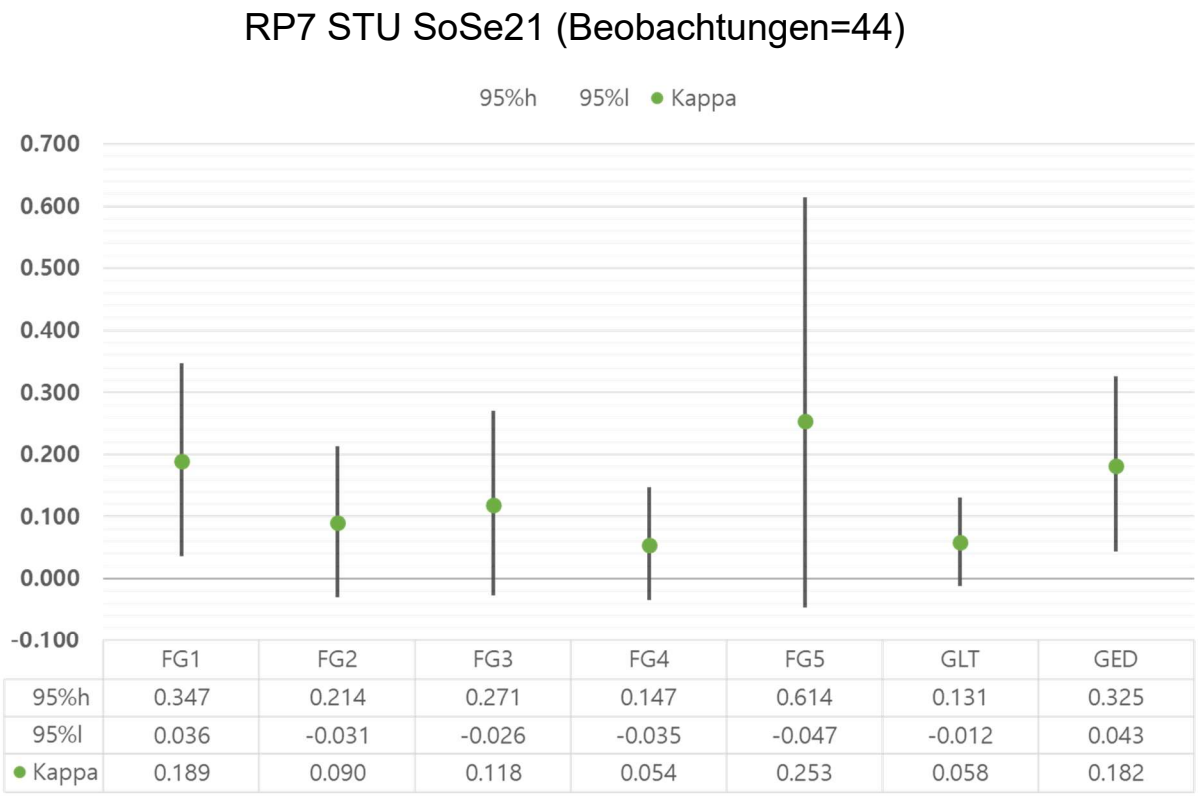
Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab. Die Werte waren eher tendenziell leicht höher als die von STU, obwohl ein zusätzlicher Bewerter hinzukam.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,03, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,18, es besteht auch eine “geringe“ Übereinstimmung. Aber das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

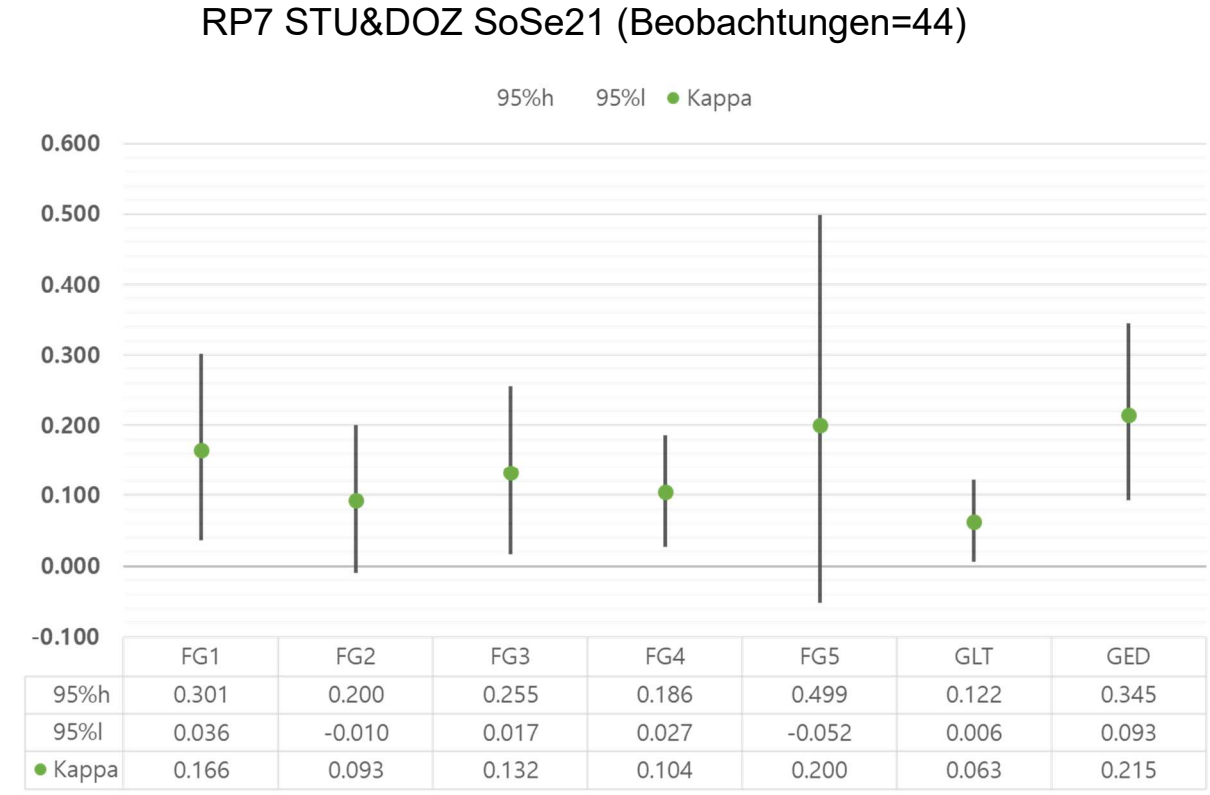
Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,29, für FG2 (Vermittlung von MI-Spirit) 0,20, für FG3 (Change Talk) 0,10, für FG4 (Geschmeidiger Umgang mit Widerstand) 0,06 und für FG5 (Qualität der Gesprächsführung) 0,05. Die FG1 und FG2 zeigten die “ausreichenden” Übereinstimmungen, FG3, FG4 und FG5 die „leichten“. Aber für FG2, FG3, FG4 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

5.2.2.7. Objektivität bei RP7 (Stressreaktion)

Beim RP7 im SoSe21 (Abbildung 14) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “ausreichenden” Übereinstimmungen zwischen den Ratern auf.

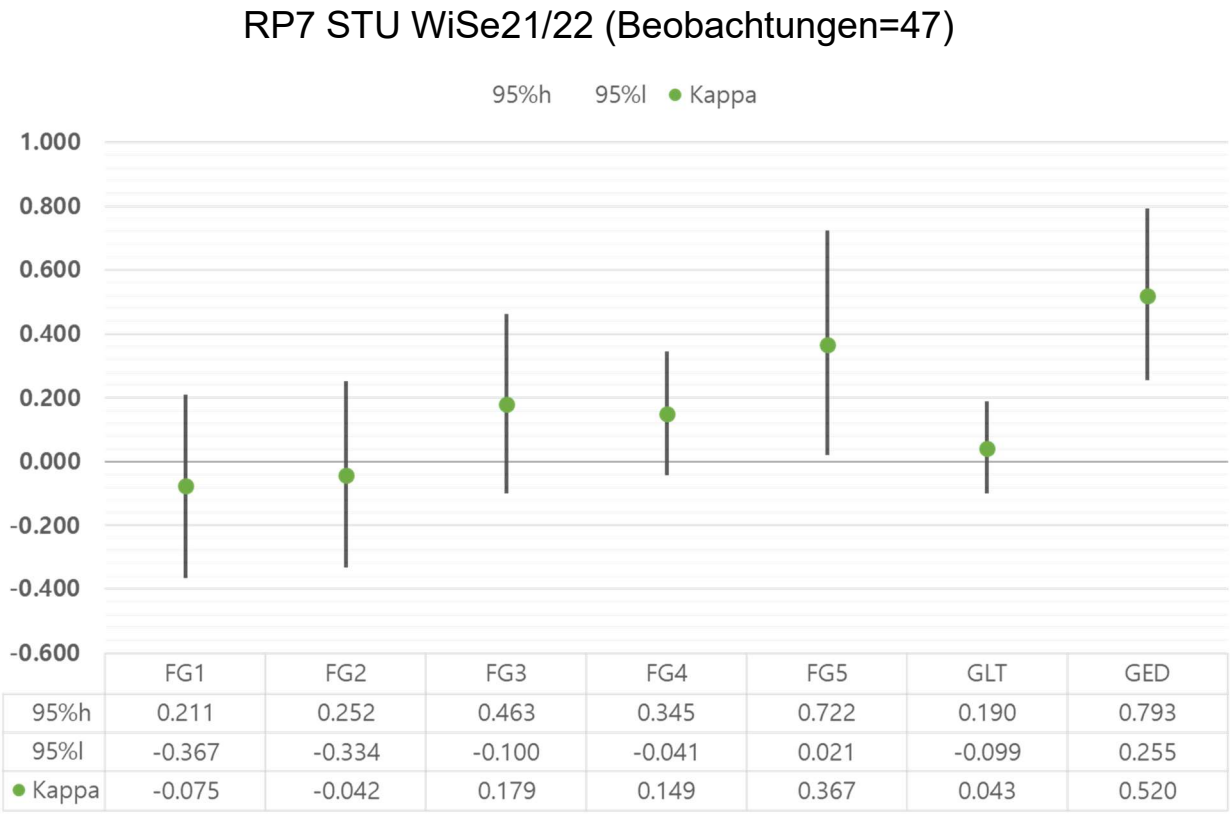


	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.347	0.214	0.271	0.147	0.614	0.131	0.325
95%l	0.036	-0.031	-0.026	-0.035	-0.047	-0.012	0.043
Kappa	0.189	0.090	0.118	0.054	0.253	0.058	0.182

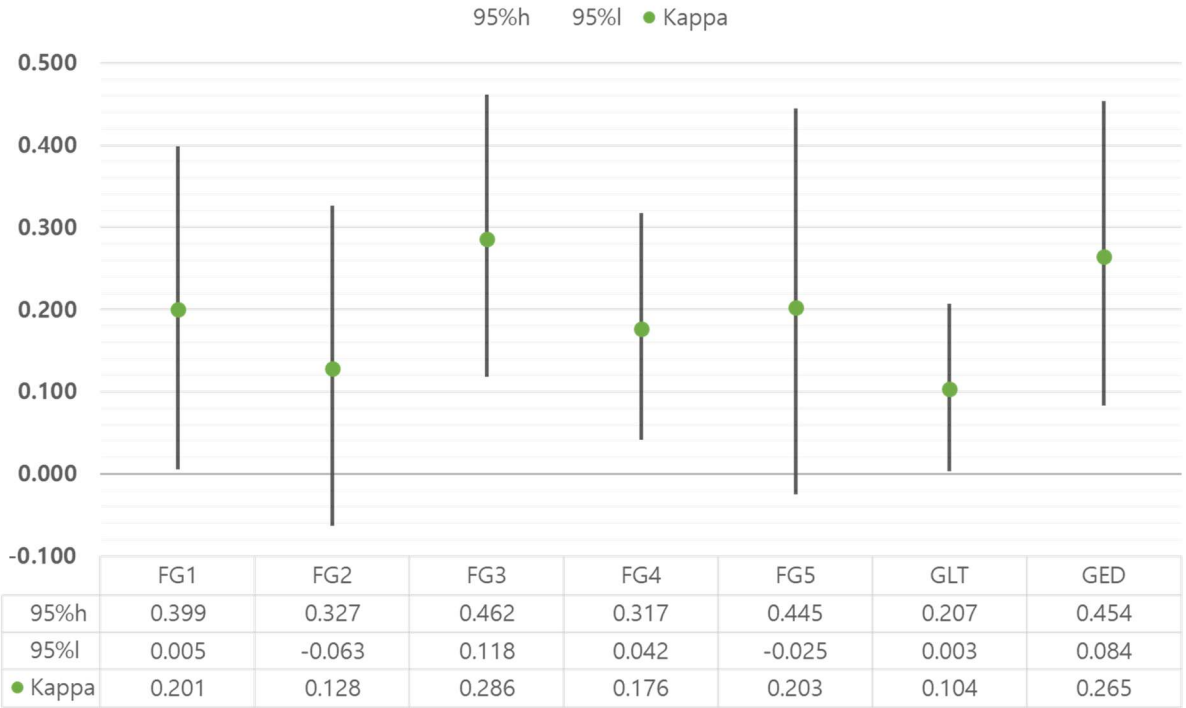


	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.301	0.200	0.255	0.186	0.499	0.122	0.345
95%l	0.036	-0.010	0.017	0.027	-0.052	0.006	0.093
Kappa	0.166	0.093	0.132	0.104	0.200	0.063	0.215

Abbildung 14 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP7 SoSe21



	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.211	0.252	0.463	0.345	0.722	0.190	0.793
95%l	-0.367	-0.334	-0.100	-0.041	0.021	-0.099	0.255
Kappa	-0.075	-0.042	0.179	0.149	0.367	0.043	0.520



	FG1	FG2	FG3	FG4	FG5	GLT	GED
95%h	0.399	0.327	0.462	0.317	0.445	0.207	0.454
95%l	0.005	-0.063	0.118	0.042	-0.025	0.003	0.084
Kappa	0.201	0.128	0.286	0.176	0.203	0.104	0.265

Abbildung 15 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP7 WiSe21/22

Der Kappa-Koeffizient für GLT zwischen STU war 0,06, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,18, es besteht auch eine “geringe“ Übereinstimmung. Aber das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,19, für FG2 (Stressorenanalyse, hat erfragt/herausgefunden) 0,09, für FG3 (Analyse der Stressreaktion, Ebene des) 0,12, für FG4 (Ansätze der Stressbewältigung: Stressreaktion und Stressoren) 0,05 und für FG5 (Qualität der Gesprächsführung) 0,25. Die FG1, FG2, FG3 und FG4 zeigten die „leichten“ Übereinstimmungen, und die FG5 eine “ausreichende”. Aber für die FG2, FG3, FG4 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab. Die Werte waren eher tendenziell leicht höher als die von STU, obwohl ein zusätzlicher Bewerter hinzukam.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,06, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,22, es besteht auch eine “geringe“ Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,17, für FG2 (Stressorenanalyse, hat erfragt/herausgefunden) 0,09, für FG3 (Analyse der Stressreaktion, Ebene des) 0,13, für FG4 (Ansätze der Stressbewältigung: Stressreaktion und Stressoren) 0,10 und für FG5 (Qualität der Gesprächsführung) 0,20. Alle zeigten die „leichten“ Übereinstimmungen. Aber für die FG2 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Im WiSe21/22 (Abbildung 15) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “mittelmäßigen” Übereinstimmungen zwischen den Ratern auf.

Der Kappa-Koeffizient für GLT zwischen STU war 0,04, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,52, es besteht eine “mittelmäßige” Übereinstimmung. Aber das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) -0,08, für FG2 (Stressorenanalyse, hat erfragt/herausgefunden) -0,04, für FG3 (Analyse der Stressreaktion, Ebene des) 0,18, für FG4 (Ansätze der Stressbewältigung: Stressreaktion und Stressoren) 0,15 und für FG5 (Qualität der Gesprächsführung) 0,37. Die FG1 und FG2 zeigten die „geringen“ Übereinstimmungen, die FG3 und FG4 die „leichten“, und die FG5 eine “ausreichende”. Aber für die FG1, FG2, FG3 und FG4 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab. Die Werte waren eher tendenziell leicht höher als die von STU, obwohl ein zusätzlicher Bewerter hinzukam.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,10, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,27, es besteht eine “ausreichende” Übereinstimmung. Die Konfidenzintervalle blieben alle im positiven Bereich.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,20, für FG2 (Stressorenanalyse, hat erfragt/herausgefunden) 0,13, für FG3 (Analyse der Stressreaktion, Ebene des) 0,29, für FG4 (Ansätze der Stressbewältigung: Stressreaktion und Stressoren) 0,18 und für FG5 (Qualität der Gesprächsführung) 0,20. Die FG1, FG3 und FG5 zeigten die “ausreichenden” Übereinstimmungen, und die FG2 und FG4 die „leichten“. Aber für die FG2 und FG5 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

5.2.2.8. Objektivität bei RP8 (Krebsaufklärung)

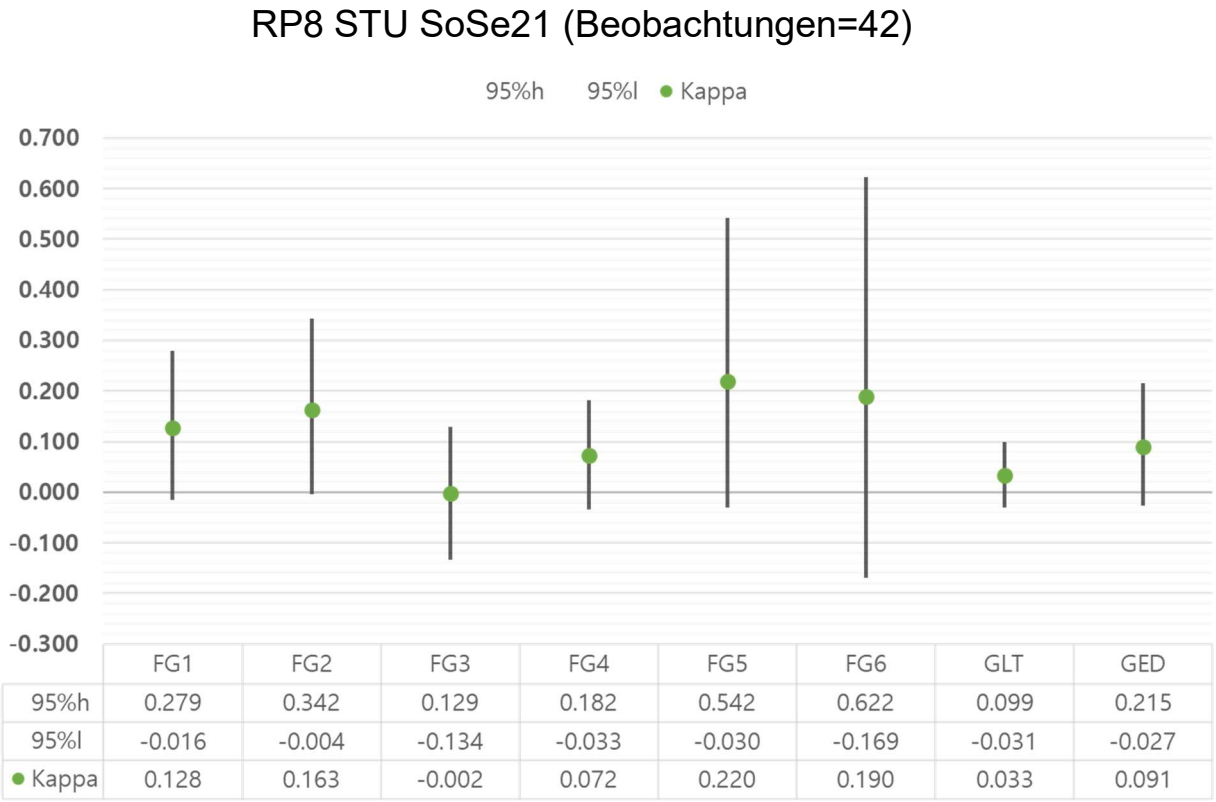
Beim RP8 im SoSe21 (Abbildung 16) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “geringen“ Übereinstimmungen zwischen den Ratern auf.

Der Kappa-Koeffizient für GLT zwischen STU war 0,03, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,09, es besteht auch eine “geringe“ Übereinstimmung. Aber für die beiden ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

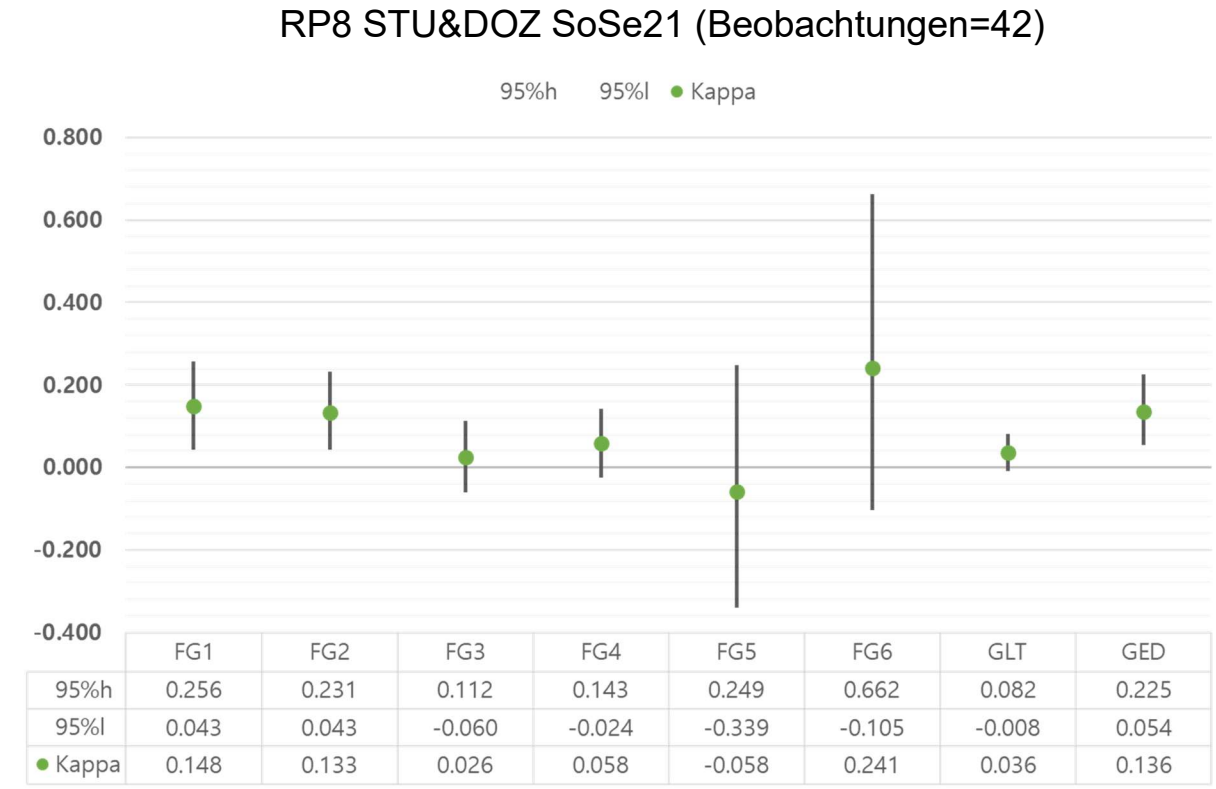
Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,13, für FG2 (Eingehen auf den Patienten) 0,16, für FG3 (Diagnosemitteilung) 0,00, für FG4 (Reaktionen des Patienten) 0,07, für FG5 (Weiterbehandlung) 0,22 und für FG6 (Qualität der Gesprächsführung) 0,19. Die FG1, FG2, FG4 und FG6 zeigten die „leichten“ Übereinstimmungen, die FG3 eine „geringe“, und die FG5 eine “ausreichende”. Aber für die allen FG ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab. Die Werte waren eher tendenziell leicht höher als die von STU, obwohl ein zusätzlicher Bewerter hinzukam.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,04, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,14, es besteht auch eine “geringe“ Übereinstimmung.

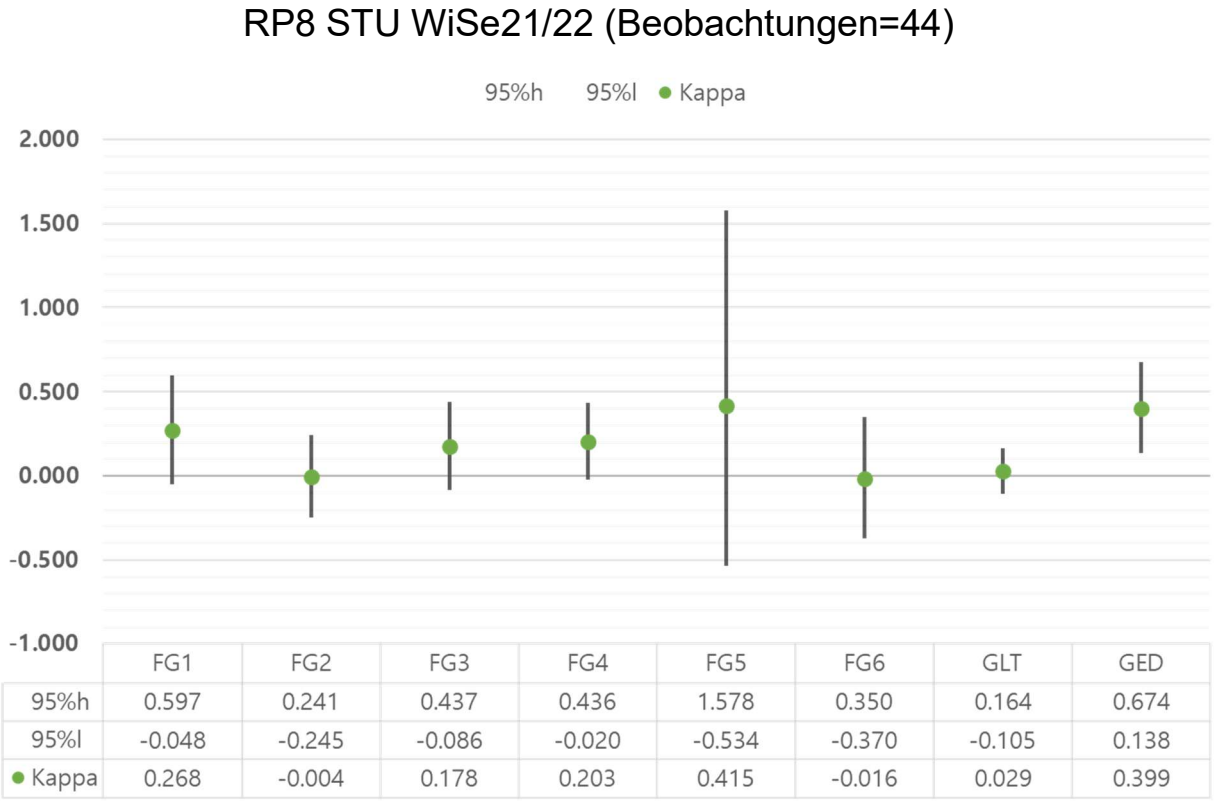


	FG1	FG2	FG3	FG4	FG5	FG6	GLT	GED
95%h	0.279	0.342	0.129	0.182	0.542	0.622	0.099	0.215
95%l	-0.016	-0.004	-0.134	-0.033	-0.030	-0.169	-0.031	-0.027
Kappa	0.128	0.163	-0.002	0.072	0.220	0.190	0.033	0.091

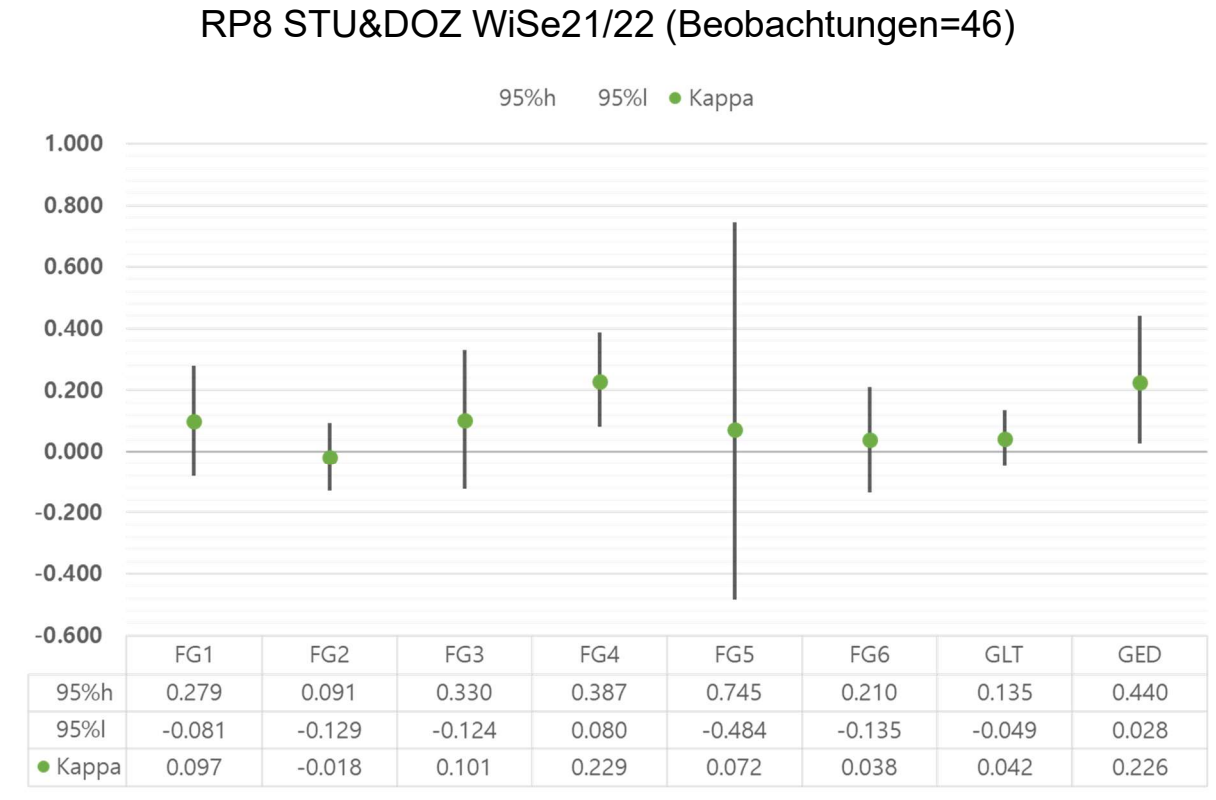


	FG1	FG2	FG3	FG4	FG5	FG6	GLT	GED
95%h	0.256	0.231	0.112	0.143	0.249	0.662	0.082	0.225
95%l	0.043	0.043	-0.060	-0.024	-0.339	-0.105	-0.008	0.054
Kappa	0.148	0.133	0.026	0.058	-0.058	0.241	0.036	0.136

Abbildung 16 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP8 SoSe21



	FG1	FG2	FG3	FG4	FG5	FG6	GLT	GED
95%h	0.597	0.241	0.437	0.436	1.578	0.350	0.164	0.674
95%l	-0.048	-0.245	-0.086	-0.020	-0.534	-0.370	-0.105	0.138
Kappa	0.268	-0.004	0.178	0.203	0.415	-0.016	0.029	0.399



	FG1	FG2	FG3	FG4	FG5	FG6	GLT	GED
95%h	0.279	0.091	0.330	0.387	0.745	0.210	0.135	0.440
95%l	-0.081	-0.129	-0.124	0.080	-0.484	-0.135	-0.049	0.028
Kappa	0.097	-0.018	0.101	0.229	0.072	0.038	0.042	0.226

Abbildung 17 Inter-Rater Korrelation der Checklisten-Bewertungen (Conger´s Kappa) RP8 WiSe21/22

Aber das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,15, für FG2 (Eingehen auf den Patienten) 0,13, für FG3 (Diagnosemitteilung) 0,03, für FG4 (Reaktionen des Patienten) 0,06, für FG5 (Weiterbehandlung) -0,06 und für FG6 (Qualität der Gesprächsführung) 0,24. Die FG1, FG2, FG3 und FG4 zeigten die „leichten“ Übereinstimmungen, die FG5 eine „geringe“, und die FG6 eine “ausreichende”. Aber für die FG3, FG4, FG5 und FG6 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Beim RP8 im WiSe21/22 (Abbildung 17) wiesen die Kappa-Koeffizienten die von „geringen“ bzw. “ausreichenden” Übereinstimmungen zwischen den Ratern auf.

Der Kappa-Koeffizient für GLT zwischen STU war 0,03, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,40, es besteht eine “ausreichende” Übereinstimmung. Aber für die GLT ragte das Konfidenzintervall in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,27, für FG2 (Eingehen auf den Patienten) 0,00, für FG3 (Diagnosemitteilung) 0,18, für FG4 (Reaktionen des Patienten) 0,20, für FG5 (Weiterbehandlung) 0,42 und für FG6 (Qualität der Gesprächsführung) -0,02. Die FG1, FG4 und FG5 zeigten die “ausreichenden” Übereinstimmungen, die FG2 und FG6 die „geringen“, und die FG3 eine „leichte“. Aber für die allen FG ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Die Ergebnisse der Kappa Analyse aus STU und DOZ zusammen wichen nicht wesentlich von denen aus STU ab.

Der Kappa-Koeffizient für GLT zwischen STU und DOZ war 0,04, es besteht eine “geringe“ Übereinstimmung. Für GED war der Kappa-Koeffizient 0,23, es besteht eine “ausreichende” Übereinstimmung. Aber das Konfidenzintervall für GLT ragte in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

Aus FG war der Kappa-Koeffizient für FG1 (Eröffnung des Gesprächs) 0,10, für FG2 (Eingehen auf den Patienten) -0,02, für FG3 (Diagnosemitteilung) 0,10, für FG4 (Reaktionen des Patienten) 0,23, für FG5 (Weiterbehandlung) 0,07 und für FG6 (Qualität der Gesprächsführung) 0,04. Die FG1, FG3, FG5 und FG6 zeigten die „leichten“ Übereinstimmungen, die FG2 eine „geringe“, und die FG4 eine “ausreichende”. Aber für die FG1, FG2, FG3, FG5 und FG6 ragten die Konfidenzintervalle in den negativen Bereich hinein, sodass man keine eindeutige Interpretation treffen kann.

5.2.3. Validität

Die Abschätzung der Validität erfolgte anhand von fünf Komponenten nach Messick: Testinhalt, Reaktionsprozess, Internale Struktur, Relation zu anderen Variablen und Konsequenz von Test.

Der Testinhalt wurde qualitativ durch den Entwicklungsprozess von RP sowie dazugehörigen Checkliste beurteilt. Basierend auf der klaren Zielsetzung wurden die RP-Szenarien und -Checkliste von Experten erstellt. Unter den Zielsetzungen, sich ein biopsychosoziales Krankheitsverständnis sowie Grundkompetenzen patientenorientierter ärztlicher Gesprächsführung zu erarbeiten und für die Situation schwerkranker und sterbender Patienten zu sensibilisieren, wurden darauf passende Fallvignetten entwickelt. Die dazugehörigen Checklisten beinhalteten die Items, die umfassend die Qualität des Gesprächs bewerten. Diese wurden ebenfalls von denselben Experten entwickelt [26,45].

Zum Reaktionsprozess wurde der gesamte Ablauf von RP (einschließlich dessen Vorbereitung) auseinandergesetzt. Die Einzelheiten wurden bereits im 4.1. vorgestellt. Die Interrater-Korrelation kann auch Hinweise auf den Reaktionsprozess geben. Diese Ergebnisse wurden in 5.2.2 präsentiert.

Die Internale Struktur ließ sich durch die Korrelation zwischen GLT und GED (s. 5.3.1) und das Cronbach's Alpha mit FG (s. 5.3.5) evaluieren. Die Ergebnisse wurden jeweils in 5.2.1 und 5.2.5 gezeigt.

Die Relation zu anderen Variablen wurden geprüft, indem wir die Bewertungen aus DOZ und die aus SP verglichen. GED aus DOZ und GNT aus SP wurden mit der Spearmann-Korrelation analysiert. Die Ergebnisse werden unten vorgestellt.

Die Konsequenz von Test konnten wir nicht messen, wie im 4.2.1.3 bereits erklärten. Die Studierenden hatten zweimalige Gespräche in Abstand von mehreren Wochen. Allerdings waren die Themen der beiden Gespräche unterschiedlich, sodass kein direkter Vergleich zwischen zwei Gesprächen möglich war. Aus diesem Grund konnten wir die Verbesserung der kommunikativen Leistung der Studierenden durch die Simulationsgespräche nicht quantifizieren.

5.2.3.1. Validität bei RP1 (Anamnese)

Beim RP1 (Tabelle 13) gab es eine "ausreichende" und eine „schwache" Korrelation zwischen GED aus DOZ und GNT aus SP. Im SoSe21 war der Rangkorrelationskoeffizient 0,34 aus 63 Beobachtungen. Der p-Wert war 0,01 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig" korrelierte. Im WiSe21/22 war der Rangkorrelationskoeffizient 0,27 aus 60 Beobachtungen. Der p-Wert war 0,04 < 0,05, sodass es im 5% Signifikanzniveau „schwach" korrelierte.

Tabelle 21 Korrelation der Bewertungen der DOZ mit Bewertungen der SP (Spearman-Korrelation) RP1

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	Rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ SP	GED GNT	63	0.338	0.007	60	0.267	0.039

5.2.3.2. Validität bei RP2 (Gesprächsförderung)

Beim RP2 (Tabelle 14) gab es “ausreichende” Korrelationen zwischen GED aus DOZ und GNT aus SP. Im SoSe21 war der Rangkorrelationskoeffizient 0,31 aus 50 Beobachtungen. Der p-Wert war 0,03 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig“ korrelierte. Im WiSe21/22 war der Rangkorrelationskoeffizient 0,36 aus 54 Beobachtungen. Der p-Wert war 0,01 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig“ korrelierte.

Tabelle 22 Korrelation der Bewertungen der DOZ mit Bewertungen der SP (Spearman-Korrelation) RP2

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ SP	GED GNT	50	0.312	0.028	54	0.364	0.007

5.2.3.3. Validität bei RP3 (Informationsvermittlung)

Beim RP3 (Tabelle 15) gab es “ausreichende” Korrelationen zwischen GED aus DOZ und GNT aus SP. Im SoSe21 war der Rangkorrelationskoeffizient 0,54 aus 49 Beobachtungen. Der p-Wert war 0,00 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig“ korrelierte. Im WiSe21/22 war der Rangkorrelationskoeffizient 0,48 aus 57 Beobachtungen. Der p-Wert war 0,00 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig“ korrelierte.

Tabelle 23 Korrelation der Bewertungen der DOZ mit Bewertungen der SP (Spearman-Korrelation) RP3

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ SP	GED GNT	49	0.544	0.000	57	0.481	0.000

5.2.3.4. Validität bei RP4 (Partizipative Entscheidung)

Beim RP4 (Tabelle 16) gab es “ausreichende” Korrelationen zwischen GED aus DOZ und GNT aus SP. Im SoSe21 war der Rangkorrelationskoeffizient 0,48 aus 52 Beobachtungen. Der p-Wert war 0,00 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig“ korrelierte. Im WiSe21/22 war der Rangkorrelationskoeffizient 0,37 aus 40 Beobachtungen. Der p-Wert war 0,02 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig“ korrelierte.

Tabelle 24 Korrelation der Bewertungen der DOZ mit Bewertungen der SP (Spearman-Korrelation) RP4

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ SP	GED GNT	52	0.475	0.000	40	0.369	0.019

5.2.3.5. Validität bei RP5 (Compliance)

Beim RP5 (Tabelle 17) gab es “ausreichende” Korrelationen zwischen GED aus DOZ und GNT aus SP. Im SoSe21 war der Rangkorrelationskoeffizient 0,32 aus 53 Beobachtungen. Der p-Wert war

0,02 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig“ korrelierte. Im WiSe21/22 war der Rangkorrelationskoeffizient 0,57 aus 42 Beobachtungen. Der p-Wert war 0,00 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig“ korrelierte.

Tabelle 25 Korrelation der Bewertungen der DOZ mit Bewertungen der SP (Spearman-Korrelation) RP5

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ SP	GED GNT	53	0.320	0.019	42	0.568	0.000

5.2.3.6. Validität bei RP6 (Motivationales Interview)

Beim RP6 (Tabelle 18) gab es “ausreichende” Korrelationen zwischen GED aus DOZ und GNT aus SP. Im SoSe21 war der Rangkorrelationskoeffizient 0,38 aus 50 Beobachtungen. Der p-Wert war 0,01 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig“ korrelierte. Im WiSe21/22 war der Rangkorrelationskoeffizient 0,52 aus 51 Beobachtungen. Der p-Wert war 0,00 < 0,05, sodass es im 5% Signifikanzniveau „mittelmäßig“ korrelierte.

Tabelle 26 Korrelation der Bewertungen der DOZ mit Bewertungen der SP (Spearman-Korrelation) RP6

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ SP	GED GNT	50	0.380	0.006	51	0.519	0.000

5.2.3.7. Validität bei RP7 (Stressreaktion)

Beim RP7 (Tabelle 19) gab es eine “ausreichende” Korrelation zwischen GED aus DOZ und GNT aus SP. Im SoSe21 war der Rangkorrelationskoeffizient 0,32 aus 31 Beobachtungen. Der p-Wert war 0,08 > 0,05, sodass sich die “ausreichende” Korrelation im 5% Signifikanzniveau nicht annehmen ließ. Im WiSe21/22 war der Rangkorrelationskoeffizient 0,42 aus 41 Beobachtungen. Der p-Wert war 0,01 < 0,05, sodass es im 5% Signifikanzniveau doch “ausreichend” korrelierte.

Tabelle 27 Korrelation der Bewertungen der DOZ mit Bewertungen der SP (Spearman-Korrelation) RP7

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ SP	GED GNT	31	0.317	0.082	41	0.420	0.006

5.2.3.8. Validität bei RP8 (Krebsaufklärung)

Beim RP8 (Tabelle 20) gab es eine “mittelmäßige” Korrelation zwischen GED aus DOZ und GNT aus SP. Im SoSe21 war der Rangkorrelationskoeffizient 0,25 aus 31 Beobachtungen. Der p-Wert war 0,18 > 0,05, sodass sich die „schwache“ Korrelation im 5% Signifikanzniveau nicht annehmen ließ. Im WiSe21/22 war der Rangkorrelationskoeffizient 0,66 aus 43 Beobachtungen. Der p-Wert war 0,00 < 0,05, sodass es im 5% Signifikanzniveau “mittelmäßig” korrelierte.

Tabelle 28 Korrelation der Bewertungen der DOZ mit Bewertungen der SP (Spearman-Korrelation) RP8

Stichprobe	Daten	SoSe21			WiSe21/22		
		Obs	rho	Prob > |t|	Obs	rho	Prob > |t|
DOZ SP	GED GNT	31	0.245	0.184	43	0.661	0.000

5.2.4. Aufgabenschwierigkeit (AS)

Die FG wurden als Item zur Aufgabenschwierigkeit geprüft. Der ermittelte Mittelwert aus DOZ wurde in einen Prozentsatz gegen den maximal erreichbaren Punkt berechnet. Items mit einem AS-Index außerhalb des Bereichs von 20-80% werden in der Regel ausgeschlossen.

5.2.4.1. AS bei RP1 (Anamnese)

Beim RP1 (Tabelle 21) sowohl im SoSe21 als auch WiSe21/22 brachten die STU sehr gute Leistungen. So hatten alle FG in den beiden Semestern den AS-Index über 80%, was allerdings nicht zur guten testtheoretischen Qualität beitragen sollte. Alle Aufgaben waren zu leicht für die Studierenden. Zudem zwischen den AS-Indizes ließ sich nur ein enges Spektrum beobachten, was zwecks Auseinandersetzung unterschiedlicher Leistungen ungünstig ist.

Tabelle 29 Aufgabenschwierigkeit (AS-Index) RP1

Item	SoSe21				WiSe21/22			
	Obs	Mean	Max	AS	Obs	Mean	Max	AS
FG1	70	1.450	1.5	96.67%	58	1.474	1.5	98.28%
FG2		4.721	5	94.43%		4.672	5	93.45%
FG3		3.121	3.5	89.18%		2.991	3.5	85.47%
FG4		3.657	4	91.43%		3.888	4	97.20%

5.2.4.2. AS bei RP2 (Gesprächsförderung)

Beim RP2 (Tabelle 22) sowohl im SoSe21 als auch WiSe21/22 brachten die STU sehr gute Leistungen. So hatten abgesehen von FG3 im WiSe21/22 alle FG in den beiden Semestern den AS-Index über 80%, was allerdings nicht zur guten testtheoretischen Qualität beitragen sollte. Diese Aufgaben waren zu leicht für die Studierenden. Zudem zwischen den AS-Indizes ließ sich nur ein enges Spektrum beobachten, was zwecks Auseinandersetzung unterschiedlicher Leistungen ungünstig ist.

Tabelle 30 Aufgabenschwierigkeit (AS-Index) RP2

Item	SoSe21				WiSe21/22			
	Obs	Mean	Max	AS	Obs	Mean	Max	AS
FG1	59	1.890	2	94.49%	57	1.772	2	88.60%
FG2		3.881	4.5	86.25%		3.711	4.5	82.46%
FG3		1.246	1.5	83.05%		1.132	1.5	75.44%
FG4		2.627	3	87.57%		2.614	3	87.13%
FG5		2.822	3	94.07%		2.877	3	95.91%

5.2.4.3. AS bei RP3 (Informationsvermittlung)

Beim RP3 (Tabelle 23) sowohl im SoSe21 als auch WiSe21/22 brachten die STU gute Leistungen. So hatten FG1, FG3 und FG5 in den beiden Semestern den AS-Index über 80%, was allerdings nicht zur guten testtheoretischen Qualität beitragen sollte. Diese Aufgaben waren zu leicht für die Studierenden. Die FG2 hatte den AS-Index von 77,19% im SoSe21 und 62,86% im WiSe21/22, es war leicht für sie. Die FG4 hatte den AS-Index von 54,97% im SoSe21 und 46,03% im WiSe21/22, es war mittelschwer für sie.

Tabelle 31 Aufgabenschwierigkeit (AS-Index) RP3

Item	SoSe21				WiSe21/22			
	Obs	Mean	Max	AS	Obs	Mean	Max	AS
FG1	57	1.921	2	96.05%	63	1.905	2	95.24%
FG2		1.930	2.5	77.19%		1.571	2.5	62.86%
FG3		4.561	5	91.23%		4.508	5	90.16%
FG4		0.825	1.5	54.97%		0.691	1.5	46.03%
FG5		2.772	3	92.40%		2.897	3	96.56%

5.2.4.4. AS bei RP4 (Partizipative Entscheidung)

Beim RP4 (Tabelle 24) sowohl im SoSe21 als auch WiSe21/22 brachten die STU sehr gute Leistungen. So hatten alle FG in den beiden Semestern den AS-Index über 80%, was allerdings nicht zur guten testtheoretischen Qualität beitragen sollte. Alle Aufgaben waren zu leicht für die Studierenden. Zudem zwischen den AS-Indizes ließ sich nur ein enges Spektrum beobachten, was zwecks Auseinandersetzung unterschiedlicher Leistungen ungünstig ist.

Tabelle 32 Aufgabenschwierigkeit (AS-Index) RP4

Item	SoSe21				WiSe21/22			
	Obs	Mean	Max	AS	Obs	Mean	Max	AS
FG1	57	1.465	1.5	97.66%	45	1.444	1.5	96.30%
FG2		8.667	10.5	82.54%		9.078	10.5	86.46%
FG3		2.886	3	96.20%		2.900	3	96.67%

5.2.4.5. AS bei RP5 (Compliance)

Beim RP5 (Tabelle 25) sowohl im SoSe21 als auch WiSe21/22 brachten die STU gute Leistungen. So hatten FG1, FG2 und FG5 in den beiden Semestern und FG3 im WiSe21/22 den AS-Index über 80%, was allerdings nicht zur guten testtheoretischen Qualität beitragen sollte. Diese Aufgaben waren zu leicht für die Studierenden. Die FG3 im SoSe21 hatte den AS-Index von 79,31%. FG4 hatte den Wert jeweils 71,84% im SoSe21, 71,21% im WiSe21/22 und FG5 78,16% im SoSe21, 78,79% im WiSe21/22. Diese Aufgaben waren leicht für sie. Zudem zwischen den AS-Indizes ließ sich nur ein enges Spektrum beobachten, was zwecks Auseinandersetzung unterschiedlicher Leistungen ungünstig ist.

Tabelle 33 Aufgabenschwierigkeit (AS-Index) RP5

Item	SoSe21				WiSe21/22			
	Obs	Mean	Max	AS	Obs	Mean	Max	AS
FG1	58	1.422	1.5	94.83%	44	1.375	1.5	91.67%
FG2		1.845	2	92.24%		1.659	2	82.95%
FG3		2.379	3	79.31%		2.432	3	81.06%
FG4		2.155	3	71.84%		2.136	3	71.21%
FG5		1.172	1.5	78.16%		1.182	1.5	78.79%
FG6		2.922	3	97.41%		2.830	3	94.32%

5.2.4.6. AS bei RP6 (Motivationales Interview)

Beim RP6 (Tabelle 26) sowohl im SoSe21 als auch WiSe21/22 brachten die STU gute Leistungen. So hatten FG1, FG2 und FG5 in den beiden Semestern den AS-Index über 80%, was allerdings nicht zur guten testtheoretischen Qualität beitragen sollte. Diese Aufgaben waren zu leicht für die Studierenden. Die FG3 hatte den AS-Index von 62,58% im SoSe21 und 66,83% im WiSe21/22, es war leicht für sie. Die FG4 hatte den AS-Index von 77,73% im SoSe21 und 76,44% im WiSe21/22, es war ebenso leicht. Zudem zwischen den AS-Indizes ließ sich nur ein enges Spektrum beobachten, was zwecks Auseinandersetzung unterschiedlicher Leistungen ungünstig ist.

Tabelle 34 Aufgabenschwierigkeit (AS-Index) RP6

Item	SoSe21				WiSe21/22			
	Obs	Mean	Max	AS	Obs	Mean	Max	AS
FG1	55	1.455	1.5	96.97%	52	1.413	1.5	94.23%
FG2		1.318	1.5	87.88%		1.288	1.5	85.90%
FG3		3.755	6	62.58%		4.010	6	66.83%
FG4		1.555	2	77.73%		1.529	2	76.44%
FG5		2.873	3	95.76%		2.904	3	96.79%

5.2.4.7. AS bei RP7 (Stressreaktion)

Beim RP7 (Tabelle 27) sowohl im SoSe21 als auch WiSe21/22 brachten die STU gute Leistungen. So hatten FG1, FG2 und FG5 in den beiden Semestern und FG3 im SoSe21 den AS-Index über 80%, was allerdings nicht zur guten testtheoretischen Qualität beitragen sollte. Diese Aufgaben waren zu leicht für die Studierenden. Die FG3 im WiSe21/22 hatte den AS-Index von 75,00%, es war leicht für sie. Die FG4 hatte den AS-Index von 76,69% im SoSe21 und 77,89% im WiSe21/22, es war ebenso leicht. Zudem zwischen den AS-Indizes ließ sich nur ein enges Spektrum beobachten, was zwecks Auseinandersetzung unterschiedlicher Leistungen ungünstig ist.

Tabelle 35 Aufgabenschwierigkeit (AS-Index) RP7

Item	SoSe21				WiSe21/22			
	Obs	Mean	Max	AS	Obs	Mean	Max	AS
FG1	38	1.342	1.5	89.47%	42	1.369	1.5	91.27%
FG2		2.829	3	94.30%		2.738	3	91.27%
FG3		2.421	3	80.70%		2.250	3	75.00%
FG4		2.684	3.5	76.69%		2.726	3.5	77.89%
FG5		2.868	3	95.61%		2.917	3	97.22%

5.2.4.8. AS bei RP8 (Krebsaufklärung)

Beim RP8 (Tabelle 28) sowohl im SoSe21 als auch WiSe21/22 brachten die STU gute Leistungen. So hatten FG1, FG3, FG5 und FG6 in den beiden Semestern, FG2 im SoSe21 und FG4 im WiSe21/22 den AS-Index über 80%, was allerdings nicht zur guten testtheoretischen Qualität beitragen sollte. Diese Aufgaben waren zu leicht für die Studierenden. Die FG2 im WiSe21/22 hatte den AS-Index von 79,09%, es war leicht für sie. Die FG4 im SoSe21 hatte den AS-Index von 78,24%, es war ebenso leicht. Zudem zwischen den AS-Indizes ließ sich nur ein enges Spektrum beobachten, was zwecks Auseinandersetzung unterschiedlicher Leistungen ungünstig ist.

Tabelle 36 Aufgabenschwierigkeit (AS-Index) RP8

Item	SoSe21				WiSe21/22			
	Obs	Mean	Max	AS	Obs	Mean	Max	AS
FG1	36	1.806	2	90.28%	44	1.773	2	88.64%
FG2		2.222	2.5	88.89%		1.977	2.5	79.09%
FG3		2.292	2.5	91.67%		2.341	2.5	93.64%
FG4		2.347	3	78.24%		2.432	3	81.06%
FG5		0.917	1	91.67%		0.977	1	97.73%
FG6		2.847	3	94.91%		2.830	3	94.32%

5.2.5. Trennschärfe

Mit der Trennschärfe ist es möglich, die statistische Beziehung zwischen den Items zu analysieren. Die Ergebnisse wurden mit Cronbach's Alpha ermittelt. Bei der Interpretation wurden zwei Werte in Betracht gezogen, nämlich Item-Rest-Correlation und Alpha. Item-Rest-Correlation, sogenannte Trennschärfe, gibt an, wie gut das Item mit anderen Items korreliert. Dies sollte über 0,3 liegen. Zudem musste der Alpha-Wert auch berücksichtigt werden. Er gibt an, wie hoch die Reliabilität ist, wenn das Item von der Checkliste ausgeschlossen wird. Ist ein Alpha-Wert eines Items deutlich größer als der Alpha-Wert von Test-scale, hat die Checkliste ohne dieses Item noch bessere Reliabilität als die Checkliste mit diesem Item. Und Dieser Alpha-Wert sollte über 0,5 sein, um akzeptiert zu werden [46,50].

5.2.5.1. Trennschärfe bei RP1 (Anamnese)

Beim RP1 im SoSe21 (Tabelle 29) hatte nur die FG3 die akzeptierbare Trennschärfe von 0,35, die FG1, FG2 und FG4 hatten den Wert unter 0,3. Selbst der Wert von FG3 lag jedoch nicht weit von der Grenze entfernt. Das Test-scale sowie Alpha-Werte von allen FG waren kleiner als 0,5, sodass keine weitere Interpretation sinnvoll ist.

Tabelle 37 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP1 SoSe21

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	70	+	0.320	0.17	7.414	0.363
FG2		+	0.453	0.061	7.676	0.436
FG3		+	0.702	0.353	1.396	0.109
FG4		+	0.783	0.288	1.724	0.185
Test scale					4.552	0.367

Im WiSe21/22 (Tabelle 30) hatte nur die FG2 die akzeptierbare Trennschärfe von 0,37, die FG1, FG3 und FG4 hatten den Wert unter 0,3. Selbst der Wert von FG2 lag jedoch nicht weit von der Grenze entfernt. Das Test-scale sowie Alpha-Werte von allen FG waren kleiner als 0,5, sodass keine weitere Interpretation sinnvoll ist.

Tabelle 38 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP1 WiSe21/22

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	58	+	0.185	0.076	6.375	0.336
FG2		+	0.780	0.368	-0.870	.
FG3		+	0.723	0.151	3.047	0.320
FG4		+	0.396	0.104	5.165	0.310
Test scale					3.429	0.312

5.2.5.2. Trennschärfe bei RP2 (Gesprächsförderung)

Beim RP2 im SoSe21 (Tabelle 31) hatten die FG2, FG4 und FG5 die akzeptierbare Trennschärfe von 0,33, 0,37 und 0,39. Die FG1 und FG3 hatten den Wert unter 0,3. Selbst die Werte von FG2, FG4 und FG5 lagen jedoch nicht weit von der Grenze entfernt. Das Test-scale sowie Alpha-Werte von allen FG waren kleiner als 0,5, sodass keine weitere Interpretation sinnvoll ist.

Tabelle 39 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP2 SoSe21

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	59	+	0.339	0.165	7.980	0.485
FG2		+	0.741	0.329	3.641	0.399
FG3		-	0.459	0.154	7.220	0.494
FG4		+	0.695	0.371	3.715	0.343
FG5		+	0.576	0.390	5.523	0.388
Test scale					5.616	0.486

Im WiSe21/22 (Tabelle 32) hatten die FG1, FG2, FG4 und FG5 die akzeptierbare Trennschärfe von 0,39, 0,43 0,56 und 0,45. Die FG3 hatte den Wert unter 0,3. Selbst der Wert von FG1 lag jedoch nicht weit von der Grenze entfernt. Das Test-scale war 0,62. Es gab kein Alpha-Wert von FG beobachtet, das mit großem Abstand höher als Test-scale war.

Tabelle 40 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP2 WiSe21/22

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	57	+	0.568	0.390	11.136	0.574
FG2		+	0.761	0.431	7.041	0.562
FG3		-	0.511	0.228	12.061	0.637
FG4		+	0.776	0.555	6.417	0.460
FG5		+	0.579	0.450	11.629	0.574
Test scale					9.657	0.621

5.2.5.3. Trennschärfe bei RP3 (Informationsvermittlung)

Beim RP2 im SoSe21 (Tabelle 33) hatten die FG3 und FG5 die akzeptierbare Trennschärfe von 0,32 und 0,35. Die FG1, FG2 und FG4 hatten den Wert unter 0,3. Selbst die Werte von FG3 und FG5 lagen jedoch nicht weit von der Grenze entfernt. Das Test-scale sowie Alpha-Werte von allen FG waren kleiner als 0,5, sodass keine weitere Interpretation sinnvoll ist.

Tabelle 41 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP3 SoSe21

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	57	+	0.348	0.173	6.031	0.409
FG2		+	0.618	0.234	3.822	0.359

FG3		+	0.706	0.317	2.522	0.278
FG4		+	0.440	0.080	6.195	0.469
FG5		+	0.599	0.351	3.446	0.290
Test scale					4.403	0.423

Im WiSe21/22 (Tabelle 34) hatten die FG2 und FG4 die akzeptierbare Trennschärfe von 0,35 und 0,48. Die FG1, FG3 und FG5 hatten den Wert unter 0,3. Selbst der Wert von FG2 lag jedoch nicht weit von der Grenze entfernt. Das Test-scale sowie Alpha-Werte von allen FG waren kleiner als 0,5, sodass keine weitere Interpretation sinnvoll ist.

Tabelle 42 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP3 WiSe21/22

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	63	+	0.333	0.175	7.934	0.474
FG2		+	0.759	0.345	3.230	0.368
FG3		+	0.570	0.181	6.196	0.487
FG4		+	0.724	0.481	2.913	0.258
FG5		+	0.381	0.233	7.562	0.459
Test scale					5.567	0.477

5.2.5.4. Trennschärfe bei RP4 (Partizipative Entscheidung)

Beim RP4 im SoSe21 (Tabelle 35) hatten alle FG die akzeptierbare Trennschärfe von 0,63, 0,49 und 0,32. Der Wert von FG3 war aber nicht weit entfernt von der Grenze. Das Test-scale sowie Alpha-Werte von allen FG waren kleiner als 0,5, sodass keine weitere Interpretation sinnvoll ist.

Tabelle 43 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP4 SoSe21

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	57	+	0.665	0.628	20.238	0.201
FG2		+	0.987	0.488	0.932	0.246
FG3		+	0.460	0.317	15.327	0.160
Test scale					12.166	0.250

Im WiSe21/22 (Tabelle 36) hatten alle FG die akzeptierbare Trennschärfe von 0,58, 0,40 und 0,40. Der Wert von FG2 war aber nicht weit entfernt von der Grenze. Der Alpha-Wert von FG2 war 0,87, also extrem höher als das Test-scale. Allerding das Test-scale sowie Alpha-Werte von FG1 sowie FG3 waren kleiner als 0,5, sodass keine weitere Interpretation sinnvoll ist.

Tabelle 44 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP4 WiSe21/22

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	45	+	0.639	0.577	18.409	0.206

FG2		+	0.975	0.398	5.114	0.865
FG3		+	0.530	0.404	15.783	0.182
Test scale					13.102	0.292

5.2.5.5. Trennschärfe bei RP5 (Compliance)

Beim RP5 im SoSe21 (Tabelle 37) hatten die FG3, FG4 und FG5 die akzeptierbare Trennschärfe von 0,57, 0,43 und 0,32. Die FG1, FG2 und FG6 hatten den Wert unter 0,3. Selbst der Wert von FG5 lag jedoch nicht weit von der Grenze entfernt. Das Test-scale war 0,56. Es gab kein Alpha-Wert von FG beobachtet, das mit großem Abstand höher als Test-scale war.

Tabelle 45 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP5 SoSe21

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	58	+	0.236	0.100	6.770	0.574
FG2		+	0.425	0.234	5.790	0.540
FG3		+	0.820	0.569	1.991	0.332
FG4		+	0.767	0.429	2.832	0.453
FG5		+	0.539	0.322	4.949	0.505
FG6		+	0.322	0.201	6.419	0.555
Test scale					4.792	0.558

Im WiSe21/22 (Tabelle 38) hatten alle FG die akzeptierbare Trennschärfe von 0,50, 0,57, 0,69, 0,59, 0,33 und 0,56. Der Wert von FG5 war aber nicht weit entfernt von der Grenze. Das Test-scale war 0,77. Es gab kein Alpha-Wert von FG beobachtet, das mit großem Abstand höher als Test-scale war.

Tabelle 46 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP5 WiSe21/22

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	44	+	0.586	0.504	17.193	0.761
FG2		+	0.714	0.573	13.587	0.725
FG3		+	0.836	0.690	10.163	0.687
FG4		+	0.784	0.594	11.072	0.724
FG5		+	0.509	0.329	16.616	0.779
FG6		+	0.700	0.560	13.922	0.729
Test scale					13.759	0.772

5.2.5.6. Trennschärfe bei RP6 (Motivationales Interview)

Beim RP6 (Tabelle 39) im SoSe21 hatte nur die FG4 die akzeptierbare Trennschärfe von 0,50. Die FG1, FG2, FG3 und FG5 hatten den Wert unter 0,3. Das Test-scale sowie Alpha-Werte von allen FG waren kleiner als 0,5, sodass keine weitere Interpretation sinnvoll ist.

Tabelle 47 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP6 SoSe21

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	55	+	0.178	0.086	6.317	0.294
FG2		+	0.204	0.010	6.668	0.316
FG3		+	0.842	0.162	2.861	0.435
FG4		+	0.722	0.496	-0.797	.
FG5		+	0.335	0.122	5.292	0.267
Test scale					4.068	0.289

Im WiSe21/22 (Tabelle 40) hatten alle FG die akzeptierbare Trennschärfe von 0,40, 0,49, 0,51, 0,51 und 0,56. Der Wert von FG1 war aber nicht weit entfernt von der Grenze. Das Test-scale war 0,63. Es gab kein Alpha-Wert von FG beobachtet, das mit großem Abstand höher als Test-scale war.

Tabelle 48 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP6 WiSe21/22

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	52	+	0.508	0.398	17.906	0.613
FG2		+	0.628	0.491	15.006	0.568
FG3		+	0.873	0.511	7.402	0.694
FG4		+	0.686	0.512	12.665	0.533
FG5		+	0.682	0.556	14.013	0.547
Test scale					13.398	0.632

5.2.5.7. Trennschärfe bei RP7 (Stressreaktion)

Beim RP7 im SoSe21 (Tabelle 41) hatten die FG4 und FG5 die akzeptierbare Trennschärfe von 0,44 und 0,43. Die FG1, FG2 und FG3 hatten den Wert unter 0,3. Das Test-scale sowie Alpha-Werte von allen FG waren kleiner als 0,5, sodass keine weitere Interpretation sinnvoll ist.

Tabelle 49 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP7 SoSe21

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	38	-	0.242	0.078	5.026	0.423
FG2		+	0.377	0.136	4.386	0.400
FG3		+	0.519	0.094	4.267	0.462
FG4		+	0.838	0.440	0.350	0.076
FG5		+	0.623	0.425	2.312	0.247
Test scale					3.268	0.408

Im WiSe21/22 (Tabelle 42) hatten die FG2, FG3 und FG4 die akzeptierbare Trennschärfe von 0,53, 0,74 und 0,63. Die FG1 und FG5 hatten den Wert unter 0,3. Das Test-scale war 0,65. Es gab kein Alpha-Wert von FG beobachtet, das mit großem Abstand höher als Test-scale war.

Tabelle 50 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP7 WiSe21/22

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	42	+	0.205	0.085	17.119	0.687
FG2		+	0.656	0.532	11.871	0.579
FG3		+	0.910	0.740	3.540	0.362
FG4		+	0.866	0.629	4.953	0.464
FG5		+	0.343	0.231	15.972	0.663
Test scale					10.691	0.647

5.2.5.8. Trennschärfe bei RP8 (Krebsaufklärung)

Beim RP8 im SoSe21 (Tabelle 43) hatte nur die FG6 die akzeptierbare Trennschärfe von 0,36. Die FG1, FG2, FG3, FG4 und FG5 hatten den Wert unter 0,3. Selbst der Wert von FG6 lag jedoch nicht weit von der Grenze entfernt. Das Test-scale sowie Alpha-Werte von allen FG waren kleiner als 0,5, sodass keine weitere Interpretation sinnvoll ist.

Tabelle 51 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP8 SoSe21

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	36	-	0.403	0.172	2.958	0.430
FG2		+	0.517	0.223	2.478	0.404
FG3		+	0.488	0.213	2.601	0.409
FG4		+	0.647	0.274	1.871	0.374
FG5		-	0.256	0.059	3.506	0.469
FG6		+	0.675	0.364	1.514	0.302
Test scale					2.488	0.450

Im WiSe21/22 (Tabelle 44) hatten die FG1, FG2, FG3 und FG4 die akzeptierbare Trennschärfe von 0,34, 0,43, 0,42 und 0,47. Die FG5 und FG6 hatten den Wert unter 0,3. Der Wert von FG1 war aber nicht weit entfernt von der Grenze. Das Test-scale war 0,59. Es gab kein Alpha-Wert von FG beobachtet, das mit großem Abstand höher als Test-scale war.

Tabelle 52 Trennschärfe der Checklisten Items (Cronbach's Alpha) RP8 WiSe21/22

Item	Obs	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
FG1	44	+	0.529	0.339	5.196	0.549
FG2		+	0.728	0.432	3.364	0.500
FG3		+	0.597	0.420	4.738	0.521
FG4		+	0.778	0.473	2.857	0.483
FG5		-	0.237	0.152	6.794	0.605
FG6		+	0.442	0.251	5.761	0.577
Test scale					4.785	0.592

Von insgesamt 78 Ergebnissen hatten nur 44 Trennschärfe den Wert über 0,3, aber davon 16 waren grenzwertig. Vier Items waren über 0,6, aber keins hatte der Wert über 0,8. Mit dem Wert unter 0,3 waren 34 Items inakzeptabel.

Die sechs Tests von insgesamt 16 hatten Cronbach’s Alpha über 0,5. Aber die Werte waren nicht über 0,8, was man nicht als gut beurteilt werden konnte. Die anderen zehn Ergebnisse waren unter 0,5, also inakzeptabel.

5.3. Ergebnisse für Frage 2: Ersetzbarkeit durch DiFS bei RP für Arzt-Patient-Kommunikation mit PaFS

Mit dem Mann-Whitney-U-Test wurden die Daten im SoSe21 und im WiSe21/22 verglichen. Die Ergebnisse können Hinweise darauf geben, ob und inwieweit das Rückmeldesystem einen Einfluss auf die RP-Bewertungen beim RP hat. Auch die Daten aus SP wurden analysiert, sie waren in den beiden Semestern mit PaFS.

5.3.1. Ersetzbarkeit durch DiFS bei RP1 (Anamnese)

Bei den Checkliste-Bewertungen sowie Fragebogen-Bewertungen für RP1 (Tabelle 45) konnte kein signifikanter Unterschied zwischen beiden Semestern festgestellt werden.

Aus DOZ wurden die 70 Bewertungen der GLT im SoSe21 mit den 58 Bewertungen im WiSe21/22 analysiert. Der P-Wert war 0,95 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 69 Bewertungen der GED im SoSe21 wurden mit den 58 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,25 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus STU wurden die 256 Bewertungen der GLT im SoSe21 mit den 121 im WiSe21/22 analysiert. Der P-Wert war 0,49 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 250 Bewertungen der GED im SoSe21 wurden mit den 121 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,08 > 0,05 größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus SP wurden die 68 Bewertungen der GPT im SoSe21 mit den 61 im WiSe21/22 analysiert. Der P-Wert war 0,45 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 64 Bewertungen der GNT im SoSe21 wurden mit den 61 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,34 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Die SP verwendeten in den beiden Semestern PaFS, und zeigten keinen signifikanten Unterschied zwischen den beiden Semestern. Die Dozierenden und Studierenden verwendeten das PaFS im SoSe21 und das DiFS im WiSe21/22. Nichtsdestotrotz konnten wir keinen signifikanten Unterschied zwischen beiden Semestern feststellen. Deshalb lässt es sich annehmen, dass das Rückmeldesystem keinen wesentlichen Einfluss auf die RP-Bewertung beim RP1 hatte.

Tabelle 53 Korrelationen der Bewertungen zwischen SoSe21 und WiSe21/22 (Mann-Whitney-U-Test) RP1

Stichprobe	Bewertung	Beobachtungen		Z	Prob > |z|	Exact prob
		SoSe21	WiSe21/22			
DOZ	GLT	70	58	-0.070	0.9439	0.9456
	GED	69	58	-1.130	0.2583	0.2542
STU	GLT	256	121	0.691	0.4896	0.4899
	GED	250	121	-1.748	0.0805	0.0809
SP	GPT	68	61	-0.754	0.4508	0.4525
	GNT	64	61	-0.956	0.3390	0.3392

5.3.2. Ersetzbarkeit durch DiFS bei RP2 (Gesprächsförderung)

Bei den Checkliste-Bewertungen sowie Fragebogen-Bewertungen für RP2 (Tabelle 46) konnte kein signifikanter Unterschied zwischen beiden Semestern festgestellt werden.

Aus DOZ wurden die 59 Bewertungen der GLT im SoSe21 mit den 57 im WiSe21/22 analysiert. Der P-Wert war 0,21 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 58 Bewertungen der GED im SoSe21 wurden mit den 57 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,60 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus STU wurden die 216 Bewertungen der GLT im SoSe21 mit den 120 im WiSe21/22 analysiert. Der P-Wert war 0,11 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 212 Bewertungen der GED im SoSe21 wurden mit den 117 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,06 > 0,05 größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus SP wurden die 57 Bewertungen der GPT im SoSe21 mit den 59 im WiSe21/22 analysiert. Der P-Wert war 0,59 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 50 Bewertungen der GNT im SoSe21 wurden mit den 58 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,45 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Die SP verwendeten in den beiden Semestern PaFS, und zeigten keinen signifikanten Unterschied zwischen den beiden Semestern. Die Dozierenden und Studierenden verwendeten das PaFS im SoSe21 und das DiFS im WiSe21/22. Nichtsdestotrotz konnten wir keinen signifikanten Unterschied zwischen beiden Semestern feststellen. Deshalb lässt es sich annehmen, dass das Rückmeldesystem keinen wesentlichen Einfluss auf die RP-Bewertung beim RP2 keinen wesentlichen Einfluss hatte.

Tabelle 54 Korrelationen der Bewertungen zwischen SoSe21 und WiSe21/22 (Mann-Whitney-U-Test) RP2

Stichprobe	Bewertung	Beobachtungen		Z	Prob > |z|	Exact prob
		SoSe21	WiSe21/22			
DOZ	GLT	59	57	1.259	0.2080	0.2093
	GED	58	57	-0.528	0.5972	0.6006
STU	GLT	216	120	1.600	0.1096	0.1098
	GED	212	117	-1.869	0.0616	0.0623
SP	GPT	57	59	-0.548	0.5835	0.5857
	GNT	50	58	-0.787	0.4314	0.4473

5.3.3. Ersetzbarkeit durch DiFS bei RP3 (Informationsvermittlung)

Bei den Checkliste-Bewertungen sowie Fragebogen-Bewertungen für RP3 (Tabelle 47) konnte nur ein signifikanter Unterschied von sechs Analysen zwischen beiden Semestern festgestellt werden.

Aus DOZ wurden die 57 Bewertungen der GLT im SoSe21 mit den 63 im WiSe21/22 analysiert. Der P-Wert war 0,12 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 55 Bewertungen der GED im SoSe21 wurden mit den 62 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,31 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus STU wurden die 212 Bewertungen der GLT im SoSe21 mit den 130 im WiSe21/22 analysiert. Der P-Wert war 0,01 < 0,05 kleiner als 0,05, sodass ein signifikanter Unterschied mit 5% Irrtumswahrscheinlichkeit besteht. Die 205 Bewertungen der GED im SoSe21 wurden mit den 127 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,15 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus SP wurden die 56 Bewertungen der GPT im SoSe21 mit den 63 im WiSe21/22 analysiert. Der P-Wert war 0,19 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 52 Bewertungen der GNT im SoSe21 wurden mit den 61 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,07 > 0,05 größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Die SP verwendeten in den beiden Semestern PaFS, und zeigten keinen signifikanten Unterschied zwischen den beiden Semestern. Die Dozierenden und Studierenden verwendeten das PaFS im SoSe21 und das DiFS im WiSe21/22. Bei den GLT aus STU wurde ein signifikanter Unterschied zwischen den beiden Semestern nachgewiesen. Aber bei den GLT und GED aus DOZ bzw. den GED aus STU konnten wir keinen signifikanten Unterschied zwischen beiden Semestern feststellen. Deshalb lässt es sich annehmen, dass das Rückmeldesystem keinen wesentlichen Einfluss auf die RP-Bewertung beim RP3 hatte.

Tabelle 55 Korrelationen der Bewertungen zwischen SoSe21 und WiSe21/22 (Mann-Whitney-U-Test) RP3

Stichprobe	Bewertung	Beobachtungen		Z	Prob > |z|	Exact prob
		SoSe21	WiSe21/22			
DOZ	GLT	57	63	1.551	0.1208	0.1214
	GED	55	62	-1.018	0.3086	0.3093
STU	GLT	212	130	2.703	0.0069	0.0068
	GED	205	127	-1.456	0.1454	0.1461
SP	GPT	56	63	-1.306	0.1916	0.1927
	GNT	52	61	-1.824	0.0681	0.0747

5.3.4. Ersetzbarkeit durch DiFS bei RP4 (Partizipative Entscheidung)

Bei den Checkliste-Bewertungen sowie Fragebogen-Bewertungen für RP4 (Tabelle 48) konnte nur ein signifikanter Unterschied von sechs Analysen zwischen beiden Semestern festgestellt werden.

Aus DOZ wurden die 57 Bewertungen der GLT im SoSe21 mit den 45 im WiSe21/22 analysiert. Der P-Wert war 0,16 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 57 Bewertungen der GED im SoSe21 wurden mit den 44 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,37 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus STU wurden die 231 Bewertungen der GLT im SoSe21 mit den 67 im WiSe21/22 analysiert. Der P-Wert war 0,90 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 224 Bewertungen der GED im SoSe21 wurden mit den 67 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,02 < 0,05 kleiner als 0,05, sodass ein signifikanter Unterschied mit 5% Irrtumswahrscheinlichkeit besteht.

Aus SP wurden die 58 Bewertungen der GPT im SoSe21 mit den 51 im WiSe21/22 analysiert. Der P-Wert war 0,64 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten

Unterschied bestand. Die 58 Bewertungen der GNT im SoSe21 wurden mit den 53 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,93 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Die SP verwendeten in den beiden Semestern PaFS, und zeigten keinen signifikanten Unterschied zwischen den beiden Semestern. Die Dozierenden und Studierenden verwendeten das PaFS im SoSe21 und das DiFS im WiSe21/22. Bei den GED aus STU wurde ein signifikanter Unterschied zwischen den beiden Semestern nachgewiesen. Aber bei den GLT und GED aus DOZ bzw. den GLT aus STU konnten wir keinen signifikanten Unterschied zwischen beiden Semestern feststellen. Deshalb lässt es sich annehmen, dass das Rückmeldesystem keinen wesentlichen Einfluss auf die RP-Bewertung beim RP4 hatte.

Tabelle 56 Korrelationen der Bewertungen zwischen SoSe21 und WiSe21/22 (Mann-Whitney-U-Test) RP4

Stichprobe	Bewertung	Beobachtungen		Z	Prob > |z|	Exact prob
		SoSe21	WiSe21/22			
DOZ	GLT	57	45	-1.422	0.1549	0.1558
	GED	57	44	-0.892	0.3725	0.3728
STU	GLT	231	67	-0.125	0.9006	0.9031
	GED	224	67	-2.446	0.0144	0.0154
SP	GPT	58	51	-0.472	0.6368	0.6390
	GNT	58	53	0.079	0.9370	0.9347

5.3.5. Ersetzbarkeit durch DiFS bei RP5 (Compliance)

Bei den Checkliste-Bewertungen sowie Fragebogen-Bewertungen für RP5 (Tabelle 49) konnten die Hälfte der sechs Analysen zwischen beiden Semestern festgestellt werden.

Aus DOZ wurden die 58 Bewertungen der GLT im SoSe21 mit den 44 im WiSe21/22 analysiert. Der P-Wert war 0,84 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 58 Bewertungen der GED im SoSe21 wurden mit den 43 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,54 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus STU wurden die 247 Bewertungen der GLT im SoSe21 mit den 81 im WiSe21/22 analysiert. Der P-Wert war 0,01 < 0,05 kleiner als 0,05, sodass ein signifikanter Unterschied mit 5% Irrtumswahrscheinlichkeit besteht. Die 243 Bewertungen der GED im SoSe21 wurden mit den 81 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,01 < 0,05 kleiner als 0,05, sodass ein signifikanter Unterschied mit 5% Irrtumswahrscheinlichkeit besteht.

Aus SP wurden die 63 Bewertungen der GPT im SoSe21 mit den 55 im WiSe21/22 analysiert. Der P-Wert war 0,03 < 0,05 kleiner als 0,05, sodass ein signifikanter Unterschied mit 5% Irrtumswahrscheinlichkeit besteht. Die 61 Bewertungen der GNT im SoSe21 wurden mit den 55 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,11 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Die SP verwendeten in den beiden Semestern PaFS, und zeigten bei GPT einen signifikanten Unterschied zwischen den beiden Semestern. Die Dozierenden und Studierenden verwendeten das PaFS

im SoSe21 und das DiFS im WiSe21/22. Bei den GLT und GED aus STU wurden signifikante Unterschiede zwischen den beiden Semestern nachgewiesen. Aber bei den GLT und GED aus DOZ konnten wir keinen signifikanten Unterschied zwischen beiden Semestern feststellen. Beim RP5 lässt es sich deshalb keinen aussagekräftigen Schluss annehmen, ob das Rückmeldesystem einen wesentlichen Einfluss auf die RP-Bewertung beim RP5 hatte.

Tabelle 57 Korrelationen der Bewertungen zwischen SoSe21 und WiSe21/22 (Mann-Whitney-U-Test) RP5

Stichprobe	Bewertung	Beobachtungen		Z	Prob > |z|	Exact prob
		SoSe21	WiSe21/22			
DOZ	GLT	58	44	0.201	0.8409	0.8428
	GED	58	43	-0.615	0.5386	0.5423
STU	GLT	247	81	2.502	0.0124	0.0123
	GED	233	81	-2.675	0.0075	0.0078
SP	GPT	63	55	-2.115	0.0345	0.0343
	GNT	61	55	-1.586	0.1127	0.1139

5.3.6. Ersetzbarkeit durch DiFS bei RP6 (Motivationales Interview)

Bei den Checkliste-Bewertungen sowie Fragebogen-Bewertungen für RP6 (Tabelle 50) konnte nur ein signifikanter Unterschied von sechs Analysen zwischen beiden Semestern festgestellt werden.

Aus DOZ wurden die 55 Bewertungen der GLT im SoSe21 mit den 52 im WiSe21/22 analysiert. Der P-Wert war 0,44 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 55 Bewertungen der GED im SoSe21 wurden mit den 51 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,02 < 0,05 kleiner als 0,05, sodass ein signifikanter Unterschied mit 5% Irrtumswahrscheinlichkeit besteht.

Aus STU wurden die 221 Bewertungen der GLT im SoSe21 mit den 115 im WiSe21/22 analysiert. Der P-Wert war 0,91 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 210 Bewertungen der GED im SoSe21 wurden mit den 115 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,63 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus SP wurden die 58 Bewertungen der GPT im SoSe21 mit den 53 im WiSe21/22 analysiert. Der P-Wert war 0,60 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 57 Bewertungen der GNT im SoSe21 wurden mit den 53 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,18 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Die SP verwendeten in den beiden Semestern PaFS, und zeigten keinen signifikanten Unterschied zwischen den beiden Semestern. Die Dozierenden und Studierenden verwendeten das PaFS im SoSe21 und das DiFS im WiSe21/22. Bei den GED aus DOZ wurde ein signifikanter Unterschied zwischen den beiden Semestern nachgewiesen. Aber bei den GLT aus DOZ bzw. den GLT und GED aus STU konnten wir keinen signifikanten Unterschied zwischen beiden Semestern feststellen.

Deshalb lässt es sich annehmen, dass das Rückmeldesystem keinen wesentlichen Einfluss auf die RP-Bewertung beim RP6 hatte.

Tabelle 58 Korrelationen der Bewertungen zwischen SoSe21 und WiSe21/22 (Mann-Whitney-U-Test) RP6

Stichprobe	Bewertung	Beobachtungen		Z	Prob > |z|	Exact prob
		SoSe21	WiSe21/22			
DOZ	GLT	55	52	-0.776	0.4378	0.4403
	GED	55	51	2.291	0.0220	0.0217
STU	GLT	221	115	-0.119	0.9054	0.9062
	GED	210	115	0.490	0.6240	0.6259
SP	GPT	58	53	0.524	0.6003	0.6027
	GNT	57	53	1.341	0.1798	0.1775

5.3.7. Ersetzbarkeit durch DiFS bei RP7 (Stressreaktion)

Bei den Checkliste-Bewertungen sowie Fragebogen-Bewertungen für RP7 (Tabelle 51) konnte kein signifikanter Unterschied zwischen beiden Semestern festgestellt werden.

Aus DOZ wurden die 38 Bewertungen der GLT im SoSe21 mit den 42 im WiSe21/22 analysiert. Der P-Wert war 0,65 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 38 Bewertungen der GED im SoSe21 wurden mit den 42 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,48 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus STU wurden die 163 Bewertungen der GLT im SoSe21 mit den 85 im WiSe21/22 analysiert. Der P-Wert war 0,48 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 153 Bewertungen der GED im SoSe21 wurden mit den 85 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,08 > 0,05 größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus SP wurden die 41 Bewertungen der GPT im SoSe21 mit den 42 im WiSe21/22 analysiert. Der P-Wert war 0,75 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 37 Bewertungen der GNT im SoSe21 wurden mit den 42 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,14 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Die SP verwendeten in den beiden Semestern PaFS, und zeigten keinen signifikanten Unterschied zwischen den beiden Semestern. Die Dozierenden und Studierenden verwendeten das PaFS im SoSe21 und das DiFS im WiSe21/22. Nichtsdestotrotz konnten wir keinen signifikanten Unterschied zwischen beiden Semestern feststellen. Deshalb lässt es sich annehmen, dass das Rückmeldesystem keinen wesentlichen Einfluss auf die RP-Bewertung beim RP7 keinen wesentlichen Einfluss hatte.

Tabelle 59 Korrelationen der Bewertungen zwischen SoSe21 und WiSe21/22 (Mann-Whitney-U-Test) RP7

Stichprobe	Bewertung	Beobachtungen		Z	Prob > |z|	Exact prob
		SoSe21	WiSe21/22			
DOZ	GLT	38	42	-0.461	0.6446	0.6480

	GED	38	42	-0.720	0.4714	0.4773
STU	GLT	163	85	0.714	0.4754	0.4763
	GED	153	85	-1.753	0.0797	0.0804
SP	GPT	41	42	0.325	0.7454	0.7483
	GNT	37	42	1.494	0.1351	0.1394

5.3.8. Ersetzbarkeit durch DiFS bei RP8 (Krebsaufklärung)

Bei den Checkliste-Bewertungen sowie Fragebogen-Bewertungen für RP8 (Tabelle 52) konnte nur ein signifikanter Unterschied von sechs Analysen zwischen beiden Semestern festgestellt werden.

Aus DOZ wurden die 36 Bewertungen der GLT im SoSe21 mit den 44 im WiSe21/22 analysiert. Der P-Wert war 0,72 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 36 Bewertungen der GED im SoSe21 wurden mit den 44 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,33 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus STU wurden die 155 Bewertungen der GLT im SoSe21 mit den 81 im WiSe21/22 analysiert. Der P-Wert war 0,01 < 0,05 kleiner als 0,05, sodass ein signifikanter Unterschied mit 5% Irrtumswahrscheinlichkeit besteht. Die 151 Bewertungen der GED im SoSe21 wurden mit den 81 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,15 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Aus SP wurden die 40 Bewertungen der GPT im SoSe21 mit den 44 im WiSe21/22 analysiert. Der P-Wert war 0,82 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand. Die 37 Bewertungen der GNT im SoSe21 wurden mit den 44 im WiSe21/22 analysiert. Der P-Wert ergab sich 0,90 > 0,05 deutlich größer als 0,05, sodass es im 5% Signifikanzniveau keinen signifikanten Unterschied bestand.

Die SP verwendeten in den beiden Semestern PaFS, und zeigten keinen signifikanten Unterschied zwischen den beiden Semestern. Die Dozierenden und Studierenden verwendeten das PaFS im SoSe21 und das DiFS im WiSe21/22. Bei den GLT aus STU wurde ein signifikanter Unterschied zwischen den beiden Semestern nachgewiesen. Aber bei den GLT und GED aus DOZ bzw. den GED aus STU konnten wir keinen signifikanten Unterschied zwischen beiden Semestern feststellen. Deshalb lässt es sich annehmen, dass das Rückmeldesystem keinen wesentlichen Einfluss auf die RP-Bewertung beim RP8 hatte.

Tabelle 60 Korrelationen der Bewertungen zwischen SoSe21 und WiSe21/22 (Mann-Whitney-U-Test) RP8

Stichprobe	Bewertung	Beobachtungen		Z	Prob > |z|	Exact prob
		SoSe21	WiSe21/22			
DOZ	GLT	36	44	-0.366	0.7141	0.7173
	GED	36	44	-0.960	0.3372	0.3295
STU	GLT	155	81	2.566	0.0103	0.0103
	GED	151	81	-1.436	0.1511	0.1527
SP	GPT	40	44	-0.228	0.8198	0.8227
	GNT	37	44	-0.142	0.8870	0.8975

5.3.9. Fehlerrate

In den einzelnen Checkliste-Bewertungen wurden Fehler bei der Bewertung festgestellt, wie z.B. das Übersehen einer Frage oder die fehlende Vergabe der GED. Wir untersuchten daher, ob das Rückmeldesystem einen Einfluss auf die Häufigkeit des Fehlers bei der Checkliste-Bewertung hatte.

Es wurde als ein Fehler erfasst, wenn auf einem Bewertungsbogen keine GED gezeichnet wurde. Die STU im SoSe21 machten 65 Fehler in 1701 Bewertungen, die STU im WiSe21/22 machten aber nur sechs Fehler in 800 Bewertungen. Die Fehlerraten betrugen jeweils 3,82% und 0,75%, zeigten mit p-Wert 0,00 einen deutlich signifikanten Unterschied zwischen beiden Rückmeldesystemen.

Tabelle 61 Fehlerrate bei Bewertungen der STU

	RP1	RP2	RP3	RP4	RP5	RP6	RP7	RP8	Summe	Fehlerrate
SoSe21 PaFS	7	4	8	7	14	11	10	4	65	3,82% (63/1701)
WiSe21/22 DiFS	0	3	3	0	0	0	0	0	6	0,75% (6/800)

Bei DOZ waren die Ergebnisse anders als die von STU. Im SoSe21 gab es nur vier Fehler in 430 Bewertungen, und im WiSe21/22 auch vier in 405 Bewertungen. Die Fehlerraten betrugen jeweils 0,93% und 0,99%, also mit p-Wert 0,27 zeigten keinen signifikanten Unterschied.

Tabelle 62 Fehlerrate bei Bewertungen der DOZ

	RP1	RP2	RP3	RP4	RP5	RP6	RP7	RP8	Summe	Fehlerrate
SoSe21 PaFS	1	1	2	0	0	0	0	0	4	0,93% (4/430)
WiSe21/22 DiFS	0	0	1	1	1	1	0	0	4	0,99% (4/405)

6. Diskussion

Die oben gezeigten Ergebnisse werden hier interpretiert. Zudem werden die Anmerkungen genannt.

6.1. Diskussion für Frage1: Testtheoretische Qualität der Checkliste bei RP für Arzt-Patient-Kommunikation

Um die testtheoretische Qualität der Checklisten zu schätzen, wurden ihre Reliabilität, Objektivität, Validität, Aufgabenschwierigkeit, Trennschärfe und Dimensionalität geprüft.

6.1.1. Reliabilität

Das Cronbach's Alpha, was zur Abschätzung der Reliabilität häufig verwendet wird, war für unsere Datensätze nicht optimal anwendbar, um die Reliabilität einzuschätzen [68]. Es kann zu Unterschätzung der Reliabilität führen, wenn die Trennschärfe der Items schwach ist, genauso wie unsere Ergebnisse [63,66]. Daher wurde die Reliabilität der internen Konsistenz zur Schätzung der Reliabilität herangezogen. Dazu ließen sich die Checklisten-Bewertungen der Dozierenden und der Studierenden in derselben RP-Beobachtung durch die Spearmann-Korrelation vergleichen.

Wie in der Studie von Setyonugroho W. et al. erwähnt wurde, kann die Kombination der Checkliste mit einem Global-Rating-Scales die Konstruktvalidität sowie die Inhaltsvalidität des Testverfahrens verbessern[27]. Die in unserer Studie eingesetzten Checkliste beinhalten sowohl die Checklistefragen, deren Ergebnis als gesamte Leistung (GLT) berechnet wurde, als auch das globale Rating in Note (GED). Anhand dieser Struktur der Checkliste konnten wir eine bessere Validität sicherstellen.

Eine hohe Korrelation zwischen der Leistung (GLT) und der Note (GED) deutet darauf hin, dass die Studierenden umso bessere Note erhalten, je mehr einzelne Aufgaben der Checkliste sie erfolgreich erledigen, also eine gute Leistung aufweisen. Umgekehrt erhalten die Studierenden eine schlechte Note, wenn sie eine mangelnde Leistung erbrachten. Eine niedrige Korrelation zeigt hingegen einen geringen oder keinen Zusammenhang zwischen der Leitung (GLT) und der Note (GED), was auf eine schwache Reliabilität der Checkliste hinweist.

Bei jedem Rollenspiel (RP) bekamen wir vier Ergebnisse, jeweils aus Studierenden bzw. Dozierenden in den beiden Semestern. Alle 32 Ergebnisse zeigten eine gute Korrelation zwischen der Note (GED) und der Leistung (GLT). Also, das globale Rating (GED) korrelierte mit dem gewichteten Summenpunkte (GLT). Somit war die Bewertung mit Checkliste internal konsistent.

6.1.2. Objektivität

Die Kappa Analyse diente zur Abschätzung der Objektivität. Das Conger's Kappa wurde angewendet, weil die Bewertungen von mehr als zwei Ratern analysieren sollten. Zudem bekamen wir das Konfidenzintervall aus dem Conger's Kappa, sodass wir die Ergebnisse ausführlich interpretieren konnten [51,53,55,56].

Theoretisch könnte man durch das Matching zwischen der Checklisten-Bewertung der Dozierenden und dem Mittelwert von denen der Studierenden die zwei-Rater-Analyse mit dem häufig angewendeten Cohen's Kappa bzw. t-Test ausführen [57,58]. Allerdings schwächt die Ermittlung von Mittelwert die Auswirkung der einzelnen Bewertungen der Studierenden auf das Ergebnis ab. Dies

führt möglicherweise zu einer Überschätzung der Objektivität. Daher wurde auf solche Analysen durch das Matching verzichtet [51,53,54].

Die Interpretation des Kappa-Werts im Bereich von 0 bis 0,2 lautet eine “geringe“ Übereinstimmung zwischen den Ratern, von 0,2 bis 0,4 eine “ausreichende”, von 0,4 bis 0,6 eine “mittelmäßige”, von 0,6 bis 0,8 eine „beachtliche“ und von 0,8 bis 1 eine „vollkommene“. Wenn der Wert kleiner als 0 ist, ist es eine geringere Übereinstimmung als Zufall, sodass man schwer interpretieren kann. [51,53]

Das Conger’s Kappa gibt neben den Kappa-Koeffizienten das Konfidenzintervall an, was uns noch strengere Interpretation ermöglicht. Wenn das Konfidenzintervall einen negativen Bereich umfasst, kann man das Ergebnis nicht als zuverlässig gelten lassen, obwohl der Koeffizient auf eine gute Übereinstimmung zwischen den Ratern hinweist.

Bei jedem RP wurde zwei Analysen durchgeführt, einmal nur zwischen den Studierenden (STU), dann Studierende und Dozierende (STU&DOZ). Die Note (GED) sowie die Leistung (GLT) der Checklisten-Bewertungen wurden zwischen den Ratern verglichen. Darüber hinaus wurden die Teilbewertungen (FGs) in der Checkliste auch analysiert.

In den ermittelten Ergebnissen gab es keinen Kappa-Koeffizienten größer als 0,6, der als „beachtliche“ oder „vollkommene“ Übereinstimmung zwischen den Ratern zu interpretieren war. Die Ergebnisse deuteten überwiegend auf eine „geringe“ oder “geringe“ Übereinstimmung hin. Von 32 Kappa-Koeffizienten für GLT gab es zwei negative Werte, die nicht in die Interpretation einbezogen werden konnte. Von 32 Kappa-Koeffizienten für GED gab es keinen negativen Wert. Von 152 Kappa-Koeffizienten für einzelne FG gab es 12 negative Werte. Insgesamt zeigten 202 von 214 Analyse (93,52%) eine „geringe“ bis “mittelmäßige” Übereinstimmung zwischen zwei bis sechs Ratern.

Allerdings wurden die Ergebnisse unter Berücksichtigung des Konfidenzintervalls anders interpretiert. Vielen Werten enthielten einen negativen Wert im 95% Konfidenzintervall. Von den 32 Kappa-Koeffizienten für GLT gab es 20 Werte mit dem Konfidenzintervall im negativen Bereich. Von den 32 Kappa-Koeffizienten für die Note (GED) gab es neun Werte, Von 152 Kappa-Koeffizienten für die Teilbewertungen (FG) gab es sogar 95 Werte, die einen negativen Wert im 95% Konfidenzintervall einschlossen. Somit waren mehr als Hälfte der Ergebnisse „nicht interpretierbar“.

Diese Interpretation mit dem Konfidenzintervall scheint aber zu streng. Tatsächlich wird es immer noch diskutiert, wie den Kappa-Koeffizienten genau interpretieren soll [51,53,56]. Das Konfidenzintervall ist eine Möglichkeit davon, was jedoch noch nicht als Standard für die Kappa-Analyse etabliert ist. Ohne Berücksichtigung des Konfidenzintervalls zeigte die Interpretation mit den Kappa-Koeffizienten allein immerhin eine akzeptable Objektivität. Die Kappa-Koeffizienten zeigten trotz der großen Anzahl der Rater durchschnittlich eine “geringe“ Übereinstimmung zwischen den Ratern. Somit würde die Objektivität der Checkliste bestätigt.

Noch auffällig ist, dass die Kappa-Koeffizienten aus Studierenden und Dozierenden zusammen (STU&DOZ) nicht schlechter als die aus Studierenden allein (STU) waren. Eher zeigten sie tendenziell sogar bessere Ergebnisse, obwohl sie noch einen zusätzlichen Rater in der Analyse hatten. Dies ist ein überraschendes Ergebnis, betrachtend auf die Rechenformel von Conger’s Kappa [51,53]. Das legt die Vermutung nahe, dass die Dozierenden kein Ausreißer für die Interrater-Analyse waren, sondern, dass sich die Dozierenden und die Studierenden eher stark einig waren.

Tendenziell zeigten die Note (GED) noch höheren Kappa-Koeffizienten als die Leistung (GLT). Es könnte daran liegen, dass die Note (GED) den Wert von 1 bis 5 vergeben wurde, während die Leistung (GLT) den Wert von 0 bis 14 hatte, was somit noch breiteres Spektrum von Werten haben konnte.

Trotzdem ist es merkwürdig, weil die Note (GED) eine subjektive Bewertung ist, während die Leistung (GLT) eine Summe von objektiven Bewertungen darstellt. Das bedeutet, dass sich die Rater in ihrem subjektiven Eindruck miteinander einig waren, obwohl sie die einzelnen Aufgaben unterschiedlich bewerteten. Die deskriptiven Statistiken (5.1) zeigten, dass sich die Mittelwerte der Note (GED) fast an „1“ annäherten, und die höchste Säule auf dem Wert „1“ stand. Dies könnte durch den Mildeeffekt erklärt werden, dass die Rater die Neigung haben, bessere Note zu vergeben. Die Studierenden kannten sich schon im Kurs persönlich, was möglicherweise sie beeinträchtigte, ihre Kommilitonin und Kommilitonen streng zu bewerten. Es wurde bereits in einigen Studien als eine mögliche Fehlerquelle genannt [12,16,76].

Wir haben keine Interrater Analyse zwischen den Dozierenden, weil bei jedem Gespräch lediglich eine Dozierende oder ein Dozierender anwesend war. Natt N. et al. nahmen in ihrer Studie die Gespräche in Video auf [36]. Dadurch konnten sie die Interrater-Objektivität analysieren, indem die nicht beim Gespräch anwesenden Rater auch durch das Video die Gespräche bewerteten. In unserer Studie fand aufgrund von Datenschutz keine Video-Aufnahme statt, sodass die nachträglichen mehr-Rater-Bewertungen zwischen den Dozierenden wie bei Natt N. et al. nicht möglich waren. Das ist eine Einschränkung unserer Studie, es wird im 6.3. weiter diskutiert.

6.1.3. Validität

Es gibt noch keinen festen Goldstandard für Validitätsprüfung eines psychometrischen Messverfahrens [27]. Aus diesem Grund konnten wir keine Korrelationsprüfung mit einem bereits als "valid" anerkannten Messverfahren durchführen. Zum Alternativ wurde die Validität mit dem multimodalen Schema nach Messick geprüft. Es besteht aus fünf Komponenten, und zwar Testinhalt, Reaktionsprozess, internale Struktur, Relation zu anderen Variablen und Konsequenz von Test [41,44,70–72]. Somit kann man ausführlich einschätzen, wie „valid“ das Messverfahren ist.

Jedoch bleibt es noch offen, ob durch diese Prüfung die Validität eines Messverfahrens endgültig als „valid“ akzeptiert werden kann. Es liegt aber nicht an der Struktur des Messisks Schemas, sondern eher allgemein an den Eigenschaften des psychometrischen Messverfahrens: Es gibt keinen validen Goldstandard, mit dem die Validität eines neuen Messverfahrens geprüft werden kann. Zudem erfolgen die Messungen meistens durch eine subjektive Bewertung [36,40,69].

Testinhalt: Für den Testinhalt wurde der Entwicklungsprozess der Checkliste qualitativ analysiert. Basierend auf der klaren Zielsetzung wurden die RP-Szenarien und -Checkliste von den Experten erstellt. Unter Berücksichtigung der Zielsetzungen, sich ein biopsychosoziales Krankheitsverständnis sowie Grundkompetenzen patientenorientierter ärztlicher Gesprächsführung zu erarbeiten und für die Situation schwerkranker und sterbender Patienten zu sensibilisieren, wurden entsprechende Fallvignetten entwickelt. Die dazugehörigen Checklisten enthielten umfassende Bewertungskriterien für die Gesprächsqualität, was ebenfalls von denselben Experten entwickelt wurde [26,45]. Diese qualifizierte Entwicklung diente als Nachweis für die Validität im Rahmen von Testinhalt.

Reaktionsprozess: Den Reaktionsprozess konnte man durch den Ablauf der Messung betrachten. Im Abschnitt 4.1. wurden der gesamte Ablauf des Simulationsgesprächs, der Bewertung mit der

Checkliste und des Feedbacks detailliert beschrieben. Dies zeigt, dass die Messung gemäß einem klaren Plan objektiv erfolgte.

Durch die Interrater Korrelation kann man auch den Reaktionsprozess einschätzen, wie im Abschnitt 5.2.2. dargestellt. Trotz der strengen Prüfung mit dem Conger's Kappa galten die Interrater Korrelationen aller acht Checklisten insgesamt als "gering", also waren sie akzeptabel. Somit kann der Reaktionsprozess als objektiv angesehen werden.

Internale Struktur: Für die Prüfung der internalen Struktur wurde die internale Konsistenz ermittelt. Die Ergebnisse sind im Abschnitt 5.2.1. gezeigt. Alle acht Checklisten wiesen eine „gute" Internal-Konsistenz auf. Somit kann die internale Struktur als gut betrachtet werden.

Relation zu anderen Variablen: Zur Relation zu anderen Variablen wurde die Korrelation der Bewertung der Dozierenden mit der Bewertung der Simulationspatienten und -innen ermittelt. Diese statistische Analyse eignet sich zur Abschätzung der Validität im engeren Sinne. Denn es wurde untersucht, ob die Checkliste-Bewertung tatsächlich messen konnte, was eigentlich durch die Checkliste gemessen werden sollte. Die kommunikative Leistung konnte so indirekt eingeschätzt werden, indem man maß, wie sich die Patienten und -innen mit dem Gespräch anfühlten, und wie sie das Gespräch auf der kommunikativen Ebene bewerteten. Deshalb wurde die Korrelation der Bewertungen der Dozierenden (GED) mit denen der Simulationspatienten und -innen (GNT) ermittelt, um die Checkliste-Bewertung mit der Bewertung von den Patienten zu vergleichen.

Die Ergebnisse sind im Abschnitt 5.2.3. gezeigt. Durch die Spearmann-Korrelation bekamen wir insgesamt 16 Koeffizienten bei allen Rollenspielen (RPs) in den beiden Semestern. 13 von 16 Koeffizienten zeigten eine „mittelmäßige" Korrelation zwischen den Bewertungen der Dozierenden und der Simulationspatienten und -innen. Beim RP1 (Anamnese) im WiSe21/22 bestand eine "geringe" Korrelation mit dem Koeffizienten 0,27. Beim RP7 (Stressreaktion) im SoSe21 ergab sich ein "ausreichender" Rangkorrelationskoeffizient 0,32. Aber es wurde mit einem p-Wert von 0,08 angelehnt. Beim RP8 im SoSe21 gab es einen „schwachen" Rangkorrelationskoeffizienten 0,25, ebenfalls wurde mit einem p-Wert von 0,18 abgelehnt. Es scheint auf die zu geringe Anzahl der Beobachtungen zurückzuführen zu sein. RP7 (Stressreaktion) im SoSe21 sowie RP8 (Krebsaufklärung) im SoSe21 hatten wir jeweils nur 31 Beobachtungen. Mit einer kleinen Stichprobe ist es schwierig, ein zuverlässiges Ergebnis herzuleiten. Für die Spearmann-Korrelation wird eine Anzahl der Beobachtungen mindestens 40 empfohlen [66]. Abgesehen von diesen zwei Ausnahmen zeigten die Analysen aber durchschnittlich eine "ausreichende" Korrelation. So kann die Relation zu anderen Variablen als mittelmäßig gelten.

Konsequenz von Test: In unserer Studie wurden keine passenden Daten erhoben, aus denen Konsequenz von Test abgeleitet werden konnten. Durch die wiederholten Gespräche mit zeitlichem Abstand könnten wir eventuell eine Verbesserung ihrer Kommunikationstechnik beobachten haben. Die Studierenden führten tatsächlich zweimal das Gespräch durch, wobei aber aus didaktischen Gründen die zwei Gespräche unterschiedliche Rollenspiel-Thema hatten. Daher konnten wir nicht analysieren, ob die Studierenden beim zweiten Gespräch die verbesserte Leistung erbrachten im Vergleich zum ersten Gespräch. Somit konnten wir die genaue Konsequenz von Test nicht einschätzen.

Es gibt noch ein anderes Vorgehen zur Einschätzung der Konsequenz von Test. In der Studie von Natt N. et al. sowie den Studien, die von Phillips A. et al. analysiert wurden, wurden die Teilnehmer zur Selbsteinschätzung gefragt, über die Verbesserung der kommunikativen Technik sowie Zufriedenheit usw., und diese Umfragen wurden als Konsequenz von Test analysiert [30,36]. Allerdings ist es fraglich, ob die Zufriedenheit der Studierenden tatsächlich für die Verbesserung ihrer Fähigkeiten gleichgesetzt werden kann. Man kann annehmen, dass die Studierenden durch die Übungen etwas Erfahrungen sammeln und Selbstvertrauen aufbauen, dies kann zur Verbesserung ihrer kommunikativen Fähigkeiten führen. Solche Selbsteinschätzung ist jedoch eine subjektive Bewertung, dass ihre kommunikativen Fähigkeiten verbessert worden sind. Dadurch wird aber keine tatsächliche Verbesserung ihrer kommunikativen Fähigkeiten gemessen. Es soll stattdessen die kommunikativen Leistungen der Studierenden bei ärztlichen Tätigkeiten direkt gemessen werden, um die Konsequenz von Test für unser Testverfahren genau einzuschätzen. Die Umstände dafür ist leider sehr aufwendig, sodass in unserer Studie leider nicht realistisch durchgeführt werden konnte [73].

Aus diesen Gründen konnten wir die Konsequenz von Test von unserer Checkliste nicht beurteilen.

Zusammenfassend war der Testinhalt qualifiziert, der Reaktionsprozess war objektiv, die internale Struktur hatte eine gute internale Konsistenz, die Relation zu anderen Variablen war mittelmäßig, aber die Konsequenz von Test konnte nicht geprüft werden. Aus diesen Ergebnissen ließ es sich annehmen, dass die Checkliste als Feedback-System „valid“ ist.

6.1.4. Aufgabenschwierigkeit

Mit den Bewertungen der Dozierenden wurde die Aufgabenschwierigkeit eingeschätzt. Die Punkte einzelner Fragengruppen in der Checkliste (FG) wurden als Item analysiert. Jede Fragengruppen in der Checkliste besteht aus ein bis zehn Checklistefragen, die gemeinsame Aspekte betrachten. Es gibt insgesamt 39 Fragengruppen aus acht Checklisten. Aus den beiden Semestern hatten wir 78 Ergebnisse ermittelt. Die 59 von den 78 Fragengruppen bei allen Rollenspielen waren „zu leicht“, sodass sie nicht zur Differenzierung unterschiedlicher Leistungen geeignet waren. Dies betraf alle enthaltenen Fragengruppen im RP1 (Anamnese), RP4 (Partizipative Entscheidung) und RP8 (Krebsaufklärung). Nur insgesamt 17 von den 78 Fragengruppen hatten die leichte Schwierigkeit. Aber 12 davon hatten einen AS-Index von fast 80%, was doch grenzwertig war. Allein die Fragengruppe 4 beim RP3 (Informationsvermittlung) in den beiden Semestern hatte eine mittlere Schwierigkeit. Da fast alle Aufgaben als zu leicht bzw. leicht beurteilt wurden, konnte man kein breites Spektrum an Aufgabenschwierigkeit beobachten.

Diese durchschnittlich sehr guten Leistungen sind durch den Mildeeffekt zu erklären [12,16,76]. Da die Checkliste-Bewertungen keinen Einfluss auf den Leistungsüberprüfung im Kurs hatten, mussten die Leistungen der Studierenden nicht streng differenziert werden. Dies könnte zu einer Neigung zur besseren Bewertung geführt haben.

Andererseits, es ist durch Schulungseffekt zu erklären: Dank hoher Qualität des Kurses bereiteten sich die Studierenden sehr gut auf die Simulationsgespräche vor. Die Simulationsgespräche fanden nicht spontan ohne Vorbereitung statt, sondern es gab genügend Zeit für Vorbereitungen. Im Laufe des Semesters lernten sie den Lernzielen entsprechend die verschiedenen Kommunikationstechniken, und konnten vor dem angekündigten Termin das Gespräch ausreichend üben. Also, der Kurs war effektiv organisiert, daher war die Lehre des Kurses erfolgreich. Das könnte zu den

durchschnittlich sehr guten Leistungen der Studierenden geführt haben. Um diese Erklärung zu bestätigen, ob die gute Lehre zur guten Qualität der Simulationsgespräche führte, musste der Notenspiegel der Studierenden in diesem Kurs mit ihren Noten für Simulationsgespräche analysiert werden. Da die Daten zu den Klausurergebnissen für diesen Kurs nicht erhoben wurden, konnte dies allerdings nicht weiter beurteilt werden.

6.1.5. Trennschärfe

Die Trennschärfen einzelner Fragengruppen (FG) in der Checkliste wurden durch Cronbach's Alpha analysiert. Wir interpretierten zwei Werte, nämlich die Item-Rest-Korrelation und den Alpha-Wert [46,50].

Die 34 von den 78 ermittelten Item-Rest-Korrelationen waren kleiner als 0,3, sodass diese Items von der Messung ausgeschlossen werden sollten. Die 50 von den 78 ermittelten Alpha-Werten waren kleiner als 0,5, also waren sie nicht akzeptabel.

Beim RP5 (Compliance) im WiSe21/22 und RP6 (Motivationales Interview) im WiSe21/22 gab es keine Fragengruppe, die ausgeschlossen werden sollten. Sie waren "ausreichend" bzw. "gering" gegeneinander ausdifferenzierbar. Auch die Fragengruppe 1, 2 und 4 beim RP2 (Gesprächsförderung) im WiSe21/22, die Fragengruppe 5 beim RP5 (Compliance) im SoSe21, die Fragengruppe 2 beim RP7 (Stressreaktion) im WiSe21/22 und die Fragengruppe 1, 2 und 3 beim RP8 (Krebsaufklärung) im WiSe21/22 hatten die akzeptablen Item-Rest-Korrelationen und auch die akzeptablen Alpha-Werte. Insgesamt durften nur 19 von den 78 Fragengruppen im Testverfahren bleiben, aber sie hatten den Wert knapp über dem Grenzwert.

Die überwiegend inakzeptabel niedrigen Trennschärfen lassen sich auf die "zu leichten" Aufgabenschwierigkeiten zurückführen. Die Mehrheit der Fragengruppen hatte eine Aufgabenschwierigkeit über 80%. So konnten die verschiedenen Fragengruppen in einer Checkliste schwer auseinander differenziert werden, die sich mit unterschiedlichen Aspekten beschäftigten.

6.1.6. Dimensionalität

Beim Simulationsgespräch müssen die Studierenden ihre Kommunikationstechnik anwenden, um die in RP-Szenarien gegebenen Aufgaben zu erledigen. Die Aufgaben umfassen verschiedene Aspekte. Es wird geprüft, ob die Hauptaufgaben behandelt wurden, z.B. bestimmte Informationen oder Diagnose mitzuteilen (RP3 Informationsvermittlung und RP8 Krebsaufklärung), gemeinsam mit Patienten eine Entscheidung zu treffen (RP4 Partizipative Entscheidung) usw. Das lässt sich anhand detaillierter Fragen bewerten, ob der Umgang mit der Aufgabe, Erklärung, Begründung, Empfehlung und das Zuhören angemessen war. Es wird aber nicht nur in einem inhaltlichen Aspekt geprüft. Den Studierenden ist gefordert, das Gespräch mit den Patientinnen und Patienten möglichst reibungslos und angenehm zu gestalten. Daher wird neben dem Hauptthema des Gesprächs auch die Qualität des Gesprächs bewertet, so die Eröffnung des Gesprächs, Einstieg ins Thema und Non-Verbal Kommunikationen, z.B. Körperposition, Körpersprache, Blickkontakt usw. All diese Aspekte sind in der Checkliste aufgeführt, sodass der Bewerter die einzelnen Kriterien nicht übersieht.

Die Checkliste besteht aus mehreren Fragengruppen, die sich entweder mit Erfüllung der Aufgabe oder mit der Qualität der Kommunikation befassen. Diese beiden Komponenten gehören gemeinsam

zur Kommunikationsfähigkeit bei ärztlichen Tätigkeiten. So handelt es sich eine eindimensionale Prüfung, und das Testverfahren hat eine homogene Dimensionalität.

Die testtheoretische Qualität der Checklisten von Fischbeck et al. für die Simulationsgespräche wurde so nachgewiesen: gute Internal-Konsistenz, leichte Interrater Objektivität, akzeptable Validität, zu leichte Aufgabenschwierigkeit, zu schwache Trennschärfe und homogene Dimensionalität. Diese Ergebnisse wurden in den beiden Semestern unabhängig von dem Rückmeldesystem stabil beobachtet.

6.2. Diskussion für Frage2: Ersetzbarkeit durch DiFS bei RP für Arzt-Patient-Kommunikation mit PaFS

Die Digitalisierung ermöglicht erheblich verbesserte Effizienz beim Informationshandeln [30,31,42]. Dadurch wird der Einsatz des Digitalsystems erweitert, sodass jetzt kaum zu finden ist, wo man immer noch nur manuell und analog arbeitet. In allen Bereichen wurde schon ein digitales System eingeführt und es wird weiterhin versucht die Effizienz zu verbessern sowie den Anwendungsbereich zu erweitern. Für die Arbeit mit psychometrischem Messverfahren gilt dies auch. Die konventionellen Fragebögen in Papierform werden in ein digitales System überführt. Dadurch kann man schnell und korrekt die Daten erfassen, integrieren, verarbeiten und statistisch analysieren. Aus diesem Grund versuchten wir in dieser Studie die Einsetzbarkeit eines digitalbasierten Rückmeldesystem (EvaSys) zur Beurteilung der kommunikativen Kompetenz bei Medizinstudierenden zu prüfen.

Der Unterschied zwischen den verschiedenen Rückmeldesystemen kann möglicherweise ein Störfaktor für das Messverfahren sein [30,31,42]. Mit einem vertrauten und gewöhnlichen System kann sich der Prüfer besser auf die Bewertung konzentrieren. Im Gegensatz dazu kann ein ungewohntes und kompliziertes System die Konzentration des Prüfers stören, sodass sich wichtige Bewertungskriterien eventuell übersehen lassen. So sollte es geprüft werden, ob ein neu einzusetzendes Rückmeldesystem genauso zuverlässige Ergebnisse liefern kann, wie das bereits eingesetzte System [29–31,42]. In unserer Studie wurde das digitale Rückmeldesystem EvaSys (DiFS) mit der papierbasierten Checkliste (PaFS) zur Bewertung des Simulationsgesprächs im Rahmen des Kurses (Teil II) der Medizinischen Psychologie und Medizinischen Soziologie verglichen.

In den beiden darauffolgenden Semestern wurden die Checkliste-Bewertungen jeweils mit dem papierbasierten Rückmeldesystem (PaFS, SoSe21) bzw. dem digitalen Rückmeldesystem (DiFS, WiSe21/22) erhoben. Da die Ergebnisse nicht normalverteilt waren, wurden diese beiden Datensätze durch Mann-Whitney-U-Test analysiert (s. 5.1.1) [18,46,62]. Die Ergebnisse sind unter 5.3. dargestellt.

Die gesamten Punkte in Checkliste (GLT) der Dozierenden (DOZ) zeigte in allen acht Rollenspielen keinen signifikanten Unterschied zwischen beiden Rückmeldesystem (PaFS und DiFS). Die gesamten Punkte in Checkliste (GLT) der Studierenden (STU) hatte in RP3 (Informationsvermittlung), RP5 (Compliance) und RP8 (Krebsaufklärung) einen signifikanten Unterschied, in den restlichen fünf Rollenspielen gab es keinen signifikanten Unterschied.

Die gesamte Note in der Checkliste (GED) der Dozierenden (DOZ) hatte nur in RP6 (Motivationales Interview) einen signifikanten Unterschied, in den anderen sieben Rollenspielen wurde kein signifikanter Unterschied beobachtet. Die gesamte Note in der Checkliste (GED) der Studierenden (STU)

hatte in RP4 (Partizipative Entscheidung) und RP5 (Compliance) einen signifikanten Unterschied, aber in den anderen sechs Rollenspielen zeigte keinen signifikanten Unterschied.

Dozierenden hatte von insgesamt 16 Ergebnissen nur einmal einen signifikanten Unterschied. Studierenden hatte von insgesamt 16 Ergebnissen fünfmal einen signifikanten Unterschied.

Die 26 von insgesamt 32 Ergebnissen konnten wir keinen signifikanten Unterschied zwischen den beiden Rückmeldesystem beobachten. So können wir annehmen, dass das Rückmeldesystem (PaFS und DiFS) keinen signifikanten Einfluss auf die Checkliste-Bewertung zum Simulationsgespräch für die kommunikative Kompetenz ausübte. Dies entspricht den Ergebnissen von Phillips A. et al., Daniels V. et al., Markland A. et al. und Li J. et al. [29,30,32,43].

Neben diesen direkten statistischen Analysen untersuchten wir auch die Bewertungen der Simulationspatienten und -innen (SP), um die zufällige Übereinstimmung beider Rückmeldesysteme auszuschließen. Sie wurden in den beiden Semestern auf dem Papierbogen bearbeitet. Daher in diesem Vergleich von beiden Semestern war der Einfluss unterschiedlicher Rückmeldesysteme ausgeschaltet. Durch den Vergleich der Analyse aus Simulationspatienten und -innen (SP) mit den Analysen aus Dozierenden (DOZ) und Studierenden (STU) konnte der Einfluss des Rückmeldesystems quantifiziert werden. Bei Simulationspatienten und -innen (SP) wurde nur einmal im RP5 (Compliance) bei gesamten Punkten im Fragebogen (GPT) ein signifikanter Unterschied beobachtet. Die gesamten Punkte in Fragebogen (GPT) in den restlichen sieben Rollenspielen und die gesamte Note in Fragebogen (GNT) in allen acht Rollenspielen zeigten keinen signifikanten Unterschied. Dieses Ergebnis entspricht dem Ergebnis von Dozierenden (DOZ). Dies verstärkt die Annahme, dass die Übereinstimmung der Checkliste-Bewertung zwischen beiden Rückmeldesystemen (PaFS und DiFS) nicht durch einen Zufall beobachtet wurde.

Wenn das digitale Rückmeldesystem (DiFS) die Checkliste-Bewertung nicht beeinträchtigt, und zuverlässig das gleiche Ergebnis liefern kann, wie das papierbasierte Rückmeldesystem (PaFS), ist der Einsatz des digitalen Rückmeldesystem (DiFS) klar zu empfehlen. Diese digitale Arbeit bietet den Bewertern eine praktische Benutzerschnittstelle [27]. Dies führt zur höheren Zufriedenheit der Bewerter, die eine deutliche Präferenz für das digitale Rückmeldesystem (DiFS) gegenüber dem papierbasierten Rückmeldesystem (PaFS) haben [27,30,31,42]. Zudem ist es auch sehr praktisch die Daten zu bearbeiten bzw. statistisch zu analysieren. Durch Minimierung manueller Datenarbeit verringert das digitale Rückmeldesystem (DiFS) die Fehlerquote erheblich. Diese Vorteile des digitalen Rückmeldesystem (DiFS) wurden sowohl in unserer Studie als auch in der Studie von Felicitas L. et al. und von Schmitz F. et al. bestätigt [31,42].

Um diese Vorteile des digitalen Rückmeldesystem (DiFS) zu quantifizieren, beschäftigten wir uns mit der Fehlerrate beider Rückmeldesysteme. Es gab einige Bewertungen, bei denen einzelne Fragen der Checkliste angekreuzt aber keine Note für die Gesamtnote (GED) vergeben wurde. Aus diesen Bewertungen konnte man nur die Gesamtpunkte (GLT) ermitteln aber keine Daten für die Gesamtnote (GED) erheben. Daher bekamen wir eine unterschiedliche Anzahl der Beobachtungen von Gesamtpunkte (GLT) und Gesamtnote (GED) (s. Tabelle 3). Wenn alle Bewerter in der Checkliste nicht übersehen hätten, die Gesamtnote (GED) zu bewerten, sollten die Anzahl der Beobachtungen von Gesamtpunkte (GLT) und die von Gesamtnote (GED) gleich sein. Dieser Fehler wurde aufgezählt und daraus die Fehlerquote ermittelt. Die Ergebnisse waren im Abschnitt 5.3.9. gezeigt.

Die Dozierenden hatten unabhängig vom Rückmeldesystem eine sehr geringe Anzahl von Fehlern (4/430 mit PaFS, 4/405 mit DiFS). Die Studierenden hatten aber mit dem papierbasierten Rückmeldesystem (PaFS) eine deutlich höhere Anzahl von Fehlern (63/1701). Mit dem digitalen Rückmeldesystem (DiFS) hatten die Studierenden doch sehr geringe Anzahl von Fehlern wie die Dozierenden

(6/800). Während die Dozierenden mit den beiden Rückmeldesystemen fast identische Fehlerraten hatten, hatten die Studierenden mit dem papierbasierten Rückmeldesystem (PaFS) die ca. fünffach höhere Fehlerrate als mit dem digitalen Rückmeldesystem (DiFS) (3,82% mit PaFS, 0,75% mit DiFS). In der Studie von Phillips A. et al. und von Schmitz F. et al. wurde ebenfalls eine signifikant unterschiedliche Fehlerrate zwischen den beiden Rückmeldesystemen beobachtet [30,42]. Die Dozierenden hatten bereits in ihrer bisherigen Berufstätigkeit ausreichend Erfahrung mit den beiden Rückmeldesystemen, sodass sie unabhängig von dem Rückmeldesystem selten Fehler machten. Für die Studierenden sollten die Bewertungen mit beiden Rückmeldesystemen keine routinemäßige Arbeit sein, sodass sie bei der Bewertung abhängig von Konvenienz und Bedienbarkeit des Rückmeldesystems einen Fehler machen können. Daher zeigten sich bei den Checklisten-Bewertungen der Studierenden die deutlich unterschiedlichen Fehlerquoten zwischen den beiden Rückmeldesystemen nach ihrer Konvenienz. Diese Analyse zeigte auf, dass das digitale Rückmeldesystem (DiFS) ein benutzerfreundlicheres und praktischeres System ist als das papierbasierte Rückmeldesystem (PaFS).

6.3. Einschränkungen

In unserer Studie konnten wir relativ große Stichproben rekrutieren und verschiedene Daten sammeln. Dies ermöglichte uns eine umfangreiche Analyse. Trotzdem gab es im Laufe der Untersuchung und der Analyse einige Einschränkungen.

Zur Homogenitätsprüfung beider Stichproben der Studierenden wurden die Verteilungen des Alters und des Geschlechts analysiert. Diese Daten wurden aus den Verwaltungsakten entnommen. Die persönlichen Daten von bewertenden Studierenden wurden bei der Checkliste-Bewertung nicht erfasst. Informationen über medizinische Vorbildungen (Einsatzbereich, Qualifikation, Zeitraum usw.) könnten auch die Stichprobe in unserer Studie gut beschreiben. Vor allem bei den Studierenden, die bereits Erfahrungen mit direkten Patientenkontakten in ihrem Berufsfeld hatten, könnte sich dies auf ihr Niveau der kommunikativen Kompetenz ausgewirkt haben. Das kann direkt auf unsere Ergebnisse beeinflussen. Daher empfehlen wir für eine bessere Qualität zukünftiger Untersuchungen, die Daten über medizinische Vorbildungen der Medizinstudierenden zu erheben.

Eine Video-Aufnahme hätte weitere Untersuchungen ermöglicht [36]. Zum Beispiel, durch wiederholte Bewertungen desselben Gesprächs mit zeitlichem Abstand kann man die Daten für Test-Retest-Reliabilität erheben. Die nicht beim Gespräch anwesenden Dozierenden können auch anhand des Videos das Gespräch beurteilen. Das würde die Analyse für Inter-Rater-Objektivität zwischen den Dozierenden ermöglichen. Allerdings aufgrund der zu großen Anzahl der Simulationsgespräche und des Datenschutzes konnte eine Video-Aufnahme in unserer Studie nicht erfolgen. Bei jedem Simulationsgespräch war nur eine Dozierende oder ein Dozierender anwesend, wodurch ein direkter Vergleich der Bewertungen zwischen den Dozierenden nicht möglich war. Daher wurden lediglich die Bewertungen der Studiereden mit einer oder einem Dozierenden analysiert, um die Objektivität einzuschätzen. Aber die objektiven Bewertungen der Studierenden könnten durch den Mildeeffekt oder sogar durch ihre persönlichen Beziehungen beeinträchtigt worden sein. Zudem hatten die Studierenden zum Zeitpunkt der Checkliste-Bewertung noch nicht den Kurs absolviert. So ist es noch unklar, ob sie über ausreichendes Fachwissen verfügten, um das Simulationsgespräch angemessen zu beurteilen. Solche Umstände können möglicherweise erklären, warum wir bei der Objektivitätsschätzung nur eine “geringe“ Übereinstimmung feststellen konnten.

Im Rahmen des Kurses führten die Studierenden jeweils zwei Gespräche an separaten Terminen aus. Die beiden Gespräche waren unterschiedliche Rollenspiel, sodass die Studierenden nicht das gleiche

Gespräche zweimal wiederholten. Da jedes Rollenspiel eine unterschiedliche Checkliste hatte, konnten wir die zwei Gespräche eines Studierenden nicht vergleichen. Wenn die Studierenden ein gleiches Rollenspiel mit einem mehrwöchigen Zeitabstand wiederholt hätten, hätten wir dadurch quantifizieren können, wie sich ihre kommunikative Leistung innerhalb dieses Zeitraums verbessert hat. Damit hätte dann Konsequenz von Test, also die fünfte Komponente vom Schema nach Messick zur Abschätzung der Validität, überprüft werden können. In den weiteren Studien soll auf dieses Problem geachtet werden, und sollte einen verbesserten Studienplan erstellt werden.

Die Aufgabenschwierigkeit der Checkliste war zu leicht, da die überwiegende Mehrheit der Studierenden in allen RP über 80% der maximal erreichbaren Punkte erzielte. Das führte zur schwachen Trennschärfe. Diese Ergebnisse lassen sich durch den Mildeeffekt erklären. Die Studierenden bewerteten die Gespräche von ihren Mitstudierenden in ihrer Kursgruppe, die aus maximal 16 Studierenden bestand. Diese Situation erschwerte den Studierenden eine strenge Beurteilung bzw. eine Kritik an ihren Mitstudierenden, die sie möglicherweise auch persönlich kannten. Die Dozierenden vergaben ihren Studierenden ebenfalls tendenziell sehr positive Bewertungen. Sie waren die Dozierenden, die in Rahmen des Kurses ihre Studierenden betreuten, und gleichzeitig bewerteten die Gespräche ihrer Kursteilnehmenden. Zudem nahmen die Bewertungen keinen Einfluss auf die Abschlussnote für den Kurs, sodass die Leistungen der Studierenden nicht stark voneinander differenziert werden mussten. So könnte der Mildeeffekt die Checkliste-Bewertung beeinträchtigt haben.

Neben Mildeeffekt scheinen die organisatorischen Bedingungen des Kurses zu allgemein guten Bewertungen beigetragen zu haben. Die Studierenden erhielten die Einteilung von RP-Themen bereits am Anfang des Semesters, also wann und welches RP sie vorbereiten sollten. Ca. 14 Tage vor dem Termin erhielten sie die konkreten Fallvignetten und die Checklisten. Somit hatten sie ausreichend Zeit für die Vorbereitung, sodass sie gute Leistungen bei den Gesprächen bringen konnten. Daher sollte es weiter untersucht werden, ob die Checkliste unter Abschalten dieser Faktoren eine verbesserte Qualität aufweisen kann, insbesondere in Bezug auf Aufgabenschwierigkeit und Trennschärfe.

7. Konklusion

Die testtheoretische Qualität der Checkliste zur Beurteilung der kommunikativen Kompetenz bei Simulationsgesprächen wurde untersucht. Dabei wurde die Checkliste im papierbasierten System (PaFS) bzw. im digitalbasierten EvaSys (DiFS) angewendet. Es wurde eine gute Internal-Konsistenz, eine leichte Interrater Objektivität, eine akzeptable Validität, eine zu leichte Aufgabenschwierigkeit, eine zu schwache Trennschärfe und eine homogene Dimensionalität nachgewiesen. Diese Qualität war unabhängig von dem Rückmeldesystem, sowohl im PaFS als auch im DiFS.

Weiterhin wurde die Einsetzbarkeit des DiFS für die Checkliste zur Beurteilung der kommunikativen Kompetenz bei Simulationsgespräch überprüft. Zwischen den Ergebnissen mit DiFS und denen mit PaFS wurden keine signifikanten Unterschiede festgestellt. Zudem zeigte DiFS eine geringere Fehlerrate als PaFS. Diese Ergebnisse begründen, dass EvaSys als Standard-Rückmeldesystem für die Checkliste zur Beurteilung der kommunikativen Kompetenz bei Simulationsgesprächen etabliert werden kann.

8. Ethische Aufklärung

Das Forschungsprojekt wurde in Übereinstimmung mit den ethischen Standards der Deklaration von Helsinki und dem Genfer Gelöbnis durchgeführt. Eine Ethikkommission prüfte und genehmigte das Forschungsprojekt (Antragsnummer: 2017-JGU-psychEK-007).

Interessenkonflikte: Keine.

9. Literaturverzeichnis

[1] Bundesministerium für Gesundheit. Referentenentwurf des Bundesministeriums für Gesundheit Verordnung zur Neuregelung der ärztlichen Ausbildung A. Problem und Ziel 2020. https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/A/Referentenentwurf_AEApprO.pdf (accessed November 6, 2025).

[2] Nestel D, Tierney T. BMC Medical Education Role-play for medical students learning about communication: Guidelines for maximising benefits 2007. https://doi.org/10.1186/1472-6920-7-3.

[3] Charlton RC. Using role-plays to teach palliative medicine. Med Teach 1993;15:187–93. https://doi.org/10.3109/01421599309006713.

[4] Steinert Y. Twelve tips for using role-plays in clinical teaching. Med Teach 1993;15:283–91. https://doi.org/10.3109/01421599309006651.

[5] Hargie O, Dickson D, Boohan M, Hughes K. A survey of communication skills training in UK Schools of Medicine: present practices and prospective proposals. Med Educ 1998;32:25–34.

[6] Nestel D, Muir E, Plant M, Kidd J, Thurlow S. Modelling the lay expert for first-year medical students: the actor-patient as teacher. Med Teach 2009:562–4. https://doi.org/10.1080/0142159021000042649.

[7] Henderson P. Assisting medical students to conduct empathic conversations with patients from a sexual medicine clinic. Sex Trans Infect 2002;78:246–9. https://doi.org/10.1136/sti.78.4.246.

[8] Joyner B, Young L. Medical Teacher Teaching medical students using role play: Twelve tips for successful role plays Teaching medical students using role play: Twelve tips for successful role plays. Med Teach 2006;28:225–9. https://doi.org/10.1080/01421590600711252.

[9] al Odhayani A, Ratnapalan S, Bs MB, Faap MF. Teaching communication skills. Canadian Family Physician • Le Médecin de Famille Canadien | 2011;57.

[10] Ouldali N, Le Roux E, Faye A, Leblanc C, Angoulvant F, Korb D, et al. Early formative objective structured clinical examinations for students in the pre-clinical years of medical education: A nonrandomized controlled prospective pilot study. PLoS One 2023;18. https://doi.org/10.1371/journal.pone.0294022.

[11] Ho C, Weisleder P, Ream MA, Albert DVF. Education Research: A Qualitative Analysis of Communication-Focused Feedback Provided to Child Neurology Residents During an Objective Structured Clinical Examination. Neurology Education 2025;4. https://doi.org/10.1212/NE9.0000000000200187.

[12] Raski B, Eissner A. Implementation of online peer feedback for student self-reflection-first steps on the development of a feedback culture at a medical faculty. GMS J Med Educ 2019;36.

[13] Keifenheim KE, Teufel M, Ip J, Speiser N, Leehr EJ, Zipfel S, et al. Teaching history taking to medical students: a systematic review 2015. https://doi.org/10.1186/s12909-015-0443-x.

[14] Makoul G, Altman M, Kogan JR, Bellini LM, Shea JA. Early Assessment of Medical Students' Clinical Skills. Academic Medicine 2002;77:1156–76.

[15] Salander P. Patients with cancer react differently – Training in breaking bad news can therefore not be reduced to learning pre-defined behaviours. Patient Educ Couns 2017;100:1955–6. https://doi.org/10.1016/j.pec.2017.05.025.

[16] Bosse HM, Schultz JH, Nickel M, Lutz T, Möltner A, Jünger J, et al. The effect of using standardized patients or peer role play on ratings of undergraduate communication training: A randomized controlled trial. Patient Educ Couns 2012;87:300–6. https://doi.org/10.1016/j.pec.2011.10.007.

[17] Sperling JD, Clark S, Kang Y. Teaching medical students a clinical approach to altered mental status: simulation enhances traditional curriculum. Med Educ Online 2013;18. https://doi.org/10.3402/meo.v18i0.19775.

[18] Ahmady S, Shahbazi S, Khajeali N. Comparing the effect of traditional and role-play training methods on nursing students' performance and satisfaction in the principles of patient education course. J Educ Health Promot 2021;10. https://doi.org/10.4103/JEHP.JEHP_722_20.

[19] Allenbaugh J, Corbelli J, Rack L, Rubio D, Spagnoletti C. A Brief Communication Curriculum Improves Resident and Nurse Communication Skills and Patient Satisfaction. J Gen Intern Med 2019;34:1167–73. https://doi.org/10.1007/s11606-019-04951-6.

[20] Huang LJ, Huang HC, Chuang CL, Chang SL, Tsai HC, Lu DY, et al. Role-play of real patients improves the clinical performance of medical students. Journal of the Chinese Medical Association 2021;84:183–90. https://doi.org/10.1097/JCMA.0000000000000431.

[21] Nestel D, Tierney T. Role-play for medical students learning about communication: Guidelines for maximising benefits. BMC Med Educ 2007;7. https://doi.org/10.1186/1472-6920-7-3.

[22] Unrue EL, Ii O, White G, Cheng N, Lindsey T. Medical Education Brief Report Effect of a standardized patient encounter on first year medical student confidence and satisfaction with telemedicine. J Osteopath Med 2021;121:733–7. https://doi.org/10.1515/jom-2020-0277.

[23] May W. Training Standardized Patients for a High-Stakes Clinical Performance Examination in the California Consortium for the Assessment of Clinical Competence. 2008. https://doi.org/10.1016/S1607-551X(09)70029-4.

[24] Wilbur K, Elmubark A, Shabana S. Systematic review of standardized patient use in continuing medical education. Journal of Continuing Education in the Health Professions 2018;38:3–10. https://doi.org/10.1097/CEH.0000000000000190.

[25] Mounsey AL, Bovbjerg V, White L, Gazewood J. Do students develop better motivational interviewing skills through role-play with standardised patients or with student colleagues? Med Educ 2006;40:775–80. https://doi.org/10.1111/j.1365-2929.2006.02533.x.

[26] Fischbeck S, Mauch M, Leschnik E, Beutel ME, Laubach W. Überprfung ärztlicher kommunikativer Kompetenz mittels einer OSCE bei Studierenden der Medizin im ersten Studienjahr. PPmP Psychotherapie Psychosomatik Medizinische Psychologie 2011;61:465–71. https://doi.org/10.1055/s-0031-1291277.

[27] Setyonugroho W, Kennedy KM, Kropmans TJB. Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. Patient Educ Couns 2015;98:1482–91. https://doi.org/10.1016/j.pec.2015.06.004.

[28] Piumatti G, Cerutti B, Perron NJ. Assessing communication skills during OSCE: need for integrated psychometric approaches. BMC Med Educ 2021;21. https://doi.org/10.1186/s12909-021-02552-8.

[29] Markland AD, Burgio KL, Beasley TM, David SL, Redden DT, Goode PS. Psychometric evaluation of an online and paper accidental bowel leakage questionnaire: The ICIQ-B questionnaire. Neurourol Urodyn 2017;36:166–70. https://doi.org/10.1002/nau.22905.

[30] Phillips AC, Mackintosh SF, Gibbs C, Ng L, Fryer CE. A comparison of electronic and paper-based clinical skills assessment: Systematic review. Med Teach 2019;41:1151–9. https://doi.org/10.1080/0142159X.2019.1623387.

[31] Felicitas L. Wagner, Sabine Feller, Felix M. Schmitz, Philippe G. Zimmermann, Rabea Krings, Sissel Guttormsen, et al. Usability and preference of electronic vs. paper and pencil OSCE checklists by examiners and influence of checklist type on missed ratings in the Swiss Federal Licensing Exam. GMS J Med Educ 2022;39.

[32] Li J, Svicher A, Silva W, Choi K-H, Kim S-H, Park K, et al. A Brief Online and Offline (Paper-and-Pencil) Screening Tool for Generalized Anxiety Disorder: The Final Phase in the Development and Validation of the Mental Health Screening Tool for Anxiety Disorders (MHS: A). Front Psychol 2021;12:639366. https://doi.org/10.3389/fpsyg.2021.639366.

[33] Wang CC, Wang YCL, Hsu YH, Lee HC, Kang YC, Monrouxe LV, et al. Digitizing Scoring Systems With Extended Online Feedback: A Novel Approach to Interactive Teaching and Learning in Formative OSCE. Front Med (Lausanne) 2022;8. https://doi.org/10.3389/fmed.2021.762810.

[34] Saeed E, Hamad MH, Alhuzaimi AN, Aljamaan F, Elsenterisi H, Assiri H, et al. Virtual Objective Structured Clinical Examination (OSCE) Training in the Pandemic Era: Feasibility, Satisfaction, and the Road Ahead. Cureus 2024. https://doi.org/10.7759/cureus.61564.

[35] evasys GmbH. Webbasierte Befragungssoftware für effizientes Feedbackmanagement n.d. https://evasys.de/evasys/ (accessed November 6, 2025).

[36] Natt N, Starr SR, Reed DA, Park YS, Dyrbye LN, Leep Hunderfund AN. High-Value, Cost-Conscious Communication Skills in Undergraduate Medical Education: Validity Evidence for Scores Derived from Two Standardized Patient Scenarios. Simulation in Healthcare 2018;13:316–23. https://doi.org/10.1097/SIH.0000000000000316.

[37] Fox MP, Lash TL, Bodnar LM. Common misconceptions about validation studies. IEA International Epidemiological Association International Journal of Epidemiology 2020:1392–6. https://doi.org/10.1093/ije/dyaa090.

[38] Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ 2003;37:830–7.

[39] Souza AC de, Alexandre NMC, Guirardello E de B. Propriedades psicométricas na avaliação de instrumentos: avaliação da confiabilidade e da validade. Epidemiol Serv Saude 2017;26:649–59. https://doi.org/10.5123/S1679-49742017000300022.

[40] Messick S. Validity of Psychological Assessment. American Psychologist 1995. https://doi.org/10.1037/0003-066X.50.9.741.

[41] Hubley AM, Zumbo BD. Validity and the Consequences of Test Interpretation and Use. Springer Science+Business Media 2011;103:219–30. https://doi.org/10.1007/s11205-011-9843-4.

[42] Schmitz FM, Zimmermann PG, Gaunt K, Stolze M, Guttormsen Schär S. Electronic Rating of Objective Structured Clinical Examinations: Mobile Digital Forms Beat Paper and Pencil Checklists in a Comparative Study. Lecture Notes in Computer Science (including subseries

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7058 LNCS, 2011, p. 501–12. https://doi.org/10.1007/978-3-642-25364-5_35.

[43] Daniels VJ, Strand AC, Lai H, Hillier T. Impact of tablet-scoring and immediate score sheet review on validity and educational impact in an internal medicine residency Objective Structured Clinical Exam (OSCE). Med Teach 2019;41:1039–44. https://doi.org/10.1080/0142159X.2019.1615609.

[44] Lane S. Validity evidence based on testing consequences. Psicothema 2014;26:127–35. https://doi.org/10.7334/psicothema2013.258.

[45] Fischbeck S, Mauch M, Leschnik E, Laubach W. Entwicklung und Evaluation einer OSCE für die Uberprüfung kommunikativer ärztlicher Kompetenz im Kursus der Medizinischen Psychologie und Medizinischen Soziologie. Zeitschrift Für Medizinische Psychologie 2010:94–6.

[46] Pospeschill, Markus. Empirische Methoden in der Psychologie. München: Reinhardt; 2013.

[47] Colliver JA, Conlee MJ, Verhulst SJ. From test validity to construct validity … and back? Med Educ 2012;46:366–71. https://doi.org/10.1111/j.1365-2923.2011.04194.x.

[48] Schreier M. Fallauswahl in der qualitativ-psychologischen Forschung. Handbuch Qualitative Forschung in der Psychologie, Springer Fachmedien Wiesbaden; 2017, p. 1–21. https://doi.org/10.1007/978-3-658-18387-5_19-1.

[49] Bortz J, Döring N. Forschungsmethoden und Evaluation. Heidelberg: Springer; 2005.

[50] Nunnally, Jum C. Psychometric theory. 3rd ed. New York: McGraw Hill; 1994.

[51] Mart In Andr Es A, Hern Andez A. Hubert's multi-rater kappa revisited. British Journal of Mathematical and Statistical Psychology 2020;73:1–22. https://doi.org/10.1111/bmsp.12167.

[52] Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004;38:1006–12. https://doi.org/10.1046/j.1365-2929.2004.01932.x.

[53] Stoyan D, Pommerening A, Hummel M, Kopp-Schneider A. Multiple-rater kappas for binary data: Models and interpretation. Biometrical Journal 2018;60:381–94. https://doi.org/10.1002/bimj.201600267.

[54] Morris R, MacNeela P, Scott A, Treacy P, Hyde A, O'Brien J, et al. Ambiguities and conflicting results: The limitations of the kappa statistic in establishing the interrater reliability of the Irish nursing minimum data set for mental health: A discussion paper. Int J Nurs Stud 2008;45:645–7. https://doi.org/10.1016/j.ijnurstu.2007.07.005.

[55] Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy 2013;9:330–8. https://doi.org/10.1016/j.sapharm.2012.04.004.

[56] Mandrekar JN. Measures of Interrater Agreement. Biostatistics For Clinicians 2010. https://doi.org/10.1097/JTO.0b013e318200f983.

[57] Shirazi M, Labaf A, Monjazebi F, Jalili M, Mirzazadeh M, Ponzer S, et al. Assessing medical students' communication skills by the use of standardized patients: Emphasizing standardized patients' quality assurance. Academic Psychiatry 2014;38:354–60. https://doi.org/10.1007/s40596-014-0066-2.

[58] Wollney EN, Vasquez TS, Stalvey C, Close J, Markham MJ, Meyer LE, et al. Are evaluations in simulated medical encounters reliable among rater types? A comparison between

standardized patient and outside observer ratings of OSCEs. PEC Innovation 2023;2. https://doi.org/10.1016/j.pecinn.2023.100125.

[59] Noble H, Smith J. Issues of validity and reliability in qualitative research. Evid Based Nurs 2015;18:34–5. https://doi.org/10.1136/eb-2015-102054.

[60] Walters K, Osborn D, Raven P. The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. Med Educ 2005;39:292–8. https://doi.org/10.1111/j.1365-2929.2005.02091.x.

[61] Price EG, Windish DM, Magaziner J, Cooper LA. Assessing validity of standardized patient ratings of medical students' communication behavior using the Roter interaction analysis system. Patient Educ Couns 2008;70:3–9. https://doi.org/10.1016/j.pec.2007.10.002.

[62] Rosner B, Grove D. Use of the Mann-Whitney U-Test for Clustered Data. STATISTICS IN MEDICINE Statist Med 1999;18:1387–400.

[63] Cipresso P, Lorenzo O, Astivia O, Chun C, Kam S, Anselmi P, et al. A Comparison of Classical and Modern Measures of Internal Consistency. Front Psychol 2019;10. https://doi.org/10.3389/fpsyg.2019.02714.

[64] Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychol Bull 1979;86:420–8.

[65] Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15:155–63. https://doi.org/10.1016/j.jcm.2016.02.012.

[66] Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med 2018;18:91–3. https://doi.org/10.1016/j.tjem.2018.08.001.

[67] Vester Thorsen S, Bue Bjorner J. Reliability of the Copenhagen Psychosocial Questionnaire. Scand J Public Health 2010;38:25–32. https://doi.org/10.1177/1403494809349859.

[68] Sijtsma K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. Psychometrika 2009;74:107–20. https://doi.org/10.1007/S11336-008-9101-0.

[69] Ark T, Kalet A, Tewksbury L, Altshuler L, Crowe R, Wilhite J, et al. Validity evidence for the clinical communication skills assessment tool (CCSAT) from 9 years of implementation in a high stakes medical student OSCE. Patient Educ Couns 2024;127. https://doi.org/10.1016/j.pec.2024.108323.

[70] Padilla JL, Benítez I. Validity evidence based on response processes. Psicothema 2014;26:136–44. https://doi.org/10.7334/psicothema2013.259.

[71] Sireci S, Faulkner-Bond M. Validity evidence based on test content. Psicothema 2014;26:100–7. https://doi.org/10.7334/psicothema2013.256.

[72] Rios J, Wells C. Validity evidence based on internal structure. Psicothema 2014;26:108–16. https://doi.org/10.7334/psicothema2013.260.

[73] Schrauth M, Schäfer S, Zipfel S, Sammet I. Praxisorientierte prüfungen in den psychosozialen fächern. Psychosom Konsiliarpsychiatr 2007;1:229–31. https://doi.org/10.1007/s11800-007-0032-x.

[74] Tomak L, Bek Y, Cengiz MA. Graphical modeling for item difficulty in medical faculty exams. Niger J Clin Pract 2016;19:58–65. https://doi.org/10.4103/1119-3077.173701.

[75] Hwang JE, Kim NJ, Kim SY. Comparisons of item difficulty and passing scores by test equating in a basic medical education curriculum. Korean Journal of Medicine Education 2019;31:147–57. https://doi.org/10.3946/kjme.2019.126.

[76] Talwalkar JS, Murtha TD, Prozora S, Fortin AH, Morrison LJ, Ellman MS. Assessing Advanced Communication Skills via Objective Structured Clinical Examination: A Comparison of Faculty Versus Self, Peer, and Standardized Patient Assessors. Teach Learn Med 2020;32:294–307. https://doi.org/10.1080/10401334.2019.1704763.

10. Danksagung

Bei der Untersuchung und der Verfassung dieser Dissertation erhielt ich von vielen Leuten große Unterstützung. Für all diese Hilfe möchte ich mich herzlich bedanken.

Von der Fragestellung über die wissenschaftliche Recherche bis hin zum organisatorischen Verfahren kümmerte sich *** mit großer Leidenschaft um alles. Dank ihrer Unterstützung konnte ich mich intensiv auf diese Studie konzentrieren. Ich kann mir nicht vorstellen, diese Dissertation ohne ihre Hilfe erfolgreich abgeschlossen zu haben.

Die Unterstützung von *** darf ich ebenfalls nicht vergessen. Insbesondere danke ich dafür, dass ich die Chance bekam, mich mit dieser Studie zu beschäftigen.

Vor Beginn dieser Studie hatte ich keinerlei Erfahrung mit STATA17. Dank der Hilfe von *** konnte ich mich schnell mit STATA17 vertraut machen. Dadurch konnte ich nicht nur die hochspezifische Analysenergebnisse erzielen, sondern hatte auch viel Spaß daran, eine neue digitale Sprache zu erlernen.

In dieser Studie wurden großen Mengen an Daten gesammelt und analysiert. Bei dieser Datenauswertung halfen mir *** und *** sehr, wodurch die Zeit für Datenerfassung sowie-übertragung erheblich verkürzt wurde. Sonst hätte ich wohl noch immer an Wochenenden am PC sitzen und mühsam einzelne Werte eintippen müssen.

Die statistische Beratung von *** bei IMBEI war äußerst hilfreich, um einen konkreten Analyseplan zu erstellen. Besonders dachte ich lange darüber nach, wie die Analysen für die Objektivität und Validität in unserer Studie verbessert werden können. Bei der Beratung gab er uns viele Vorschläge, die die vertiefte Analyse ermöglichte.

Im Jahr 2018 kamen meine Familie und ich nach Deutschland, ohne hier Verwandte oder Freunde zu haben. Doch bald darauf hatten wir das Glück, die *** und *** kennenzulernen. Von ihnen lernten wir die deutsche Sprache und Kultur, und sie halfen uns bei sämtlichen Angelegenheiten, um uns schnell und leicht in Deutschland einzuleben. Und nun erhielt ich erneut ihre große Unterstützung: Das Korrekturlesen dieses Textes. Sätze so zu formulieren, dass sie klar ausdrücken, was man sagen möchte, ist für Nicht-Muttersprachler eine große Herausforderung. Mit viel Geduld halfen sie mir dabei, meine Sätze klarer und verständlicher zu formulieren. Dank ihrer Hilfe konnte ich schaffen, einen kompakten und strukturieren Text zu verfassen. Ich bin ganz herzlich dankbar für ihre Liebe, Fürsorge und Hilfe. Besonders möchte ich *** und *** für ihre intensive Unterstützung danken!

Abschließend möchte ich mich bei meiner Familie bedanken. Meine Schwiegereltern und meine Eltern, die unsere Entscheidung für Einwanderung nach Deutschland unterstützen, und auch meine Brüder und Schwester, habe ich vielen Dank und herzliche Liebe an euch! Meine Tochter und meine Frau, die zusammen mit mir nach Deutschland kommen, sind stets meine größte Motivation im Leben, und ihr Dasein ist bereits die stärkste Unterstützung für all meine Tätigkeiten. Es ist für mich unglaublich, dass ich nun am Tisch die Danksagung schreibe. Eure Liebe und Unterstützung ließen mich alle Schwierigkeiten überwinden. Vielen Dank und ich liebe euch, *** und ***!

11. Anhang

[Checkliste für RP1 „Anamnese“]

Beurteiler/in: _____	0 Studierender 0 Dozierender	**Datum:** **_____._____.2021**

Bitte notieren Sie, inwiefern Sie die einzelnen Elemente während des Anamnesegesprächs bei dem Studierenden in der Arztrolle haben beobachten können, es bedeutet **+** = ja/eher ja, - = nein/eher nein):
* Punkt geben, wenn mindestens ein Aspekt vorhanden

1. Eröffnung des Gesprächs, hat …

hat Patient freundlich und höflich begrüßt	0,5
hat sich mit Namen (und ggf. Funktion) vorgestellt	0,5
hat mit echtem Interesse eine Eröffnungsfrage gestellt	0,5

2. Exploration der jetzigen Symptome

Seit wann/wie lange? (Beginn und Dauer der Symptome)	0,5
Wie? (Qualität: etwa Druckschmerz, dumpfer Schmerz)	1
Wie stark? (Intensität)	1
Wo? (Lokalisation)	0,5
jetzige Begleitsymptome (z. B. Übelkeit, Schwindel)/ Umstände, unter denen die jetzigen Symptome auftreten, sich mildern oder intensivieren (z. B. körperliche Belastung) *	1
ergänzende Informationen (z. B. frühere Behandlungsversuche, früheres Auftreten der Symptome)	1

* Punkt geben, wenn mindestens ein Aspekt vorhanden

3. Ergänzende Anamnesen (im Erstgespräch notwendig)

Eigenanamnese (Vorerkrankungen des Patienten), Sozial-/Familienanamnese (Erkrankungen von Angehörigen, berufliche, familiäre Situation) *	1
Vegetative Anamnese (Stuhlgang, Wasserlassen, Appetit, Durst)	1
Medikamentenanamnese (Einnahme), Genussmittelanamnese (z. B. Alkohol, Zigaretten)	1
Überleitung zur Erhebung der ergänzenden Anamnesen für Pat. war nachvollziehbar und zusammenhängend	0,5

4. Qualität der Gesprächsführung

nahm offene Körperposition ein	0,5
hatte Blickkontakt zum Patienten	0,5
hat vom Pat. Gesagtes wahrgenommen und angesprochen (Aktives Zuhören/Empathie)	1
hat sich nicht floskelhaft oder widersprüchlich geäußert (Echtheit)	0,5
hat Pat. nicht unbegründet unterbrochen, nicht bagatellisiert, nicht vorschnell interpretiert oder Ratschläge gegeben, hat Patient ernst genommen (Wertschätzung)*	0,5
Exploration wurde strukturiert vorgenommen und war dem Patienten angepasst	1

Gesamteindruck, Note (1= „sehr gut“ bis 5 = „mangelhaft“): _____

Was war gut, was könnte die künftige Ärztin/der künftige Arzt besser machen?

[Checkliste für RP2 „Gesprächsförderung“]

Beurteiler/in: _____	0 Studierender 0 Dozierender	**Datum:** _____._____.2021

Bitte notieren Sie, inwiefern Sie die einzelnen Elemente während des Anamnesegesprächs bei dem Studierenden in der Arztrolle haben beobachten können, es bedeutet **+** = ja/eher ja, - = nein/eher nein):

1. Eröffnung des Gesprächs, hat …

Patient freundlich und höflich begrüßt	0,5
Zielsetzung des Gesprächs genannt	1
mit echtem Interesse eine Eröffnungsfrage gestellt	0,5

2. Welche Gesprächsförderer haben Sie beobachtet?

Ermuntern	0,5
Umschreiben	1
Wiederholen	0,5
Offenes Nachfragen	1
Zusammenfassen	0,5
Emotionale Inhalte ansprechen	1

3. Waren Gesprächsstörer zu beobachten?

Bagatellisieren (Punkt geben, wenn nicht vorhanden)	0,5
Diagnostizieren/Interpretieren (Punkt geben, wenn nicht vorhanden)	0,5
Vorschnell Ratschläge/Lösungen anbieten (Punkt geben, wenn nicht vorhanden)	0,5

4. Aufrollen der Hintergrundproblematik, hat …

durch geschicktes Fragen/Argumentieren die Hintergrundproblematik offengelegt	1
souverän den Krankschreibungswunsch abgelehnt	1
eine umsetzbare Alternative zur Krankschreibung mit dem Patienten vereinbart	1

5. Qualität der Gesprächsführung

nahm offene Körperposition ein	0,5
hatte Blickkontakt zum Patienten	0,5
hat vom Pat. Gesagtes wahrgenommen und angesprochen (Aktives Zuhören/Empathie)	1
hat sich nicht floskelhaft oder widersprüchlich geäußert (Echtheit)	0,5
hat Pat. nicht unbegründet unterbrochen, nicht bagatellisiert, nicht vorschnell interpretiert oder Ratschläge gegeben, hat Patient ernst genommen (Wertschätzung)*	0,5

Gesamteindruck, Note (1= „sehr gut“ bis 5 = „mangelhaft“): _____

Was war gut, was könnte die künftige Ärztin/der künftige Arzt besser machen?

[Checkliste für RP3 „Informationsvermittlung“]

Beurteiler/in: _____	0 Studierender 0 Dozierender	**Datum:** _____._____.2021

Bitte notieren Sie, inwiefern Sie die einzelnen Elemente während des Anamnesegesprächs bei dem Studierenden in der Arztrolle haben beobachten können, es bedeutet **+ = ja/eher ja, - = nein/eher nein**:
* Punkt geben, wenn mindestens ein Aspekt vorhanden

1. Eröffnung des Gesprächs, hat …

Patient freundlich und höflich begrüßt	0,5
Zielsetzung des Gesprächs benannt	1
mit echtem Interesse eine Eröffnungsfrage gestellt	0,5

2. Vorbereitung, hat …

Kenntnisstand ermittelt	1
Erwartungen, Befürchtungen, subjektive Krankheitstheorie ergründet*	0,5
zum Fragen ermutigt	1

3. Verständliche Vermittlung der Information, Gesagtes

war einfach formuliert: kurze Sätze, keine Fremdwörter, Fachwörter erklärt	1
war geordnet und gegliedert	1
war kurz und prägnant	1
hat anregende Zusätze benutzt: z.B. sprachliche Bilder, Vergleiche, Abbildungen, Modelle*	1
ist auf Zwischenfragen eingegangen	1

4. Rückversicherung, hat …

Self blaming gezeigt„Habe ich mich Ihnen verständlich machen können?“ (Cave: Nicht „Haben Sie das verstanden?“	1
das Wichtigste nochmal wiederholen lassen	0,5

5. Qualität der Gesprächsführung

nahm offene Körperposition ein	0,5
hatte Blickkontakt zum Patienten	0,5
hat vom Pat. Gesagtes wahrgenommen und angesprochen (Aktives Zuhören/Empathie)	1
hat sich nicht floskelhaft oder widersprüchlich geäußert (Echtheit)	0,5
hat Pat. nicht unbegründet unterbrochen, nicht bagatellisiert, nicht vorschnell interpretiert oder Ratschläge gegeben, hat Patient ernst genommen (Wertschätzung)*	0,5

Gesamteindruck, Note (1= „sehr gut“ bis 5 = „mangelhaft“): _____

Was war gut, was könnte die künftige Ärztin/der künftige Arzt besser machen?

[Checkliste für RP4 „Partizipative Entscheidung“]

Beurteiler/in: _____	0 Studierender 0 Dozierender	**Datum:** _____._____.2021

Bitte notieren Sie, inwiefern Sie die einzelnen Elemente während des Anamnesegesprächs bei dem Studierenden in der Arztrolle haben beobachten können, es bedeutet **+** = ja/eher ja, - = nein/eher nein):

1. Eröffnung des Gesprächs, hat …

Patient freundlich und höflich begrüßt	0,5
Zielsetzung des Gesprächs genannt	0,5
mit echtem Interesse eine Eröffnungsfrage gestellt	0,5

2. Schritte der partizipativen Entscheidungsfindung, hat …

mitgeteilt, dass eine Entscheidung ansteht	1
die Gleichberechtigung der Partner (A-P) beim Entscheiden formuliert	1
über Wahlmöglichkeiten informiert (Equipoise*)	1,5
Information über Vor- und Nachteile der Therapieoption 1 gegeben	1
Information über Vor- und Nachteile der Therapieoption2 gegeben	1
Information über Vor- und Nachteile der Therapieoption 3 gegeben	1
Verständnis gezeigt, Gedanken und Erwartungen erfragt	1
Präferenzen des Patienten ermittelt	1
ausgehandelt/gemeinsame Entscheidung herbeigeführt	1
Vereinbarung zur Umsetzung der Entscheidung getroffen	1

*Equipoise-Statement: Die Formulierung von Ungewissheit durch den Arzt soll den Patienten einen Mitentscheidungsraum eröffnen. «Bei Ihren Beschwerden gibt es keine festgelegte Form der Behandlung, aber es gibt verschiedene Wege, ihnen etwas entgegen zu setzen.» (vgl. Elwyn et al., 2000)

3. Qualität der Gesprächsführung

nahm offene Körperposition ein	0,5
hatte Blickkontakt zum Patienten	0,5
hat vom Pat. Gesagtes wahrgenommen und angesprochen (Aktives Zuhören/Empathie)	1
hat sich nicht floskelhaft oder widersprüchlich geäußert (Echtheit)	0,5
hat Pat. nicht unbegründet unterbrochen, nicht bagatellisiert, nicht vorschnell interpretiert oder Ratschläge gegeben, hat Patient ernst genommen (Wertschätzung)	0,5

Gesamteindruck, Note (1= „sehr gut“ bis 5 = „mangelhaft“): _____

Was war gut, was könnte die künftige Ärztin/der künftige Arzt besser machen?

[Checkliste für RP5 „Compliance“]

Beurteiler/in: _____	0 Studierender 0 Dozierender	**Datum:** _____._____.2021

Bitte notieren Sie, inwiefern Sie die einzelnen Elemente während des Anamnesegesprächs bei dem Studierenden in der Arztrolle haben beobachten können, es bedeutet **+ = ja/eher ja, - = nein/eher nein**:

1. Eröffnung des Gesprächs

hat Patient freundlich und höflich begrüßt	0,5
hat Zielsetzung des Gesprächs genannt	0,5
hat mit echtem Interesse eine Eröffnungsfrage gestellt	0,5

2. Informierung, hat …

Gründe für Non-Compliance sensibel ergründet	1
über die Notwendigkeit der Maßnahme erneut Informiert	1

3. Gewinnen, hat …

sich Bedenken und Handlungshemmnisse des Pat. angehört und relativiert	1
Wirkungen/Nebenwirkungen der Therapie angesprochen	1
überzeugend die Notwendigkeit der Therapie vermittelt	1

4. Erleichtern

förderte Selbstkontrolle und Eigeninitiative des Patienten	1
gestaltete Behandlungsplan möglichst einfach	1
hat vereinbarte Therapie für den Pat. schriftlich notiert	0,5
hat nächste(n) Termin(e) klar vereinbart	0,5

5. Unterstützen, hat …

Familienmitglieder in den Behandlungsplan einbezogen	1
Patient für gewünschtes Verhalten gelobt	0,5

6. Qualität der Gesprächsführung

nahm offene Körperposition ein	0,5
hatte Blickkontakt zum Patienten	0,5
hat vom Pat. Gesagtes wahrgenommen und angesprochen (Aktives Zuhören/Empathie)	1
hat sich nicht floskelhaft oder widersprüchlich geäußert (Echtheit)	0,5
hat Pat. nicht unbegründet unterbrochen, nicht bagatellisiert, nicht vorschnell interpretiert oder Ratschläge gegeben, hat Patient ernst genommen (Wertschätzung)	0,5

Gesamteindruck, Note (1= „sehr gut“ bis 5 = „mangelhaft“): _____
Was war gut, was könnte die künftige Ärztin/der künftige Arzt besser machen?

[Checkliste für RP6 „Motivational Interview“]

Beurteiler/in: _____	0 Studierender 0 Dozierender	**Datum:** _____._____.2021

Bitte notieren Sie, inwiefern Sie die einzelnen Elemente während des Anamnesegesprächs bei dem Studierenden in der Arztrolle haben beobachten können, es bedeutet **+** = ja/eher ja, - = nein/eher nein):

1. Eröffnung des Gesprächs

hat Patient freundlich und höflich begrüßt	0,5
hat Zielsetzung des Gesprächs genannt	0,5
hat mit echtem Interesse eine Eröffnungsfrage gestellt	0,5

2. Vermittlung von MI-Spirit

hat die Beweggründe für ungesundes Verhalten herausgefunden/-gearbeitet	1
hat die Schwierigkeit, etwas zu ändern angesprochen	0,5

3. Change Talk

hat Widersprüchlichkeit von Gesundheitsinteresse und -verhalten herausgearbeitet	1
hat die Gesundheitsmotivation (Stage of Change) analysiert	1
hat bisherige (fehlgeschlagene) Versuche einer Verhaltensänderung analysiert	1
hat das Nachdenken über die Folgen mangelnden Gesundheitsverhalten evoziert	1
hat Erlaubnis angefragt, eine Verhaltensänderung vorzuschlagen	0,5
hat angeregt, über Vor- und Nachteile einer Gesundheitsmaßnahme nachzudenken	0,5
hat die Wichtigkeit einer Verhaltensänderung bewerten lassen	0,5
hat eine Zielvereinbarung getroffen	0,5

4. Geschmeidiger Umgang mit Widerstand

hat ungesundes Gesundheitsverhalten nicht verurteilt	0,5
hat das Recht auf Autonomie zugestanden/betont	1
hat Einwände positiv umgedeutet	0,5

5. Qualität der Gesprächsführung

nahm offene Körperposition ein	0,5
hatte Blickkontakt zum Patienten	0,5
hat vom Pat. Gesagtes wahrgenommen und angesprochen (Aktives Zuhören/Empathie)	1
hat sich nicht floskelhaft oder widersprüchlich geäußert (Echtheit)	0,5
hat Pat. nicht unbegründet unterbrochen, nicht bagatellisiert, nicht vorschnell interpretiert oder Ratschläge gegeben, hat Patient ernst genommen (Wertschätzung)	0,5

Gesamteindruck, Note (1= „sehr gut“ bis 5 = „mangelhaft“): _____
Was war gut, was könnte die künftige Ärztin/der künftige Arzt besser machen?

[Checkliste für RP7 „Stressreaktion“]

Beurteiler/in: _____	0 Studierender 0 Dozierender	**Datum:** _____._____.2021

Bitte notieren Sie, inwiefern Sie die einzelnen Elemente während des Gesprächs bei dem Studierenden in der Arztrolle haben beobachten können. Es bedeutet **+** = ja/eher ja, - = nein/eher nein:

* Punkt geben, wenn mindestens ein Aspekt vorhanden

1. Eröffnung des Gesprächs

hat Patient freundlich und höflich begrüßt	0,5
hat Zielsetzung des Gesprächs genannt	0,5
hat mit echtem Interesse eine Eröffnungsfrage gestellt	0,5

2. Stressorenanalyse, hat erfragt/herausgefunden …

Arbeitslosigkeit	1
Abzahlung Haus/finanzieller Druck	1
Vorstellungsgespräch, möglicher Ortswechsel	0,5
Unzufriedenheit Partner/Partnerin	0,5

2. Analyse der Stressreaktion, Ebene des …

Körpers (Ärger schlägt vermutlich auf den Magen, Magenschmerzen, Anspannung)*	1
Gedanken (Sorgen, Grübeln, Angstgedanken, keine Lösung finden)*	0,5
Gefühle (Sorgen, Ärger, Wut, Angst aussprechen)*	1
Verhalten (Computerspielen, Rauchen, Alkohol)*	0,5

3. Ansätze der Stressbewältigung: Stressreaktion und Stressoren

körperlich: Ulcusbehandlung thematisiert, Entspannungstraining vorgeschlagen*	1
Kognitiv: gedankliche Falle der Hilflosigkeit aufgelöst	0,5
Verhalten: inadäquate Stressbewältigung in Frage gestellt, zur Änderung motiviert*	1
Emotionen: sich Aussprechen mit Partner/im Stressbewältigungskurs angeregt*	0,5
Plan zur Belastungsreduktion angeregt: Jobveränderung angehen, Hilfe von Beratern annehmen u.a.*	0,5

4. Qualität der Gesprächsführung

nahm offene Körperposition ein	0,5
hatte Blickkontakt zum Patienten	0,5
hat vom Pat. Gesagtes wahrgenommen und angesprochen (Aktives Zuhören/Empathie)	0,5
hat sich nicht floskelhaft oder widersprüchlich geäußert (Echtheit)	0,5
hat Pat. nicht unbegründet unterbrochen, nicht bagatellisiert, nicht vorschnell interpretiert oder Ratschläge gegeben, hat Patient ernst genommen (Wertschätzung)	1

Gesamteindruck, Note (1= „sehr gut“ bis 5 = „mangelhaft“): _____
Was war gut, was könnte die künftige Ärztin/der künftige Arzt besser machen?

_____

[Checkliste für RP8 „Krebsaufklärung“]

Beurteiler/in: _____	0 Studierender 0 Dozierender	**Datum:** _____._____.2021

Bitte notieren Sie, inwiefern Sie die einzelnen Elemente während des Gesprächs bei dem Studierenden in der Arztrolle haben beobachten können. Es bedeutet **+ = ja/eher ja, - = nein/eher nein**:

1. Eröffnung des Gesprächs (Setting up the Interview), hat …

Patienten freundlich und höflich begrüßt	0,5
für günstige Gesprächssituation gesorgt (z. B. nach bequemem Sitzen gefragt; Distanz sichtlich angenehm)	0,5
Anlass des Termins genannt/ klargestellt	1

2. Eingehen auf den Patienten (Perception und Invitation), hat …

aktuelles Befinden des Patienten erfragt und aufrichtiges Interesse dafür gezeigt	0,5
Erfahrungen in der Untersuchung (Darmspiegelung) erfragt	0,5
Informationsstand des Patienten ermittelt („Was hat man Ihnen dort mitgeteilt?“)	0,5
nach Informationsbedürfnis gefragt („Kann mir vorstellen, dass Sie jetzt wissen wollen, …“)	1

3. Diagnosemitteilung (Knowledge), hat …

Diagnosemitteilung vorbereitet (mit einem einleitenden Satz)	0,5
Diagnose eindeutig mitgeteilt (z. B. Krebserkrankung des Darms, bösartiger Tumor)	1
behutsam und zeitnah die gute Prognose genannt	1

4. Reaktionen des Patienten (Emotion), hat …

Reaktionen des Patienten abgewartet und Raum dafür gegeben	1
die emotionale Reaktion des Patienten erfasst und angesprochen	1
die Vorstellungen des Patienten von der Krankheit erfragt	0,5
negative Vorstellungen von der Krankheit relativiert	0,5

5. Weiterbehandlung (Strategy)

hat einen Plan für die nächsten Schritte der Behandlung mitgeteilt	1

6. Qualität der Gesprächsführung

nahm offene Körperposition ein	0,5
hatte Blickkontakt zum Patienten	0,5
hat vom Pat. Gesagtes wahrgenommen und angesprochen (Aktives Zuhören/Empathie)	1
hat sich nicht floskelhaft oder widersprüchlich geäußert (Echtheit)	0,5
hat Pat. nicht unbegründet unterbrochen, nicht bagatellisiert, nicht vorschnell interpretiert oder Ratschläge gegeben, hat Patient ernst genommen (Wertschätzung)	0,5

Gesamteindruck, Note (1= „sehr gut“ bis 5 = „mangelhaft“): _____

Was war gut, was könnte die künftige Ärztin/der künftige Arzt besser machen?