Messung der statistischen Kompetenz in der Hochschulausbildung am Beispiel des Statistical Reasoning Assessment

Inauguraldissertation
zur Erlangung des Doktorgrades
der Humanwissenschaftlichen Fakultät der Universität zu Köln
nach der Promotionsordnung vom 10.05.2010

vorgelegt von

Alla Sawatzky
aus Shdanowka, Russland

im September 2019

angefertigt bei Prof. Dr. Christian Rietz

1. Berichterstatter: Prof. Dr. Christian Rietz (Pädagogische Hochschule Heidelberg)
2. Berichterstatter: Prof. Dr. Matthias Grünke (Universität zu Köln)

Diese Dissertation wurde von der Humanwissenschaftlichen Fakultät der Universität zu Köln im Januar 2020 angenommen.

Tag der Disputation: 31.01.2020

Danksagung

Ein – für mich – äußerst glücklicher Zufall sorgte dafür, dass ich Prof. Dr. Christian Rietz kennenlernte. Er hat mir Vertrauen geschenkt, Türen geöffnet und mich fachlich und persönlich unterstützt, nicht nur während meiner Promotion in guten und in nicht besonders guten Zeiten. Er hat mir viele neue Impulse gegeben und meinen Blick darauf, was Forschung im Kern ausmacht, enorm erweitert. Dafür, und für vieles mehr, bedanke ich mich von Herzen bei Prof. Dr. Rietz. Ohne ihn wäre mein Promotionsvorhaben sicher noch immer in der Planungsphase.

Auch bei Prof. Dr. Matthias Grünke möchte ich mich für das entgegengebrachte Vertrauen und die bereichernden Gespräche herzlich bedanken. Durch ihn habe ich viele, für mich vollkommen neue empirische Fragestellungen kennengelernt und bin daran gewachsen.

Meinen Kolleginnen und Kollegen an der Pädagogischen Hochschule Heidelberg und an der Hochschule Fresenius spreche ich meinen aufrichtigen Dank für ihre Geduld und ihr Verständnis für mich besonders am Ende meiner Promotion aus (besondere Erwähnung verdient Prof. Dr. Katja Mierke). Tamara Lautenschläger gilt besonderer Dank für das Korrekturlesen von großen Teilen der Arbeit. Ebenso bedanke ich mich herzlich bei meinen ehemaligen Kollegen an der Universität zu Köln. Ohne ihre Unterstützung wäre die Datenerhebung nicht möglich gewesen. Besonderer Dank gebührt hier Ann-Kathrin Hennes.

Verena und Nilu, meine Kolleginnen und Freundinnen, haben mich während der letzten Jahre in mehr Situationen unterstützt, als ich es zu Papier bringen kann. Ich danke Euch für Eure Hilfe, Euer Verständnis und Euren Zuspruch!

Schließlich gelten mein Dank, meine Liebe und Verbundenheit meiner Familie und meinen engsten Vertrauten (ich führe Euch nicht namentlich auf – Ihr wisst, wer Ihr seid!). Ich danke Euch für Euer unermessliches Verständnis für meine seit Jahren „momentan stressige“ Situation und meine häufig fragwürdige Prioritätensetzung. Ihr seid mir Trost, Unterstützung und sehr nötige Ablenkung gewesen, nicht nur während der Promotion.

Die Freude am Lernen, Verstehen, Hinterfragen verdanke ich in Gänze meinem intellektuellen Vorbild, meiner Mutter. Alles, was ich heute bin, ist ihr Verdienst.

Das Dissertationsvorhaben ist meinem Großvater David Sawatzky (1922-2012) gewidmet, der in Zeiten der Entkulakisierung, des stalinistischen Terrors und der Zwangsarbeitslager des zweiten Weltkriegs für sein Überleben und das Überleben seiner Familie sorgen musste. Im jungen Alter verließ er die Schule wider seinen Wunsch. „Ich sah, wie die anderen Kinder in die Schule gingen, während ich arbeitete und weinte, weil ich so gerne Mathematik gemacht hatte“. Bildung ist das höchste Gut und ich bin meinem Schicksal zutiefst dankbar, dass ich die Möglichkeit habe, lebenslang zu lernen, nicht nur Mathematik.

Vorwort

„*Hail Science!*“
Professor Farnsworth,
Futurama, 7. Staffel, Folge 20 „Calculon 2.0“ (Groening & Cohen, 2013)

Im Laufe des letzten Jahrzehnts habe ich vielen Studierenden und im Wissenschaftsbetrieb Forschenden in den Humanwissenschaften methodische oder statistische Fragen beantwortet und so manches Mal einen ungebetenen Ratschlag gegeben:
„Rechne lieber etwas ganz Einfaches. Eine Korrelationstabelle oder die Übersicht von Mittelwertdifferenzen, am besten mit ein paar Grafiken reicht für viele Zwecke vollkommen aus.“
„Lass den *p*-Wert (bei den üblichen Verfahren) ruhig weg, der sagt dir ohnehin nur, wie groß dein Effekt ist und wie viele Personen du untersucht hast.“
„Schreib lieber nicht über ‚Ursachen‘ oder ‚Folgen‘ oder ‚Wirkungen‘ usw., wenn du kein Experiment oder Quasi-Experiment durchgeführt hast.“
„Streiche das Wort ‚Beweis‘ aus deinem Vokabular, wenn du Ergebnisse von Datenanalysen berichtest.“
„Ich weiß, dass du gern Strukturgleichungsmodelle berechnen würdest, aber hast du eine Theorie oder ein Modell, am besten zwei oder mehr, anhand deren du sinnvolle, am besten miteinander konkurrierende mathematische Strukturen aufstellen kannst? Wenn nein, dann ist ein Strukturgleichungsmodell vielleicht nicht der richtige Weg.“
„Die Theorie und die (begründete) Fragestellung sind das wichtigste Element in einer Untersuchung, auch und vor allem für eine sinnvolle Datenanalyse – versuche, da den größten Wert drauf zu legen.“

Wie oft habe ich dabei festgestellt, dass sich mein Gegenüber, nach einem verstohlenen Blick links und rechts, outete: „Du, um ganz ehrlich zu sein, habe ich nicht mehr ganz parat, was genau bei einer Korrelation gerechnet wird…“, „Was bedeutet denn ein Konfidenzintervall? Ich hab‘ das nie so richtig verstanden…“ oder „Es war mir wirklich nicht klar, dass Falsifikation mehr mit der Versuchsplanung als mit Inferenzstatistik zu tun hat“. Mit Anekdoten dazu könnte ich viele Seiten füllen.

Es ist *the elephant in the room*, der in der Öffentlichkeit, auch und gerade in der wissenschaftlichen Öffentlichkeit sehr selten angesprochen wird, obwohl es

allen gut bekannt ist: Viele von uns sind nicht besonders gut in – und auch nicht besonders interessiert an – Forschungsmethoden, vor allem der Statistik. Das führt dazu, dass wir uns nach dem Studium selten forschungsmethodisch weiterbilden und bestimmte Wissens- und Verständnislücken oder Missverständnisse weiter bestehen bleiben.

Meine Erfahrung ist die: Die allermeisten Studierenden, ganz gleich, ob im ersten Semester des Bachelors oder im letzten Mastersemester, wollen „nur durch", d. h. die Modulprüfung bestehen, was die Statistik betrifft. Einige dieser Studierenden werden zu Doktoranden, aber der Status der Statistik ändert sich auch hier nicht nennenswert – ich paraphrasiere eine/n Doktoranden/in, dem/der ich vorgeschlagen hatte, das genutzte Datenanalyseverfahren zu erklären: „Du, ehrlich gesagt, interessiert mich das nicht. Ich will nur wissen, welche Schritte ich da durchführen muss und wie ich das berichte." Auch viele Fachdozierende und Berufsforschende sind oft eher an Arbeitsteilung interessiert, wenn es um Datenauswertung, vor allem mit unbekannten Verfahren geht, als am Statistiklernen. Die ureigenste Aufgabe der Wissenschaft, ihre Existenzberechtigung – das Schaffen von Wissen, auch auf individueller Ebene – gerät aus dem Blick.

Das hat gravierende Konsequenzen. Viele der genutzten statistischen Methoden werden zu einer magischen Blackbox, deren Verständnis uns, einfachen Sterblichen, nicht möglich scheint und so wird die Datenauswertung zu einem Ritual (siehe dazu z. B. Gigerenzer, 2004 oder Gigerenzer & Marewski, 2015). In der Folge werden Verfahren nicht nach inhaltlich-rationalen Überlegungen angewendet, sondern aus Traditionsgründen: „Das haben die meisten bisher so gemacht" (siehe z. B. Fabriger, Wegener, MacCallum & Strahan, 1999 zum Einsatz von explorativen Faktorenanalysen, Miller & Chapman, 2001 zur Kovarianzanalyse, Nachtigall, Kroehne, Funke & Steyer, 2003 zu Strukturgleichungsmodellen, Sijtsma, 2009 zur internen Konsistenz oder Cohen, 1994 und Gigerenzer, 2004, 2018 zum Signifikanztest). Mir scheint es nur folgerichtig, dass das Wissenschaftliche der Wissenschaft dadurch abhandenkommt.

Dass das nicht bloß meine subjektive Ansicht ist, zeigt ein Blick auf die Forschungslandschaft. Die „Replikationskrise" oder neutraler ausgedrückt, der Mangel an gelungenen Replikationen stellt (mindestens) einige Ergebnisse der Humanwissenschaften, allen voran die der Psychologie in Frage, wie einige große

Replikationsprojekte eindrücklich zeigen (Open Science Collaboration, 2015, Hagger et al., 2016). Vor mehr al 20 Jahren bemängelte Gerd Gigerenzer den Umstand, dass viele unserer Theorien in der Psychologie einiges zu wünschen übriglassen (Gigerenzer, 1998, 2010). Etwa zehn Jahre zuvor übte Jacob Cohen – nicht zum ersten Mal – Kritik an den in der Psychologie genutzten typischen Forschungsvor- (oder -ver-?) -gehen in einem seiner berühmten Artikel „The earth is round ($p < .05$)“ (1994). Im Laufe seines Lebens veröffentlichte Paul Meehl mehrere kritische Artikel zu den häufig genutzten Methoden in der (akademischen wie praktischen) Psychologie (z. B. Meehl, 1954 zu klinischen Urteilen und 1978 zum Nullhypothesensignifikanztest) und noch früher erklärte Lewin, dass und warum „nichts so praktisch sei, wie eine *gute* Theorie“ (Lewin, 1951, S. 169, übersetzt durch die Autorin, Hervorhebung hinzugefügt). An guten Ratschlägen fehlt es zumindest in der Psychologie also nicht. Trotzdem hat sich die Situation in den letzten (mindestens) sechzig Jahren scheinbar kaum geändert.

Was können wir tun?

Wie bei allen komplexen Problemen gibt es hier keine einfache Lösung. Aber – und das ist der große Vorteil der Anwendung wissenschaftlicher Methoden – es gibt zumindest eine vernünftige Blaupause, die wir nutzen können, um das Problem anzugehen. Hierbei sollten wir *wissenschaftlich begründete* Antworten finden auf Fragen, wie „Lehren wir in der statistischen Hochschulausbildung sinnvolle Inhalte?“, „Lehren wir Inhalte auf sinnvolle/effektive Weise?“ und damit zu allererst „Wie können wir herausfinden, was wir lehren sollten und wie?“. Ich glaube fest daran, dass es sich lohnen wird, diese lange Reise zu unternehmen und am Ende die, meines Erachtens, das wissenschaftliche Prinzip am besten wiedergebende Forderung „Sapere aude!“ (Kant, 1784) zu erfüllen.

Zusammenfassung

In der Arbeit wird das ins Deutsche übersetzte Statistical Reasoning Assessment (SRA, Garfield, 2003) hinsichtlich seiner Eignung zur Erfassung statistischer Kompetenz bewertet. Da die Messintention des SRA (*statistical reasoning*) als Konstrukt, Lehrziel oder kognitiver Prozess verstanden werden kann, erfolgt die Validierung hinsichtlich dieser drei Aspekte. Das Instrument wurde in vier Stichproben (Psychologiestudierende im ersten und zweiten Semester, $n = 31$ und $n = 51$, Sonderpädagogikstudierende, $n = 277$, sowie Masterstudierende mit wenig statistischem Vorwissen, $n = 34$) jeweils zu Beginn und am Ende einer einsemestrigen Statistiklehrveranstaltung eingesetzt. Zusätzlich zum SRA wurden kognitive Leistungsmaße (mathematische Fertigkeiten, deduktives Schließen, Figurreihenfortsetzungsaufgaben) sowie Einstellungen zu Statistik (Survey of Attitudes Towards Statistics) erhoben. Die explorativen und konfirmatorischen faktoriellen Analysen des SRA für die 8 Stichproben ergeben keine klaren Hinweise auf eine erwartete ein-, vier- oder achtdimensionale Struktur für das correct statistical reasoning. Der Vergleich der Summe gelöster Items und Itemschwierigkeiten zwischen Stichproben (Psychologie vs. Sonderpädagogik, Master) und im Zeitverlauf (Beginn vs. Ende des Semesters) ergeben nur für Studierende der Psychologie im ersten Semester niedrige bis moderate Effekte. Korrelations- und Regressionsanalysen zeigen für den SRA-Wert am Ende des Semesters geringe inkrementelle Varianzaufklärung durch Figurreihenfortsetzungsaufgaben über andere kognitive Maße, Einstellungen zu Statistik sowie den SRA-Wert zu Beginn des Semesters hinaus. Insgesamt lassen sich wenige Hinweise auf die Inhaltsvalidität, die konvergente und diskriminante Konstrukt- sowie Lehrzielvalidität beobachten. Auf Basis der Ergebnisse wird eine Empfehlung für die Konstruktion eines validen Instruments zur Erfassung von Statistikkompetenz aufgezeigt.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1. Einleitung

Was die Wissenschaft betrifft, kann man über vieles hitzig diskutieren: wie man Wissenschaft am besten betreibt, was als Wissenschaft gelten kann und was nicht, ob sich Wissenschaft der Ethik unterordnen sollte und vieles mehr (z. B. Poser, 2012). Was dagegen üblicherweise als unstrittig gilt, ist die Einschätzung, dass Wissenschaft erfolgreich[1] ist, zumindest wenn man als Erfolg bzw. als die Voraussetzung für Erfolg das wiederholte Herbeiführen von Veränderungen in Paradigmen, dem Wissensstand und den menschlichen Tätigkeiten versteht und die Wissenschaft dabei anderen Denk- und Handlungssystemen, wie denen der Religion oder der Politik, gegenüberstellt. Der Erfolg der Wissenschaft zeigt sich sowohl in Grundlagenwissenschaften, wie der Physik (etwa in Form von wiederholten grundlegenden Paradigmenwechsel, Kuhn, 1970) als auch in angewandten Wissenschaften, wie der Medizin (etwa in Form von wiederholten Empfehlungen von Änderungen ärztlicher Handlungen, wie denen in Diagnostik und Therapie, Eckart, 2009).

Der Erfolg des wissenschaftlichen Denk- und Handlungssystems wird in der Regel der in diesem System genutzten Methodik zugeschrieben, den wissenschaftlichen Methoden oder Forschungsmethoden[2] (z. B. Faust & Meehl, 1992). Das Merkmal Forschungsmethoden ist für die Wissenschaft dabei so prominent, dass einige Texte die Begriffe „science" und „scientific method[s]" stellvertretend verwenden (z. B. Chalmers, 2013, Andersen & Hepburn, 2016, Faust & Meehl, 1992). Auch finden in den Beschreibungen der meisten, wenn nicht aller, Einzelwissenschaften formale oder empirische Forschungsmethoden ausdrückliche

[1] Der Autorin ist bewusst, dass der Versuch einer stichhaltigen Definition von „Wissenschaft" und „wissenschaftlichem Erfolg" den Rahmen dieser Arbeit sprengen würde und möglicherweise grundsätzlich nicht widerspruchsfrei gelingen kann. Für einen Überblick zum wissenschaftlichen Fortschritt als Erfolg der Wissenschaften siehe z. B. Niilinuoto (2015). Ein Überblick über grundlegende Fragen der Wissenschaftsphilosophie findet sich z. B. bei Chalmers, 2013 und der Wissenschaftstheorie bei Poser, 2012. Vereinfachend und aus rein pragmatischen Gesichtspunkten sollen im Folgenden die an Hochschulen studierbaren und dort beschriebenen Fächer als (Einzel-) Wissenschaften verstanden werden und Erfolg als wiederholte Änderungen im Wissensstand

[2] Im Folgenden werden die beiden Begriffe „wissenschaftliche Methoden" und „Forschungsmethoden" synonym verwendet. Im englischsprachigen Raum werden sowohl wissenschaftliche Methoden als auch Forschungsmethoden als „scientific method(s)" bezeichnet. Im deutschen Sprachraum scheint sich die Verwendung des Begriffs „Forschungsmethoden" insbesondere in human- und sozialwissenschaftlichen Fächern eher verbreitet zu haben (z. B. Döring & Bortz, 2016).

und prominente Erwähnung (siehe etwa Einführungslehrbücher, Modulbeschreibungen oder Beschreibungen der jeweiligen Wissenschaften auf Universitätswebseiten).

Die Definition dessen, was wissenschaftliche Methoden auszeichnet und von unwissenschaftlichen Methoden abgrenzt, erweist sich allerdings bei genauerem Hinsehen als äußerst schwierig und beschäftigt Wissenschaftsphilosophen als explizites Problem bereits seit mindestens fünf Jahrzehnten und in impliziter Form seit Beginn der Wissenschaften selbst (Poser, 2012, einen der stärksten Angriffe auf spezifische Methoden als Garant wissenschaftlichen Erfolgs führt Feyerabend, z. B. 1993). Allen wissenschaftlichen Methoden, sowohl den nicht-empirischen (z. B. formalen) als auch den empirischen, ist jedoch gemeinsam, dass sie in sich widerspruchsfrei, also logisch konsistent sein und einer rationalen Auseinandersetzung mit dem interessierenden Gegenstand dienen sollen (Haack, 1999). Die zentrale Funktion der wissenschaftlichen Methoden besteht darin, dass mit ihrer Hilfe bestimmte a priori getroffene wahrheitsfähige Aussagesätze (z. B. Hypothesen) oder Aussagensatzsysteme (z. B. Theorien) auf ihren Wahrheitsgehalt hin geprüft werden *können* sollen, mit anderen Worten, dass Aussagen sich zumindest *prinzipiell* durch Anwendung von Forschungsmethoden widerlegen lassen sollten. Diese Eigenschaft grenzt Forschungsmethoden deutlich von anderen Methoden, etwa Methoden der Rhetorik oder der in der Religion genutzten Methoden (z. B. Autoritäts- oder Glaubensbegründungen) ab[3] (Döring & Bortz, 2016). Hier soll aus pragmatischen Gründen eine stark vereinfachte Begriffsanalyse der für die vorliegende Arbeit zentralen *empirischen Forschungsmethoden* angeboten werden. So können Methoden der Versuchsplanung, der systematischen Datengenerierung (oder Datenerhebung) und der Datenmodellierung (oder Datenauswertung, oft synonym mit „Statistik" bezeichnet, z. B. Döring, 2019), die es ermöglichen, falsifikationsfähige Theorien und Hypothesen aufzustellen und zu prüfen, zumindest aus einer pragmatisch-praktischen Perspektive derzeit als wissenschaftlich betrachtet werden (Mayo & Spanos, 2010).

[3] Die Charakterisierung der Forschungsmethoden muss an dieser Stelle in ihrer äußerst vereinfachten und reduzierten Form erfolgen, da – ebenso, wie bei dem Begriff Wissenschaft – eine detaillierte Auseinandersetzung den Rahmen der Arbeit sprengen würde (für einen kurzen Überblick siehe Anderson & Hepburn, 2016)

Wenn man Forschungsmethoden als Alleinstellungsmerkmal von Wissenschaft akzeptiert, so folgt daraus, dass eine erfolgreiche Tätigkeit von individuellen Wissenschaftlern die erfolgreiche Anwendung und damit die Beherrschung von Forschungsmethoden voraussetzt. Damit erweist sich die Beherrschung von Forschungsmethoden in direkter Weise als zentral für die wissenschaftliche oder, allgemeiner formuliert, forscherische Betätigung.

Da die Ergebnisse wissenschaftlicher Tätigkeit die Grundlage für die praktische Tätigkeit in vielen hochqualifizierten Berufen darstellen, können Forschungsmethoden als (mindestens) in indirekter Weise auch zentral für die jeweilige berufliche Tätigkeit angesehen werden. Die enge Beziehung zwischen Wissenschaft und beruflicher Praxis ergibt sich natürlich[4]: innerhalb der Landwirtschaft, dem Ingenieurs- und Bauwesen, der Verwaltung, der Medizin, der Schifffahrt, dem Kriegswesen, um nur einige Berufsfelder zu nennen, werden Ergebnisse der wissenschaftlichen Betätigung bereits seit Jahrtausenden aufgegriffen und genutzt oder es wird innerhalb der Berufe selbst Forschung betrieben (Überblick bei Sommer, Müller-Wille & Reinhardt, 2017). In den letzten Jahrhunderten sind aber auch weitere berufliche Felder, wie beispielsweise die Diagnostik im Personal- und Schulwesen dazugekommen, die sich unmittelbar auf wissenschaftliche Ergebnisse beziehen und regelmäßig die Anwendung von Forschungsmethoden enthalten (etwa die Konstruktion eines Personalauswahlinstruments oder eines Instruments zur Leistungsstandmessung in der Schule). Die Relevanz der Forschungsmethoden für forschende Praktiker scheint daher offensichtlich. Weniger offensichtlich, aber ebenso folgerichtig, ist die große Bedeutung der Forschungsmethoden für diejenigen Praktiker, die in ihren Berufen wissenschaftliche Erkenntnisse nutzen: Da sich die Wissenschaft gerade durch beständige (und oft unerwartete) Änderungen in Paradigmen, dem aktuellen Wissenstand und den Tätigkeiten auszeichnet, müssen Praktiker auf dem aktuellen Stand der Forschung bleiben und in der Lage sein, wissenschaftliche Positionen zu verstehen und kritisch zu bewerten, um ihre berufliche Tätigkeiten optimal ausüben zu können (und um etwa Modeerscheinungen von ernstzunehmenden neuen Entwicklungen abgrenzen zu können). Die Kommunikation wissenschaftlicher

[4] und lässt sich keineswegs erst durch den in den letzten Jahrzehnten populär gewordenen Vorsatz „evidence based" erkennen

Positionen und neuer Erkenntnisse geschieht in aller Regel in der Sprache der Forschungsmethoden des betreffenden Fachs, so dass Rezipienten die jeweiligen Forschungsmethoden beherrschen müssen, um an der Kommunikation erfolgreich teilnehmen zu können. Nur mit Hilfe von forschungsmethodischem Knowhow können hochqualifizierte Berufstätige daher ihre Aufgaben in der Bildung oder Medizin und vielen anderen Berufen bestmöglich ausüben, auch wenn sie selbst nicht forschend tätig sind.

Damit erweist sich das Beherrschen wissenschaftlicher Methoden als zentral sowohl für die wissenschaftliche bzw. forscherische Tätigkeit (in direkter Weise) als auch für die angewandt-praktische Tätigkeit in vielen hochqualifizierten Berufen (in mindestens indirekter Weise).

Wenn man nun die häufig unausgesprochene Forderung, dass sowohl forscherische als auch angewandt-praktische Tätigkeiten so erfolgreich, wie möglich betrieben werden sollten, explizit als Annahme hinzunimmt, so folgt aus den Überlegungen weiter oben, dass das Beherrschen von Forschungsmethoden sowohl für Forschende als auch für angewandt-praktisch Tätige ein sinnvoll zu forderndes Attribut ist. Um den geforderten Beherrschungsgrad von Forschungsmethoden zu erreichen gibt es verschiedene Wege: die in Forschung oder Praxis Tätigen könnten die Beherrschung von Forschungsmethoden etwa on the job durch learning by doing erwerben. Dieser Weg hat, wie alle anderen on-the-job-Optionen, den Nachteil, dass vor allem bei den ersten Projekten mit suboptimaler Qualität bzw. mangelndem Erfolg der Aufgabenbearbeitung zu rechnen ist. Da einige der hochqualifizierten Berufe jedoch Aufgaben enthalten, die unmittelbar über Leben und Tod von Personen entscheiden können (etwa Ärzte und Ärztinnen, Pharmazeutinnen und Pharmazeuten in der klinischen Forschung, Ingenieure und Ingenieurinnen im Brückenbau und viele andere), scheinen on-the-job-Optionen aufgrund des damit einhergehenden Risikos von lebensbedrohlichen Situationen für eine potenzielle Vielzahl von Menschen nicht verantwortbar zu sein. Eine andere, derzeit praktizierte Variante besteht darin, dass Forschungsmethoden einen Teil der beruflich qualifizierenden Hochschulausbildung darstellen. Damit wird das Beherrschen der Forschungsmethoden für Studierende zu einem wichtigen Studienziel. Dies spiegelt sich auch in der Rechtslage der Ausbildung an Hochschulen wider. So lautet ein Auszug aus dem bundesländerübergreifenden Hochschulrahmengesetz (HRG,

1999): „Lehre und Studium sollen [den Studierenden] […] Kenntnisse, Fähigkeiten und Methoden […] so vermitteln, dass [sie] zu wissenschaftlicher […] Arbeit und zu verantwortungsvollem Handeln […] befähigt [werden]" (§ 7 HRG, 1999). Mit guten Gründen kann also gefordert werden, dass Forschungsmethoden von Studierenden spätestens am Ende der Hochschulausbildung beherrscht werden. Dabei sollte die Lehre an der Hochschule mindestens zum Teil dafür sorgen, dass sich der Beherrschungsgrad der Forschungsmethoden bei Studierenden im Verlauf des Studiums verbessert.

Während die entscheidende Rolle der Forschungsmethoden für viele hochqualifizierte Tätigkeiten in Forschung und Praxis sowie der Rahmen, in dem das Knowhow der Forschungsmethoden idealerweise erworben werden sollte, gewissermaßen „im Lehnstuhl" geklärt werden kann, muss die Frage danach, ob die Hochschule den beiden Zielen „Sicherstellen der Beherrschung von Forschungsmethoden bei Studierenden" und „Verbesserung der Beherrschung von Forschungsmethoden bei Studierenden durch Lehre" gerecht wird, unter Zuhilfenahme empirischer Daten beantwortet werden. Mit anderen Worten muss dem normativ-theoretischen Soll (in Form der beiden angeführten Ziele) ein empirischer Ist-Zustand gegenübergestellt werden:

1) Beherrschen Studierende am Ende ihres Studiums Forschungsmethoden (in ausreichendem Ausmaß)?
2) Trägt die Hochschullehre (in ausreichendem Ausmaß) zu der Beherrschung von Forschungsmethoden bei?

Diese beiden Fragen lassen sich nur beantworten, wenn Informationen über den Beherrschungsgrad von Forschungsmethoden bei Studierenden vorliegen. Damit wird ein Instrument benötigt, das den Beherrschungsstand von Forschungsmethoden bei Studierenden zuverlässig erfasst.

Eine Herausforderung bei der Erstellung eines solchen Instruments stellt die große Vielfalt der Forschungsmethoden dar. Von formalen Methoden der Logik und Beweisführung über die statistischen Methoden der Gruppierung von Hirnaktivitätsmessungen bis zu den verschiedenen Verfahren der Textinterpretation: die Heterogenität der Forschungsmethoden ist nahezu so groß, wie die Anzahl der sie nutzenden Fächer. Es kann daher vermutlich nicht gelingen, Aussagen über alle

Gruppen von Forschungsmethoden anhand der Ergebnisse nur eines Erhebungsinstruments zu treffen.

Hinzu kommt, dass, auch wenn manche grundlegenden und häufig implizit genutzten Methoden (etwa basale Aussagen- und/oder Prädikatenlogik) allen Einzelwissenschaften gemeinsam sind, eine große Menge der Methoden dem jeweiligen Fach oder einer Fächergruppe eigen ist. So werden in mathematischen Wissenschaften zwar zahlreiche formale Methoden genutzt, aber nur in seltenen Fällen (wenn überhaupt) empirische. In der Physik und verwandten Fächern werden sowohl formale (mathematische) als auch empirische Methoden angewendet. In den Geschichtswissenschaften und einigen anderen Geisteswissenschaften werden hingegen fast ausschließlich empirische qualitative Methoden genutzt. Ein Instrument zur Erhebung der Beherrschung von Forschungsmethoden sollte daher vermutlich spezifisch für eine verwandte Gruppe von Einzelwissenschaften zugeschnitten sein.

In der vorliegenden Arbeit richtet sich der Fokus auf die beiden Fächer Psychologie und Sonderpädagogik als Lehramt. Gemeinsam ist den zwei Wissenschaften, dass sowohl für Psychologen als auch für Sonderpädagogen die Erhebung, Auswertung und Interpretation von Messungen psychologischer Konstrukte, wie Intelligenz und Aggressivität, eine große Rolle spielen, ebenso wie die Diagnostik von Kompetenzen, Auffälligkeiten oder Störungen. Damit scheint es zulässig zu sein, das gleiche Erhebungsinstrument bei Studierenden beider Fächer einzusetzen. Während jedoch das Fach Psychologie (Bachelor of Science) an den meisten Hochschulen mit etwa 40 Leistungspunkten in Forschungsmethoden (darunter Statistik, Diagnostik und das empirisch-experimentelle Praktikum, Abele-Brehm et al., 2014) ein eher methodenstarkes Fach ist, stellt die Sonderpädagogik als Lehramt (Bachelor of Arts) mit maximal 18 Leistungspunkten in Forschungsmethoden ein Fach mit einem vergleichsweise geringen forschungsmethodischen Anteil dar (unter den 18 Leistungspunkten vor allem Diagnostik und ein Teilmodul mit zwei Leistungspunkten in Statistik; 44% der Studiengänge deutschlandweit führen in den Modulbeschreibungen kein eigenständiges Modul oder Teilmodul zu Forschungsmethoden auf [Sawatzky & Rietz, 2016]). Aufgrund der relativen inhaltlichen Ähnlichkeit auf der einen Seite und des deutlichen Unterschieds in dem Umfang der gelehrten Forschungsmethoden

auf der anderen Seite sollten sich die beiden Fächer daher gut für einen Gruppenvergleich im Sinne diskriminanter Validität des Erhebungsinstruments eignen.

Die in der Psychologie und der Sonderpädagogik genutzten empirischen[5] Forschungsmethoden lassen sich in der Regel den Methoden der Versuchsplanung, der Datenerhebung oder der Datenauswertung bzw. Statistik zuordnen. Da die Vielfalt und der Umfang dieser Forschungsmethoden immer noch sehr groß sind, soll im Folgenden noch eine weitere Einschränkung vorgenommen werden. Als empirische Wissenschaften sind beide Fächer auf Methoden der aussagekräftigen Analyse von (numerischen) Daten angewiesen. Daher wird der Fokus in der vorliegenden Arbeit auf Statistik (im Sinne der Datenanalyse) als eine der für die beiden Fächer relevanten Forschungsmethoden gelegt.

Die Frage danach, ob bzw. wie gut Studierende statistische Inhalte beherrschen, beschäftigt Forscher und Lehrende bereits seit einiger Zeit, und so ist es nicht weiter verwunderlich, dass in den letzten Jahrzehnten einige Erhebungsinstrumente entwickelt wurden, um statistische Kompetenzen zu erfassen (siehe Abschnitt 3.2). Ein solches Instrument ist das Statistical Reasoning Assessment (SRA, Garfield, 1991, 2003). Obgleich zu dem Instrument einige Studien durchgeführt wurden, liegt bisher noch keine Beurteilung des SRA vor dem Hintergrund der weiter oben formulierten Anforderungen vor. Es ist damit noch offen, wie gut das SRA geeignet ist, um einerseits den Beherrschungsstand von Statistik im Sinne des Ziels einer statistischen Hochschulausbildung und andererseits Effekte von Lehrinterventionen auf die Beherrschung von Statistik zu erfassen.

[5] Beide Wissenschaften nutzen auch nicht-empirische Methoden, z. B. die der Theoriekonstruktion, die selbstverständlich wichtige Implikationen für das empirische Vorgehen haben. Im vorliegenden Zusammenhang liegt der Fokus jedoch der Machbarkeit halber auf den „rein“ empirischen Methoden.

1.1. Zielsetzung der Arbeit

Die vorliegende Arbeit setzt sich damit das Ziel, zu klären, ob das Instrument SRA geeignet ist, um

1) die Beherrschung statistischer Inhalte bei Studierenden der Psychologie und Sonderpädagogik und
2) Effekte der Hochschullehre auf die Beherrschung statistischer Inhalte bei Studierenden der Psychologie und Sonderpädagogik

zu erfassen.

Hierfür soll zum einen offengelegt werden, wie die Beherrschung statistischer Inhalte im SRA konzeptionell verstanden wird. Da im Laufe der letzten 40 Jahre verschiedene Vorschläge dazu formuliert worden sind, wie Beherrschung statistischer Inhalte konzipiert werden kann, sollen auch die Unterschiede der konzeptionellen Grundlage des SRA zu anderen Konzepten der Beherrschung statistischer Inhalte aufgezeigt werden. In diesem Zusammenhang soll ferner beleuchtet werden, wie sich die konzeptionelle Grundlage des SRA in die Art der angenommenen psychologischen Variablen einordnen lässt, d. h. ob es sich hierbei um ein psychologisches Konstrukt, ein Kriterium oder einen kognitiven Prozess handelt. Relevant ist eine solche Betrachtung vor dem Hintergrund, dass sich, je nach intendierter Art der Variablen, unterschiedliche Anforderungen an die Konstruktion und die psychometrischen Eigenschaften des Instruments ergeben.

Zum anderen soll geklärt werden, ob das Instrument in der Lage ist, Interventionseffekte in Form von Hochschullehre abzubilden. Diese lassen sich einerseits durch Veränderungen in der Beherrschung statistischer Inhalte durch Lehrinterventionen und andererseits durch Unterschiede in den Veränderungen in Abhängigkeit vom Umfang sowie von den Inhalten der Lehrinterventionen operationalisieren.

Insgesamt soll die vorliegende Arbeit damit dazu dienen, zunächst sowohl aus theoretisch-inhaltlichen als auch empirischen Gesichtspunkten die Eignung des SRA zur Erfassung der Beherrschung statistischer Inhalte bei Studierenden zu beurteilen. Schließlich sollen auf Basis der Beurteilung Kriterien für Messinstrumente statistischer Beherrschung abgeleitet und Empfehlungen für die Konstruktion und Validierung solcher Messinstrumente gegeben werden.

1.2. Aufbau der Arbeit

Die Inhalte der vorliegenden Arbeit lassen sich in neun Sinnabschnitte einteilen. Zunächst erfolgt die Darstellung der prominentesten Konzeptualisierungen der Beherrschung statistischer Inhalte mit einem kurzen vorangestellten historischen Überblick. Wo möglich, wird dabei die theoretische Grundlage der Konzeptualisierungen erläutert. Abgeschlossen wird dieser Abschnitt mit bisher vorgelegten Systematisierungsvorschlägen und einem Vergleich der Konzeptualisierungen.

Anschließend wird das Erhebungsinstrument SRA detailliert beschrieben und auf der Basis der bisher vorliegenden Forschung als Messinstrument der Beherrschung statistischer Inhalte im Sinne eines Konstrukts, eines Kriteriums sowie eines kognitiven Prozesses beurteilt. Hierbei werden noch offene Fragen an die Eigenschaften des Erhebungsinstruments aufgeführt. Der Abschnitt schließt mit einem Überblick über weitere Messinstrumente der Beherrschung statistischer Inhalte.

Unter Bezugnahme auf die im vorangegangenen Abschnitt identifizierten offenen Fragen hinsichtlich der Eignung des SRA für die Erfassung der Beherrschung statistischer Inhalte erfolgt die Ableitung der Fragestellungen und Hypothesen im nächsten Sinnabschnitt. Die Abschnitte „Methode“ und „Ergebnisse“ widmen sich dem empirischen Teil der Arbeit. Obwohl für die empirischen Untersuchungen im vorliegenden Fall vier unterschiedliche Stichproben hinsichtlich zum Teil unterschiedlicher Variablen erhoben wurden, genau genommen hier daher vier Studien vorliegen, werden die Ergebnisse nicht nach Stichproben (d. h. nach einzelnen Studien), sondern nach den im vorangegangenen Abschnitt abgeleiteten Fragestellungen unterteilt. Damit soll der Aspekt der Replikation auf der einen und der von Gruppen- (oder Populations-)Unterschieden auf der anderen Seite deutlich werden.

Im darauffolgenden Sinnabschnitt („Zusammenfassung und kritische Interpretation der Ergebnisse“) werden die empirischen Ergebnisse nach den drei Fragestellungen zusammengefasst, interpretiert und hinsichtlich der Grenzen der Aussagekraft diskutiert. Abschließend erfolgt in diesem Zusammenhang eine gesamtheitliche Beurteilung des SRA.

Der Abschnitt „Einschränkungen der Forschungsbemühungen allgemein“ dient der Beurteilung der bisherigen Forschungsbemühungen zur Erfassung der Beherrschung statistischer Inhalte und zur Statistik-Hochschullehre insgesamt. Es werden hier insbesondere Aspekte der Zielsetzung, der theoretischen Grundlagen sowie der methodischen Umsetzung behandelt.

Auf der Basis der identifizierten Optimierungsmöglichkeiten werden abschließend Empfehlungen für die Forschung zur Statistiklehre in der Hochschule allgemein sowie für Kriterien an ein Messinstrument zur Erfassung der Beherrschung statistischer Inhalte und damit verbunden für die Konstruktion und die Validierung eines solchen Instruments formuliert.

2. Konzepte der Statistikkompetenz

Bevor die in der Literatur bisher formulierten Vorschläge zur Konzipierung und Erfassung der Beherrschung statistischer Inhalte im Einzelnen angeführt werden, soll zunächst ein kurzer (und sicher nicht vollständiger) historischer Überblick über die Anlässe und die Entwicklung einiger Konzepte gegeben werden.

Hierbei wird statt *Beherrschung statistischer Inhalte* aus Gründen der besseren Lesbarkeit *Statistikkompetenz* als Ober- oder Containerbegriff in einem allgemeinen Sinne als Fähigkeit und/oder Fertigkeit, erfolgreich mit statistischen Konzepten umzugehen, verwendet. Zu beachten ist dabei, dass der Begriff stets lediglich eine umreißende Platzhalterfunktion und keine definitorisch-genaue Bedeutung haben soll (entgegen des Gebrauchs bei z. B. Weinert, 2001/2014 oder Klauer, 1987 und eher im Sinne von Weinert, 1999).

Eine weitere sprachliche Besonderheit verdient eine kurze Erläuterung. So werden die vorgeschlagenen Entwürfe der Statistikkompetenz durchgängig als *Konzepte*, *Begriffe* oder *Konzipierungen* u. ä. bezeichnet, da größtenteils unklar bleibt, ob sich die entsprechenden Autoren auf Konstrukte im Sinne latenter Variablen (etwa Eigenschaften bzw. *traits*), Kriterien im Sinne von Lehrzielen, kognitive Prozesse oder andere psychologische Variablen beziehen. Eine Bezeichnung der einzelnen Konzeptvorschläge als z. B. Konstrukt stünde damit einerseits möglicherweise im Widerspruch zu der Intention der Autoren und würde andererseits einen höheren theoretischen Reifegrad der Konzepte suggerieren.

Für die einzelnen Konzeptvorschläge werden für eine höhere Genauigkeit und einen besseren Lesefluss ihre ursprünglichen Bezeichnungen, meist in englischer Sprache, genutzt. So gehen semantische Feinheiten bei einer Übersetzung häufig verloren (siehe z. B. Nuerk, Engel & Martignon, 2015). *Literacy* etwa kann zwar in die Wortneuschöpfung *Alphabetismus* als Antonym zu Analphabetismus übertragen werden, allerdings meint literacy auch Grundbildung, häufig die Grundbildung in einem bestimmten Bereich, wie etwa *statistical literacy*. Auch der Begriff *reasoning* meint mehr als die direkte Übersetzung *schlussfolgern*, so dass eine sinngemäße Übertragung ins Deutsche mehrere Wörter (*schlussfolgern*, *argumentieren*, *begründen*, *beurteilen*) erfordern würde.

2.1. Historische Einordnung

Das Interesse an und die wahrgenommene Wichtigkeit von Statistikkompetenz hat eine lange Tradition, die in das 19. Jahrhundert zurückreicht (kurzer Überblick bei Zieffler, Garfied und Fry, 2018). Eine der ersten expliziten Empfehlungen, den Umgang mit quantitativen Informationen im Rahmen der Schulbildung unter einer eigenständigen Bezeichnung zu berücksichtigen, stammt aus dem Crowther Report (Central Advisory Council For England & Crowther, 1959), der im Auftrag des englischen Bildungsministeriums erstellt wurde. Crowther nimmt die folgende Einschätzung vor:

> Statistical ignorance and statistical fallacies are quite as widespread and quite as dangerous as the logical fallacies which come under the heading of illiteracy. The man who is innumerate is cut off from understanding some of the relatively new ways in which the human mind is now most busily at work. [...] The educated man, therefore, needs to be numerate as well as literate. (S. 270-271)

Er führt im gleichen Rahmen den Begriff *numeracy* für einen kompetenten Umgang mit Zahlen und quantitativen Informationen sowie den forschungsmethodischen Konzepten, wie Datenerhebung, Hypothese, Experiment und Verifikation als Äquivalent zum Begriff der literacy ein. In dem späteren Cockcroft Report an die englischen und walisischen Staatssekretäre für Bildung und Wissenschaft wird der Begriff numeracy in seiner Bedeutung etwas eingegrenzt und umfasst keine forschungsmethodischen Konzepte mehr (Cockcroft, 1982). Jedoch wird in diesem Rahmen nun der eigenständige Begriff *statistical numeracy* für Kompetenzen im Umgang mit statistischen Inhalten auf der Basis der Ergebnisse des Schools Council Project on Statistical Education (zitiert nach Cockcroft) formuliert:

> Statistical numeracy requires a feel for numbers, and appreciation of appropriate levels of accuracy, the making of sensible estimates, a commonsense approach to the use of data in supporting an argument, the awareness of variety of interpretation of figures, and a judicious understanding of widely used concepts such as means and percentages. (Cockcroft, 1982, S. 236)

Dieser Bericht diente als eine der Grundlagen für das Konzept der *quantitative literacy* in den Vereinigten Staaten von Amerika im Rahmen des durch die National Science Foundation (NSF) geförderten Quantative Literacy Projekts der American Statistical Association (ASA) und des National Council of Teachers of Mathematics (NCTM) in den 1980er Jahren (Schaeffer, 1990). Dieses groß angelegte Projekt diente der Entwicklung und Umsetzung von Lerneinheiten zu den Themen „Daten explorieren", „Wahrscheinlichkeit", „Simulation" sowie „Informationen aus Stichproben" für Lehrer (Schaeffer).

Die Zusammenarbeit der ASA und des NCTM begann in 1967 im Rahmen des Joint Committee on the Curriculum in Statistics and Probability (Zieffler et al. 2018) und führte später zur Formulierung von Bildungsstandards für das Schulfach Mathematik unter besonderer Berücksichtigung statistischer und stochastischer Inhalte (NCTM, 1989).

Im Jahr 1993 schlug die damalige Präsidentin der ASA, Katherine Wallman zusätzlich das Konzept statistical literacy in ihrer präsidialen Rede vor:

> 'Statistical Literacy' is the ability to understand and critically evaluate statistical results that permeate our daily lives – coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions. (S. 1)

Gleichzeitig entstanden in den 1970er Jahren intensive internationale Bestrebungen für ein eigenes Forum zur Lehre von Statistik, die im Jahr 1982 in der ersten internationalen Konferenz *First International Conference on Teaching Statistics* mündeten. Eigene Gesellschaften und Zeitschriften, wie das *International Statistical Institute* (ISI) *Education Committee* (1981) und später die *International Association for Statistical Education* (IASE, 1991), die *IASE Review* (1994), das *Statistics Education Research Journal* (ab 2002) sowie die *International Collaboration for Research on Statistical Reasoning, Thinking and Literacy* (1999) wurden im Rahmen dieser Bemühungen gegründet. Innerhalb dieser Entwicklungen nahm und nimmt (unter anderen) Joan Garfield eine zentrale Rolle ein. Sie war und ist, teilweise zusammen mit Dani Ben-Zvi, in großem Umfang an der Gründung und Pflege einiger Zeitschriften und Gesellschaften, Handbücher sowie Projekte zur Statistikausbildung (wie z. B. der Empfehlungen zur statistischen Hochschulausbildung) beteiligt und entwickelte im Rahmen des ChancePlus-

Projekts basierend auf den NCTM-Bildungsstandards das erste objektive Messinstrument zur Erfassung von Statistikkompetenz. Die konzeptionelle Grundlage für das Instrument bildet dabei das *statistical reasoning* und bezieht sich auf typische Fehldeutungen und Missverständnisse in Abgrenzung zu korrektem Verständnis statistischer Konzepte.

Parallel zu und teilweise in Verbindung mit den Bemühungen der ASA-NCTM-Kooperationen begannen in den 1990er Jahren durch die Mathematical Association of America (MAA) organisierte Bestrebungen für die Beschreibung von Empfehlungen zur statistischen Hochschulausbildung, ebenfalls in den Vereinigten Staaten von Amerika. Der resultierende sogenannte Cobb-Bericht (Cobb, 1992) bildete die Grundlage für die Formulierung von (unverbindlichen) Leitlinien zur Lehre und Leistungsmessung in Statistik an Schulen und Hochschulen (Guidelines for Assessment and Instruction in Statistics Education GAISE, ASA, 2005, Franklin et al., 2005). Cobb (1992) führt in diesem Zusammenhang den Begriff *statistical thinking* ein, den er abgrenzt von Statistikkompetenz als Verwendung von kochrezeptartigen Prozeduren zum einen und Statistikkompetenz als Verwendung ausschließlich formal-mathematischer Methoden zum anderen. Im GAISE College Bericht (ASA, 2005, 2016) wird die Empfehlung für *statistical thinking* als Ziel einer erfolgreichen Statistikhochschulausbildung von Cobb (1992) übernommen und im Bericht aus 2005 ergänzt durch die Empfehlung, einen Fokus auf *statistical literacy* zu setzen.

Gleichzeitig zu den Entwicklungen in Großbritannien und den Vereinigten Staaten von Amerika erfolgten auch in Australien, Neuseeland und Israel durch Veröffentlichungen und offizielle Empfehlungen für Lehrinhalte sichtbare Auseinandersetzungen mit dem Thema Statistikkompetenz (z. B. Pfannkuch, 1996, 1997, zitiert nach Wild & Pfankuch, 1999, Gal, 1995, Watson, 1997). Für die gegenwärtige Fragestellung sind hier vor allem die Konzepte statistical literacy (Watson, 1997; Gal, 2002; Watson & Callingham, 2003) sowie statistical thinking (Wild & Pfannkuch, 1999) von Relevanz.

Die Bemühungen in Deutschland beschränken sich bisher hauptsächlich auf die Integration statistischer Konzepte aus mathematischer Perspektive unter dem Begriff *Leitidee „Daten und Zufall“* innerhalb der Bildungsstandards (Eichler & Vogel, 2013, Biehler & Engel, 2015). Die statistischen Inhalte sind dabei vor allem

durch Stochastik abgedeckt. Bisher wurden entsprechende Bildungsstandards für den Primarbereich (KMK, 2004a), den Hauptschulabschluss (KMK, 2004b), den Mittleren Schulabschluss (KMK, 2003) sowie die Allgemeine Hochschulreife (KMK, 2012) formuliert.

Im historischen Verlauf der Beschäftigung mit Konzepten der Statistikkompetenz kristallisieren sich vor allem drei zentrale Begriffe heraus: statistical literacy, statistical thinking und statistical reasoning. Diese Konzepte sollen im Folgenden in größerem Detail zunächst beschrieben und anschließend verglichen werden.

2.2. Statistical literacy

Der Ausgangspunkt für die Entstehung des Begriffs statistical literacy lässt sich in der präsidialen Rede von Katherine Wallman (1993) bestimmen. Wie aus der Definition weiter oben zu erkennen ist, wird der Fokus hier auf das Verständnis und die kritische Beurteilung statistischer Informationen im Rahmen von Alltagssituationen gelegt. Wallman bezieht sich in ihrer Ansprache vor allem auf statistische Informationen, die im Rahmen staatlicher und wirtschaftlicher Bemühungen gewonnen werden und somit für die gesamte Gesellschaft in recht direkter Weise von Relevanz sind, im Gegensatz zu statistischen Informationen im Rahmen von akademischer Forschung. In diesem Zusammenhang stellt sie heraus:

> The concerns [...] [of citizens] stem from what I would charactarize as a series of 'mis-es' – misunderstandings, misperceptions, mistrust, and misgivings: misunderstandings about the sources of statistical data; misperceptions about the willingness of citizens to provide information; mistrust by some members of our population about how statistical data will be used; and misgivings about the value of statistics for guidance in public and private choices. These mis-es, I contend, are rooted in society's lack of 'statistical literacy'. (S. 1)

Wallman (1993) sieht also den Grund für Widerstände gegenüber der Erfassung und Nutzung von statistischen Informationen in einer defizitären statistischen Grundbildung. Damit wird statistical literacy hier in den Kontext des öffentlichen und privaten, nicht aber des akademischen bzw. forschungsbezogenen

Lebens gesetzt. Wallman versteht statistical literacy augenscheinlich daher als die Kompetenz, mit im öffentlichen (und privaten) Leben genutzten statistischen Informationen, wie Prozentangaben, Wahrscheinlichkeiten, Risiken, Maßen zentraler Tendenz etc. umzugehen.

Die eher globale Beschreibung von Statistikkompetenz der erwachsenen Allgemeinbevölkerung als Fähigkeit, statistische Ergebnisse zu verstehen und kritisch zu beurteilen sowie die im vorangegangenen Abschnitt skizzierten Empfehlungen für Ziele und Inhalte statistischer Schulbildung wurden in der Form von statistical literacy später von vor allem zwei Autoren aufgegriffen und in mehreren Veröffentlichungen (z. B. Watson, 1997; Gal, 2002; Watson & Callingham, 2003, 2005; Callingham & Watson, 2017) verfeinert, systematisiert und ergänzt. Die Arbeit der beiden Autoren (Iddo Gal und Jane Watson) soll in den nächsten zwei Abschnitten vorgestellt werden.

2.2.1. Vorschlag von Gal (2002).

Auf der Basis vorliegender Auseinandersetzungen mit dem Begriff statistical literacy (Wallman, 1993, Watson, 1997) und scientific literacy (z. B. Shamos, 1995) schlägt Gal (2002) die folgende Definition vor:

> (a) people's ability to *interpret and critically evaluate* statistical information, data-related arguments, or stochastic phenomena, which they may encounter in diverse contexts, and when relevant
> (b) their ability to *discuss or communicate* their reactions to such statistical information, such as their understanding of the meaning of the information, their opinions about the implications of this information, or their concerns regarding the acceptability of given conclusions. (S. 2-3, Hervorhebungen im Original)

Diese Definition enthält gegenüber der Definition von Wallman (1993) einen zusätzlichen Aspekt. So wird auch die Fähigkeit, sich über die Inhalte und die Bedeutung statistischer Informationen austauschen zu können als Teil von statistical literacy angesehen. Wie Wallman auch, sieht Gal hierbei als relevanten Kontext vor allem öffentliche Medien, wie Zeitungen, Zeitschriften, Nachrichtensendungen sowie Verbraucherinformationen (z. B. medizinischer Art).

Im gleichen Artikel schlägt Gal (2002) ein Modell für das Konzept statistical literacy vor. Er unterscheidet hierbei zwischen „'data consumers' [...] [and] 'data producers' or 'data analyzers'" (S. 3), die statistischen Informationen in „*[r]eading contexts*" in Abgrenzung zu „*enquiry contexts*" (beide S. 3, Hervorhebungen im Original) begegnen und beschränkt sein Modell auf den Lesekontext, d. h. auf die Rezeption statistischer Informationen. Das Modell nimmt zwei Befähigungskomponenten für eine erfolgreiche Auseinandersetzung mit statistischen Informationen an: eine Wissenskomponente und eine Dispositionskomponente. Zu den fünf Bestandteilen der Wissenskomponente zählt Gal literacy skills im Sinne grundlegender Lesekompetenzen, statistisches, mathematisches und kontextbezogenes Wissen sowie kritische Fragen. Die Dispositionskomponente enthält zum einen Überzeugungen und Einstellungen und zum anderen eine kritische Haltung. Die Komponenten und Bestandteile der Komponenten nimmt Gal dabei als abhängig von einem jeweiligen Kontext und in dynamischer Wechselwirkung stehend an.

Zu den Lesekompetenzen (literacy skills) als Bestandteil von statistical literacy zählt Gal (2002) nicht nur die Fertigkeit, mit allgemeinen sprachlichen Informationen zu operieren, sondern auch die Fertigkeit, die Bedeutung bestimmter statistischer Begriffe von ihrem umgangssprachlichen oder alltäglichen Sinn abgrenzen zu können. Des Weiteren zählt er auch *document literacy* zu dem allgemeinen sprachlichen Grundbildungsbestandteil von statistical literacy. Document literacy bezieht sich auf die Fertigkeit, Informationen in tabellarischen oder grafischen Darstellungen zu verorten und zu integrieren sowie neue Informationen zu generieren und Schlussfolgerungen zu ziehen.

Für das statistische Wissen schlägt Gal (2002) auf der Basis von Forschungsarbeiten zu mathematisch-statistischer Lehre fünf Inhaltsbereiche vor. Erstens sollten Personen wissen, warum Daten benötigt werden und wie Daten produziert werden können. Hierzu gehört es, anzuerkennen, dass Daten anekdotischer Evidenz vorzuziehen sind, dass aufgrund von Variation eine Reduktion von Daten notwendig ist und schließlich, dass Versuchspläne und die Art der Stichprobengewinnung die Qualität der Aussagekraft (im Sinne interner Validität) und Verallgemeinerbarkeit (im Sinne externer Validität) von Daten entscheidend beeinflussen. Zweitens sollten Personen mit grundlegenden

Begrifflichkeiten und Ideen der deskriptiven Statistik, wie etwa Prozentangaben oder Maßen zentraler Tendenz, vertraut sein. Drittens sollten Personen mit grundlegenden Begriffen und Ideen der tabellarischen und grafischen Darstellung von Daten vertraut sein. Hierzu gehört etwa das Bewusstsein darüber, dass unterschiedliche grafische Darstellungen unterschiedliche Datenmuster produzieren. Viertens sollten Personen grundlegende Ideen der Wahrscheinlichkeit, wie Prozente, Odds oder Chancen, verstehen und schließlich sollten Personen fünftens wissen, wie statistische Inferenzen (oder informelle Schlüsse) gezogen werden. Bei diesem letzten Punkt ist es nach Gal von großer Bedeutung, dass Rezipienten sensibel sind für verschiedene Möglichkeiten von systematischen und unsystematischen Fehlern bei der Stichprobengewinnung, der Messungen sowie beim Ziehen von Schlussfolgerungen.

Was das mathematische Wissen betrifft, verweist Gal (2002) auf nennenswerte Unterschiede zwischen der Statistik und der Mathematik. Während der inhaltliche Kontext in der Mathematik ausschließlich eine Beispielfunktion hat (und auch das lediglich in der Didaktik in den frühen Schulstufen) und für die mathematische Verfahren keinerlei Bedeutung hat, spielt der Inhalt in der Statistik die alles entscheidende Rolle (siehe dazu z. B. Moore, 1992, Cobb & Moore, 1997). Vor diesem Hintergrund erfordert statistical literacy nach Gal lediglich grundlegendes mathematisches Wissen, z. B. das Wissen darüber, wie Prozentwerte oder Mittelwerte bestimmt werden.

Aufgrund der großen Abhängigkeit statistischer Informationen von einem spezifischen Kontext spielt kontextbezogenes Wissen eine entscheidende Rolle für statistical literacy. Das Wissen über den inhaltlichen Kontext stellt laut Gal (2002) die Hauptdeterminante für das Erkennen oder Vermuten von Einschränkungen, Erklärungen und Fehlerquellen bei der Beurteilung statistischer Informationen dar.

Eng im Zusammenhang damit steht schließlich das Wissen über relevante kritische Fragen, die eine Person sich bei der Interpretation und Beurteilung statistischer Informationen stellen sollte. Gal (2002) schlägt hier zehn sogenannte *worry questions* vor, Fragen, die anzeigen, über was man sich bei statistischen Informationen gegebenenfalls Sorgen machen könnte (Gal, S. 16, sinngemäße Übertragung ins Deutsche durch die Autorin):

1. Wo kommen die Daten, auf deren Basis die Aussage zustande kommt, her? Welche Art von Studie war das? Ist diese Art von Studie in diesem Kontext angemessen und sinnvoll?
2. Wurde eine Stichprobe genutzt? Wie erfolgte die Stichprobenziehung? Wie viele Personen nahmen schlussendlich teil? Ist die Stichprobe groß genug? Enthielt die Stichprobe Personen/Objekte/Einheiten, die die Grundgesamtheit repräsentieren? Ist die Stichprobe in irgendeiner Weise verzerrt? Insgesamt betrachtet: Könnte diese Stichprobe vernünftigerweise valide Schlüsse über die Zielpopulation ermöglichen?
3. Wie reliabel und valide waren die Messinstrumente (Tests, Fragebögen, Interviews usw.), die für die Erhebung der vorliegenden Daten genutzt wurden?
4. Welche Form hat die Verteilung der Daten (auf deren Basis die vorliegende statistische Kennzahl bestimmt wurde)? Spielt es eine Rolle, wie die Form der Verteilung aussieht?
5. Sind die berichteten statistischen Kennzahlen angemessen für diese Art der Daten, wurde z. B. ein Mittelwert zur Zusammenfassung von ordinalen Daten genutzt; liefert der Modus in diesem Zusammenhang eine sinnvolle Zusammenfassung? Könnten Ausreißer zu einer Fehldarstellung der tatsächlichen Situation durch eine statistische Kennzahl geführt haben?
6. Wurde die dargestellte grafische Darstellung angemessen erstellt oder verzerrt sie die Muster in den Daten?
7. Wie wurde die vorliegende Wahrscheinlichkeitsaussage hergeleitet? Liegen genügend glaubhafte Daten vor, um die getroffene Wahrscheinlichkeitsaussage zu rechtfertigen?
8. Insgesamt betrachtet: Sind die Behauptungen in der Quelle sinnvoll und durch die Daten belegt? Wurde z. B. Korrelation mit Kausalität gleichgesetzt oder ein kleiner Unterschied als großer Unterschied dargestellt?
9. Sollten mir zusätzliche Informationen oder Verfahren (-sergebnisse) zur Verfügung gestellt werden, damit ich die Sinnhaftigkeit der Argumente

beurteilen kann? Fehlt etwas? Z. B., „vergaß“ der Autor, die Basis für die Änderung in Prozent anzugeben oder die Stichprobengröße?

10. Gibt es alternative Interpretationen der Bedeutung der Ergebnisse oder andere Erklärungen für deren Ursache, z. B. Mediations- oder Moderationseffekte? Gibt es zusätzliche oder andere Implikationen, die nicht genannt wurden? (Gal, 2002, S. 16)

Die von Gal (2002) vorgeschlagenen Dispositionen kennzeichnen die Bereitschaft, die fünf Bestandteile der Wissenskomponente zu aktivieren. Dazu gehört die kritische Haltung als Neigung, quantitative Botschaften ohne Aufforderung von außen zu hinterfragen, indem etwa die weiter oben beschriebene Liste der worry questions abgerufen wird. Außerdem fördern, Gal zufolge, positive Einschätzungen der eigenen statistischen Kompetenzen und die Überzeugung, dass Statistik sinnvoll und nützlich ist, die Bereitschaft, statistisch gebildetes Verhalten zu zeigen.

Gals (2002, Gal et al., 1995, zitiert nach Wild & Pfannkuch, 1999) Konzeptualisierung der statistical literacy bezieht sich, ebenso wie die von Wallman (1993) vorgeschlagene, auf den erfolgreichen Umgang mit statistischen Informationen im Alltag durch die Allgemeinbevölkerung. Seine Veröffentlichungen fanden in späteren Auseinandersetzungen mit statistischen Kompetenzen großen Anklang (z. B. Watson & Callingham, 2003, Wild & Pfannkuch, 1999, Pfannkuch & Wild, 2004) und haben zweifelsohne einen hohen pragmatischen Wert, allerdings wurde sein Modell (Gal, 2002) bisher noch keiner gründlichen empirischen Prüfung unterzogen. Damit ist noch offen, wie statistical literacy in Gals (2002) Sinne operationalisiert werden kann und welche Beziehungen zwischen statistical literacy und literacy skills, statistischem, mathematischen und kontextbezogenem Wissen, dem Nutzen kritischer Fragen sowie Überzeugungen, Einstellungen und Haltungen bestehen.

2.2.2. Vorschlag von Watson und Callingham (2003).

Watson und Callingham (2003) setzen sich mit dem Begriff statistical literacy im Rahmen des Schulunterrichts auseinander. Sie bieten keine verbindliche und abschließende Definition von statistical literacy an, vielmehr schlagen sie vor, statistical literacy als eine hierarchisch strukturierte Fähigkeit zu begreifen, die

durch „not just knowing curriculum-based formulas and definitions but integrating these with an understanding of the increasingly sophisticated and often subtle settings within which statistical questions arise“ (S. 20) gekennzeichnet ist. Hinsichtlich dieser Fähigkeit können zwei Aspekte des Verständnisses, strukturelle Komplexität und statistische Angemessenheit, beschrieben werden. Watson und Callingham greifen bei ihrer Auseinandersetzung mit dem Begriff statistical literacy die von Gal (2002) vorgeschlagene Definition auf und ziehen die auf der von Piaget formulierten Theorie der kognitiven Entwicklung gründende Taxonomie von Biggs und Collis (1982) sowie das von Watson (1997) vorgeschlagene Drei-Ebenen-Modell heran, um die hierarchische Struktur von statistical literacy empirisch zu untersuchen.

Bevor die empirischen Untersuchungen der hierarchischen Struktur von statistical literacy (Watson & Callingham, 2003, Callingham & Watson, 2017) dargestellt werden, sollen daher zunächst die *Structure of Observed Learning Outcomes Taxonomie* (SOLO, Biggs & Collis, 1982) und das Drei-Ebenen-Modell von Watson (1997) beschrieben und beurteilt werden.

SOLO-Taxonomie.

Biggs und Collis (1982) setzten sich die Aufgabe, die Erfassung der Stufen kognitiver Entwicklung von Piaget (Collis, 1975, zitiert nach Biggs & Collis, 1982), für die in der Regel Items genutzt werden, die keine spezifischen Schulfachinhalte (z. B. Bruchrechnung oder Gedichtanalyse) enthalten, auf Inhalte aus Schulfächern zu übertragen. Ihr Ziel bestand darin, eine Methode zu entwickeln, mittels derer Lehrende ihre Schülerinnen und Schüler anhand von Schulleistungsmessungen kognitiven Entwicklungsstufen zuordnen können. Eine solche Einstufung sollte den Lehrkräften einerseits das Setzen realistischer Erwartungen an die Leistungsfähigkeit von Lernenden und andererseits das Ableiten von Ansatzpunkten für eine effektive Anpassung der Lehre ermöglichen (d. h. für eine formative Evaluation). Hierfür entwickelten Biggs und Collis ein System zur Kodierung von offenen und geschlossenen Antworten auf typische Aufgaben zu Schulfachinhalten als prästrukturell, unidimensional, multidimensional, relational sowie erweitert-abstrakt (siehe Tabelle 1). Die Analyse der Antworten von mehreren Hundert Lernenden aus dem Primar-, Sekundar- und tertiärem Bereich in verschiedenen Fächern zeigte jedoch Inkonsistenzen der Antwortkategorien im Sinne der

Piagetschen Stufen sowohl über verschiedene Fächer als auch über verschiedene Aufgaben innerhalb derselben Fächer hinweg. So zeigten dieselben Personen zu einem Zeitpunkt unterschiedliche Stufen kognitiver Entwicklung sowohl in verschiedenen Fächern als auch zu verschiedenen Inhalten desselben Fachs. Dieses Ergebnismuster ließ sich nicht mit dem Konzept der kognitiven Stufen von Piaget vereinbaren. Biggs und Collis änderten aufgrund dessen ihre Zielsetzung von der Einordnung der Lernenden in eine Entwicklungsstufe auf die Einordnung der spezifischen Antwort hinsichtlich der Lernqualität in fünf SOLO-Ebenen (*levels*, Übersicht in Tabelle 1).

Tabelle 1

Ebenen der SOLO-Taxonomie innerhalb eines Modus (nach Biggs & Collis, 1991, S. 65, übersetzt durch Autorin)

Ebene		Beschreibung
1	Prästrukturell (*prestructural*)	Beschäftigung mit der Aufgabe, aber Lernende/r ist abgelenkt oder in die Irre geführt durch einen irrelevanten Aspekt, der zu einem vorherigen Modus bzw. einer niedrigeren Stufe gehört
2	Unistrukturell (*unistructural*)	Lernende/r fokussiert sich auf die relevante Domäne und greift einen Aspekt auf, mit dem sie/er arbeitet
3	Multistrukturell (*multistructural*)	Lernende/r greift mehr und mehr relevante oder korrekte Merkmale auf, integriert diese jedoch nicht
4	Relational (*relational*)	Lernende/r integriert die einzelnen Teile, so dass das Ganze eine kohärente Struktur und Bedeutung hat
5	Erweitert-abstrakt (*extended-abstract*)	Lernende/r generalisiert die Struktur und berücksichtigt somit neue und abstraktere Merkmale. Das repräsentiert eine neue, höhere Funktionsweise (*mode of operation*)

Die Ebenen der Lernqualität (SOLO-Ebenen) werden innerhalb eines jeweiligen *Modus der Repräsentation* (*mode of representation*, Biggs & Collis, 1991, Bezeichnung ursprünglich bei Bruner, 1964) zum einen als zyklisch und zum anderen als hierarchisch von prästrukturell bis erweitert-abstrakt angeordnet

angenommen. Der Modus der Repräsentation bezieht sich dabei auf die Art, wie gelernte Inhalte intern repräsentiert werden, ist typisch für eine bestimmte kognitive Entwicklungsstufe (im Piagetschen Sinne) und produziert jeweils eine spezifische Art von Wissen. Die fünf von Biggs und Collis vorgeschlagenen Modi sind der sensorimotorische Modus (produziert implizites Wissen) im Säuglingsalter, der ikonische Modus (produziert intuitives Wissen) in der frühen Kindheit, der konkret-symbolische Modus (produziert deklaratives Wissen) in der mittleren und späten Kindheit, sowie der formale Modus in der Adoleszenz und der postformale Modus im frühen Erwachsenenalter (produzieren theoretisches Wissen[6]). Ein sich später entwickelnder Modus ersetzt frühere Modi nach dem Verständnis der Autoren nicht, sondern ergänzt diese. Innerhalb eines jeden Modus durchlaufen Lernende nun die hierarchisch verstandenen Ebenen der Lernqualität von unistrukturell bis relational. Die prästrukturelle Ebene kennzeichnet hierbei eine Lernqualität, die dem jeweiligen Modus noch nicht entspricht, während die erweitert-abstrakte Ebene die Überwindung des jeweiligen Modus und damit den Eintritt in den nächsten Modus kennzeichnet.

Während unter Nutzung des Ansatzes von Piaget also Personen einer kognitiven Entwicklungsstufe zugeordnet werden, soll die SOLO-Taxonomie eine Einstufung von Antworten Lernender in eine Lernqualitätsebene innerhalb eines Repräsentationsmodus erlauben (Biggs & Collis, 1982). Um den Unterschied ihres Modells zu Neo-Piaget-Ansätzen herauszustellen, d. h. den Unterschied zwischen der Fokussierung auf Lernqualität und der Einordnung in Entwicklungsstufen zu verdeutlichen, schreiben die Autoren:

> The difficulty [der Inkonsistenzen in den Antworten], from a practical point of view, can be resolved simply by shifting the label from the *student* to his *response* to a particular task. There is nothing inconsistent about saying that on a particular math task the student gave a response one day that looked like the sort of response that would be expected from a formal operational student, and on another day gave a response that looked like a typical response from a middle concrete operational.“ (Biggs & Collis, 1982, S. 22)

[6] Biggs und Collis (1991) unterscheiden zwischen deklarativem und theoretischem Wissen.

und „[t]his distinction, between describing *responses* and describing *people*, is important […]“ (S. 23). Wie aus den Zitaten deutlich wird, sehen die Autoren es daher für ihre neue Zielsetzung nicht als problematisch an, dass Antworten auf intraindividueller Ebene Inkonsistenzen über (zeitlich eng beieinander liegende) Messgelegenheiten oder Items zu denselben Inhalten (das wird aus der Darstellung der Autoren nicht vollends deutlich) aufweisen. Dies stellt allerdings eine Problematik für die Reliabilität und damit für die Validität des genutzten Verfahrens dar. Wenn das Ziel der Klassifizierung von Antworten darin besteht, zu ermitteln, welche Lernqualitätsebene eine Lernende oder ein Lernender in einem spezifischen Lehrstoff erreicht hat, um daraus Änderungen von Erwartungen an die Leistungsfähigkeit der oder des Lernenden oder Änderungen des didaktischen Vorgehens für die spezifische Lernende bzw. den spezifischen Lernenden abzuleiten, dann ist vermutlich die Lernqualitätsebene einer spezifischen Aufgaben*klasse*, nicht einer spezifischen Aufgabe von Interesse (siehe hierzu auch Osburn, 1968). So wird bei einer Aufgabe, wie

> In einem Behälter befinden sich 100 Kugeln, 30 sind weiß, 70 sind schwarz. Wie groß ist die Wahrscheinlichkeit dafür, eine weiße Kugel per Zufall aus dem Behälter zu ziehen?

sowohl für formative Evaluationszwecke als auch die Erwartungshaltung vermutlich weniger von Interesse sein, welcher Ebene von Lernqualität die Antwort von Lernenden auf *diese spezifische Aufgabe* zugeordnet wird. Vielmehr interessiert vor dem Hintergrund der oben genannten Zielsetzungen vermutlich, ob *Aufgaben dieses Formats*, unabhängig von der konkreten Anzahl der Kugeln (z. B. 30 und 70 vs. 40 und 60) und der Farbe der Kugeln (z. B. weiß und schwarz vs. rot und blau) auf einer bestimmten Ebene der Lernqualität gelöst werden. Möglicherweise wird man hier die Klasse der Aufgaben noch weiter fassen wollen und etwa die Art der Objekte (Kugeln vs. Murmeln, Tennisbälle, Gummibärchen usw.), den Umfang der Population (100 vs. 50, 200 usw.), den Ergebnisraum (zwei Merkmalsausprägungen vs. drei oder mehr) sowie die Art des Merkmals (Farbe vs. Gewicht, Muster, Geschmack usw.) als von der Nutzung einer bestimmten Lernqualitätsebene unabhängig ansehen. Beispielsweise aus der Forschung zu deduktivem Schließen ist allerdings bekannt, dass die Änderung bestimmter Elemente von Aufgaben die Lösungswahrscheinlichkeit für eine Aufgabe mit demselben strukturellen Inhalt (z.

B. modus tollens) beeinflusst (z. B. Byrne, 1989). Es ist daher nicht auszuschließen, dass auch die strukturelle Qualität der Antwort je nach Aufgabenmerkmal unterschiedlich ausfällt (z. B. unistrukturell bei schwarz-weißen Kugeln vs. relational bei männlich-weiblichen Kursteilnehmern für das Aufgabenbeispiel weiter oben). In einem solchen Fall wäre es daher fraglich, inwiefern die betreffende Aufgabenantwort die Lernqualität und nicht etwa automatisierte Prozesse, Vertrautheit mit dem Kontext usw. abbildet.

Auch wenn es also so scheint, dass sich das Problem der inkonsistenten Antworten löst, wenn sich das Messobjekt (dem ersten Blick nach) von der Person auf die Antwort der jeweiligen Person ändert, so verschiebt sich die Anforderung der Konsistenz damit lediglich von einer größeren Klasse von Antworten (Antworten auf Aufgaben aus unterschiedlichen Fächern oder zu unterschiedlichen Inhalten desselben Fachs) auf eine enger definierte Antwortklasse (z. B. Antworten auf Aufgaben desselben Fachs zu demselben Inhalt). Die Messintention verschiebt sich hier zwar von dem Merkmal *kognitive Entwicklungsstufe* auf das Merkmal *Lernqualitätsebene*, jedoch ist das Ziel der Messung in beiden Fällen die Person – die „response to a particular task“ (Biggs & Collis, 1982, S. 22) ist nur insofern von Interesse, als dass sie einen Indikator für die erreichte Lernqualitätsebene der Person darstellt. Dieses Problem der Inhaltsvalidität einerseits und Reliabilität andererseits ist innerhalb der kriteriumsorientierten Messtradition bekannt und in zahlreichen Veröffentlichungen behandelt worden (detaillierte Behandlung bei Klauer, 1987).

Vor diesem Hintergrund kann daher nicht abschließend beurteilt werden, ob die SOLO-Taxonomie (Biggs & Collis, 1982, 1991) eine reliable Erfassung von Ebenen der Lernqualität erlaubt, auch wenn man sich hierbei auf spezifische, eng umrissene Inhalte eines bestimmten Fachs begrenzt.

Zwei weitere offene Fragen betreffen zum einen das zugrundeliegende Merkmal, das innerhalb der SOLO-Taxonomie (Biggs & Collis, 1982) erhoben wird, und zum anderen die Mechanismen der Veränderungen. Die Tatsache, dass die Autoren die Repräsentationsmodi als altersgebunden annehmen, könnte darauf hindeuten, dass reifungsgebundene oder/und entwicklungsaufgabenbezogene Prozesse als ursächlich für das Fortschreiten entlang des Lernzyklus und der Repräsentationsmodi angenommen werden, dies wird jedoch nicht vollkommen deutlich. Ähnlich unklar ist das Merkmal, das sich während eines Lernzyklus bzw.

im Rahmen verschiedener Repräsentationsmodi verändert. So ist es möglich, dass die Autoren hier domänenspezifisches Wissen (z. b. mathematisches Wissen oder historisches Wissen) oder aber domänenübergreifende Fertigkeiten oder Fähigkeiten (z. B. allgemeine Problemlösefähigkeiten oder Intelligenz) annehmen. Auch der Bezug der beiden Aspekte aufeinander ist nicht vollumfänglich klar, d. h. es ist nicht ganz klar, wie sich domänenspezifisches Wissen in verschiedenen Domänen in unterschiedlicher Geschwindigkeit entwickeln könnte, wenn Repräsentationsmodi als Altersstufen gebunden sind.

Drei-Ebenen-Modell von Watson (1997).

Watson (1997) schlug (zunächst unter dem Begriff statistical thinking) eine Hierarchie der Statistikkompetenz von Schülerinnen und Schülern auf drei Ebenen vor. Die Hierarchie gründet auf dem Neo-Piaget-Ansatz von Case (1985, zitiert nach Case, 1987) und auf der oben beschriebenen SOLO-Taxonomie (Biggs & Collis, 1982), die durch Watson, Collis, Callingham und Moritz (1995) auf statistische Inhalte angewandt wurde. Watsons Modell enthält drei Ebenen von mit statistischem Verständnis[7] assoziierten Fertigkeiten. Diese Ebenen sind (a) das grundlegende Verständnis stochastischer und statistischer Terminologie (z. B. Prozentangaben oder Median), (b) das Verständnis stochastischer und statistischer Sprache, wenn diese in einen sozialen Kontext eingebettet sind (z. B. die Darstellung von Prozentangaben in einem Kreisdiagramm in einem Zeitungsartikel), und (c) eine hinterfragende Grundhaltung, die die Grundlage dafür bildet, Behauptungen ohne angemessene Begründung zu widersprechen (z. B. Erkennen von Nichtübereinstimmungen von Prozentangaben und der Fläche im Kreisdiagramm). Die drei Ebenen werden als hierarchisch aufeinander aufbauende Zyklen der Lernqualitäten (unistrukturell, multistrukturell, relational) im Sinne von Biggs und Collis (1982) verstanden (Watson et al., 1995). Das Ziel der statistischen Schulbildung besteht nach Watson dann darin, statistical literacy der dritten Ebene zu erreichen.

Die empirische Grundlage für das Drei-Ebenen-Modell besteht in der Untersuchung von Watson et al. (1995), in der ein offenes statistisches Szenario (sogenannte Data Cards, siehe Watson et al.) von Schülerinnen und Schülern im

[7] Watson (1997) gebraucht dafür den Begriff *statistical thinking*, den sie später mit statistical literacy gleichsetzt.

offenen Antwortformat bearbeitet und anschließend dem Vorgehen von Biggs und Collis (1982) entsprechend analysiert wurde. Die Autoren kamen zu dem Schluss, dass zwei Zyklen, die sich jeweils aus der Abfolge von unidimensionalen (U), multidimensionalen (M) sowie relationalen (R) Antwortqualitäten zusammensetzen, die beobachteten Ergebnisse am besten widerspiegeln. Es sollte hierbei jedoch angemerkt werden, dass die Autoren keine statistischen Angaben zu ihren Schlussfolgerungen anbieten, so dass die Güte der Passung des Modells mit zwei Zyklen und der U→ M → R Abfolge offen bleibt. Auch ist nicht ganz klar, aus welchem Grund Watson (1997) drei und nicht – wie die Untersuchung von Watson et al. nahelegt – zwei Ebenen der Statistikkompetenz annimmt.

Empirische Untersuchungen der hierarchischen Struktur von statistical literacy.

Watson und Kollegen untersuchten die auf Basis der oben beschriebenen theoretischen Annahmen formulierte hierarchische Struktur der statistical literacy als eindimensionale Fähigkeit in mehreren Studien (z. B. Watson & Callingham, 2003, Watson, Kelly, Callingham & Shaughnessy, 2003, Callingham & Watson, 2017) mit Hilfe von Partial-Credit Rasch-Modellen. Zu diesem Zweck stellten Watson und Callingham insgesamt 80 Items mit statistischen Inhalten sowohl aus bereits vorliegenden empirischen Untersuchungen (z. B. Garfield, 2003, Tversky & Kahneman, 1983) als auch aus Neukonstruktionen zusammen. Die Items wurden in einem Common-Item nichtäquivalente Gruppen-Design[8] einer großen Zahl von Schülern (mehrere Hundert bis mehrere Tausend je nach Studie) der Klassen 3 bis 9 vorgelegt und die Antworten nach dem vorher entwickelten hierarchisch angeordneten Auswertungsschema (Biggs & Collis, 1982) mit insgesamt sechs geordneten Kategorien kodiert. Die Studien zeigen eine zufriedenstellende bis gute Modellpassung für eine eindimensionale Fähigkeit, die die Autoren als statistical literacy interpretieren (z. B. Watson & Callingham, 2003, Callingham & Watson, 2017). Anhand einer qualitativen Analyse der Verteilung von Personen und Antwortkategorien entlang der latenten Dimension (Person-Item-Karte) postulieren die Autoren sechs Ebenen („levels“, Watson & Callingham, S. 14) der zugrundeliegenden Dimension statistical literacy. Diese Ebenen geben eine 1)

[8] Mehreren Substichproben von Probanden, die aufgrund von nicht-zufälliger Aufteilung entstehen, werden verschiedene Subgruppen von Items vorgegeben, wobei einige Items – sogenannte Ankeritems – von allen Probanden bearbeitet werden (de Ayala, 2009)

idiosynkratische (die unterste Ebene), 2) informelle, 3) inkonsistente, 4) konsistente nicht-kritische, 5) kritische sowie 6) kritisch-mathematische Auseinandersetzung (die oberste Ebene) mit statistischen Inhalten wieder (Watson & Callingham). Während auf der idiosynkratischen Ebene keine oder eine nicht-aufgabenbezogene, sondern lediglich anekdotische oder durch persönliche Ansichten bestimmte Auseinandersetzung mit den Inhalten der Aufgabe erfolgt, ist eine kritisch-mathematische Auseinandersetzung durch eine hinterfragende Haltung und dem Einbezug des Kontextes der Aufgabe bei der Bewertung statistischer Informationen gekennzeichnet. Für eine Aufgabe zu den Ergebnissen einer Befragung zu Waffen an Schulen in den USA wäre ein Beispiel für die idiosynkratische Ebene der Auseinandersetzung eine fehlende Antwort oder die Antwort „Menschen sollten keine Waffen haben“ (Watson & Callingham, S. 36-37). Eine kritische Ebene würde sich für die gleiche Aufgabe in der Antwort „es wurde nur in Chicago befragt“ äußern (Watson & Callingham, S. 36-37).

Nicht ganz klar ist bei der Darstellung der Autoren (Watson & Callingham, 2003, Callingham & Watson, 2017) nach welchen Kriterien die Kodierkategorien für die Auswertung der Antworten von Probanden abgeleitet werden. Bei nachfolgend dargestellten Item werden etwa die (nicht zutreffenden) Antwortoptionen (a) und (b) der zweituntersten, die (ebenfalls nicht zutreffenden) Antwortoptionen (d), (e), (f) sowie das Ankreuzen mehrerer Antwortoptionen jedoch der nächsthöheren Kategorie zugeordnet:

> Um die durchschnittliche Anzahl von Kindern pro Familie in einer Stadt zu bestimmen, zählte eine Lehrerin die gesamte Anzahl der Kinder in der Stadt. Sie teilte dann durch 50, die gesamte Anzahl der Familien. Die durchschnittliche Anzahl der Kinder pro Familie war 2,2.
> Kreuze an, was auf jeden Fall stimmt.
> (a) Die Hälfte der Familien in der Stadt hat mehr als 2 Kinder.
> (b) Mehr Familien in der Stadt haben 3 Kinder als 2 Kinder.
> (c) Es gibt insgesamt 110 Kinder in der Stadt.
> (d) Es kommen 2,2 Kinder auf jeden Erwachsenen in der Stadt.
> (e) Die häufigste Anzahl der Kinder in einer Familie ist 2.
> (f) Nichts von dem oben Aufgeführten. (Watson & Callingham, 2003, S. 31, übersetzt durch die Autorin)

Die richtige Antwort (c) wird der vierten von sechs Kategorien zugeordnet. Hierbei ist unklar, weshalb die Verwechslung des Durchschnitts mit dem Median (a) und dem Modus (e) auf unterschiedlichen Ebenen einer Hierarchie zugewiesen werden und weshalb die korrekte Antwort eine nicht-kritische Auseinandersetzung anzeigt (Watson & Callingham, S. 31).
Bei einem weiteren Beispielitem

> Eine Stadt hat zwei Krankenhäuser. In dem West-Krankenhaus werden an einem Tag 45 Kinder geboren. In dem Ost-Krankenhaus werden 15 Kinder geboren. In welchem Krankenhaus ist es wahrscheinlicher, dass 60% Jungen geboren werden oder ist die Wahrscheinlichkeit gleich? Erkläre, warum. (Callingham & Watson, 2017, S. 200, übersetzt durch die Autorin)

wird die Antwort „Im kleineren Ost-Krankenhaus, weil kleinere Stichproben mehr Variabilität haben; größere Stichproben werden Ergebnisse produzieren, die näher an den erwarteten 50% liegen“ in die drittunterste von sechs Kodierkategorien (d. h. inkonsistente Ebene) eingestuft (Callingham & Watson, S. 200, übersetzt durch die Autorin). Eine solche Einstufung ist vor dem Hintergrund unplausibel, weil die sich hier offenbarende Heuristik *Glaube an das Gesetz der kleinen Zahl* zum einen sogar für Experten, wie mathematische Psychologen, sehr schwer zu vermeiden ist (Tversky & Kahneman, 1971, siehe auch Abschnitt 2.4.1) und zum anderen die angegebene Antwort inhaltlich richtig ist. Die Antwort „60% von 15 sind 9; 60% von 45 sind 27 – es erfordert lediglich 3 Jungen mehr, um mehr als 9 zu sein, aber 6 Jungen mehr, um mehr als 27 zu sein“ (Callingham & Watson, S. 200, übersetzt durch die Autorin) wird der dritthöchsten Kategorie (d. h. konsistente nicht kritische Ebene) zugeordnet, obgleich die weiter oben dargestellte Antwort als abstrakter eingeschätzt werden könnte.

Zwei weitere offene Fragen betreffen das grundsätzliche empirisch-methodische Vorgehen bei der Erforschung der statistical literacy als hierarchisch verstandenes eindimensionales Konstrukt (z. B. Watson & Callingham, 2003, Callingham & Watson, 2017). Zum einen machen die Autoren keine Angaben zur Beurteilerübereinstimmung bei der Zuordnung der Antworten von Probanden zu den Antwortkategorien. Damit ist die Auswertungsobjektivität und in der Folge auch die Reliabilität der Kategorisierung der Antworten nicht transparent zu bewerten. Zum anderen wären Belege zur diskriminanten Validität des angenommenen Konstrukts

statistical literacy wünschenswert. Es ist denkbar, dass die hier erfasste latente Variable auch im Sinne einer allgemeinen kognitiven Fähigkeit, einer numerisch-mathematischen oder einer *critical thinking*-Fähigkeit (im Sinne einer hinterfragenden, kritischen Haltung nicht nur in Bezug auf Statistik, sondern auf Informationen bzw. Inhalte allgemein) interpretiert werden kann.

Diese letzte Frage ist eng an die theoretische Grundlage für die Untersuchungen von Watson und Kollegen (z. B. Watson, 1997, Watson & Callingham, 2003, Callingham und Watson, 2017) geknüpft. Vor dem Hintergrund des Modells von Biggs und Collis (1982, 1991) und dessen Basis in Piagetschen bzw. Neo-Piagetschen Ansätzen (z. B. Case, 1985, zitiert nach Case, 1987) ist nicht vollends klar, inwiefern die Annahme einer spezifischen Kompetenz (statistical literacy) vereinbar ist mit der Annahme genereller altersstufenspezifischer kognitiver Mechanismen und Prozesse.

Hieran schließt sich ein weiterer Aspekt, der die theoretische Grundlage der empirischen Untersuchung der statistical literacy betrifft. So scheint es, wiederum vor dem Hintergrund der Annahme altersstufenspezifischer Phänomene, wie der Repräsentationsmodi (Biggs & Collis, 1982, 1991) sinnvoll, bei der empirischen Untersuchung von Statistikfertigkeiten das Alter der untersuchten Lernenden explizit zu berücksichtigen. Den theoretischen Annahmen zufolge sollten Personen in unterschiedlichen Altersstufen (d. h. mit unterschiedlichen Repräsentationsmodi) bestimmte Aufgaben mit unterschiedlichem Erfolg (d. h. in unterschiedlichen Qualitäten) lösen. Die Auswertung der Daten scheint jedoch nicht altersstufenspezifisch zu erfolgen und es ist nicht ganz klar, inwiefern bei der Auswahl oder Vorgabe der Items die angenommenen Repräsentationsmodi berücksichtigt wurden.

Darüber hinaus wird nicht vollends deutlich, welche empirischen Erwartungen sich aus der Kombination der beiden Modelle von Biggs und Collis (1982) mit fünf Lernqualitätsebenen und Watson (1997) mit drei Ebenen sowie der von Watson und Callingham (2003) intendierten Beschreibung von zwei Aspekten des statistischen Verständnisses (strukturelle Komplexität und statistische Angemessenheit) ergeben. Zwar weisen die Autoren (Watson & Callingham, 2003) explizit auf die explorative Natur ihres empirischen Modells hin. Dennoch ist nicht ganz klar, weshalb – auch explorativ – eine eindimensional und quantitativ

verstandene spezifisch statistische Fähigkeit mit sechs Ebenen mit den theoretischen Erwartungen vereinbar sein sollte (im Gegensatz zu etwa latenten qualitativen Klassen) oder welche Konsequenzen die explorativ ermittelte Struktur für die theoretische Grundlage der Untersuchung hat.

Obwohl insgesamt also noch einige Fragen hinsichtlich der empirischen Umsetzung, der theoretischen Anbindung und der Bedeutung der Ergebnisse im Ansatz von Watson (Watson, 1997, Watson & Callingham, 2003, Callingham & Watson, 2017) offen bleiben, stellt das Vorgehen der Autoren ein Beispiel für theoriegeleitete und durch Empirie begleitete Forschung zu Statistikkompetenz dar.

2.3. Statistical thinking

In der Veröffentlichung „Heading the call for change. Suggestions for curricular action." der MAA (1992) wurde der Begriff statistical thinking im Rahmen der Entwicklung von Lehr- und Leistungsmessungsempfehlungen für statistische Inhalte an Schulen und Hochschulen eingeführt (Cobb, 1992). Statistical thinking wurde hier als Ziel der statistischen Ausbildung an Schulen und Hochschulen von einer Fokusgruppe mit 15 Statistikpädagogen formuliert. Mit dem Blick auf die Inhalte von Statistikveranstaltungen im Studium leiteten die Pädagogen drei Empfehlungen für die Lehre von Statistik ab: Statistical thinking betonen, mehr Daten und Konzepte und dafür weniger Theorie und Rezepte nutzen sowie aktives Lernen fördern. Die Hervorhebung von statistical thinking in der Lehre sollte laut der Experten dadurch gelingen, dass vier basale Elemente explizit berücksichtigt werden: das Aufzeigen 1) der Notwendigkeit von Daten als Grundlage für Entscheidungen (need for data), 2) der Bedeutung der Datengewinnung und in der Folge der Qualität von Daten (importance of data production), 3) der Allgegenwärtigkeit von Variabilität und deren Konsequenz für die Interpretation der Daten (omnipresence of variability) und 4) der Quantifizierung und Erklärung bzw. Messung und Modellierung von Variabilität als Ziel (quantification and explanation, measuring and modelling of variability) (Cobb, 1992). Trotz der Aufnahme dieser Empfehlungen für die statistische Hochschulausbildung in den USA im GAISE-Bericht (ASA, 2005, 2016) stimulierte die Veröffentlichung von Cobb kaum empirische Forschung oder tiefere theoretische Analyse der formulierten Vorschläge.

Einige Jahre später leiteten Wild und Pfannkuch (1999) unter Rückgriff auf Cobb (1992) ein Rahmenkonzept des statistical thinking aus den Ergebnissen von qualitativen Interviews mit drei Gruppen von Personen ab. Zum einen legten sie elf Studierenden verschiedene statistische Aufgaben, die von typischen Lehrbuchübungen bis zu kritischen Analysen von Zeitungsartikeln reichten, in zwei einstündigen Sitzungen vor. Während der Bearbeitung der Aufgaben wurden die Probanden nach ihren Reaktionen sowie ihrem Vorgehen bei der Lösung der Aufgaben befragt. Eine zweite Gruppe der Befragten bestand aus Studierenden, die ein Projekt im Auftrag und Zusammenarbeit mit einem Unternehmen bearbeiteten. Diese Probanden wurden in einstündigen Interviews hinsichtlich des Vorgehens bei der Projektbearbeitung befragt. Schließlich interviewten die Autoren sechs professionelle Statistiker über statistisches Denken (statistical thinking) und dessen Rolle in Projekten. Auf Basis der Erkenntnisse aus den Interviews war es das Ziel der Autoren Empfehlungen dafür abzuleiten, wie Problemlösen, das man für Probleme braucht, deren Lösung das Anwenden von Statistik erfordert, verbessert werden kann. Diese Empfehlungen mündeten in einem Rahmenkonzept mit den vier Dimensionen 1) Investigativer Kreislauf, 2) Typen des Denkens, 3) Interrogativer Kreislauf und 4) Dispositionen (siehe Abbildung 1 und Abbildung 2). Obwohl sich die Empfehlungen hier in erster Linie an Produzenten statistischer Ergebnisse, z. B. empirisch forschende Personen, richten, können sie aber auch von Rezipienten statistischer Informationen genutzt werden. Im Folgenden sollen die vier Dimensionen in Kürze erläutert werden.

Die erste Dimension des Rahmenkonzepts beschreibt den Ablauf einer Untersuchung als Problem → Planung → Daten → Analyse → Schlussfolgerungen (Problem, Planning, Data, Analysis, Conclusions oder PPDAC, Wild & Pfannkuch, 1999, Abbildung 1). Hierbei ist es entscheidend, ein vorliegendes „reales“ Problem soweit zu abstrahieren, dass es durch statistisches Problemlösen abgebildet werden kann. Dies wird unter anderem darin deutlich, dass verschiedene dynamische Aspekte des Systems (z. B. verschiedene Variablen, Beziehungen zwischen Variablen, Umgebungen und Interaktionen zwischen diesen), die das Problem betreffen, bei der statistischen Repräsentation des Problems berücksichtigt werden müssen. Nach einer solchen Problemanalyse erfolgt die Planung der Untersuchung, die Durchführung der Untersuchung, die Analyse der Daten sowohl im Hinblick auf

das definierte Problem als auch im Hinblick auf spontan generierte Fragen und schließlich die Interpretation der Ergebnisse sowie deren Kommunikation und Bewertung (siehe Abbildung 1 für weitere Details). Das Ergebnis eines investigativen Zyklus kann dann sowohl eine Wissensbasis für die Lösung des vorliegenden Problems als auch einen Ausgangspunkt für einen weiteren Untersuchungszyklus darstellen.

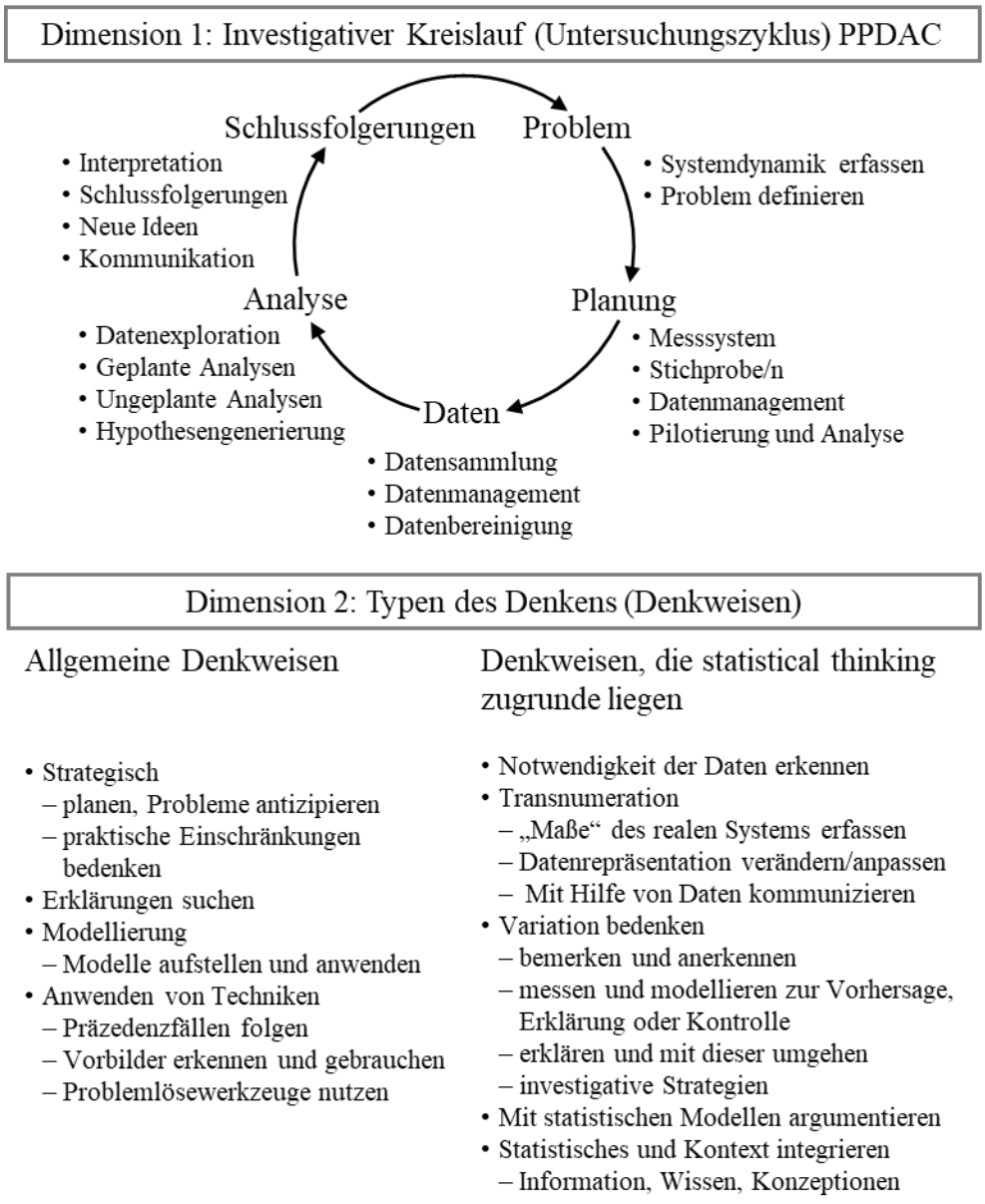


Abbildung 1. Die Dimensionen 1 und 2 des Rahmenkonzepts statistical thinking von Wild und Pfannkuch (1999, Übersetzung durch die Autorin, © International Statistical Institute, mit freundlicher Genehmigung von Maxine Pfannkuch), PPDAC: Problem, Planning, Data, Analysis, Conclusions.

Die zweite Dimension (Wild & Pfannkuch, 1999) beschreibt für das statistical thinking relevante Denkweisen. Zum einen sind dies allgemeine Aspekte des Denkens. Zentral an diesen allgemeinen, nicht spezifisch statistischen Denkweisen sind laut Autoren das Bewusstsein über verschiedene psychologische und physikalische Einschränkungen beim Denken und Urteilen sowie Konstruieren mentaler Modelle (etwa Wissenslücken, Verzerrungen der Wahrnehmung und des Denkens und zeitliche, finanzielle und materielle Ressourcenknappheit bei der Informationsbeschaffung) und die Berücksichtigung dieser Einschränkungen beim Denken. So können etwa Erklärungen und mentale Modelle durch kognitive oder Wahrnehmungsverzerrungen gefärbt sein und die Eignung von Präzedenzfällen als Vorbild für eine Untersuchung methodische Mängel aufweisen.

Zum anderen beschreiben die Autoren (Wild & Pfannkuch, 1999) Denkweisen, die spezifisch für statistical thinking sind. Diese umfassen das Anerkennen der Notwendigkeit von Daten, Transnumeration, Bewusstsein über Variation, Nutzen statistischer Modelle sowie die Integration statistischen und kontextbezogenen oder inhaltlichen Wissens. Das Anerkennen der Notwendigkeit von Daten bezieht sich vor allem auf den Nutzen von Daten gegenüber der eigenen Intuition, persönlichen Erfahrungen und anekdotischer Evidenz. Transnumeration meint die Transformation von Daten und Datenrepräsentationen (Kennzahlen oder Grafiken) zum Ziel des besseren Verständnisses des Untersuchungsinhalts. Ein typisches Beispiel hierfür ist das Nutzen unterschiedlicher Korrelationskoeffizienten und Kreuztabellen sowie eines Streudiagramms, um die Beziehung zwischen zwei Variablen besser zu verstehen. Transnumeration erfolgt dabei in der Regel dynamisch, d. h. eine Form der Datenrepräsentation (Streudiagramm) liefert den Anlass für eine weitere Form der Datenrepräsentation (Korrelationskoeffizient oder verschiedene Streudiagramme für Teilstichproben). Das Bewusstsein über Variation beinhaltet das Gewahrsein von Unsicherheit beim Treffen von Urteilen und Entscheidungen als Folge von Variabilität bei der Vorhersage, Erklärung und Kontrolle von Variablen. Zusätzlich fassen die Autoren auch Techniken zum Umgang mit Variabilität (z. B. in Form von bestimmten Versuchsplanungstechniken, wie Parallelisierung) hierzu. Die Nutzung statistischer Modelle als Argumentationsgrundlage für die Nutzung bestimmter Versuchsanordnungen sowie Datenerhebungsinstrumente und

Datenauswertungsstrategien stellt eine weitere spezifisch statistische Denkweise dar. Schließlich gehört die Berücksichtigung der hohen Bedeutung des Kontextwissens laut Ansicht der Autoren zum statistischen Denken. Insbesondere die ersten (Problemdefinition) und letzten Schritte (Interpretation und Ableiten von Schlussfolgerungen) im Untersuchungszyklus sind fast ausschließlich durch das Wissen über den Untersuchungsgegenstand bestimmt. Statistisches Wissen ist dagegen vor allem in der Mitte des Untersuchungszyklus von Bedeutung (Planung, Daten, Analyse). Wild und Pfannkuch weisen in diesem Zusammenhang darauf hin, dass insbesondere dieser letzte Aspekt nicht ausreichend in der statistischen Ausbildung berücksichtigt wird.

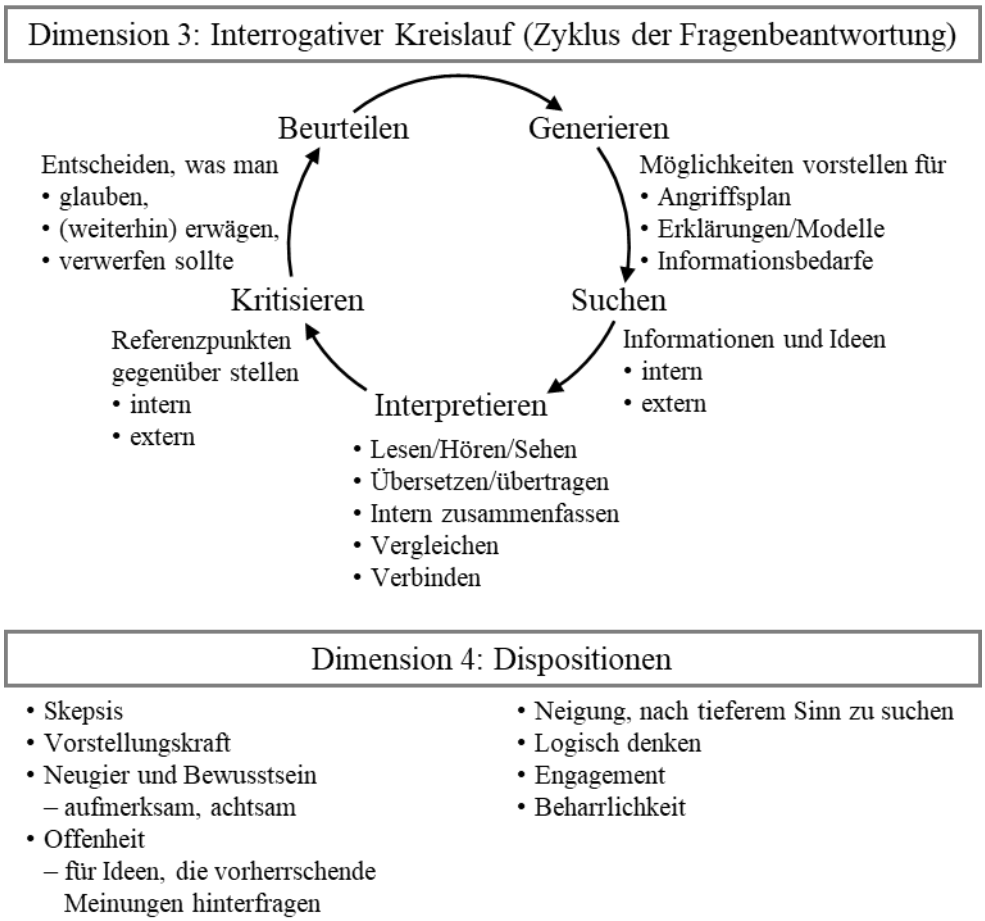


Abbildung 2. Die Dimensionen 3 und 4 des Rahmenkonzepts statistical thinking von Wild und Pfannkuch (1999, Übersetzung durch die Autorin, © International Statistical Institute, mit freundlicher Genehmigung von Maxine Pfannkuch)

Die dritte Dimension des Rahmenkonzepts bezieht sich auf den allgemeinen Denkprozess, der während des statistischen Problemlösens gebraucht wird (Abbildung 2). Der Zyklus wird von Wild und Pfannkuch (1999) dabei als idealisierte Form des tatsächlichen Denkablaufs verstanden. Das Generieren bezieht sich hier auf die Vorstellung von Möglichkeiten für beliebige das Problemlösen betreffende Bestandteile, wie mögliche Ursachen, Erklärungen, Mechanismen oder das Zusammenspiel von Phänomenen. Nach dem Generieren einer Idee hinsichtlich der Bausteine von mentalen oder statistischen Modellen beginnt die Suche nach dafür relevanten Informationen sowohl intern (d. h. Gedächtnis) als auch extern, z. B. durch Literaturrecherche oder kollegialen Austausch. Im nächsten Schritt werden die verschiedenen Informationen aufgenommen, auf das vorliegende Problem übertragen, intern zusammengefasst, miteinander verglichen und schließlich miteinander und mit bereits vorhandenen Informationen verbunden. Wild und Pfannkuch weisen hier darauf hin, dass die von ihnen beobachteten Studierenden diesen Prozess teilweise nicht abschließen, d. h. nicht alle vorhandenen Informationen miteinander verbinden, sondern direkt zu dem Schritt „Beurteilen“ übergehen. In der Phase des Kritisierens wird die Konsistenz bzw. Widerspruchsfreiheit der konsolidierten Informationen sowohl intern durch einen Vergleich zu bereits bekannten Informationen als auch extern durch einen Abgleich mit z. B. vorhandener Literatur oder anderen Personen geprüft. Referenzpunkte können hierbei die Ziele des Denkprozesses („Adressiert die Information das Ziel der statistischen Untersuchung?“), Glaubenssysteme („Werde ich durch eine ungültige vorgefasste Meinung geleitet?“) und auch emotionale Reaktionen („Werde ich von der Sorge geleitet, nicht mathematisch genug zu arbeiten?“) sein. Im letzten Schritt werden schließlich Urteile über die Nützlichkeit und Qualität der erworbenen Informationen gefällt.

Die vierte und letzte Dimension umfasst Dispositionen, die Wild und Pfannkuch (1999) auf Basis der Interviews mit Statistikfachleuten ableiteten. Diese Dispositionen umfassen die in Abbildung 2 aufgeführten Merkmale. Es geht hierbei in erster Linie darum, Anlässe für Erklärungen in einer Vielzahl von Phänomenen zu sehen und in kritischer (z. B. mit Hilfe der von Gal, 2002, formulierten worry questions), einfallsreicher, aufmerksamer und logisch konsistenter Weise Antworten zu suchen (bzw. suchen zu wollen).

Insgesamt sehen Wild und Pfannkuch (1999) statistical thinking „not [as] a separable entity [for successful problem solving,] [t]here is only holistic thinking that can and should be informed by statistical elements" (S. 244). Sie liefern eine intuitiv plausible und aus pragmatischen Gesichtspunkten sicherlich hilfreiche Beschreibung verschiedener Aspekte, die das statistische Lösen von Problemen verbessern könnten. Trotz des hohen pragmatischen Werts des Rahmenkonzepts (und der relativ hohen Anzahl an Zitationen) liegen bisher jedoch keine empirischen Daten über die Eignung des Konzepts zur Verbesserung des Problemlösens mit Hilfe statistischer Elemente vor.

2.4. Statistical reasoning

Statistical reasoning (Garfield & Gal, 1999, Garfield & Chance, 2000, Garfield, 2002, 2003) stellt eine weitere Konzeptualisierung statistischer Kompetenz dar und beschreibt die Art und Weise, wie Menschen mit statistischen Konzepten argumentieren und schlussfolgern und wie sie statistische Informationen verstehen. Hierunter ist etwa die Interpretation von Datensätzen, Datenrepräsentationen oder statistischen Zusammenfassungen sowie das Verbinden verschiedener statistischer Konzepte, wie zentraler Tendenz und Streuung oder von Daten und Zufall zu fassen. Es werden in diesem Rahmen verschiedene Typen des statistical reasoning vorgeschlagen, die das Ziel der statistischen Ausbildung darstellen sollen (z. B. Garfield & Gal, 1999). Zunächst sollen diese Typen im Einzelnen aufgeführt werden.

Das *Beurteilen von Daten* (reasoning about data) meint das Erkennen von Daten als quantitativ oder qualitativ, diskret oder stetig sowie das Wissen der Gründe für die daraus folgende Wahl einer grafischen Darstellung oder statistischen Kennzahl.

Das *Beurteilen von Datenrepräsentationen* (*reasoning about representations of data*) beinhaltet das Verstehen, welcher Aspekt einer Stichprobe in welcher Weise von einer grafischen Darstellung repräsentiert wird; Verstehen, wie eine grafische Darstellung angepasst werden kann, um einen Datensatz besser wiederzugeben sowie in der Lage sein, über zufällige Artefakte hinaus allgemeine Merkmale einer Verteilung, wie Verteilungsform, Verteilungsmitte und Streuung zu erkennen.

Das *Beurteilen statistischer Kennzahlen* (*reasoning about statistical measures*) beinhaltet das Verstehen, weshalb Maße zentraler Tendenz, Streuung und Position unterschiedliche Informationen über einen Datensatz vermitteln; das Wissen darüber, welche Maße am besten unter verschiedenen Bedingungen geeignet sind und warum die einzelnen Maße einen Datensatz repräsentieren oder nicht; Wissen, weshalb die Nutzung von statistischen Zusammenfassungen für Vorhersagen bei großen Stichproben genauer sein wird als bei kleinen Stichproben; Wissen, warum eine gute Zusammenfassung von Daten neben einem Maß zentraler Tendenz auch ein Streuungsmaß enthält und warum Zusammenfassungen von zentraler Tendenz und Streuung nützlich sind bei dem Vergleich von Daten.

Das *Schließen unter Unsicherheit* (*reasoning about uncertainty*) bezieht sich auf das korrekte Nutzen stochastischer Konzepte (randomness, chance, likelihood), um Einschätzungen über unsichere Ereignisse zu treffen; Wissen, warum nicht alle Ergebnisse gleich wahrscheinlich sind sowie das Wissen darüber, wann und warum die Wahrscheinlichkeiten verschiedener Ereignisse mit unterschiedlichen Methoden (Baumdiagramm, Simulationen mit Hilfe von Münzwürfen oder eines Computerprogramms) bestimmt werden können.

Das *Beurteilen (der Güte) von Stichproben* (*reasoning about samples*) meint das Wissen, welche Beziehung zwischen Stichproben und einer Population besteht und was aus einer Stichprobe geschlossen werden kann; Wissen, warum eine gut gewählte Stichprobe eine Population besser repräsentiert und warum es Stichprobenauswahlprozeduren gibt, die eine Population nicht angemessen repräsentieren; Wissen, dass Skepsis bei Schlüssen, die auf der Basis von kleinen oder verzerrten Stichproben gezogen wurden, angebracht ist.

Das *Beurteilen von Zusammenhängen* (*reasoning about association*) beschreibt das Wissen, wie eine Beziehung zwischen zwei Variablen beurteilt und interpretiert wird; Wissen, wie eine Vierfeldertafel oder ein Streudiagramm bei einer angenommenen bivariaten Beziehung inspiziert und interpretiert wird; Wissen, warum eine hohe Korrelation zwischen zwei Variablen nicht bedeutet, dass eine Variable die andere verursacht.

Diesen Typen des korrekten statistical reasoning werden, zum Teil symmetrisch, verbreitete Typen oder Arten von Fehlinterpretationen gegenübergestellt. Auch die von Garfield (Garfield & Gal, 1999, Garfield, 2002,

2003) vorgeschlagenen Fehlinterpretationen sollen im Folgenden einzeln aufgeführt werden.

Fehlverständnis von Mittelwerten (*misconception involving averages*): Das Verwechseln des arithmetischen Mittels mit dem Modalwert (Modus) oder dem Median; Glaube, dass ein Durchschnittswert immer aus allen Messwerten gebildet wird, ungeachtet dessen, ob Ausreißer oder Extremwerte vorliegen; Vergleich von Gruppen ausschließlich anhand der Mittelwertdifferenzen.

Orientierung am Ergebnis (*outcome orientation*, Konold, 1989): Das Heranziehen von Wahrscheinlichkeiten für dichotome ja-nein Entscheidungen für das Eintreten einzelner Ergebnisse, anstatt für Sequenzen von Ergebnissen (z. B. Bewerten einer Wahrscheinlichkeit von 70% für das Eintreten von Regen an einem Tag und nicht für das Eintreten von Regen an sieben von zehn Tagen).

Gute Stichproben müssen einen hohen Anteil der Population darstellen (*good samples have to represent a high percentage of the population*, Kahneman, Slovic & Tversky, 1982): Häufig besteht der Glaube, dass die absolute Größe einer Stichprobe (und die Auswahlprozedur der Stichprobe) für die Repräsentativität der Stichprobe eine geringere Rolle spielt, als der Anteil der Elemente der Stichprobe an der Population; eine große, zufällig gewählte Stichprobe wird als nicht repräsentativ betrachtet, wenn der Anteil der Elemente in dieser Stichprobe einen kleinen Anteil der Population darstellt.

Das *Gesetz der kleinen Zahlen* (*law of small numbers*, Kahneman et al., 1982, Tversky & Kahneman, 1971): Überschätzung der laut Stichprobentheorie zu erwartenden Ähnlichkeit von kleinen, zufällig ausgewählten Stichproben untereinander und der Ähnlichkeit kleiner, zufällig ausgewählter Stichproben mit der Population.

Repräsentativitätsheuristik (*representativeness misconception*, Kahneman et al., 1982): Die Wahrscheinlichkeit für eine Stichprobe wird anhand der Ähnlichkeit der Stichprobe mit der Population beurteilt (z. B. das Auftreten von Kopf, Zahl, Zahl, Kopf wird bei einem viermaligen Münzwurf als wahrscheinlicher im Vergleich zu dem Auftreten von Kopf, Zahl, Zahl, Zahl beurteilt; ein weiteres Beispiel für die Heuristik ist der Spielerfehlschluss, bei dem nach einer längeren Sequenz eines Ereignisses das Gegenereignis erwartet wird, etwa die Erwartung,

dass nach mehrmaligem Auftreten von Kopf die Wahrscheinlichkeit für das Auftreten von Zahl steigt).

Die *Gleichwahrscheinlichkeitsverzerrung* (*equiprobability bias*, Lecoutre, 1992): Die Wahrscheinlichkeit für das Auftreten verschiedener Ereignisse wird als gleich beurteilt trotz mathematisch unterschiedlicher Wahrscheinlichkeiten (z. B. wird die Wahrscheinlichkeit dafür, bei dreimaligem Würfeln eine Kombination von drei Fünfen und von drei verschiedenen Augenzahlen zu erhalten als gleich eingeschätzt).

Die Konzeptualisierung des statistical reasoning bezieht sich sowohl auf Produzenten als auch auf Rezipienten statistischer Informationen und kann grundsätzlich in allen Populationen genutzt werden, d. h. sowohl in der Allgemeinbevölkerung, bei Schülern und Schülerinnen (vor allem der Sekundarstufen) als auch bei Studierenden. Anders als die bisher aufgeführten Konzepte statistischer Kompetenz, setzt das statistical reasoning an konkreten statistischen Konzepten an und beansprucht es, eine Aussage über die Art und Weise, wie statistisch argumentiert wird, zu treffen.

2.4.1. Forschungsprogramm zu Heuristiken und Urteilsverzerrungen.

Wie aus den Beschreibungen der korrekten Typen des statistical reasoning und der typischen Fehlverständnisse statistischer Konzepte hervorgeht, nehmen Verzerrungen und Heuristiken hierbei eine zentrale Stellung ein. Aus diesem Grund soll die Forschung und theoretische Grundlage zu dem Ansatz der Heuristiken und (Urteils-) Verzerrungen an dieser Stelle in Kürze erläutert werden.

Die Anlässe für das Forschungsprogramm zum Urteilen unter Unsicherheit, dem einige Heuristiken und Urteilsverzerrungen zugeordnet werden, stellten unter anderem überraschende empirische Befunde zu Unterschieden zwischen klinischen und statistischen Urteilen (Meehl, 1964/1996/2003) und scheinbar irrationalen subjektiven Wahrscheinlichkeitsurteilen (Tversky & Kahneman, 1971) und Kausalattributionen dar (Kahneman et al., 1982). Eine der Studien in den frühen Anfängen der Forschung zu Heuristiken und Verzerrungen zeigte etwa, dass viele mathematische Psychologen (75 von 84 Personen) die Unsicherheiten bei der Schätzung von Parametern, die mit kleinen Stichproben verbunden sind, deutlich

unterschätzten, obwohl jede der befragten Personen die entsprechenden statistischen Kennzahlen (z. B. Standardfehler und Teststärke) ohne technische Hilfsmittel hätte ermitteln können (Tversky & Kahneman, 1971, Kahneman et al., 1982, Kahneman & Frederick, 2002; dieser Effekt der Anwendung des Gesetzes der großen Zahlen auf kleine Zahlen wurde Gesetz der kleinen Zahlen genannt). Zahlreiche weitere Untersuchungen zeigten weitere kognitive Verzerrungen für Aufgaben, die grundsätzlich unter Zuhilfenahme basaler mathematischer oder statistischer Konzepte leicht lösbar sind (Zusammenfassungen in z. B. Kahneman et al., 1982 oder Gilovich, Griffin & Kahneman, 2002), darunter auch Aufgaben zu den meisten der weiter oben aufgeführten Fehlinterpretationen statistischer Konzepte.

Zwei-System-Modell.

Zur Erklärung dieser (und anderer) insgesamt überraschender Befunde wurden mehrere Ansätze vorgeschlagen. Zum einen sind dies sogenannte „Mensch als kognitiver Geizhals"-Modelle (cognitive miser models), wie das Elaboration-Likelihood Modell (Petty & Cacioppo, 1986) und das Heuristics-Systematic Modell (Chaiken, 1980, Bohner, Moskowitz & Chaiken, 1995). Zum anderen wurden Modelle in Simons „‚approximate' rationality" Sinne (Simon, 1955, S. 114, auch bounded rationality genannt), wie etwa das Model der „fast and frugal" Heuristiken (Gigerenzer, Todd & ABC Group, 1999) und Zwei-System-Modelle (z. B. Sloman, 1996, 2002, Evans & Stanovich, 2013) vorgeschlagen. Die letzten beiden Modelle nehmen Heuristiken als Daumenregeln an, die grundsätzlich von adaptivem Nutzen sind, immer angewendet werden und lediglich in bestimmten Situationen zu Fehlschlüssen führen. So führt die Heuristik, die Ergebnisse zweier Stichproben direkt miteinander zu vergleichen, in allen Fällen, in denen die Stichprobengrößen ungefähr gleich sind, zu einem zufriedenstellenden Ergebnis. Für das bereits zuvor (Abschnitt „Empirische Untersuchungen der hierarchischen Struktur von statistical literacy.") angeführte Beispiel der zwei Krankenhäuser (S. 29) wäre also die Schätzung eines etwa gleichen Anteils an neugeborenen Jungen (bzw. die Schätzung von etwa gleichen Wahrscheinlichkeiten) in allen Fällen richtig, in denen die Krankenhäuser eine ähnliche Anzahl von Geburten verzeichnen. Lediglich wenn Stichproben deutlich unterschiedliche Größen aufweisen, führt die Heuristik zu Fehlurteilen und äußert sich in der Anwendung des Gesetzes der kleinen Zahlen. Kognitive Verzerrungen werden damit als systematische Ausnahmen bei der

Anwendung von grundsätzlich nützlichen und ständig zum Einsatz kommenden Heuristiken als „natürliche Erfassung“ einer Situation (*natural assessment*, z. B. Gilovich, 2002) betrachtet. Die anderen beiden genannten Ansätze (Elaboration-Likelihood-Modell und Heuristic-Systematic-Modell) gehen dagegen davon aus, dass Heuristiken in bestimmten Situationen angewendet werden, in denen Personen nicht hinreichend motiviert sind, eine gründliche Informationsverarbeitung zu betreiben. Lediglich in Situationen, die für Personen von hoher Bedeutung sind, werden alternative (systematische und regelgeleitete) Informationsverarbeitungsprozesse aktiviert. Da Kahneman selbst ein Zwei-System-Modell als theoretische Erklärungsgrundlage für Heuristiken und Verzerrungen heranzieht (Kahneman & Frederick, 2002, Kahneman & Frederick, 2005, Kahneman, 2011), soll im Folgenden lediglich dieses Modell und nur im Rahmen des Heuristiken-und-Verzerrungen-Ansatzes für ausgewählte Heuristiken detaillierter skizziert werden.

Kahneman und Frederick (2002, 2005) nehmen zwei Systeme von Prozessen an, die an der Beurteilung von Situationen beteiligt sind. Das System 1 produziert schnelle, automatische und intuitive Einschätzungen während das System 2 durch kontrollierte, bewusste und regelgeleitete Prozesse die Qualität der System-1-Antworten überwacht und diese bestätigt, korrigiert oder aufhebt. Die Prozesse der beiden Systeme nehmen Kahneman und Frederick (2005) als automatisiert (System 1) vs. kontrolliert (System 2) und parallel an der Aufgabenbearbeitung beteiligt an. Dies zeigt sich etwa darin, dass die gleichzeitige Bearbeitung einer Aufgabe, die Aufmerksamkeit und Konzentration, also System-2-Prozesse erfordert (z. B. das Erinnern einer zufälligen Abfolge von Zahlen), mehr intuitive, z. B. stereotype Antworten auf Fragen, also Ergebnisse der System-1-Prozesse, produziert (Gilbert, Pelham & Krull, 1988, Gilbert & Hixon, 1991). Heuristische oder intuitive Antworten zeichnen sich den Autoren zufolge dann dadurch aus, dass die im System 1 generierten Antworten nicht oder nur in sehr geringem Maße durch das System 2 modifiziert werden (Kahneman & Frederick, 2002, 2005).

Ein Bündel von System-1-Prozessen äußert sich laut Kahneman und Frederick (2002, 2005) in der Attributsubstitution (*attribute substition*). Attributsubstitution beschreibt die Ersetzung von schwerer zugänglichen Zielattributen eines einzuschätzenden Objekts mit verwandten heuristischen, leichter

zugänglichen Attributen des Objekts bei der Einschätzung. Ein klassisches Beispiel hierfür stellen Ergebnisse für das sogenannte Linda-Problem dar (Tversky & Kahneman, 1982, 1983). Bei dem Linda-Problem wird Personen eine Beschreibung von Linda vorgelegt:

> Linda ist 31 Jahre alt, single, direkt und sehr aufgeweckt. Sie hat einen Abschluss in Philosophie[9]. Als Studentin war sie sehr bewegt von Problemen der Diskriminierung und sozialen Gerechtigkeit und nahm auch an Anti-Atomkraft-Demonstrationen teil (Tversky & Kahneman, 1982, S. 92, übersetzt durch die Autorin).

Anschließend sollen die untersuchten Personen Aussagen über Lindas Zugehörigkeit zu verschiedenen Gruppen zum einen hinsichtlich der Wahrscheinlichkeit und zum anderen hinsichtlich der Ähnlichkeit zu einer typischen Vertreterin dieser Gruppen in eine Rangfolge bringen:

> Linda ist eine Grundschullehrerin.
> Linda arbeitet in einem Buchladen und macht Yoga.
> Linda ist aktiv in der Frauenbewegung.
> Linda ist eine Sozialarbeiterin im psychiatrischen Sozialdienst.
> Linda ist ein Mitglied in der League of Women Voters.
> Linda ist eine Bankangestellte.
> Linda ist eine Versicherungskauffrau.
> Linda ist eine Bankangestellte und aktiv in der Frauenbewegung. (Tversky & Kahneman, 1982, S. 92, übersetzt durch die Autorin)

Die Analyse der Einschätzungen der Wahrscheinlichkeit und der Repräsentativität für Lindas Zugehörigkeit zu den oben beschriebenen Gruppen ergab sehr hohe Korrelationen ($r = .99$ und $r = .97$ bei einer Parallelversion des Linda-Problems, dem sogenannten Lawyer-Engineer-Problem, Kahneman & Frederick, 2002). Es kann daher davon ausgegangen werden, dass die Einschätzungen der Wahrscheinlichkeit zum einen und des Ausmaßes der

[9] Die hier dargestellte originalgetreue Beschreibung von Linda und die Antwortoptionen sind für deutsche Verhältnisse nicht sinnvoll, da es in Deutschland (noch) nicht die Regel ist, einen ausbildungsfremden Beruf auszuüben (z. B. als Versicherungskauffrau zu arbeiten nach einem akademischen Abschluss in Philosophie). Tversky und Kahneman entwickelten das Linda-Problem als US-amerikanische Adaption eines ähnlichen Problems für israelische Verhältnisse.

Repräsentativität zum anderen aufgrund von ähnlichen, wenn nicht denselben Attributen erfolgten. Rein formal unterscheiden sich die Attribute der Wahrscheinlichkeit und der Repräsentativität im Linda-Problem jedoch deutlich. Während die Beschreibung von Linda *repräsentativer* oder typischer für *eine in der Frauenbewegung aktive Bankangestellte* ist, ist die Wahrscheinlichkeit für eine beliebige Person (d. h. vollkommen unabhängig von der Beschreibung) größer (oder höchstens gleich), einer Klasse mit einem Merkmal als einer Klasse mit diesem Merkmal und einem weiteren Merkmal anzugehören. Nur wenn alle (weiblichen) Bankangestellten in der Frauenbewegung aktiv sind, ist die Wahrscheinlichkeit für Linda, Bankangestellte vs. Bankangestellte und in der Frauenbewegung aktiv zu sein, gleich. Nach den üblichen Regeln der Wahrscheinlichkeitsrechnung ist es demgegenüber nicht möglich, dass Linda wahrscheinlicher eine in der Frauenbewegung aktive Bankangestellte ist als eine Bankangestellte (die in der Frauenbewegung aktiv ist oder nicht). Die richtige Wahrscheinlichkeitseinschätzung wäre in diesem Fall daher, dass Linda *wahrscheinlicher eine Bankangestellte* als eine in der Frauenbewegung aktive Bankangestellte ist. Bei der Bearbeitung dieser Aufgabe scheinen Personen also auf das einfacher verfügbare Attribut der Ähnlichkeit oder Repräsentativität zurückzugreifen, um das Zielattribut Wahrscheinlichkeit einzuschätzen und produzieren in der Folge verzerrte Antworten.

In den frühen Untersuchungen zum Linda-Problem und parallelen Versionen des Problems hat sich diese angenommene Attributsubstitution als äußerst robust gezeigt und wurde von dem Großteil der untersuchten Personen mit der Folge einer falschen Antwort verwendet, auch bei mehreren Vereinfachungen des Problems und Lösungshinweisen (z. B. Elimination der Distraktoren). Bis zu 90% der Personen beurteilten es als wahrscheinlicher, dass Linda eine in der Frauenbewegung aktive Bankangestellte ist oder unterlagen in anderen Szenarien der sogenannten conjunction fallacy (Tversky & Kahneman, 1983). Erst bei Doktoranden, die mehrere Statistikveranstaltungen absolviert hatten und denen nur die beiden kritischen Antwortoptionen („Linda ist eine Bankangestellte“ und „Linda ist eine Bankangestellte und aktiv in der Frauenbewegung“) vorgegeben wurden, reduzierte sich der Anteil falscher Antworten auf das Problem auf weniger als die Hälfte (36%, Tversky & Kahneman). Neben der Erfahrung mit statistischen Inhalten (allerdings

nur in Kombination mit einem Lösungshinweis, wie der Elimination der Distraktoren), korrelieren in Studien mit unterschiedlichen Heuristiken und Verzerrungen auch einige weitere Merkmale mit korrekten Einschätzungen. Ein solches Merkmal ist die Intelligenz (Toplak, West & Stanovich, 2011, Frederick, 2005). In bestimmten Aufgabenstellungen korrelieren darüber hinaus die Angabe von relativen Häufigkeiten statt Wahrscheinlichkeiten (z. B. „1 von 100“ statt „1%“, Gigerenzer & Hoffrage, 1995) sowie andere Manipulationen des Aufgabenformats oder der Aufmerksamkeit (z. B. Krosnick, Li & Lehman, 1990, Politzer & Macchi, 2005, Schwarz, Strack, Hilton & Naderer, 1991) mit der Tendenz zu Urteilsverzerrungen. Auch Lehr- und Trainingsinterventionen zu Konjunktionen (z. B. Anleitung zur Nutzung von Venndiagrammen, Agnoli, 1991, Agnoli & Krantz, 1989) reduzieren die Tendenz, verzerrte Urteile zu treffen. Diese Ergebnisse interpretieren Kahneman und Frederick (2005) als Hinweise darauf, dass System-2-Prozesse, wie der Vergleich der durch System 1 generierten Antworten mit Antworten anhand der Anwendung gelernter Regeln, durch Situations-, Aufgaben- und Personmerkmale aktiviert werden können und intuitive Einschätzungen als Ergebnis von System-1-Prozessen korrigieren.

Insgesamt regte das Paradigma der Heuristiken und Urteilsverzerrungen sehr viel Forschung an, die im Rahmen der vorliegenden Arbeit nur ausschnitthaft berichtet werden kann. Es soll daher lediglich noch auf die Forschungsbemühungen rund um Lecoutre verwiesen werden. Die Arbeitsgruppe (Lecoutre, Cordier, Durand, siehe Lecoutre, 1992) identifizierte in mehreren Untersuchungen die Gleichwahrscheinlichkeitsverzerrung (equiprobability bias) bei Schülerinnen und Schülern der Sekundarstufen. Die Gleichwahrscheinlichkeitsverzerrung äußert sich in der Beurteilung der Wahrscheinlichkeit für kombinatorische Ereignisse als – inkorrekterweise – gleich. Ein Beispiel hierfür ist die folgende Aufgabe (Antwortoption 1 ist korrekt, da der Wurf zweier Würfel zwei Ergebnisse mit je einer 5 und einer 6 ergeben kann, jedoch nur ein Ergebnis mit je einer 5):

> Wenn zwei Würfel gleichzeitig geworfen werden, ist es möglich, dass eins der folgenden zwei Ergebnisse eintritt: Ergebnis 1: Eine 5 und eine 6 werden gewürfelt. Ergebnis 2: Eine 5 wird zweimal gewürfelt. Welche Antwort trifft zu?
>
> (a) die Ergebnisse sind gleichwahrscheinlich.

(b) das Ergebnis 1 ist wahrscheinlicher.

(c) das Ergebnis 2 ist wahrscheinlicher.

Die Autoren untersuchten verschiedene mögliche Einflussfaktoren (z. B. Erfahrung mit Stochastik oder mit Glücksspielen) auf die Gleichwahrscheinlichkeitsverzerrung und zeigten, dass die Verzerrung äußerst robust ist. Etwa 60% der untersuchten Personen in verschiedenen Studien schätzen beide Ergebnisse als gleichwahrscheinlich ein (Lecoutre, Durand & Cordier, 1990). Sie fanden außerdem, dass eine Maskierung des Zufalls die Gleichwahrscheinlichkeitsverzerrung auf bis zu 20% reduzieren kann. Lecoutre und Kollegen (1992, Lecoutre et al., 1990) leiten aus den Ergebnissen die Vermutung ab, dass bestehende korrekte kombinatorische (gegenüber stochastischen) Repräsentationen bei Schülerinnen und Schülern verfügbar sind, jedoch durch Prompts aktiviert werden müssen. Auch wenn Lecoutre selbst keinen Bezug zu der Zwei-System-Theorie darstellt, können die Ergebnisse im Sinne von Aktivierungsbedingungen des System 2 interpretiert werden.

Ausgewählte Kritikpunkte am Forschungsprogramm zu Heuristiken und Urteilsverzerrungen.

Einige Untersuchungen (Krosnick et al., 1990, Politzer & Macchi, 2005, Schwarz et al., 1991) wurden im Rahmen einer kritischen Auseinandersetzung mit dem von Kahneman und Tversky in ihren Untersuchungen verwendeten Aufgabenformat durchgeführt (vor allem Tversky & Kahneman, 1983, Kahneman & Tversky, 1982). Diese und andere Studien kritisierten die Verletzungen der Kommunikationsmaximen von Grice (1991) im Stimulusmaterial der Heuristikforschung. Nach diesen Kommunikationsmaximen verlassen sich Rezipienten von Mitteilungen darauf, dass lediglich für das Kommunikationsziel relevante, wahre und informative Botschaften übermittelt werden. Diese Maximen werden in Aufgabenstellungen, wie dem Linda-Problem verletzt – die Beschreibung von Linda ist für das Kommunikationsziel weder relevant noch informativ und lenkt vom Kommunikationsziel sogar ab. Die weiter oben angeführten Studien zeigten entsprechend, dass durch gezielte Anpassungen des Materials eine Reduktion der Fehler auf bis zu 15% (Politzer & Macchi, 2005) gelingen kann.

Während die vor dem Hintergrund der Kommunikationsprinzipien vorgebrachte Kritik an den Aufgaben durchaus angemessen ist, wenn Urteilsfehler in Folge der Anwendung von Heuristiken und der Umgang mit

Wahrscheinlichkeiten allgemein im Rahmen alltäglicher und direkter zwischenmenschlicher Kommunikation untersucht werden, scheint die Kritik am Untersuchungsmaterial zur Erforschung von Heuristiken in anderen Lebensbereichen von Menschen jedoch grundsätzlich nicht gerechtfertigt zu sein. So zielen Werbebotschaften (wirtschaftlicher wie politischer oder sozialer Art) darauf ab, Personen zu ganz bestimmten Urteilen auf der Grundlage von mehrdeutigen und den Kommunikationsmaximen ebenfalls oft nicht folgenden Nachrichten zu bewegen. Es scheint daher nicht nur interessant, sondern gesamtgesellschaftlich von höchster Relevanz zu sein, Bedingungen für heuristische Fehlschlüsse gerade bei offenkundig irreführenden oder täuschenden Botschaften zu erforschen. Für die vorliegende Arbeit von höherer Bedeutung ist jedoch der Umstand, dass ein wichtiger Anlass für die Erforschung der Heuristiken bei unsicherem Schließen im Antwortverhalten von mathematisch hochgebildeten Fachexpertinnen und Fachexperten bestand (Tversky & Kahneman, 1971). Die Aufgaben, die in der Studie verwendet wurden, sind ökologisch höchst valide (z. B. „Stellen Sie sich vor, Sie haben in einem Experiment ein signifikantes Ergebnis auf der Grundlage von 20 Versuchspersonen erhalten. Sie wollen das Experiment nun mit 10 Versuchspersonen wiederholen. Wie groß ist die Wahrscheinlichkeit, dass das Ergebnis erneut signifikant wird?“) und stellen Fragen dar, die sich Forschende laufend stellen (sollten). Darüber hinaus zeigen Studien zu der Interpretation von p-Werten bzw. Signifikanzaussagen (Oakes, 1986, Haller & Kraus, 2002) und Konfidenzintervallen (Hoekstra, Morey, Rouder & Wagenmakers, 2014) sehr hohe Anteile an – aus stochastischer Sicht – falschen Antworten, auch bei erfahrenen Forschern. Auch hierbei handelt es sich, wenn auch in stärker eingeschränktem Maß als bei Tversky und Kahneman (1971), um ökologisch valide Items. Die Kritik an den Ergebnissen des Heuristiken-und-Verzerrungen-Forschungsprogramms aufgrund der Konversationsmaximen verletzenden Untersuchungsstimuli scheint damit für die vorliegende Fragestellung der statistischen Kompetenz bei Studierenden nicht stichhaltig.

Ein weiterer Einwand zum Heuristiken-und-Urteilsverzerrungen-Programm stammt von Gigerenzer (1991, 1996). Er argumentiert ebenfalls, dass Urteilsverzerrungen durch Modifikationen der Aufgabenstellung (z. B. die Angabe von relativen Häufigkeiten statt Prozentangaben oder eine deutlichere

Kennzeichnung von Zufallsziehungen) deutlich reduziert oder eliminiert werden können und daher nicht robust sind. Gigerenzer konzentriert sich dabei jedoch (unter anderem) stärker auf die statistisch-normativen Gesichtspunkte bei der Analyse der in dem Programm genutzten Aufgaben. So unterscheidet er verschiedene mögliche Lösungen für die Aufgaben z. B. nach frequentistischem und subjektivem (Bayesschem) Verständnis[10]. Seiner Auffassung nach ist die Definition der normativ richtigen Antworten auf die Aufgaben von Kahneman und Tversky (Kahneman et al., 1982; Tversky & Kahneman, 1974, 1983) zu eng gefasst und nicht in der Lage, die menschliche Kognition angemessen widerzuspiegeln. Auch dieser Einwand ist sicherlich mindestens teilweise berechtigt, zumindest wenn die Untersuchung des natürlichen probabilistischen Urteilens bei Menschen im Fokus steht. Jedoch soll an dieser Stelle noch einmal auf die Studien zum Verständnis (bzw. Missverständnis) statistischer Aussagen verwiesen werden (Oakes, 1986, Haller & Kraus, 2002, Hoekstra et al., 2014). Da eine richtige (hier: frequentistische) Interpretation dieser Aussagen lediglich die Unterscheidung zwischen bedingten und unbedingten Wahrscheinlichkeiten erfordert (die in aller Regel frequentistisch behandelt werden[11] und deren Kenntnis zumindest bei Statistik*lehrenden* vorausgesetzt werden sollte), ist es zulässig, die Urteilsverzerrungen und -fehler als Resultate der Anwendung von Heuristiken zu vermuten. Darüber hinaus stimmen die meisten Aussagen, denen Personen im Rahmen der Studien zustimmen (z. B. Haller & Kraus, 2002) mit keiner der verschiedenen Wahrscheinlichkeitsschulen (z. B. Frequentismus oder Bayesianismus, Romeijn, 2017) bzw. -interpretationen (logische, subjektive, propensity oder best-system, Hájek, 2012) in konsistenter Weise überein. Es scheint daher wenig plausibel, anzunehmen, dass die intuitiven Antworten von Forschenden und Studierenden eine rationale Norm darstellen. Die Kritik des unpassenden Aufgabenformats muss, unabhängig von den o. g. Punkten, für die Zielgruppe der Studierenden empirischer Wissenschaften und für Forschende in empirischen Fächern jedoch vermutlich rein prinzipiell zurückgewiesen werden. Spätestens als Forschende sollten Personen in der Lage sein, mit

[10] Aus der Sicht bestimmter frequentistischen Vertreter ist es etwa nicht zulässig, Wahrscheinlichkeiten für ein Einzelereignis zu schätzen (von Mises, 1928/1957, zitiert nach Gigerenzer, 1991, S. 5), wie dies beim Linda-Problem gefordert ist.

[11] Wie auch das Bayes-Theorem (um das zu prüfen, reicht ein Blick in die üblichen Statistiklehrbücher, z. B. Eid, Gollwitzer und Schmitt, 2017 oder Bortz und Schuster, 2010, Bortz, 2005, Hays, 1994)

Wahrscheinlichkeitsangaben, etwa in Form von Prozentangaben, umzugehen und Stichprobenarten (z. B. Zufallsstichproben aus einer Formulierung, wie „zufällig ausgewählt“) und deren statistische Konsequenzen ohne Schwierigkeiten in Studienbeschreibungen erkennen können (siehe Gigerenzer, 1991, z. B. S. 10 und 11-12)[12]. Auch die Kritik an der Wahl der normativen Lösungsschlüssel für die im Heuristiken-und-Verzerrungen-Forschungsprogramm verwendeten Aufgaben wird daher in dem vorliegenden Zusammenhang nicht als stichhaltig eingeschätzt.

Auch wenn während der fast 40-jährigen Forschung zu Heuristiken und Verzerrungen wesentlich mehr Kritik geübt worden ist, als in dem hier vorliegenden Rahmen behandelt werden kann, soll lediglich ein weiterer Kritikpunkt kurz angerissen werden, der weniger an den Ansatz selbst als an seine theoretische Grundlage der Unterscheidung zwischen System 1 und System 2 gerichtet ist. Das Zwei-Systeme Modell wurde zunächst als Verallgemeinerung zahlreicher spezialisierter Zwei-Prozess- und Zwei-System-Modelle (z. B. Schneider & Schiffrin, 1977, Fodor, 1983, Evans, 1989, alle zitiert nach Evans, 2008, Chaiken, 1980) vorgeschlagen (Evans, 2008, Sloman, 1996, Stanovich & West, 2000). Da die Unterschiede zwischen den spezifischen Modellen jedoch teilweise recht erheblich sind und die einem System als gemeinsam angenommenen Prozesse weit weniger theoretische und empirische Konsistenz aufweisen, als bei zwei distinkten Systemen zu erwarten wäre, ist eine allgemeine Zwei-Systeme-Theorie derzeit nicht haltbar (Evans, 2008). Statt von zwei getrennten Systemen auszugehen, schlägt Evans (2008) daher vor, zwei unterschiedliche Prozesstypen (*type 1* und *type 2*) anzunehmen. Auch Kahneman und Frederick (2002, 2005) unterstreichen, dass sie unter den Begrifflichkeiten „System 1“ und „System 2“, die sie von Stanovich und West (2002) übernehmen, eine Gruppe miteinander verwandter Prozesse meinen. Sie schreiben:

> [t]he placement of dividing lines between systems is arbitrary because the bases by which we characterize mental operations (difficulty of acquisition, accessibility to introspection and disruption) are all continua. However […]

[12] Der Autorin ist bewusst, dass die Anforderung, Prozentangaben (als Angabe von Wahrscheinlichkeiten und Häufigkeiten) und Stichprobenarten korrekt zu erkennen, eine normative Forderung darstellt und ihrerseits einer Begründung bedarf. An dieser Stelle wird – spekulativ und in Übereinstimmung mit typischen Statistikmodulbeschreibungen – angenommen, dass die genannten Kompetenzen sinnvolle Anforderungen darstellen.

> there is broad agreement that mental operations range from rapid, automatic, perception-like impressions to deliberate computations that apply explicit rules or external aids. (Kahneman & Frederick, 2005, S. 288).

Es scheint so, als würden Kahneman und Frederick (2005) die deutliche und willkürliche Reduktion in der Beschreibung der den Heuristiken und Verzerrungen zugrundeliegenden Prozesse in Kauf nehmen. Dieser Umstand – die Dichotomisierung von Kontinua und das Zusammenfassen distinkter kognitiver Prozesse zu einem Bündel, trotz Kontraindikation– wurde stark als Rückwärtsbewegung in der Entwicklung von psychologischen Theorien kritisiert (z. B. Gigerenzer, 2010, Gigerenzer & Regier, 1996). So führt eine solche Vereinfachung unter anderem zu unspezifischeren, teilweise tautologischen (nicht falsifizierbaren) empirischen Vorhersagen und behindert damit den Erkenntnisfortschritt im Sinne eines kumulativen Prozesses. Der Kritikpunkt stellt einen grundlegenden methodologischen Einwand dar und lässt sich der Frage nach der Beachtung und Wertschätzung theoretischer Grundlagen in der empirischen, insbesondere human- bzw. sozialwissenschaftlichen Forschung grundsätzlich zuordnen. Die Konsequenzen der (Nicht-) Nutzung von theoretischen Grundlagen für die empirische Forschung im engen Sinne (für eine einzelne Studie) und im weiteren Sinne (für die Entwicklung eines Forschungsprogramms) werden daher im Abschnitt 7.2 gesondert diskutiert. Für den vorliegenden Rahmen soll zunächst jedoch festgehalten werden, dass der wissenschaftliche Nutzen einer Integration der Forschungsergebnisse zu Heuristiken und Verzerrungen in Zwei-Prozess-Theorien derzeit noch nicht geklärt ist.

2.4.2. Relevanz des Forschungsprogramms für statistical reasoning.

Obwohl die theoretische Grundlage und die theoretische Bedeutung des Forschungsprogramms zu Heuristiken und Urteilsverzerrungen (Kahneman & Frederick, 2002, 2005, Kahneman et al., 1982; Tversky & Kahneman, 1974, 1983) derzeit als strittig gelten kann, liefert das Forschungsprogramm Ansatzpunkte für die Forschungsbemühungen zu statistischen Urteilen. So könnte etwa das Anwenden des Attributsubstitutionskonzepts auf bestimmte statistische Aufgaben kritische heuristische Attribute identifizieren, die anstelle der Zielattribute beurteilt werden. Auch die Analyse der Bedingungen, unter denen verzerrt geurteilt wird und die

Identifikation der Merkmale von Interventionen (Trainings, Lehre), die zum selteneren Einsatz von Heuristiken oder zur Automatisierung von regelgeleiteten statistischen Urteilen beitragen, könnte durch bisherige Ergebnisse des Forschungsprogramms geleitet werden.

Trotz der stärkeren Berücksichtigung von kognitionspsychologischen Erkenntnissen im Vergleich zu anderen Konzipierungen der Statistikkompetenz bleibt die Relevanz des Heuristiken-und-Verzerrungen-Forschungsprogramms für statistical reasoning jedoch offen. Zwar beinhaltet statistical reasoning das korrekte Verständnis in Abgrenzung zu typischen Verzerrungen, allerdings ist nicht vollumfänglich klar, anhand welcher Kriterien die Auswahl der statistical reasoning Typen bzw. Fehlverständnisse erfolgt ist und wie korrektes und verzerrtes statistical reasoning erklärt wird. Das betrifft Fragen, wie: Stellen die Typen des statistical reasonings und die Verzerrungen und Heuristiken eine vollständige oder teilweise (repräsentative oder nichtrepräsentative?) Auswahl dar? Erklärt Attributsubstitution verzerrte statistische Urteile? Wenn ja, welche Attribute werden bei verzerrten statistischen Urteilen herangezogen und unter welchen Bedingungen? Eine mögliche Erklärung des nicht ganz klaren Einbezugs bestimmter Aspekte des Heuristiken-und-Urteilsverzerrungen-Forschungsprogramms in das Konzept des statistical reasonings stellt die zeitliche Entwicklung der beiden Forschungsrichtungen dar. Auch die Forschung zu Heuristiken und Verzerrungen erfolgte zunächst überwiegend deskriptiv und bot wenige Erklärungsansätze über Bedingungen für die Anwendung von Heuristiken an (bis in die frühen 1980er Jahre, Kahneman & Frederick, 2002). Ungefähr zeitgleich (1980er Jahre) begann die Entwicklung des statistical reasoning Konzepts (Garfield & Ahlgren, 1988, Garfield, 1991), so dass die theoretischen oder modellhaften Überlegungen in das statistical reasoning Konzept möglicherweise keinen Eingang mehr finden konnten.[13] Insgesamt ist damit nicht ganz klar, inwiefern das Konzept statistical reasoning „*the way* people reason with statistical ideas“ (Garfield & Chance, 2000, S. 101, Hervorhebung hinzugefügt) erfasst.

Auch wenn für das Konzept des statistical reasoning von Garfield (z. B. 2003) die aufgeführten Einschränkungen festgestellt werden können, lässt sich das

[13] Allerdings wurde Attributsubstitution erstmalig bereits im Jahr 1983 (Tversky & Kahneman) als Erklärungsmechanismus für Urteilsverzerrungen vorgeschlagen.

Konzept grundsätzlich in einen theoriegeleiteten Ansatz, der mehrheitlich durch empirische Ergebnisse gut gestützt wird, einordnen. Damit stellt statistical reasoning eine möglicherweise nutzenbringende Konzeptualisierung der statistischen Kompetenz dar.

2.5. Weitere Konzepte

Neben den dargestellten Konzepten finden sich einige weitere Konzeptualisierungsversuche für die Beherrschung statistischer Inhalte, wie der Begriff der *statistischen Kompetenz* (Biehler, 2001), des *statistical thinking* im Rahmen der Qualitätskontrolle von Produktionsprozessen (Snee, 1991) sowie der *Kompetenz*begriffe innerhalb der Bildungsstandards (KMK, 2003, 2004a, 2004b, 2012) oder der internationalen Vergleichsstudien (z. B. im Organisation for Economic Co-operation and Development [OECD] Programme for International Student Assessment [PISA], 2003, 2004) für statistische Teilbereiche der Mathematik. Da diese Konzepte nur sehr allgemein gefasst wurden (Biehler, 2001) oder statistische Inhalte nur zu einem teilweise geringen Anteil in den Fokus nehmen (z. B. Leitidee Daten und Zufall, KMK, 2004, oder OECD, 2003) und sich nicht auf Statistik in der Hochschule beziehen wird auf eine gesonderte Darstellung dieser Konzepte an dieser Stelle verzichtet. Eine Übersicht der Vorkommenshäufigkeit der bisher beschriebenen Konzepte statistischer Kompetenz in der Literatur findet sich in Tabelle 23 im Anhang.

2.6. Abgrenzung und Systematisierung der Begriffe

Wie aus den Beschreibungen der Konzepte zur Beschreibung von Statistikkompetenz hervorgeht, werden die eingeführten Begriffe nicht immer trennscharf genutzt. So finden unterschiedliche Definitionen für denselben Begriff Verwendung, wie im Fall des statistical thinking (Watson, 1997 vs. Wild & Pfannkuch, 1999) und unterschiedliche Begriffe werden scheinbar synonym (statistical thinking und statistical literacy, Watson, 1997, Watson & Callingham, 2003) gebraucht oder überlappen sich in ihrer Bedeutung mehr oder weniger (statistical literacy, Wallman, 1993, und statistical thinking, Cobb, 1992 sowie statistical reasoning, Garfield, 2003). Angesichts dieser Bedeutungsvielfalt wurden

daher unterschiedliche Vorschläge für die Systematisierung und Abgrenzung der Konzepte formuliert.

2.6.1. Statistical cognition.

Einer der beiden im Folgenden skizzierten Vorschläge stammt von Beyth-Marom, Fidler und Cumming (2008). Die Autorengruppe schlägt statistical cognition „as a concept and an integrative field“ (S. 22) für die Auseinandersetzung mit den Prozessen, Repräsentationen und Aktivitäten, die am Erwerb und der Nutzung statistischer Konzepte beteiligt sind, vor. Hierbei unterscheiden die Autoren zwischen deskriptiven, normativen und präskriptiven Facetten der statistischen Kognition. Forschungsbemühungen innerhalb der deskriptiven Facette richten sich nach Ansicht der Autoren auf die Art und Weise, wie Menschen statistisches Wissen erwerben und nutzen sowie darauf, wie Menschen über statistische Konzepte denken. In diese Kategorie fällt laut der Autoren vor allem die Forschung zum Konzept statistical reasoning. Folgerichtig könnten aber auch Studien zu statistical literacy sowie statistical thinking oder anderen Konzepten die deskriptive Facette abbilden, sofern sie den Status quo der Statistikkompetenz fokussieren. Die normative Facette beinhaltet den wissenschaftlichen Diskurs über die Art und Weise, wie Menschen über statistische Konzepte denken sollten. Hierunter fallen etwa (Bildungs-)Ziele, die im Rahmen der Konzepte statistical literacy oder statistical thinking formuliert sind. Die präskriptive Facette bildet schließlich die Forschung zu den Prozessen und Methoden ab, die die Diskrepanz zwischen deskriptiven Kompetenzständen und normativen Zielen reduzieren können.

2.6.2. Statistical literacy, reasoning und thinking.

Ein anderer Vorschlag zur Abgrenzung der unterschiedlichen Konstrukt- bzw. Konzeptvorschläge für Statistikkompetenz stammt von Garfield, delMas und Chance (2003) und Ben-Zvi & Garfield (2004). Die Autoren fokussieren dabei die drei prominentesten Konzepte statistical literacy, statistical reasoning und statistical thinking. Sie schlagen vor, unter statistical literacy zum einen die basalen Fertigkeiten (skills) zu subsummieren, die für das Verständnis statistischer Informationen oder Forschungsergebnisse gebraucht werden. Hierunter fassen die

Autoren die Fertigkeit, Daten zu organisieren, Tabellen zu erstellen und zu visualisieren und mit verschiedenen Datenrepräsentationen zu arbeiten. Zum anderen umfasst statistical literacy laut der Autoren das Verständnis von statistischen Konzepten, statistischen Fachvokabulars und statistischer Symbole sowie der Wahrscheinlichkeit als Maß der Unsicherheit. Unter statistical reasoning verstehen die Autoren die Art und Weise, wie Menschen mit statistischen Konzepten argumentieren und schlussfolgern und wie sie statistische Informationen verstehen. Dies bedeutet laut der Autoren, dass statistische Prozesse verstanden und erklärt und statistische Ergebnisse in vollem Umfang interpretiert werden können. Hierunter sei etwa die Interpretation von Datensätzen, Datenrepräsentationen oder statistischer Zusammenfassungen sowie das Verbinden verschiedener statistischer Konzepte, wie zentraler Tendenz und Streuung oder von Daten und Zufall zu fassen. Statistical thinking meint, so schlagen die Autoren vor, das Verständnis über die Gründe für und die Vorgehensweise bei statistischen Untersuchungen sowie der übergeordneten Ideen („'big ideas'", Ben-Zvi & Garfield, 2004, S. 7, etwa die Allgegenwärtigkeit von Variabilität oder die Angemessenheit von Datenanalysen), die statistischen Untersuchungen zugrunde liegen. Statistical thinking umfasst laut Autoren das Verständnis über Stichprobenziehung, das Verständnis darüber, wie Schlüsse von Stichproben auf Populationen gezogen werden, und das Verständnis der Notwendigkeit von Experimenten für Kausalaussagen. Auch das Verständnis darüber, wie Modelle für die Simulation von Zufallsphänomenen gebraucht werden, wie Daten für die Schätzung von Wahrscheinlichkeiten produziert werden sowie wie, wann und warum spezifische inferenzielle Methoden in statistischen Untersuchungen von Nutzen sind, fällt den Autoren zufolge unter statistical thinking. Des Weiteren umfasst statistical thinking das Verständnis und die Nutzung des Problemkontextes für die Konzeption der Untersuchung sowie für das Ziehen von Schlussfolgerungen und insgesamt das Erkennen und Verstehen des gesamten Untersuchungsprozesses (Fragestellung, Datenerhebung, Wahl der Datenanalyseverfahren, Prüfen der Voraussetzungen usw.). Schließlich ist statistical thinking durch die Fähigkeit, Ergebnisse von statistischen Untersuchungen kritisch zu evaluieren, gekennzeichnet.

Beide Systematisierungsvorschläge (Beyth-Marom et al., 2008, Ben-Zvi & Garfield, 2004) stellen plausible, wenn auch theoretisch und empirisch nicht

eindeutig abgeleitete Empfehlungen zur Abgrenzung der Begriffe statistical literacy, statistical thinking und statistical reasoning dar.

2.7. Statistical reasoning als konzeptionelle Grundlage der Arbeit

Insgesamt stehen damit nach der Durchsicht der Literatur mehrere Konzeptualisierungen statistischer Kompetenz als Grundlage für die Untersuchung 1) der Beherrschung statistischer Inhalte durch Studierende und 2) der Effektivität der Hochschullehre bei der Vermittlung statistischer Kompetenz zur Verfügung.

Statistical literacy im Sinne von Watson (1997) und Watson und Callingham (2003) stellt eine recht ausdifferenzierte Konzeptualisierung statistischer Kompetenz dar. Sie beruht auf umfangreichen empirischen Untersuchungen (z. B. Callingham und Watson, 2017) und weist eine theoretische Grundlage auf (Biggs & Collis, 1982, 1991). Grundsätzlich wäre das Konzept statistical literacy damit für die Untersuchung des Beherrschungsstands statistischer Inhalte bei Studierenden geeignet. Es ergeben sich jedoch zwei Nachteile aus der Konzeptualisierung von statistical literacy im Sinne von hierarchisch verstandenen Ebenen für die oben genannten Ziele der vorliegenden Arbeit. Das Konzept erlaubt zwar eine Aussage über den status quo der statistischen Kompetenz auf einer der Ebenen (idiosynkratische, informelle, inkonsistente, konsistente nicht-kritische, kritische oder kritisch-mathematische), allerdings ist fraglich, inwiefern diese Positionierung von Lernenden handlungsweisend interpretiert werden kann. Wird eine Lernende oder ein Lernender etwa auf der inkonsistenten Ebene verortet, macht das Modell von Watson und Callingham (2003) keine Aussage darüber, ob es sich hierbei um einen interventionszugänglichen oder von außen nicht (bzw. schwer) beeinflussbaren Status handelt. Der Bezug auf das Modell von Biggs und Collis (1982, 1991), das in der Tradition von Piagetschen Ansätzen steht, legt nahe, dass die kognitive Entwicklung hier als nur eingeschränkt von außen modifizierbar angesehen wird und eher reifungsgebundene Prozesse für Übertritte zu höheren Ebenen sorgen. Damit erweist sich die theoretische Grundlage als eingeschränkt brauchbar für die Erforschung von Statistikkompetenz im Rahmen von Lehrbemühungen. Ein weiterer Nachteil besteht darin, dass das Konzept lediglich eine inhaltsunabhängige, d. h. nicht nach statistischen Konzepten und Inhaltsbereichen getrennte Einschätzung von Lernenden erlaubt. Erreichen Lernende

etwa nur die informelle Ebene in statistical literacy, bleibt daher zum einen offen, welche konkreten Defizite dazu geführt haben (z. B. das Konzept des Mittelwerts vs. Konzept der bedingten Wahrscheinlichkeit betreffend) und zum anderen, an welchen inhaltlichen Stellen die Instruktion geändert werden sollte (z. B. ausführlichere Behandlung von bedingten Wahrscheinlichkeiten, wenn ein nennenswerter Anteil von Lernenden die betreffenden Aufgaben nicht löst). Auch wenn der Beherrschungsstand statistischer Inhalte mit diesem Ansatz also erfasst werden könnte, eignet sich dieser für eine formative Evaluation der Hochschullehre daher nur in unzureichender Weise (wie dies auch für alle gängigen Konzeptualisierungen von *Kompetenz* im engeren Sinn, z. B. laut Weinert, 2014, der Fall ist).

Die Konzeptualisierung des statistical thinking im Sinne von Wild und Pfannkuch (1999) beschreibt Empfehlungen für gutes statistisches Handeln und weniger konkrete Komponenten dessen, was statistische Kompetenzen darstellen sollen. In der Folge erscheint es nicht ganz einfach, die Beherrschung von statistischen Inhalten für eine empirische Analyse hinreichend zu operationalisieren. Entsprechend sind auch die Möglichkeiten zur Untersuchung von Veränderungen in der statistischen Kompetenz durch Hochschullehre eingeschränkt. Diese Konzeptualisierung scheint sich daher in dem gegenwärtigen Zusammenhang ebenfalls weniger gut zu eignen.

Das Konzept statistical reasoning (Garfield & Chance, 2000, Garfield 2003) ebenso wie das statistische Wissen als Bestandteil der statistical-literacy-Wissenskomponente von Gal (2002) beziehen sich auf konkrete statistische Inhalte und stellen damit Vorschläge dafür dar, was Schülerinnen und Schüler sowie Studierende (und in der Folge die Allgemeinbevölkerung) beherrschen sollten. Statistical reasoning (Garfield, 2003) zeichnet sich gegenüber dem Ansatz von Gal (2002) darüber hinaus durch eine etwas klarere Operationalisierung der Statistikkompetenz sowie eine kognitionswissenschaftliche, empirisch relativ umfangreich gesicherte Grundlage aus (z. B. Kahneman et al., 1982, Kahneman & Frederick, 2002, 2005). In Abgrenzung zu statistical literacy (Watson, 1997, Watson & Callingham, 2003) können mit Hilfe des statistical reasoning Ansatzes zudem prinzipiell konkrete Inhalte identifiziert werden, die von Studierenden nicht

ausreichend beherrscht werden[14], wie z. B. das reasoning about statistical measures, und damit als Grundlage für formative Lehrevaluationen dienen. Für die Beurteilung der Beherrschung statistischer Inhalte zum einen und die Beurteilung der Effektivität der Hochschullehre in der Vermittlung statistischer Inhalte zum anderen wird daher das Konzept des statistical reasoning herangezogen.

[14] Hierzu wäre natürlich eine begründete Entscheidung darüber notwendig, ab bzw. bis zu welchem Ausmaß Inhalte als „beherrscht“ und „nicht beherrscht“ gelten.

3. Messung der Statistikkompetenz

Auch wenn für die Messung der Beherrschung statistischer Inhalte erst seit den 1990er Jahren eigene Verfahren vorgeschlagen wurden, müssen Beurteilungen der Leistung von Schülerinnen und Schülern sowie von Studierenden auch zuvor erfolgt sein, etwa anhand von Leistungskontrollen, wie Klassenarbeiten, Klausuren oder Projektarbeiten. Die Messung der Beherrschung statistischer Inhalte hat damit eine ebenso lange Tradition, wie die der Integration statistischer Inhalte in das Schul- oder Hochschulcurriculum.

Seit den 1990er Jahren wurden dann einige Verfahren zur Erfassung statistischer Kompetenzen vorgeschlagen, bei denen ein stärkerer Fokus auf psychometrische Kriterien gesetzt wurde. Die Verfahren lassen sich grob danach unterteilen, ob sie ein spezifisches Konzept statistischer Kompetenz erfassen (etwa statistical reasoning), die Beherrschung bestimmter oder verschiedener statistischer Inhalte (etwa *p*-Werte oder Boxplots) oder die Beherrschung bestimmter oder verschiedener statistischer Konzepte (etwa Konzepte der Variabilität oder bedingten Wahrscheinlichkeit) und ob sie spezifisches Fachvokabular für die Bearbeitung der Aufgaben voraussetzen (z. B. Nutzung von Begrifflichkeiten, wie „Boxplot“, „Korrelation“, „stochastisch unabhängig“ usw.).

Da statistical reasoning als am besten geeignetes Konzept statistischer Kompetenz für die vorliegenden Fragestellungen ausgewählt wurde, soll das zur Messung von statistical reasoning entwickelte Verfahren im Folgenden detailliert beschrieben werden. Zunächst werden die Konstruktion und der Aufbau des Verfahrens dargestellt. Im Anschluss werden einige empirische Untersuchungen zum Verfahren berichtet und die Güte des Instruments vor dem Hintergrund verschiedener Konstruktions- bzw. Messabsichten beurteilt.

Nach der Darstellung des Verfahrens zur Messung von statistical reasoning werden einige weitere vorliegende Verfahren skizziert.

3.1. Statistical Reasoning Assessment

Das Statistical Reasoning Assessment (SRA, Garfield, 1991, 2003) wurde im Rahmen des ChancePlus-Projekts für Schulen und Hochschulen in den Vereinigten Staaten von Amerika entwickelt und dient der Erfassung des statistical reasoning. Das Instrument wurde mit dem Ziel entwickelt, die Erreichung von

Bildungsstandards in statistischen Inhalten, die von dem National Council of Teachers of Mathematics in 1989 (NCTM, 1989) formuliert wurden, zu prüfen und stellt das erste (bzw. älteste) eigenständige Instrument zur Erfassung statistischer Kompetenz dar.

Für die Konstruktion des SRA (Garfield, 1991, 1998 und 2003) formulierten die Beteiligten des ChancePlus Projekts (Psychologen und Psychologinnen, Lehrende sowie Statistikerinnen und Statistiker) im ersten Schritt die weiter oben beschriebenen Typen des statistical reasoning mit entsprechenden Items. Im zweiten Schritt schätzten Expertinnen und Experten die konstruierten Items auf ihre Inhaltsvalidität hinsichtlich der anvisierten statistical reasoning Typen ein und machten Änderungs- und Ergänzungsvorschläge. Anschließend bearbeiteten verschiedene Gruppen von Lernenden die Items mit zum Teil offenen Antwortmöglichkeiten. Die Antworten auf die offen formulierten Items dienten dann als Grundlage für die (geschlossenen) Antwortoptionen der Items im Einfach- oder Mehrfachantwortformat. Vor der ersten Veröffentlichung des Instruments erfolgten schließlich mehrere Pilotierungen in unterschiedlichen Kontexten.

Das Instrument enthält in seiner finalen Version insgesamt 20 Items (Übersicht der Originalitems und der Zuordnung zu Subskalen bei Garfield, 2003 oder in der deutschen Version im Anhang in Tabelle 26), von denen 15 Verwendung in gleichzeitig zwei Kategorien, den *correct reasoning skills* und den *misconceptions* finden. Die Antworten auf diese Items lassen sich daher vierfach auswerten:

1) Wurde das Item korrekt gelöst (ohne Zuordnung zu bestimmten Typen des statistical reasoning)? (Dies kann lediglich für Item Nr. 7 der Fall sein.)
2) Wurde das Item im Sinne des correct statistical reasoning gelöst? (Dies kann für alle Items außer Nr. 7 der Fall sein.)
3) Wurde das Item im Sinne der misconceptions nicht korrekt gelöst? (Dies ist für alle Items mit Ausnahme der Items Nr. 4, 5, 8 und 18 möglich.)
4) Wurde das Item nicht korrekt gelöst (ohne Zuordnung zu bestimmten Fehlinterpretationen)? (Dies ist für die Items Nr. 4, 5, 8 und 18 möglich.)

Die Antworten auf die Items im Sinne der correct reasoning skills werden vier der sechs Typen des statistical reasoning zugeordnet (Garfield, 2003). Das Beurteilen statistischer Kennzahlen (reasoning about statistical measures) wird mit drei Items erfasst, die die Subskala „Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird“ (understands how to select an appropriate average) darstellen. Das Schließen unter Unsicherheit (reasoning about uncertainty) wird mit drei Subskalen und insgesamt zehn Items erfasst: „Wahrscheinlichkeiten korrekt interpretieren“ (correctly interprets probabilities, zwei Items), „Wahrscheinlichkeiten korrekt berechnen“ (correctly computes probabilities, fünf Items) mit den beiden Subsubskalen „Wahrscheinlichkeit als Verhältnis verstehen“ (understands probabilities as ratios, ein Item), und „kombinatorisch denken“ (uses combinatorial reasoning, vier Items) sowie „(stochastische) Unabhängigkeit verstehen“ (understands independence, drei Items). Das Beurteilen (der Güte) von Stichproben (reasoning about samples) wird mit insgesamt vier Items und den beiden Subskalen „Variablilität innerhalb von Stichproben verstehen“ (understands sampling variability, zwei Items) und „Bedeutung großer Stichproben verstehen“ (understands importance of large samples, zwei Items) abgebildet. Das Beurteilen von Zusammenhängen (reasoning about association) wird über zwei Items erfasst, die jeweils eine eigene Subskala „zwischen Korrelation und Kausalität unterscheiden“ (distinguishes between correlation and causation, ein Item) und „Vierfeldertafeln korrekt interpretieren“ (correctly interprets two-way tables, ein Item) darstellen. Die beiden verbleibenden Typen des correct statistical reasoning (das Beurteilen von Daten [reasoning about data] und das Beurteilen von Datenrepräsentationen [reasoning about representations of data]) werden nicht explizit über spezifische Items des SRA erfasst.

Auch für die Antworten auf die Items im Sinne der misconceptions stellen die Autoren eine Zuordnung zu weit verbreiteten spezifischen Fehlvorstellungen zu. Das „Fehlverständnis von Mittelwerten“ (misconception involving averages) stellt eine eigene Subskala mit vier Subsubskalen dar: „Mittelwert als am häufigsten vorkommender Messwert“, (averages are the most common number, zwei Items), „fehlende Berücksichtigung von Ausreißern bei der Bestimmung des Durchschnitts“ (fails to take outliers into consideration when computing the mean, ein Item), „Vergleich von Gruppen anhand ihrer Mittelwerte“ (compares groups based on their

averages, ein Item), „Verwechseln des Durchschnitts mit dem Median“ (confuses mean with median, ein Item). Die Fehlvorstellungen „Orientierung am Ergebnis“ (outcome orientation, fünf Items), „gute Stichproben müssen einen hohen Anteil der Population darstellen“ (good samples have to represent a high percentage of the population, zwei Items), „Gesetz der kleinen Zahl“ (law of small numbers, zwei Items), die „Repräsentativitätsheuristik“ (representativeness misconception, drei Items) und die „Gleichwahrscheinlichkeitsverzerrung“ (equiprobability bias, vier Items) bilden jeweils eigene Subskalen. Neben diesen auch bereits weiter oben beschriebenen Fehlvorstellungen werden zusätzlich das „Gleichsetzen von Korrelation mit Kausalität“ (correlation implies causation, ein Item) sowie die „Überzeugung, dass Gruppen nur verglichen werden können, wenn sie dieselbe Stichprobengröße haben“ (groups can only be compared if they are the same size, ein Item) als Subskalen der misconceptions erfasst.

Die Items des SRA enthalten nur wenige statistische Fachbegriffe (Wahrscheinlichkeit, Prozent, Stichprobe, Durchschnitt, Variation), so dass sich das Instrument zur Erfassung statistischer Kompetenz bei Lernenden mit wenig oder keiner Instruktion in statistischen Inhalten eignet und damit auch grundsätzlich für die Erfassung von statistical reasoning vor der Statistiklehre in der Hochschule. Damit besteht die Möglichkeit, Lehrbemühungen in ihrer Effektivität durch eine Vorher-Nachher-Messung zu untersuchen.

Die psychometrischen Eigenschaften des Instruments wurden in mehreren Studien untersucht. Im Folgenden werden die Ergebnisse zur Reliabilität, der internen Struktur, Zusammenhängen zu anderen Merkmalen sowie Veränderungen durch Interventionen (Validitätsaspekte) dargestellt.

3.1.1. Empirische Ergebnisse zum SRA.

Liu und Garfield (2002) bzw. Garfield (2003) setzten das Instrument am Ende einer einführenden Lehrveranstaltung zu Statistik ein und berichten eine Retestreliabilität für den Zeitraum von einer Woche von $r_{tt} = .70$ für Items des correct statistical reasoning und $r_{tt} = .75$ für Items der misconceptions bei einer Stichprobe von $n = 32$ amerikanischen Studierenden in den ersten zwei Studienjahren. Die Homogenität in Form von interner Konsistenz gibt Garfield (2003) für die Items als gering an und nicht indikativ dafür, dass das Instrument eine

einheitliche Fähigkeit erfasst. Des Weiteren berichtet Garfield (2003) sehr niedrige Korrelationen zwischen der Gesamtpunktzahl im correct reasoning bzw. in den misconceptions zu verschiedenen Leistungsmessungen am Ende eines einführenden Statistikkurses. Als Einzelkriterien dienten hierbei unter anderem die Gesamtleistung in der Lehrveranstaltung, die Leistung im Projekt sowie die Leistung in mehreren Kurztests innerhalb der Lehrveranstaltung.

Tempelaar, Gijselaers und van der Loeff (2006) erfassten in ihrer umfangreichen Untersuchung mit annähernd 2000 Probanden (Studierende der Wirtschaftswissenschaften im ersten Semester) das statistical reasoning anhand des SRA in der ersten Woche einer Statistiklehrveranstaltung. Sie geben für das correct reasoning mit den jeweiligen Subskalen als Indikatoren ein Cronbachs α von $\alpha = .29$ und für die misconceptions (ebenfalls die jeweiligen Subskalen als Indikatoren) $\alpha = .11$ für die Homogenität des Instruments an. Die Analyse der internen Struktur der 16 Subskalen des Instruments mittels einer explorativen Faktorenanalyse auf der Basis von Antworten von ca. 1300 Studierenden (Teilstichprobe von Tempelaar et al., 2006) legt nach Tempelaar (2004) eine siebenfaktorielle Struktur nahe, wobei größtenteils jeweils eine Subskala des correct reasoning und eine Subskala der misconceptions auf einen Faktor laden. Die Autoren (Tempelaar et al., 2006) analysierten zudem lineare Zusammenhänge des correct reasoning und der misconceptions zu benoteten Hausaufgaben, Kurztests sowie den Abschlusstests für jeweils drei Lehrveranstaltungsabschnitte in der Statistikveranstaltung und einer parallel stattfindenden Mathematikveranstaltung ($n = 680$). Zusätzlich erhoben Tempelaar et al. Einstellungen zu Statistik ($n = 2031$ bzw. $n = 687$) und Einstellungen zum Lernen sowie Lernverhalten ($n = 1767$), ebenfalls in der ersten Veranstaltungswoche. Die Zusammenhänge zwischen der Gesamtpunktzahl im correct reasoning und den Hausaufgaben in Statistik betragen $r = -.02$ für den ersten Lehrveranstaltungsabschnitt und $r = -.13$ für den dritten Lehrveranstaltungsabschnitt. Die Zusammenhänge zwischen dem correct reasoning und den Kurztest in Statistik sind ebenfalls sehr gering ($r = .01$ und $r = -.01$ für die ersten beiden Lehrveranstaltungsabschnitte) und überwiegend gering zu den Abschlusstests in Statistik ($r = .24$ für den ersten, $r = .06$ für den zweiten und $r = .07$ für den dritten Zeitabschnitt). Ähnliche, zum Teil etwas engere Zusammenhänge berichten Tempelaar et al. (2006) für das correct reasoning und die Leistungen in der

Mathematikveranstaltung (für Hausaufgaben $-.06 < r < -.14$ und für die Abschlusstests $.13 < r < .28$). Auch die Korrelationen zwischen dem correct reasoning und Einschätzungen bezüglich 1) der eigenen affektiven Einstellung zur Statistik, 2) der eigenen bzw. der erforderlichen kognitiven Kompetenz für das Lernen von Statistik, 3) des grundsätzlichen Werts von Statistik, 4) der Schwierigkeit des Fachs Statistik, 5) des eigenen Interesses an Statistik sowie 6) des geplanten Aufwands für die Statistikveranstaltung (gemessen über das Survey of Attitudes Towards Statistics, SATS, Schau, Stevens, Dauphinee & Del Vecchio, 1995, Schau, 2003) liegen mit $-.07 < r < .12$ im deutlich niedrigen Bereich ebenso, wie die Korrelationen zwischen dem correct reasoning und den Einstellungen zum Lernen und Lernverhalten ($-.09 < r < .10$).

Tempelaar, van der Loeff und Gijselaers (2007) berichten für eine weitere Stichprobe von $n = 1499$ Studierenden (ebenfalls im ersten Semester) eine gute Modellpassung für diese siebenfaktorielle Lösung auf Basis von konfirmatorischen Faktorenanalysen. Eine Übersicht der Ladungen der Subskalen auf die latenten Variablen und die entsprechenden Items findet sich in Tabelle 2.

Eine weitere Studie (Martin, Hughes & Fugelsang, 2017) mit $n = 199$ Studierenden mit unterschiedlich großer Erfahrung in Statistik zeigt hohe Zusammenhänge zwischen dem Gesamtwert der Subskalen des correct reasoning und dem Cognitive Reflection Test (CRT, erfasst Inhibition intuitiver Antworten auf mathematische Aufgaben, Frederick, 2005, $r = .56$) sowie dem Wonderlic Personnel Test (WPT, erfasst allgemeine kognitive Fähigkeiten, Wonderlic Inc., 1999, zitiert nach Martin, 2017, $r = .55$). Etwas niedrigere Zusammenhänge ergaben sich in der Studie zwischen dem correct reasoning zu der Preference for Numerical Information Skala (PNI, Viswanathan, 1993, $r = .38$) und zur Numeracy Skala (NS, erfasst Interpretation von Wahrscheinlichkeitsangaben, Lipkus, Samsa & Rimer, 2001, $r = .44$). Eher geringe Zusammenhänge berichten die Autoren zwischen correct reasoning zum Bedürfnis nach kognitiver Beanspruchung (need for cognition, NFC, Cacioppo, Petty & Kao, 1984, $r = .26$), der Actively Open-Minded Thinking Skala (AOT, erfasst epistemische Regulation, Sá, West & Stanovich, 1999, $r = .20$) sowie zum rezeptiven Wortschatzwissen (Stanovich & West, 1997, Zimmerman, Broder, Shaughnessy & Underwood, 1977, $r = .14$). Für die misconceptions in Form des Gesamtwerts der entsprechenden Subskalen berichten

Tabelle 2

Subskalen, die jeweils einer latenten Variablen zugeordnet werden mit den entsprechenden Ladungen und Items aus Tempelaar et al. (2007)

LV	Subskalen	Ladungen	Items
1	Wahrscheinlichkeit korrekt berechnen	.80	8, 13, 18, 19, 20
	Gleichwahrscheinlichkeitsverzerrung	.90	13, 18, 19, 20
2	Variabilität innerhalb von Stichproben verstehen	.81	14, 15
	Gesetz der kleinen Zahl	.85	12, 14
3	(stochastische) Unabhängigkeit verstehen	.88	9, 10, 11
	Repräsentativitätsheuristik	.77	9, 10, 11
4	Wahrscheinlichkeiten korrekt interpretieren	.88	2, 3
	Vierfeldertafeln korrekt interpretieren	.13	5
	Bedeutung großer Stichproben verstehen	.51	6, 12
	Orientierung am Ergebnis	-.52	2, 3, 11, 12
5	Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird	.85	1, 4, 17
	Fehlverständnis von Mittelwerten	-.51	1, 15, 17
6	Zwischen Korrelation und Kausalität unterscheiden	.50	16
	Gleichsetzen von Korrelation mit Kausalität	-.93	16
7	Gute Stichproben müssen einen hohen Anteil der Population darstellen	-.09	7, 16
	Überzeugung, dass Gruppen nur verglichen werden können, wenn sie dieselbe Stichprobengröße haben	-.95	6

Anmerkungen. LV = latente Variable

die Autoren eine moderate bis hohe negative Korrelation zum CRT ($r = -.41$), niedrige bis moderate negative Zusammenhänge zum NFC ($r = -.25$), PNI ($r = -.24$), WPT ($r = -.22$) und NS ($r = -.22$) und einen niedrigen negativen Zusammenhang zum rezeptiven Wortschatzwissen ($r = -.14$). In einem Strukturgleichungsmodell klärten die Denkdispositionen als latente unabhängige Variable (gemessen durch NFC, PNI und AOT) und die kognitiven Fähigkeiten als latente Mediatorvariable (gemessen durch CRT, WPT und NS) 50% der Varianz in der latenten abhängigen Variablen statistical reasoning (gemessen durch die Indikatoren correct statistical reasoning und misconceptions) auf.

Olani, Hoekstra, Harskamp und van der Werf (2011) und Gundlach, Richards, Nelson und Levesque-Bristol (2015) führten für einen Teil der SRA-Items Wiederholungsmessungen zu Beginn und am Ende einer Statistiklehrveranstaltung durch. In beiden Studien wurden durch die Lehrenden (Olani et al.) bzw. die Autoren (Gundlach et al.) die Items ausgewählt, die den Inhalt der Lehrveranstaltung am besten abbilden. Gundlach et al. wählten insgesamt acht Items[15] mit je einem Item aus den Subskalen „Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird", „Variabilität innerhalb von Stichproben verstehen", „zwischen Korrelation und Kausalität unterscheiden", „Bedeutung großer Stichproben verstehen" und drei Items aus der Subskala „(stochastische) Unabhängigkeit verstehen" aus und gaben diese zu Beginn und am Ende von jeweils drei unterschiedlichen Lehrveranstaltungsformaten (traditionell, $n = 193$, online, $n = 43$, sowie inverted classroom, $n = 25$) Bachelor-Studierenden („undergraduates") vor. Die Autoren berichten einen Anstieg von 4% (traditionelle und *inverted classroom* Statistiklehrveranstaltung) bzw. 9% (online Statistiklehrveranstaltung) der durchschnittlich korrekt gelösten Items zum Ende der Lehrveranstaltung (der besseren Interpretierbarkeit halber wurden die Angaben von gewichteten Punktwerten von Gundlach et al. durch die Autorin in Prozent umgerechnet). Bei den misconceptions ließ sich ein Rückgang von 1% (traditionelle Lehrveranstaltung), 3% (Veranstaltung im inverted classroom Format) sowie 6% (Veranstaltung im online Format) beobachten. Olani et al. wählten sieben Items aus dem SRA sowie acht Items aus dem Statistical Concept Inventory (SCI, Allen,

[15] Das achte Item ging lediglich in eine Subskala der misconceptions ein, daher ergeben die Items für die Subskalen des correct reasoning in der Summe sieben.

2006) aus. Zum Ende des Semesters berichten die Autoren einen Anstieg von insgesamt 11% der korrekt gelösten Items durch die Studierenden im ersten Studienjahr, allerdings schränken sie ein, dass lediglich sieben der 15 Items einen signifikanten Anstieg in ihren Schwierigkeiten aufweisen (d. h., der Anteil der Studierenden, die das Item korrekt lösten, stieg zum zweiten Messzeitpunkt lediglich für sieben der 15 Items).

3.1.2. Beurteilung des SRA auf der Basis bisheriger Forschung.

Eine grundlegende Herausforderung bei der Erforschung und Beurteilung der psychometrischen Eigenschaften des SRA liegt darin, dass die Konstruktionsintention bisher nicht explizit ausformuliert worden ist. So ergeben sich unterschiedliche Anforderungen an bzw. Einsatzbereiche für das Instrument, je nachdem, ob dieses ein (psychologisches) Konstrukt, ein lehrzielorientiertes Kriterium oder einen kognitiven Prozess erfassen soll.

Statistical reasoning als Konstrukt.

Wird statistical reasoning als Konstrukt, etwa im Sinne eines Persönlichkeitsmerkmals verstanden, so sollte das Messinstrument unter anderem das Konstrukt in seiner internen Struktur, d. h. Dimensionalität abbilden, mit anderen relevanten Merkmalen in einem sinnvollen Muster zusammenhängen und plausible Gruppenunterschiede aufweisen (siehe z. B. Cronbach & Meehl, 1955 und Messick, 1995, zu Begriffen der Validität bzw. Konstruktvalidität und deren empirischer Nachweisbarkeit).

Statistical reasoning als Lehrziel.

Wenn statistical reasoning dagegen ein Lehrziel beschreibt, dann spielt die interne Struktur des Messinstruments keine entscheidende Rolle: Testbestandteile, d. h. Items oder Gruppen von Items, müssen nicht notwendigerweise in bestimmter Weise zusammenhängen, da für die Lösung der einzelnen Inhalte unterschiedliche kognitive Prozesse benötigt werden könnten (etwa die Auswahl eines angemessenen Maßes der zentralen Tendenz in Item Nr. 1 vs. die Bestimmung von Wahrscheinlichkeit in Item Nr. 19). Daneben führt die Abwesenheit von Streuung im eigentlich angestrebten (extremen) Fall, dass alle Lernenden das Lehrziel erreichen (d. h. die Höchstpunktzahl im Test erzielen) sogar dazu, dass die Schätzung von Zusammenhängen sowohl zwischen den Items als auch zwischen

dem Testwert und anderen Merkmalen mathematisch unmöglich wird[16] (Klauer, 1987).

Nomologisch sinnvollerweise kann man dagegen erwarten, dass das Instrument einen Lernzuwachs in Gruppen von Lernenden abbildet, die die Inhalte, die das Messinstrument erfasst, behandelt haben, die Effektivität der Lehrintervention vorausgesetzt (Hambleton, Swaminathan, Algina & Coulson, 1978). Mindestens sollten Lernendengruppen, die die Inhalte, die das Messinstrument erfasst, behandelt haben, eine höhere Anzahl von Testpunkten erzielen als die Lernendengruppen, die die entsprechenden Inhalte nicht behandelt haben (konvergente und diskriminante Validität, siehe hierzu z. B. Campbell & Fiske, 1959). Auch kann erwartet werden, dass Testaufgaben nach einer Lehrintervention leichter zu lösen sind, als vor der Intervention (Hambleton et al., 1978).

Solche empirischen Hinweise auf Konstruktvalidität sind jedoch der zentralen Anforderung der Inhaltsvalidität oder Inhaltsrepräsentativität (Messick, 1987) an einen lehrzielorientierten (bzw. kriteriumsorientierten) Test nachgeordnet (erste Hinweise hierauf finden sich bei Hammock, 1960 und Glaser, 1963). Inhaltsvalide ist das Instrument dann, wenn es im Sinne eines Lehrziels „die Gesamtheit einer Menge von Aufgaben [d. h., den Lehrstoff] enthält oder repräsentiert“ (Klauer, S. 12). Das Lehrziel hat dabei einen qualitativen und einen quantitativen Aspekt. Der quantitative Aspekt beschreibt das geforderte Ausmaß der Beherrschung. Der qualitative Aspekt bezieht sich auf den Lehrstoff – die Aufgabengrundmenge („universe of content“, Osburn, 1968, S.95). Damit definiert eine Aufgabengrundmenge den qualitativen Aspekt des Lehrziels, den Lehrstoff und auch den Test, der das Lehrziel erfassen soll. Ein Beispiel hierfür wäre etwa das Lehrziel „Addition natürlicher Zahlen“. Dieses Lehrziel ist, so wie es hier formuliert ist, nicht konkret genug und kann daher in dieser Form nicht valide getestet werden. Es bleibt offen, ob zwei, drei oder unendlich viele natürliche Zahlen addiert werden können sollen (dies würde deutlich unterschiedliche Kompetenzen widerspiegeln) und ggf. ob ein bestimmter Bereich der natürlichen Zahlen, etwa 0 bis 100 oder aber 0 bis ∞, relevant ist (was ebenfalls dramatische Unterschiede von Kompetenzen

[16] Dies gilt ungeachtet der verwendeten Testtheorie, d. h. sowohl für die Klassische Testtheorie als auch die Item Response Theorie. Wenn lediglich der wahre Wert als ideal angesehen wird, betragen die Zusammenhänge sowie die Reliabilität 0 (Klauer, 1987, S. 7)

reflektieren würde). Sobald man das Lehrziel hinreichend genau bestimmt hat, ergibt sich aus der Definition gleichzeitig die *mögliche* Menge von Aufgaben, die das Lehrziel erfassen: „all items that could possibly appear in the test should be specified in advance“ (Osburn, S. 96). Mit anderen Worten, zu sagen, dass das Lehrziel darin besteht, dass eine Lernende die Addition von zwei natürlichen Zahlen im Zahlenbereich von 0 bis 100 beherrscht, ist äquivalent zu der Aussage „die Lernende kann Aufgaben zur Addition zweier natürlicher Zahlen im Zahlenbereich von 0 bis 100 lösen“ (die genaue Angabe, darüber, wie viele der Aufgaben gelöst werden können sollen, wäre dann die Beschreibung des quantitativen Lehrzielaspekts). Gleichzeitig liefert die Formulierung auch eine konkrete Angabe darüber, was gelehrt werden soll, d. h. über den Lehrstoff. Damit wird eine weitere Anforderung an ein lehrzielorientiertes Verfahren deutlich: der Lehrstoff (die Aufgabenmenge) sollte Gegenstand der Lehre sein. Zentral hierbei ist aber der Umstand, dass eine Aussage der Art „die Lernende *beherrscht* die Addition von zwei natürlichen Zahlen im Zahlenbereich von 0 bis 100“ (bzw. „die Lernende *löst Aufgaben* zur Addition von zwei natürlichen Zahlen im Zahlenbereich von 0 bis 100“) nur dann getroffen werden kann, wenn die Aufgaben, die die Lernende tatsächlich bearbeitet hat, eine repräsentative (oder vollständige) Stichprobe der Aufgabengrundmenge, über die man eine Aussage macht, darstellen. Hierfür ist es daher erforderlich, dass die Aufgaben, die man in einem lehrzielorientierten Test zusammenstellt, eine Zufallsstichprobe aus der Aufgabengrundmenge darstellen. Damit erlaubt es ein lehrzielorientiertes Verfahren, die Leistung von Lernenden in Bezug auf eine wohldefinierte Verhaltensdomäne (wie z. B. das Maß erfolgreicher Bearbeitung von vollständig definierten Additions- oder Statistikproblemen) zu beschreiben (Popham, 1974).

Die zentralen Anforderungen an die Messung des statistical reasoning im Sinne eines Lehrziels (oder Kriteriums) sind damit zunächst eine wohldefinierte Aufgabengrundmenge und eine repräsentative Auswahl von Items aus der Aufgabengrundmenge. Zusätzlich sollten sich erwartungsgemäße Unterschiede zwischen und innerhalb von Gruppen von Personen sowie Testaufgaben in

Abhängigkeit von der Behandlung des Lehrstoffs (d. h. der Aufgabengrundmenge) zeigen.[17]

Statistical reasoning als kognitiver Prozess.

Wird statistical reasoning als kognitiver Prozess (oder als ein Bündel von Prozessen) aufgefasst, ähnlich wie deduktives Schließen oder Problemlösen, so stehen hier die einzelnen Items und deren leichte Abwandlungen im Vordergrund. So kann etwa untersucht werden, ob und wenn ja, weshalb (d. h. unter welchen Bedingungen) das Erkennen von stochastischer Unabhängigkeit unterschiedlich gut gelingt, je nach Formulierung im Aufgabenstamm (z. B. „Welche der folgenden Reihenfolgen ist am wahrscheinlichsten das Resultat vom fünfmaligen Werfen einer fairen Münze?“ vs. „Welche der folgenden Reihenfolgen ist das am wenigsten wahrscheinliche Resultat vom fünfmaligen Werfen einer fairen Münze?“, siehe Items Nr. 9 und Nr. 11). Auch die Rolle von Kontexteffekten (z. B. Relevanz oder Vorwissen) für einen Iteminhalt im statistical reasoning als kognitiven Prozess kann hierbei ein Untersuchungsgegenstand sein (z. B. ob der Inhalt von Item Nr. 3 leichter gelöst werden kann, wenn der Aufgabenkontext die Wahrscheinlichkeit des Bestehens einer Klausur statt die Regenwahrscheinlichkeit darstellt). Prominente Forschungsbemühungen zu diesem möglichen Verständnis von statistical reasoning wurden im Rahmen des Heuristiken-und-Urteilsverzerrungen Forschungsprogramms rund um Kahneman und Tversky (z. B. Tversky & Kahneman, 1983, siehe auch Abschnitt 2.4.1) unternommen.

In der Tradition der experimentellen Allgemeinen Psychologie steht die gegenseitige Beziehung zwischen Manipulationsparadigma (z. B. ein bestimmter Aufgabentyp) und Theorie (z. B. Theorie mentaler Modelle, z. B. Johnson-Laird [2012] oder Zwei-Prozess- bzw. Zwei-System-Theorien, z. B. Evans [2008]) bei der Untersuchung von kognitiven Prozessen im Zentrum der Forschungsbemühungen. Letztendlich sind daher solche Items als valide zu beurteilen, die einen Beitrag zur Weiterentwicklung von Theorien leisten (siehe z. B. Prinz, Müsseler & Rieger, 2017 oder Funke & Spering, 2006).

[17] Das recht komplexe Thema der Reliabilitätsprüfung von kriteriumsorientierten Messverfahren wird an dieser Stelle nicht ausgeführt, sondern auf die Beschreibung von z. B. der Generalisierbarkeitstheorie von Cronbach (Cronbach, Rajaratnam & Gleser, 1963, Anwendungsbeispiel bei Hively, Patterson & Page, 1968), Modellen der Klassischen Testtheorie oder Modellen der Item Response Theorie verwiesen (Klauer, 1987). Die zentrale Bedeutung der Inhaltsvalidität für kriteriumsorientierte Messverfahren bleibt davon unberührt.

SRA als Messinstrument des statistical reasoning als Konstrukt.

Um das Instrument aus konstruktorientierter Sicht beurteilen zu können, muss zunächst offengelegt werden, wie statistical reasoning als Konstrukt definiert wird. Garfield (2003) lässt hier unterschiedliche Möglichkeiten zu. Zunächst ist denkbar, dass statistical reasoning (im Sinne des korrekten Verständnisses), wie es im SRA gemessen wird, ein eindimensionales Konstrukt darstellt. Dafür spricht der Umstand, dass Garfield (2003) sowie Liu und Garfield (2002) die Gesamtsumme des correct statstical reasoning (d. h. die Summe der Subskalen des correct reasoning) in ihren Analysen verwenden und die interne Konsistenz über alle Items bestimmen. Auch eine vierdimensionale Struktur des SRA ist mit den Darstellungen von Garfield vereinbar. Sie nennt vier Typen des (correct) statistical reasoning, die anhand des SRA erfasst werden können. Garfield führt hierbei nicht an, ob Zusammenhänge zwischen den verschiedenen Typen des statistical reasoning zu erwarten sind, so dass beliebige Beziehungen (positiv, negativ und null) mit der Struktur vereinbar sind. Schließlich ist auch eine achtdimensionale interne Struktur des (correct) statistical reasoning im Sinne der von Garfield beschriebenen Subskalen denkbar. Hierbei lässt Garfield explizit auch negative Zusammenhänge zwischen den Subskalen zu. So führt sie etwa an: „even people who can correctly compute probabilities tend to apply faulty reasoning when asked to make an inference or judgment about an uncertain event, relying on incorrect intuitions" (S. 26). Damit sind auch für die Beziehungen zwischen den correct reasoning Subskalen des SRA beliebige Zusammenhänge denkbar. Analog lässt sich für das Fehlverständnis statistischer Konzepte im Sinne der misconceptions jeweils eine ein- (misconceptions), sechs- (Arten von misconceptions) oder achtdimensionale (Subskalen der misconceptions) Struktur laut Garfield postulieren.

Eine Besonderheit im Aufbau des SRA besteht darin, dass überwiegend dieselben Items (15 von 20, davon drei im Mehrfachwahlformat, siehe Tabelle 25 im Anhang) gleichzeitig Indikatoren für die correct reasoning und die misconceptions Subskalen darstellen. Jeweils zwei der correct reasoning und misconceptions Subskalen enthalten identische Items (siehe z. B. [stochastische] Unabhängigkeit verstehen und Repräsentativitätsheuristik, Tabelle 25), ein weiteres Subskalenpaar enthält vier gleiche Items und bis auf eine correct reasoning Subskala enthalten alle Subskalen mindestens ein gleiches Item (Tabelle 25 im Anhang). Für die Analyse

des correct reasoning und der misconceptions innerhalb eines gemeinsamen Modells ergibt sich damit das Problem der abhängigen Messungen: Hat eine Person eines dieser Items korrekt gelöst, erhält sie einen Punkt in einer Subskala des correct reasoning und kann keinen Punkt mehr für eine oder mehrere misconceptions Subskalen erhalten. In der Folge sind Zusammenhänge zwischen den correct reasoning und entsprechenden misconceptions Subskalen auf statistische Artefakte zurückzuführen (mindestens zum Teil bei Subskalenpaaren, die nur teilweise aus denselben Items oder aus Items im Mehrfachantwortformat generiert werden, und im vollen Umfang bei Subskalenpaaren, die aus denselben Items im Einfachwahlformat generiert werden). Aus statistischer Sicht scheint es daher am sinnvollsten, die Struktur des correct statistical reasoning und der misconceptions in getrennten Modellen zu bestimmen. Auch aus inhaltslogischer Sicht gibt es ein gutes Argument für die getrennte Betrachtung des correct reasoning und der misconceptions: da das korrekte statistical reasoning definitionsgemäß voraussetzt, dass mögliche Fehlverständnisse überwunden wurden, bildet es gleichzeitig das Nichtvorliegen der misconceptions ab. Nur in dem Fall, wenn korrektes statistical reasoning in bestimmten Bereichen nicht erfolgt, ist es von Interesse, die Art des Defizits zu untersuchen und hierbei gegebenenfalls spezifische gängige Fehlverständnisse aufzudecken (hierbei sollte allerdings berücksichtigt werden, dass auch andere Defizite als misconceptions im Sinne der heuristischen Urteilsverzerrungen im correct statistical reasoning vorliegen können). Damit können das correct statistical reasoning und die misconceptions gewissermaßen als verschiedene Konstrukte begriffen werden, auch wenn sie natürlich als in einem Zusammenhang stehend angenommen werden sollten, analog zu der Beziehung zwischen z. B. Leistungen in der Satzsemantik und Defiziten in der Syntax im Rahmen der Sprachkompetenzforschung.

Für die Erfassung der Beherrschung von statistischen Konzepten einerseits und der Wirksamkeit von Lehrbemühungen im Fach Statistik andererseits genügt aus den oben genannten Gründen die Messung des correct statistical reasoning. Im Folgenden beschränken sich die Darstellungen daher auf das correct statistical reasoning als Konstrukt, Lehrziel oder kognitiver Prozess, wie es mit dem SRA erhoben wird. Die Bezeichnungen „correct statistical reasoning“ und „statistical reasoning“ werden dabei synonym gebraucht, sofern nicht anders erkenntlich.

Reliabilität und interne Struktur.

Die bei Garfield (2003) berichtete Retestreliabilität des Instruments kann vorerst als zufriedenstellend eingeordnet werden. Hinsichtlich der internen Struktur des statistcal reasoning als Konstrukt lässt sich die Befundlage als nicht eindeutig bewerten. Garfield (2003) und Liu und Garfield (2002) legen zwar eine geringe interne Konsistenz für die Gesamtmenge der Items nahe, jedoch geben sie nicht an, welche Koeffizienten für die Analyse herangezogen wurden (z. B. Cronbach α oder McDonald ω). So wird die tatsächliche interne Konsistenz durch das Cronbach α unterschätzt, wenn das Instrument mehrere (ggf. hierarchisch angeordnete) Dimensionen erfasst (z. B. Sijtsma, 2009, Revelle & Zinbarg, 2009, Dunn, Baguley & Brunsden, 2014), was im vorliegenden Fall inhaltslogisch erwartet werden kann. Darüber hinaus ist nicht erkennbar, wie hoch die resultierenden Kennwerte genau ausfallen und ob tatsächlich alle Items bei der Analyse berücksichtigt wurden oder nur Items, die in das correct reasoning eingehen (d. h. unter Ausschluss des Items Nr. 7). Auch für die berichteten geringen Iteminterkorrelationen fehlen konkretere Angaben, wie die Art des verwendeten Korrelationskoeffizienten und die genaue Höhe der Zusammenhänge (z. B. ist die Höhe des Produkt-Moment-Koeffizienten abhängig von den Randverteilungen der beteiligten Variablen, Bortz, 2005, Eid, Gollwitzer & Schmitt, 2017) sowie die Mittelwerte bzw. Streuungen der Items (bei Varianzeinschränkung ist ebenfalls mit in ihrer Höhe eingeschränkten Produkt-Moment-Korrelationen zu rechnen, Bortz, 2005, Eid et al., 2017). Damit bleibt offen, ob die Items gegebenenfalls Zusammenhänge im Sinne einer vier- oder achtdimensionalen Struktur aufweisen.

Tempelaar et al. (2007) berichten ebenfalls eine geringe interne Konsistenz anhand des Cronbach α für das correct statistical reasoning auf der Basis der Subskalen des correct statistical reasoning. Die Autoren formulieren daher ein alternatives siebenfaktorielles Messmodell für das statistical reasoning und geben für dieses einen guten Modellfit an. Jedoch ist hierbei zu berücksichtigen, dass sowohl die correct reasoning als auch die misconceptions Subskalen in das Modell eingehen und daher statistische Artefakte für die Struktur verantwortlich sein könnten (siehe Tabelle 2). Auch wenn das Modell frei von statistischen Artefakten wäre, bliebe allerdings unklar, welche inhaltliche Bedeutung die ermittelte siebenfaktorielle Struktur für das angenommene Konstrukt statistical reasoning

hätte. Schließlich bleibt auch in dieser Untersuchung offen, ob das Instrument möglicherweise eine vier- oder achtdimensionale Struktur im correct statistical reasoning abbildet.

Weitere Aspekte der Validität.

Garfield (2003) berichtet sehr geringe Zusammenhänge des SRA zu akademischen Leistungsmaßen in bzw. nach einer Statistiklehrveranstaltung. Interpretiert man die Zusammenhänge als Hinweise auf konvergente (Kriteriums-) Validität, so liefern sie nur einen eingeschränkten Beleg dafür, dass das SRA zur Erklärung der Leistungen in einer Statistiklehrveranstaltung beiträgt. Erklärungen für die niedrigen Korrelationen könnten zusätzlich in der Art und im Inhalt der verwendeten Leistungsmaße (Aufgaben in den Kurztests und eventuellen Tests sowie die Bewertungskriterien für die Projekte) gesucht werden, allerdings fehlen zu der Erfassung der Leistungsmaße genauere Informationen.

Tempelaar et al. (2006) fanden ebenfalls geringe Zusammenhänge zwischen dem correct statistical reasoning und akademischen Leistungsmaßen zu verschiedenen Zeitpunkten der Statistik- und Mathematiklehrveranstaltungen. Die Autoren setzten das SRA allerdings lediglich zu Beginn des Semesters ein, so dass die niedrigen Korrelationen auch dadurch erklärt werden könnten, dass sich das correct statistical reasoning während des Semesters sowohl im Niveau als auch in der Reihenfolge bei den Studierenden verändert haben könnte. Damit bleibt die Möglichkeit höherer Zusammenhänge zwischen dem SRA (wenn es am Ende der Lehrveranstaltung erhoben würde) und den akademischen Leistungsmaßen und damit von Belegen konvergenter (Kriteriums-) Validität bestehen.

Zunächst kontraintuitiv scheinen die von Tempelaar et al. (2006) berichteten Richtungen der Korrelationen zwischen dem correct statistical reasoning und den Leistungen in den Hausaufgaben in den Statistik- und Mathematikveranstaltungen. Die (sehr schwachen) negativen Korrelationen bedeuten, dass Studierende, die eine geringere Punktezahl im SRA am Anfang des Semesters erzielten, bessere Bewertungen in den Hausaufgaben erreichten. Eine mögliche Erklärung für dieses Muster liefert die Art der Bewertung der Hausaufgaben: Tempelaar et al. erläutern, dass vor allem der Fleiß (d. h. die Mühe, Anstrengung und der Aufwand) bei der Bearbeitung, nicht aber die Korrektheit der Lösungen die Basis für die Beurteilung der Hausaufgaben darstellten. Zudem erhielten Studierende Bonuspunkte für die

Bearbeitung der Hausaufgaben, die für die Gesamtleistung in der Lehrveranstaltung angerechnet wurden. Damit scheint es plausibel zu sein, dass vor allem die Studierenden, die in statistischen Inhalten im Sinne des statistical reasoning unsicher sind (was sich unter anderem an niedrigeren SRA-Werten zu Beginn der Lehrveranstaltung zeigen könnte), mehr in Ausgleichsleistungen investieren als Studierende, die ein höheres Ausmaß des statistical reasoning aufweisen (was sich entsprechend an einem höheren SRA-Wert zu Beginn der Lehrveranstaltung zeigen könnte) und sich dementsprechend in Bezug auf die Abschlussprüfungen sicherer fühlen und daher weniger in Ausgleichsleistungen während des Semesters investieren. Ein solches Ergebnismuster ist vor diesem Hintergrund mit einer konstruktorientierten Interpretation des SRA gut vereinbar, auch wenn die teilweise sehr geringen Effekte der Zusammenhänge zum einen und die explorative Art der Daten zum anderen zu großer Vorsicht bei einer solchen (spekulativen) Deutung mahnen.

Überraschender und zunächst weniger gut mit einer konstruktorientierten Interpretation des SRA vereinbar sind die von Tempelaar et al. (2006) berichteten überwiegend sehr ähnlichen Zusammenhänge des SRA mit den Hochschulleistungen in Statistik und denen in Mathematik mit teilweise engeren Korrelationen zwischen dem SRA und Mathematikleistungen (siehe Tabelle 24 im Anhang). Eine mögliche Erklärung für diese Ergebnisse wäre der Einfluss einer (oder mehreren) Drittvariablen. Eine allgemeine Schlussfolgerungsfähigkeit (z. B. deduktives Schließen oder Intelligenz) könnte sich einen nennenswerten Anteil der Varianz sowohl mit mathematischen als auch mit statistischen akademischen Leistungen sowie mit dem SRA teilen. Es ist vorstellbar, dass statistical *reasoning* sich gleichermaßen oder sogar mehr Varianz mit mathematischen Leistungen (im Vergleich zu statistischen) teilt, da hier Schlussfolgerungsfähigkeiten bzw. -fertigkeiten ebenfalls eine große oder sogar noch größere Bedeutung haben. Auch diese Erklärung bleibt allerdings im Lichte der bisher vorliegenden Studien spekulativ, so dass diese Ergebnisse von Tempelaar et al. sowohl als fehlender Beleg für diskriminante Validität als auch als Hinweis auf konvergente Validität interpretiert werden könnten.

Die bei Garfield (2003) und Tempelaar et al. (2006) berichteten niedrigen Zusammenhänge verdienen noch eine weitere mögliche Erklärung. Da in beiden

Studien Studierende im ersten Studienjahr untersucht wurden, ist nicht auszuschließen, dass die Homogenität der Stichprobe zu einer Varianzeinschränkung geführt hat und in der Folge die tatsächlichen Zusammenhänge unterschätzt werden.

Die Studie von Martin et al. (2017) liefert einige interessante Ergebnisse hinsichtlich der Zusammenhänge des SRA und verschiedenen kognitiven Variablen. Eine der beiden höchsten Korrelationen berichten Martin et al. für den CRT ($r = .56$), ein aus drei Items bestehendes Instrument, das die Tendenz von Personen erfasst, intuitive (falsche) Antworten auf mathematische Textaufgaben zu unterdrücken (Frederick, 2005). Ein Beispielitem aus dem Test lautet: „Ein Schläger und ein Ball kosten zusammen 1,10 Dollar. Der Schläger kostet einen Dollar mehr als der Ball. Wie viel kostet der Ball?“ Die intuitive und von den meisten Probanden gegebene Antwort lautet „10 Cent“, obwohl die richtige Antwort „5 Cent“ schnell und einfach ermittelt werden kann. Das Verfahren wird in der Literatur als Messinstrument für die Dominanz des System 1 im Sinne der Zwei-System-Theorie diskutiert (Toplak, et al., 2011). Vor diesem Hintergrund kann der hohe Zusammenhang zwischen dem korrekten statistical reasoning und der Anzahl der korrekt gelösten Aufgaben im CRT als Hinweis auf konvergente Validität des SRA angesehen werden. In Kombination mit der ebenfalls hohen Korrelation des correct statistical reasoning mit dem WPT ($r = .55$), einem Maß für allgemeine kognitive Fähigkeit (McKelvie, 1989, Hicks, Harrison & Engle, 2015) und der um ca. einen kleinen Effekt (Cohen $q = .14$[18], Cohen, 1988) geringeren Korrelation mit dem korrekten Interpretieren und Berechnen von Prozentangaben in der Form von Wahrscheinlichkeiten bzw. Risiken (Numeracy Scale, Lipkus et al., 2001) sowie einer sehr viel geringeren Korrelation mit dem Wortschatzwissen (Cohen $q = .48$[19], Cohen, 1988) liegen damit auch Hinweise auf diskriminante Validität vor. Im Sinne der Bezeichnung statistical *reasoning* könnte das Instrument tatsächlich eher übergeordnete kognitive Fähigkeiten oder Fertigkeiten als konkrete numerische Fertigkeiten, wie sie in der Numeracy Skala (Lipkus et al.) abgefragt werden, erfassen. Berücksichtigt werden muss allerdings bei dieser Lesart der Ergebnisse, dass die Numeracy Skala nur sehr eng begrenzte numerische bzw. mathematische

[18] berechnet auf https://www.psychometrica.de/effektstaerke.html, Lenhard und Lenhard, 2016
[19] berechnet auf https://www.psychometrica.de/effektstaerke.html, Lenhard und Lenhard, Einordnung des Effekts abweichend von Lenhard & Lenhard als „large“ (nach Cohen, 1988, S. 115)

Fertigkeiten erfasst (ausschließlich Be-, und Umrechnung sowie Interpretation von Risiken und Wahrscheinlichkeiten in Form von Prozentangaben oder absoluten Häufigkeiten, z. B. „bei 20 von 100 Personen“). Es wäre daher wünschenswert, den Zusammenhang des correct statistical reasoning zu breiter gefassten numerischen Fertigkeiten (z. B. mathematischen Fertigkeiten) zu untersuchen.

Insgesamt ergeben die vorliegenden Studien ein uneinheitliches und lückenhaftes Bild über die interne Struktur des SRA im Sinne des Konstrukts statstical reasoning, wie es nach Garfield (2003) verstanden werden kann. Zwar scheint das SRA nach bisherigem Forschungsstand (correct) statistical reasoning nicht als eindimensionales Konstrukt zu erfassen, jedoch bleibt offen, ob die Items des *correct* statistical reasoning gegebenenfalls eine vier- oder achtdimensionale Struktur aufweisen.

Hinsichtlich der konvergenten Validität des SRA zur Messung von statistical reasoning in Bezug auf die akademischen Leistungen in der Statistikveranstaltung liefern die empirischen Ergebnisse wenige Belege. Es ist hierbei zusätzlich zu beachten, dass die Wahl der Leistungen in der Statistikveranstaltung als Kriterium möglicherweise mit Problemen verbunden ist (es kommt hier erheblich auf die Form und die Inhalte der entsprechenden Aufgaben an). Möglicherweise würden für das Konstrukt *statistical* reasoning andere Variablen, wie z. B. die kritische Analyse einer Ergebnisdarstellung in einem Fachartikel oder Empfehlungen für die Stichprobengenerierung bei gegebener Fragestellung, sinnvollere Kriterien darstellen.

Für die konvergente und diskriminante Validität des SRA im Sinne des statistical *reasoning* sprechen die ermittelten Zusammenhänge zwischen dem correct statistical reasoning und mathematischen Leistungen, der kognitiven Reflexion (gemessen über den CRT), den allgemeinen kognitiven Fähigkeiten sowie den leistungs- und fleißbezogenen Maßen (gemessen über Abschlusstests und Hausaufgaben), auch wenn eine solche Interpretation, wie weiter oben erläutert, nur mit großer Vorsicht vorgenommen werden kann.

SRA als Messinstrument des statistical reasoning als Lehrziel.

Wie weiter oben ausgeführt, bestehen die beiden zentralen Anforderungen an ein lehrzielvalides Messinstrument darin, dass zum einen das Lehrziel als Aufgabengrundmenge expliziert wird und zum anderen, dass die Items, die das

Lehrziel abbilden sollen, die Aufgabenmenge – durch Zufallsauswahl der Items aus der Aufgabengrundmenge oder eine vollständige Aufnahme – repräsentieren (z. B. Klauer, 1987). Um beurteilen zu können, ob das SRA statistical reasoning im Sinne eines Lehrziels erfasst, muss daher in erster Linie die Konstruktion des Instruments beleuchtet werden.

Das SRA wurde mit dem Ziel entwickelt, die Effektivität der im Jahr 1989 eingeführten NCTM Standards (NCTM, 1989) und der damit verbundenen Reform der Lehre statistischer Inhalte zu überprüfen (Garfield, 2003, delMas, Garfield, Ooms & Chance, 2007). Garfield (1991, 1998, 2003) berichtet von einer multidisziplinär zusammengesetzten Expertengruppe, die hierfür relevante Typen des statistical reasoning sowie relevante Arten von Fehlverständnissen identifizierte. Damit kann das Vorliegen von relevanten Domänen als gegeben angesehen werden.

Die Definition der Domänen (der Typen des correct reasoning und der Fehlverständnisse) als auch der Subdomänen (der Subskalen des SRA) besteht aus Beschreibungen der (Lehr-) Ziele dessen, was Lernende nach Instruktion im Fach Statistik beherrschen können sollen. Die Beschreibungen sind jedoch nicht konkret genug, um eine direkte Ableitung der Aufgabengrundmenge zu erlauben. Beispielsweise lässt die Beschreibung „Wissen darüber, welche Maße am besten unter verschiedenen Bedingungen geeignet sind und warum die einzelnen Maße einen Datensatz repräsentieren oder nicht“ als Teil der Domäne „Beurteilen statistischer Kennzahlen“ offen, welche Aufgaben eine Aussage über die Leistung in diesem Bereich treffen würden. Auch die Beschreibung der zugehörigen Subdomäne im SRA, die Subskala „Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird“ ist recht allgemein formuliert und lässt nicht unmittelbar erkennen, welche Grundmenge von Items diesen Inhalt abbilden würde. Daher scheint das Lehrziel im Rahmen der SRA-Konstruktion nicht als Aufgabengrundmenge expliziert worden zu sein. Hierbei muss allerdings eingeschränkt werden, dass das konkrete Vorgehen bei der Konstruktion der Items (d. h. welche Technologie oder welche Erzeugungsvorschriften bei der Konstruktion der Items angewandt wurden) durch die Literatur zum SRA nicht vollends deutlich wird. Die Beschreibungen können jedoch zumindest als *behavioral objectives* (wie sie sich in verschiedenen Modulbeschreibungen auch unter den Bezeichnungen „Lernergebnisse [learning outcomes]“ oder „zu erwerbende Kompetenzen“ finden)

betrachtet werden (Hambleton et al., 1978). Alle entwickelten Items wurden zudem Experten vorgegeben, die die Aufgaben im Hinblick auf Aspekte der Inhaltsvalidität (Zugehörigkeit des Items zu der intendierten Domäne oder dem behavioral objective) einschätzten. Auch die Entwicklung von Distraktoren (d. h. falschen Antwortalternativen) erfolgte anhand empirischer Ergebnisse. Die so erstellten Items fanden vermutlich alle Eingang in das finale Messinstrument (dies geht aus der Darstellung der Autoren nicht eindeutig hervor).

Trotz der dadurch erfüllten zweiten Anforderung an ein inhaltsvalides lehrzielorientiertes Verfahren, der repräsentativen Itemstichprobe in der Form der vollständigen Aufnahme der Aufgabengrundmenge, ist die Inhaltsvalidität oder Inhaltsrepräsentativität beim SRA nur in eingeschränktem Maß gegeben. Insgesamt ist damit nicht sichergestellt, dass die Leistung von Lernenden in Bezug auf eine wohldefinierte Verhaltensdomäne (wie z. B. das Maß erfolgreicher Bearbeitung von vollständig definierten statistical reasoning Aufgaben) durch das SRA Verfahren beschrieben wird (siehe hierzu z. B. Popham, 1974).

Garfield (2003) berichtet nur geringe Zusammenhänge zwischen dem SRA und verschiedenen Leistungsmaßen der statistischen Kompetenz in den Statistiklehrveranstaltungen. Wenn man annimmt, dass das SRA, die Leistung im Projekt sowie die Leistungen in den Kurztests das gleiche Lehrziel abbilden, wären höhere Korrelationen zwischen den Maßen zu erwarten. Die Ergebnisse liefern daher nur begrenzt Belege für eine konstruktbezogene Validität des SRA als Messinstrument des statistical reasoning im Sinne eines Lehrziels. Wie weiter oben bereits angeführt, ist jedoch nicht auszuschließen, dass Varianzeinschränkung zu den beobachteten niedrigen Zusammenhängen geführt hat. Offen bleibt zudem, ob die Inhalte der SRA Items explizite Bestandteile der Statistiklehrveranstaltungen darstellten. So ist zu erwarten, dass lediglich implizit behandelte Inhalte eine Transferleistung für die Bearbeitung des SRA erfordern würden und damit gegebenenfalls eine Aussage über andere Variablen als den Instruktionserfolg treffen.

Gundlach et al. (2015) zeigen für eine Auswahl von SRA Items, dass Studierende nach einer Instruktion mehr Aufgaben korrekt lösen konnten (4% bzw. 9%). Auch wenn die Effekte der Intervention insgesamt als eher schwach einzuschätzen sind, kann dieses Ergebnis als Hinweis auf vorliegende konvergente

Konstruktvalidität interpretiert werden. Unterstützt wird eine solche Interpretation zudem durch die differenziellen Ergebnisse der unterschiedlichen Instruktionsformate (4% für traditionell und inverted classroom sowie 9% für online), auch wenn eine solche Deutung nur vorsichtig und provisorisch vorgenommen werden sollte, bis Replikationen diese Ergebnisse bestätigen. Da die Autoren in der Studie allerding lediglich SRA Items einsetzten, deren Inhalt sich mit Inhalten der Lehrveranstaltung so sehr wie möglich deckte, bleiben Fragen zu diskriminanter Validität offen. Die Autoren geben zwar an, dass manche der Inhalte der SRA Items nicht in der Lehrveranstaltung behandelt worden sind, berichten hierfür allerdings keine gesonderten Kennzahlen. Ein Vergleich von Items oder Itemgruppen, deren Inhalte Bestandteil vs. nicht (expliziter) Bestandteil der Lehre sind, hinsichtlich der korrekten Lösung nach einer Instruktion würde hierüber Aufschluss geben können.

Olani et al. (2011) nutzten im Rahmen ihrer Studie sieben Items des SRA sowie zusätzlich acht Items aus einem anderen Messinstrument der Statistikkompetenz (SCI, Allen, 2006). Da die Autoren ihre Ergebnisse nicht nach dem Ursprungsmessinstrument differenzieren, ist eine spezifische Aussage über die Items des SRA daher leider nicht möglich. Nimmt man an, dass sich die Items im Hinblick auf die Aufgabengrundmenge nicht unterscheiden, dann zeigt die Studie ein heterogenes Bild über die Konstruktvalidität der 15 Items zur Erfassung eines Lehrziels. Einen Hinweis auf konvergente Validität liefert der beobachtete Anstieg von 11% korrekt gelöster Aufgaben nach der Instruktion. Einschränkungen der konvergenten Konstruktvalidität liegen jedoch in Form der beobachteten Schwierigkeiten der Items vor: acht Items wurden am Ende des Semesters von weniger Studierenden gelöst als zu Beginn des Semesters. Ein solches Ergebnis ist erwartungskonträr, wenn man annimmt, dass die Lehre effektiv gewesen ist. Hierbei ist natürlich zu beachten, dass eine Aussage darüber, inwiefern diese Ergebnisse die SRA Items oder die SCI Items betreffen, nicht gemacht werden kann. Auch Olani et al. (2011) verwendeten für ihre Untersuchung Items, die zu den in der Lehre behandelten Inhalten deckungsgleich waren, daher bleibt auch hier eine Aussage über diskriminante Validität offen.

Zusätzlich zu der Analyse der Veränderungen im Lösungsverhalten von Studierenden vor und nach einer Lehrinstruktion sollten auch die absoluten

Leistungen im SRA betrachtet werden. Die angeführten Studien decken sich hier in den Angaben überwiegend[20]. Garfield (2003) berichtet bei Studierenden am Ende einer Statistikveranstaltung im Schnitt 57% für Studierende aus den USA und 61% für Studierende aus Taiwan im correct reasoning (d. h. vermutlich unter Ausschluss des Items Nr. 7). Die Studie von Gundlach et al. (2015) zeigt für acht SRA Items im Anschluss an eine Lehrveranstaltung 58% (traditionelles Veranstaltungsformat), 60% (online Veranstaltung), 54% (inverted classroom Veranstaltung) gelöste Items. Einen höheren Anteil der 15 SRA- und SCI-Items lösten Studierende in der Untersuchung von Olani et al. (2011) nach einer Lehrintervention mit 71%. Nimmt man die Ergebnisse als Indikator für die Zielerreichung einer Statistiklehrveranstaltung ernst, weisen sie auf deutliche Defizite in der Statistiklehre hin. Geht man dagegen davon aus, dass die Lehre grundsätzlich effektiv ist, müssen die Ergebnisse als Hinweise auf geringe Validität der genutzten Instrumente gewertet werden. Tempelaar et al. (2006) berichten für verschiedene Teilstichproben 54% bis 59% und Tempelaar et al. (2007) 58% korrekt gelöster Items, beide zu Beginn des ersten Semesters. Zieht man hier zum Vergleich (siehe aber Fußnote 20) den Anteil gelöster Aufgaben aus den Untersuchungen von Olani et a. (2011) mit 60% und Grundlach et al. (2015) mit 55%, 50% und 50% korrekt gelöster Aufgaben heran, scheinen die Ergebnisse von Tempelaar et al. (2006, 2007) vor allem denen von Olani et al. (2011) zu ähneln. Dies kann gegebenenfalls auf ähnlichere Strukturen im Bildungssystem zurückgeführt werden, da die Studien von Garfield (2003, US-amerikanische Teilstichprobe) und Gundlach et al. (2015) an US-amerikanischen Universitäten und die Studien von Olani et al. (2011) und Tempelaar et. al. (2006, 2007) an niederländischen Universitäten durchgeführt wurden. Das könnte als weiterer, vorläufiger Hinweis auf Konstruktvalidität interpretiert werden. Da Olani et al. (2011) Studierende der Pädagogik und Tempelaar et al. (2006, 2007), ebenso wie Garfield (2003, Taiwan-Teilstichprobe) Studierende der Wirtschaftswissenschaften untersuchten, scheint eine Erklärung

[20] Zur besseren Vergleichbarkeit wurden die Angaben in den Studien durch die Autorin in Prozentangaben umgerechnet. Hierbei muss beachtet werden, dass durch den Bezug auf Subskalen mit teilweise sehr unterschiedlicher Itemanzahl eine Gewichtung stattfand. Die Angaben aus Studien, die das gesamte Instrument nutzten, können daher unmittelbar miteinander verglichen werden. Angaben aus Studien, die ausgewählte Items nutzten, entsprechen einander nicht direkt und sollten nur als Annäherung verstanden werden.

durch Unterschiede oder Ähnlichkeiten in der Affinität zur Mathematik dagegen eher nicht plausibel.

Schließlich ist es für die Beurteilung des SRA als Messinstrument des statistical reasoning als Lehrziel relevant, die Nähe der Testitems zu den in der Lehre explizit behandelten Inhalten zu berücksichtigen. Es ist zu erwarten, dass bei einer konkreteren Behandlung der im SRA erfassten Inhalte in der Lehre, entsprechend eine höhere Wahrscheinlichkeit für die Lernenden, das betreffende Item zu lösen, einhergeht und damit die Lehrzielvalidität des Instruments steigt. Wenn etwa der Umgang mit Ausreißern bei der Bestimmung der Maße zentraler Tendenz explizit behandelt sowie die genauen Bedingungen für das Entfernen von Ausreißern und des Einsatzes des Mittelwerts vs. der Verwendung von Median oder Modus besprochen werden, werden Studierende die Items 1 und 4 im SRA erwartungsgemäß eher lösen, als wenn lediglich darauf hingewiesen wird (ggf. mit einem rechnerischen Beispiel), dass der Mittelwert sensibel für Ausreißer ist, währen der Median und der Modus dies nicht sind. Wenn sich die Lehre und Items dagegen nicht oder nur geringfügig entsprechen, müssen alternative Erklärungen einer geringen Leistung im lehrzielorientierten Test gesucht werden. Mit anderen Worten, die geringe Leistung kann sich als Folge der Defizite im genutzten Messinstrument oder aber der zu hohen Transferanforderung (oder weiterer Variablen) äußern. Damit kann keine zuverlässige Aussage über die lehrzielorientierte Validität des verwendeten Messinstruments gemacht werden. Da für die Beurteilung der Entsprechung von Lehrinhalten und Items in den aufgeführten Studien jedoch nicht genügend Informationen vorliegen, bleibt dieser Aspekt offen.

Insgesamt muss bei dem SRA als Messinstrument eines Lehrziels aufgrund der Vorgehensweise bei der Konstruktion mit erheblichen Einschränkungen der Inhaltsvalidität gerechnet werden. Es kann keine Aussage darüber getroffen werden, auf welchen universe of content, d. h. welche Aufgabenmenge, sich das Verfahren bezieht. Entsprechend können Leistungen von Studierenden in den Aufgaben des SRA nicht in Bezug zu einer allgemeiner gefassten Domäne des statistical reasoning gesetzt werden.

Wenn man die Items des SRA zur Erfassung der sehr allgemein definierten statistical reasoning Kompetenzen (im Sinne von behavioral objectives) akzeptiert,

zeigen sich für eine Auswahl von Items recht konsistente, wenn auch mehrheitlich schwache Belege der Konstruktvalidität in Form von höheren Werten nach einer Lehrintervention. Einschränkend muss hierbei beachtet werden, dass die Studien lediglich wenige Items des SRA einer Untersuchung unterzogen, so dass Angaben über alle Items des SRA vor und nach einer Lehrintervention sicherlich wünschenswert wären. Weitere Einschränkungen der Konstruktvalidität liegen möglicherweise in erwartungskonträren Veränderungen der Schwierigkeit von SRA- bzw. SCI Items vor und nach einer Lehrintervention vor.

Insbesondere vor dem Hintergrund der unbefriedigenden absoluten Leistungen im SRA nach einer Lehrintervention scheint es daher von hohem Interesse, die Eignung des SRA für die Erfassung von Lehrzielen in Erfahrung zu bringen. Hierfür ist es notwendig, unter Beachtung der Entsprechung von Lehrinhalten und Items diskriminante Aspekte der Konstruktvalidität des Instruments zur Erfassung des statistical reasoning als Lehrziel zu untersuchen.

SRA als Messinstrument des statistical reasoning als kognitiver Prozess.

Die Beurteilung der Eignung des SRA zur Messung des statistical reasoning als kognitiven Prozess kann lediglich skizzenhaft und spekulativ erfolgen. Bisher liegen kaum Studien vor, die einzelne Items (oder Itemformen) hinsichtlich der spezifischen theoriegeleiteten Bedingungen von Antwortmustern untersuchen. Da fünf Items aus der Forschung zu probabilistischem Schließen in leicht modifizierter Form Eingang in das SRA finden (Item 9 bzw. 11 und Item 14 aus Kahneman et al., 1982 bzw. Tversky & Kahneman, 1983; Items 18, 19, 20 aus Lecoutre, 1992), wird für diese hier auf entsprechende Literatur bzw. Abschnitt 2.4.1 verwiesen.

Die Studie von Martin et al. (2017) liefert Hinweise darauf, dass das SRA ähnliche Inhalte erfasst, wie der CRT, ein Verfahren, das die Tendenz, intuitive Urteile zu korrigieren und damit die Aktivierung möglicher System 2-Prozesse zu messen beansprucht. Damit wäre, zumindest vorläufig, denkbar, dass auch das SRA kognitive Prozesse im Sinne der Zwei-System-Theorie teilweise abbildet. Eine solche Interpretation sollte allerdings durch weitere Untersuchungen bestätigt werden.

Im Rahmen einer studentischen, von der Autorin betreuten Projektarbeit (Kunzi, Müller, Pewinsky, Rath & Sawatzky, 2018) wurden mehrere Items des SRA modifiziert und im experimentellen Between-Subjects-Design Personen auf einer

sozialen Plattform vorgegeben (unter den untersuchten Personen waren vor allem Studierende unterschiedlicher Fächer). Die Ergebnisse zu einer der Aufgaben sollen an dieser Stelle kurz berichtet werden. Den Probandinnen und Probanden wurde das Item Nr. 3 des SRA in seiner Originalversion und zwei modifizierten Versionen vorgelegt. Die Originalversion des Items findet sich im Anhang (Tabelle 25), soll jedoch hier noch einmal der besseren Übersicht halber aufgeführt werden:

> Die Mitarbeiter des Meteorologischen Zentrums in Neustadt wollten die Genauigkeit ihrer Wettervorhersage bestimmen. Sie durchsuchten ihre Aufzeichnungen nach den Tagen, an denen die Vorhersage eine 70%ige Regenwahrscheinlichkeit angab. Sie verglichen diese Vorhersagen mit Meldungen, ob es an den jeweiligen Tagen wirklich geregnet hatte oder nicht. Die Vorhersage von 70%iger Regenwahrscheinlichkeit kann als sehr genau angesehen werden, wenn es
> (a) an 95%-100% dieser Tage regnete.
> (b) an 85%-94% dieser Tage regnete.
> (c) an 75%-84% dieser Tage regnete.
> (d) an 65%-74% dieser Tage regnete.
> (d) an 55%-64% dieser Tage regnete.

Von den 29 Personen, die die Aufgabe bearbeiteten, wählten lediglich 11 (38%) die richtige Antwortoption (d).

Eine zweite Gruppe von Personen beantwortete das modifizierte Item:

> In der letzten Vorlesung gab Sarahs Dozentin die Information, dass die Wahrscheinlichkeit, in der Klausur durchzufallen, für eine/n Studierende/n bei 70% liegt. Sarah möchte nun überprüfen, ob die Dozentin vielleicht über- oder untertrieben hat. Zu diesem Zweck fragt sie beim Prüfungsamt die Übersicht der Klausurergebnisse für die letzten drei Semester an. Der Übersicht kann Sarah entnehmen, wie viele Studierende bei den letzten Prüfungen anteilig durchgefallen sind. In welchem Fall hat die Dozentin weder unter- noch untertrieben?
> (a) wenn 95% bis 100% der Studierenden in den letzten drei Semestern durchgefallen sind.
> (b) wenn 85% bis 94% der Studierenden in den letzten drei Semestern

durchgefallen sind.
(c) wenn 75% bis 84% der Studierenden in den letzten drei Semestern durchgefallen sind.
(d) wenn 65% bis 74% der Studierenden in den letzten drei Semestern durchgefallen sind.
(e) wenn 55% bis 64% der Studierenden in den letzten drei Semestern durchgefallen sind.

In dieser Gruppe wählten 33 von 40 Probandinnen und Probanden (83%) die korrekte Option (d).

Eine letzte modifizierte Version bestand darin, dass den Personen vor der Bearbeitung des Originalitems ein zusätzliches Item vorgegeben wurde:

Die Wettervorhersage gibt für morgen eine Regenwahrscheinlichkeit von 70% an. Was bedeutet diese Angabe?
(a) Es wird morgen an 70% des Tages regnen.
(b) Es hat bisher an 7 von 10 Tagen mit ähnlichen Bedingungen, wie morgen geregnet.
(c) Es wird morgen regnen.
(d) Es wird morgen nicht regnen.
(e) Nichts von dem oben Aufgeführten

Von den 38 Personen, die zunächst diese Frage bearbeiteten, wählten 16 (42%) die korrekte Antwortoption im Originalitem Nr. 3 (siehe weiter oben). Die Hälfte der Personen (50%, $n = 11$), die die Frage nach der Bedeutung der Regenwahrscheinlichkeit von 70% richtig beantworteten (Option [b], $n = 21$), wählten im nachfolgenden Originalitem die korrekte Antwortoption (d). Von den 17 Personen, die die vorangestellte Frage falsch beantworteten, wählten lediglich 5 (30%) die richtige Antwortoption im nachfolgenden Originalitem.

Es kann auf Basis der Ergebnisse vermutet werden, dass Defizite in grundlegenderen statistischen Konzepten (Bedeutung von Regenwahrscheinlichkeit) für die Defizite in der Bearbeitung hierarchisch komplexerer Aufgaben (Genauigkeit als Eigenschaft der Regenwahrscheinlichkeit oder die Relation zwischen Genauigkeit und Wahrscheinlichkeitsangaben) verantwortlich sind. Da jedoch die Hälfte der Probandinnen und Probanden, die die Regenwahrscheinlichkeit korrekt

interpretieren, die Genauigkeit der Regenwahrscheinlichkeit nicht korrekt erkennen, spielen weitere Aspekte offenbar eine Rolle. Die Ergebnisse weisen (vorläufig) darauf hin, dass Relevanz- oder Vertrautheitseffekte des Iteminhalts (Regenwahrscheinlichkeit vs. Wahrscheinlichkeit, eine Klausur nicht zu bestehen) einen deutlichen Einfluss auf das Lösungsverhalten von (hier mehrheitlich) Studierenden nehmen. In der Konsequenz muss bei der Interpretation des Itemlösungsverhaltens von Personen das Zusammenspiel zwischen den probabilistischen und den nicht-probabilistischen (z. B. Vertrautheit) Kognitionen beachtet werden. Mit anderen Worten, die Frage „worüber gibt die Antwort der Person auf das Item Auskunft, Wahrscheinlichkeitsverständnis oder Vertrautheit mit dem Kontext?“ kann ohne Weiteres nicht beantwortet werden.

Einschränkungen der Aussagekraft der berichteten empirischen Untersuchung bestehen zunächst in der recht kleinen Fallzahl. Von der Notwendigkeit einer Replikation abgesehen, bleiben auch einige andere Punkte ungeklärt. Zwar bestand die Intention darin, lediglich den Inhalt des Items Nr. 3, nicht aber seine Form bzw. Struktur in seiner modifizierten Version abzuändern. Dies kann hier aber selbstverständlich nicht ohne weiteres angenommen werden (z. B. der Bezug auf drei Semester in den Antwortoptionen könnte zu einer Formatänderung von Anteilshäufigkeiten auf natürliche Häufigkeiten geführt haben, siehe hierzu Gigerenzer und Hoffrage [1995]). Auch kann anhand der Ergebnisse nicht entschieden werden, ob etwa eine höhere Relevanz (Wetter ist Studierenden ggf. unwichtiger als Klausuren) und damit eine Form von höherer Motivation oder aber eine höhere Vertrautheit (Studierende wissen, wie „Durchfallquoten“ berechnet werden, nicht jedoch, wie Regenwahrscheinlichkeit bestimmt wird) und damit eine Form von leichterer kognitiver Zugänglichkeit auch bei geringer Motivation (z. B. laut der *spreading activation theory*, Anderson, 1983) für die bessere Lösung der Aufgabe verantwortlich ist.

Auf der Basis der vorliegenden Studienergebnisse kann insgesamt vermutet werden, dass zumindest ausgewählte Items des SRA bestimmte kognitive Prozesse (z. B. heuristische) erfassen. Jedoch liegen bisher nur wenige Ergebnisse zu den im SRA neu konzipierten Items und den kognitiven Prozessen, die hierbei involviert sind vor. Die Frage danach, welche kognitiven Prozesse durch Items des SRA abgebildet werden, ist daher für viele Items noch offen.

Da die differenzielle Beteiligung bestimmter kognitiver Prozesse jeweils unterschiedliche Konsequenzen für die Lösungswahrscheinlichkeit von statistischen Aufgaben hätte, scheint es von hoher Bedeutung, einen besseren Einblick in das Zusammenspiel dieser Variablen zu gewinnen. Wenn die Vertrautheit mit den Inhalten von Problemen (z. B. das Wissen, wie Regenwahrscheinlichkeit bestimmt wird) eine Voraussetzung für die Beurteilung der Genauigkeit darstellt (z. B. in Form von Modellen), müsste im Rahmen der Hochschullehre weniger an der Lehre der statistischen Konzepte selbst, sondern vielmehr an dem Inhaltswissen von Studierenden angesetzt werden. Das hätte auch Konsequenzen für die Anordnung von statistischen Fächern im Curriculum. So wäre es gegebenenfalls nötig, statistische Module (mindestens teilweise) erst nach den ersten Semestern im Studienverlaufsplan anzuordnen, wenn Studierende bereits grundlegendes Wissen über verschiedene Versuchs- und Variablenanordnungen in fach-inhaltlichen Modulen erworben haben.

Fazit.

Wie aus den obigen Darstellungen hervorgeht, finden sich in einigen Studienergebnissen vorläufige Hinweise auf die Eignung des SRA zur Messung des statistical reasoning als Konstrukt, in begrenztem Maß als Lehrziel und, ebenfalls in eingeschränktem Maß, als kognitiven Prozess. Einige Aspekte zur Reliabilität und Validität des SRA sind derzeit jedoch noch offen und bedürfen einer Klärung, wenn die Eignung des Instruments zur Erhebung des Beherrschungsstands von Statistik einerseits und Effekten der Hochschullehre in Statistik andererseits beurteilt werden soll. Die zentralen offenen Fragen sollen an dieser Stelle noch einmal zusammenfassend aufgeführt werden.

Ein Aspekt, der genauerer Untersuchung bedarf, ist die interne Struktur des Instruments. Versteht man das correct statistical reasoning als Konstrukt im Sinne einer homogenen ein-, vier- oder achtdimensionalen Fähigkeit (z. B. als Persönlichkeitsmerkmal) und setzt man voraus, dass das SRA dieses Konstrukt zu messen beansprucht, so betrifft die Frage nach der internen Struktur sowohl Reliabilitäts- als auch Validitätsgesichtspunkte. Zum einen kann die Übereinstimmung der erwarteten internen Struktur mit der empirisch ermittelten Struktur im Sinne der Konstruktvalidität interpretiert werden. Zum anderen liefert die Güte der erwarteten und empirischen Passung der Teilstrukturen Hinweise auf

die Homogenität oder Konsistenz der als zusammengehörig angenommenen Itemgruppen.

Eine weitere offene Frage betrifft die konvergente und diskriminante Konstruktvalidität des SRA zur Messung des correct statistical reasoning als Lehrziel. Zwar wurde das Instrument nicht den Anforderungen an ein inhaltsvalides lehrzielorientiertes Instrument entsprechend entwickelt, jedoch besteht dennoch die Möglichkeit, dass mit Hilfe des Instruments erwartungsgemäße Zuwächse und Unterschiede in den Leistungen im Rahmen von Statistiklehre differenziert abgebildet werden können. Hierfür ist die Vorgabe aller SRA Items vor und nach einer Statistiklehrveranstaltung notwendig. Von Vorteil für die Analyse diskriminanter Validität wäre hierbei, wenn die Inhalte der Lehrveranstaltung den Inhalt einzelner Items nicht abbilden würden.

Schließlich würden Analysen der Zusammenhänge des SRA mit anderen akademischen Leistungsmaßen in der Lehrveranstaltung, allgemeinen kognitiven Maßen sowie breiteren mathematischen Fertigkeiten und weiteren Variablen, wie Einstellungen zu Statistik, Aufschluss über die konvergente und diskriminante Validität des Instruments geben.

Die Klärung dieser offenen Fragen bildet den empirischen Teil der vorliegenden Arbeit. Da eine zusätzliche detaillierte empirische Auseinandersetzung mit dem statistical reasoning als kognitiven Prozess im Rahmen dieser Arbeit nicht zufriedenstellend gelingen kann, wird dieser Aspekt bei der nachfolgenden empirischen Auseinandersetzung mehrheitlich außer Acht gelassen[21]. In der Diskussion wird jedoch die Rolle kognitiver Prozesse für die Erhebung und Lehre statistischer Kompetenz wieder aufgegriffen und ein Vorschlag unterbreitet, wie kognitionspsychologische Modelle und Ergebnisse Eingang in die Forschung zur Statistiklehre an der Hochschule finden können.

Bevor im nächsten Teil der Arbeit die genannten Aspekte weiter konkretisiert und in empirisch prüfbare Begriffe überführt werden, sollen im nächsten Abschnitt noch einige alternative Möglichkeiten der Erhebung statistischer Kompetenzen skizziert werden.

[21] Stattdessen wird auf die weiter oben dargestellten Ergebnisse der studentischen Projektarbeit unter Betreuung der Autorin (Kunzi et al., 2018) verwiesen.

3.2. Weitere Messinstrumente zur Erfassung statistischer Kompetenz

Die Entwicklung und Veröffentlichung der GAISE Empfehlungen für Statistiklehre an Schulen und Hochschulen zum einen (ASA, 2005, 2016) und eines ersten Messinstruments zur Erfassung von Statistikkompetenzen (SRA, Garfield, 1991, 2003) zum anderen stießen weitere breit gefächerte Auseinandersetzungen mit der Thematik der Statistikkompetenz in der (Hoch-) Schullehre an. Diese Auseinandersetzungen äußerten sich sowohl in der Konzeption von Lehrmaterialien (Change Agents for Teaching and Learning Statistics Projekt, CATALST, Garfield, delMas, Zieffler, 2012, Zieffler & Catalysts for Change, 2019) als auch in der Konstruktion von weiteren Messinstrumenten. Eine (unvollständige) Übersicht der Messinstrumente, die (ausschließlich) statistische Kompetenz zum Gegenstand haben und deren Vorkommenshäufigkeit in wissenschaftlicher Literatur findet sich in Tabelle 3.

Bis auf ein Instrument in Tabelle 3 stammen alle Verfahren von Autoren rund um Garfield (Assessment of Inferential Reasoning in Statistics, AIRS, Park, 2012, Basic Literacy In Statistics, BLIS, Ziegler, 2015, Comprehensive Assessment of Outcomes in a First Statistics Course, CAOS, delMas et al., 2007, Goals and Outcomes Associated with Learning Statistics, GOALS-2, Sabbag & Zieffler, 2015, REasoning and Literacy Instrument, REALI, Sabbag, Garfield & Zieffler, 2018, Reasoning about P-Values and Statistical Significance, RPASS, Lane-Getaz, 2007) und wurden nach ähnlichen Prinzipien konstruiert. Für die Konstruktion aller genannten Instrumente wurden die zu erfassenden Konzepte anhand der Literatur identifiziert, auf dieser Basis verhaltenszielorientierte (objective based) Domänen entwickelt und durch Experteneinschätzungen inhaltsvalidiert. Zu den einzelnen Domänen wurden passende einzelne konkrete Items formuliert oder den anderen Instrumenten bzw. Itemsammlungen (z. B. ARTIST, Garfield et al., 2003) entnommen. Die Items aller Instrumente wurden anhand von Experteneinschätzungen sowie empirischen Pilotierungen in den Zielpopulationen einer Prüfung der Inhaltsvalidität zum Teil in mehrfachen Iterationen unterzogen. Für alle Instrumente erfolgte eine Analyse der internen Konsistenz, für manche Instrumente wurden auch Faktorenanalysen sowie Item-Response-Modell-Prüfungen durchgeführt (z. B. GOALS-2, Sabbag & Zieffler, 2015, REALI, Sabbag et al., 2018).

Tabelle 3

Treffer für Erhebungsinstrumente statistischer Kompetenz in verschiedenen Datenbanken

	PsycINFO und PSYNDEX	Scopus	Web of Science
AIRS	0	0	0
ARTIST	4	20	0
BLIS	1	1	0
CAOS	6	3	0
GOALS-2	2	2	1
REALI	0	1	0
RPASS	2	5	0
SCI	3	93	3
SRA	11	37	4

Anmerkungen. AIRS: Assessment of Inferential Reasoning in Statistics (Park, 2012), ARTIST: Assessment Resource Tools for Improving Statistical Thinking (Sammlung von Items und Skalen zur Erfassung der Beherrschung statistischer Inhalte, Garfield et al., 2003), BLIS: Basic Literacy In Statistics (Ziegler, 2015), CAOS: Comprehensive Assessment of Outcomes in a First Statistics Course (delMas et al., 2007), GOALS-2: Goals and Outcomes Associated with Learning Statistics (Sabbag & Zieffler, 2015), REALI: REasoning and Literacy Instrument (Sabbag et al., 2018), RPASS: Reasoning about P-Values and Statistical Significance (Lane-Getaz, 2007, 2013), SCI: Statistics Concept Inventory (Allen, 2006), SRA: Statistical Reasoning Assessment (Garfield, 2003). Die Suche erfolgte mit: „[Name Messinstrument]“ als *basic search* und den Datenbanken PsycINFO und PSYNDEX mit dem EBSCOhost, in *all fields* in der Scopus Datenbank und in *all databases* in *topic* im Web of Science (Datum: 30.08.2019).

Fünf der Instrumente (BLIS, Ziegler, 2015, CAOS, delMas et al., 2007, GOALS-2, Sabbag & Zieffler, 2015, REALI, Sabbag et al., 2018, Statistics Concept Inventory, SCI Allen, 2006) haben den Anspruch, alle Inhalte, die innerhalb einer ersten Studienveranstaltung in Statistik gelehrt werden, zu erfassen und damit eine Auskunft über die Lehrzielerreichung der Veranstaltung insgesamt zu erteilen. Alle vier Instrumente nutzen statistisches Fachvokabular (z. B. „Stichprobenkennwerteverteilung“) und eignen sich daher in erster Linie für eine

Messung von statistischen Kompetenzen im Anschluss an eine Lehrveranstaltung. BLIS erfasst laut Autorin (Ziegler, 2015) mit 37 Items grundlegende Fertigkeiten im Umgang mit statistischen Konzepten (im Sinne der eingeschränkten Definition von statistical literacy durch Garfield & Ben-Zvi, 2004, siehe Abschnitt 2.6.2) für die Inhalte „Datenproduktion“, „Grafiken“, „deskriptive Statistiken“, „empirische Stichprobenverteilungen“, „Konfidenzintervalle“, „Wahrscheinlichkeitsverteilungen“, „Hypothesentests“, „Interpretationsgrenzen“, „Regression“ und „Korrelation“. Das CAOS-Verfahren hat den Anspruch, das Konzept der Variabilität mit 40 Items zu erfassen (delMas et al., 2007). Das GOALS-2 Instrument (Sabbag & Zieffler, 2015) hat zum Ziel, Effekte der Lehre mit Nutzung bzw. Schwerpunktsetzung von Simulationen und den Themengebieten Versuchsplanung, Variabilität, Stichprobenziehung, Konfidenzintervalle und p-Werte, statistische Inferenz sowie Modellierung und Simulation mit 27 Items zu erfassen. Das REALI (40 Items) erfasst die Bereiche „Datenrepräsentationen“, „Maße zentraler Tendenz“, „Maße der Streuung“, „Versuchsplanung,“ „Konfidenzintervalle“, „Hypothesentesten“ und „p-Werte“, „Wahrscheinlichkeit“ und „bivariate Daten“ (Sabbag et al., 2018). Das Verfahren von Allen (2006) beruht auf den gleichen Konstruktionsprinzipen, wie den oben beschriebenen und setzt sich zum Ziel korrektes Verständnis und verschiedene Missverständnisse (ähnlich dem SRA, Garfield, 2003) von 16 verschiedenen statistischen Konzepten, die nicht einzeln aufgeführt werden sollen, anhand von 25 Items zu erfassen. Die zwei übrigen Verfahren aus Tabelle 3, AIRS (34 Items, Park, 2012) und RPASS (15 Items, Lane-Getaz, 2007) sind auf die Erhebung statistischer Inferenz ausgerichtet.

Zu den Einschränkungen, die bereits für das SRA Instrument beschrieben wurden, kommt bei den meisten hier aufgeführten Verfahren hinzu, dass nur sehr wenige empirische Studien vorliegen, wie aus Tabelle 3 hervorgeht. Des Weiteren ist kritisch anzumerken, dass bei der Entwicklung aller Verfahren implizit oder explizit Bezug auf Lehrziele genommen wird, jedoch keine lehrzielorientierte Konstruktion der entsprechenden Instrumente erfolgt. In diesem Zusammenhang ist auch die Itemauswahl (vor allem Elimination) anhand empirischer Kriterien (z. B. anhand von Schwierigkeiten oder Diskriminationsindizes) zu kritisieren. Die Inhaltsvalidität eines lehrzielorientierten Tests wird eingeschränkt, wenn Items anhand von relativen Referenzstatistiken (d. h. Kennwerten, die Mittelwerte und

Streuungen als Referenz heranziehen) eliminiert werden: ein sehr leichtes Item indiziert in einem lehrzielvaliden Verfahren, dass sehr viele Lernende den Aspekt des Lehrziels erreicht haben, der in dem Item repräsentiert wird. Analoges gilt für die Diskrimination: unterscheidet ein Item nicht zwischen „guten“ und „weniger guten“ Lernenden, spricht das bei einem validen lehrzielorientierten Verfahren dafür, dass die Lernenden hinsichtlich der Lehrzielerreichung homogen sind. Nur wenn in bewusst heterogen zusammengestellten Gruppen (z. B. Novizen vs. Experten bzw. vor und nach der Instruktion) erwartungskonträre Muster in Schwierigkeiten und Diskriminationskennwerten auftreten, kann das als Indiz für mangelhafte Items gedeutet werden (auch dies trifft natürlich unter der Bedingung zu, wenn es sich um ein lehrzielorientiertes Verfahren handelt).

Selbstverständlich wurden nur einige wenige Aspekte der in Tabelle 3 aufgeführten Instrumente aufgeführt und es wird hierbei kein Anspruch auf eine vollständige Beleuchtung und Bewertung der Verfahren erhoben. Die ausschnitthafte Betrachtung der Instrumente zeigt jedoch einige Gemeinsamkeiten aller in der vorliegenden Arbeit behandelten Messverfahren statistischer Kompetenz. Diese Gemeinsamkeiten bestehen zum einen in einer eher unklaren theoretischen Basis, zum zweiten in einem (in der Folge) nicht hinreichend deutlich definierten Erhebungsziel (z. B. Konstrukt im Sinne eines Persönlichkeitsmerkmals vs. Lehrziel), zum dritten in den resultierenden methodischen Problemen der Konstruktion der Items und des Tests und schließlich zum vierten bei dem empirischen Vorgehen zur Untersuchung der Instrumente.

4. Fragestellungen und Hypothesen

Da die Beherrschung statistischer Inhalte eine hohe Bedeutung für effektives forscherisches und professionelles Handeln in allen empirisch begründeten Tätigkeitsbereichen hat, besteht eines der Ziele der Hochschullehre in den entsprechenden Fächern, eine solche Beherrschung sicherzustellen. Hieraus ergeben sich zwei Fragen:

1) Beherrschen Studierende statistische Inhalte?
2) Verbessert die Hochschullehre die Beherrschung statistischer Inhalte?

Um diese Fragen klären zu können, muss die Beherrschung statistischer Inhalte in konkrete theoretische und operationale Begriffe überführt und damit einer Messung zugänglich gemacht werden. Ein Vorschlag für die theoretisch-inhaltliche Konzipierung der Beherrschung statistischer Inhalte liegt in Form des statistical reasoning vor. Dieses Konzept wird mit dem Messverfahren SRA operationalisiert.

Hieraus ergibt sich die Frage, inwiefern das SRA geeignet ist, um statistical reasoning zu erfassen und in der Konsequenz inwiefern das SRA genutzt werden kann, um die Beherrschung statistischer Inhalte und ihre Veränderung durch Lehre bei Studierenden zu erheben. Aus bisheriger Forschung zu dem Messinstrument liegen einige vorläufige Belege über die Reliabilität und Validität des SRA vor. Einige Fragen bleiben jedoch offen und sollen in der vorliegenden Arbeit geklärt werden:

I. Bildet das Instrument die angenommene Struktur des statistical reasoning ab? Diese Frage betrifft die Validität und Reliabilität des SRA zur Messung des statistical reasoning als Konstrukt (z. B. im Sinne eines Persönlichkeitsmerkmals).
II. Zeigt das Instrument Sensitivität für Lehrinterventionen? Diese Frage betrifft die Validität des SRA zur Messung des statistical reasoning als Lehrziel.
III. Weist das Instrument plausible Zusammenhänge zu anderen Merkmalen auf? Diese Frage betrifft die Validität des SRA zur Messung des statistical reasoning als Konstrukt, Lehrziel und als kognitiven Prozess.

4.1. Bildet das SRA die Struktur des statistical reasoning ab?

Diese erste Fragestellung lässt sich empirisch durch die Analyse der Items und der Beziehungen zwischen den Items beantworten. Klassischerweise steht hier die Analyse der internen Struktur der Messwerte auf den Items des Instruments im Fokus. Folgt man der Argumentation der Autoren (Konold, 1990, zitiert nach Garfield, 2003, Garfield, 1991, 1998, 2003) unter Verwendung des *principle of charity*[22] (Feldman, 1998), so kann das Konstrukt statistical reasoning auf mehrere Weisen aus statistischem Gesichtspunkt verstanden werden (siehe Abbildung 3 auf S. 94).

Erstens wäre eine eindimensionale Struktur der Daten im Sinne der Autoren. Wenn alle Items, die statistical reasoning erfassen sollen, positiv miteinander korrelieren und/oder ein Faktor (vs. mehr als ein Faktor) die Korrelationsstruktur am besten reproduziert, so kann das als Hinweis angesehen werden, dass das Konstrukt statistical reasoning dem Instrument tatsächlich zugrunde liegt.

Zweitens wäre auch eine vierdimensionale Struktur der Daten mit den Darstellungen der Autoren vereinbar. Statistical reasoning würde dabei verstanden als das Schließen über verschiedene statistische Aspekte von Daten – Typen des statistical reasoning. Im Einzelnen sind dies 1) das Beurteilen statistischer Kennzahlen, 2) das Beurteilen (der Güte) von Stichproben, 3) das Schließen unter Unsicherheit und 4) das Beurteilen von Zusammenhängen. Ein Beleg für diese Art der Struktur wäre gegeben bei höheren (positiven) Korrelationen zwischen Items, die einem Typ des statistical reasoning zugeordnet werden als zwischen Items, die erwartungsgemäß unterschiedlichen Typen des statistical reasoning zugehören. Ergänzend oder alternativ sollten in diesem Fall vier Faktoren die Zusammenhangsstruktur der Daten am besten reproduzieren können. Dabei muss die Art und Höhe der Kovariation zwischen den Faktoren nicht explizit berücksichtigt werden: sowohl positive als auch negative Zusammenhänge beliebiger Höhe (d. h. auch statistische Unabhängigkeit) können als vereinbar mit der Struktur des SRA, wie sie von Garfield (2003) beschrieben wird, angesehen werden.

[22] Das Prinzip der wohlwollenden Interpretation bezieht sich auf eine Interpretation der Aussagen anderer, die die Wahrheit oder Rationalität der betreffenden Aussage maximiert.

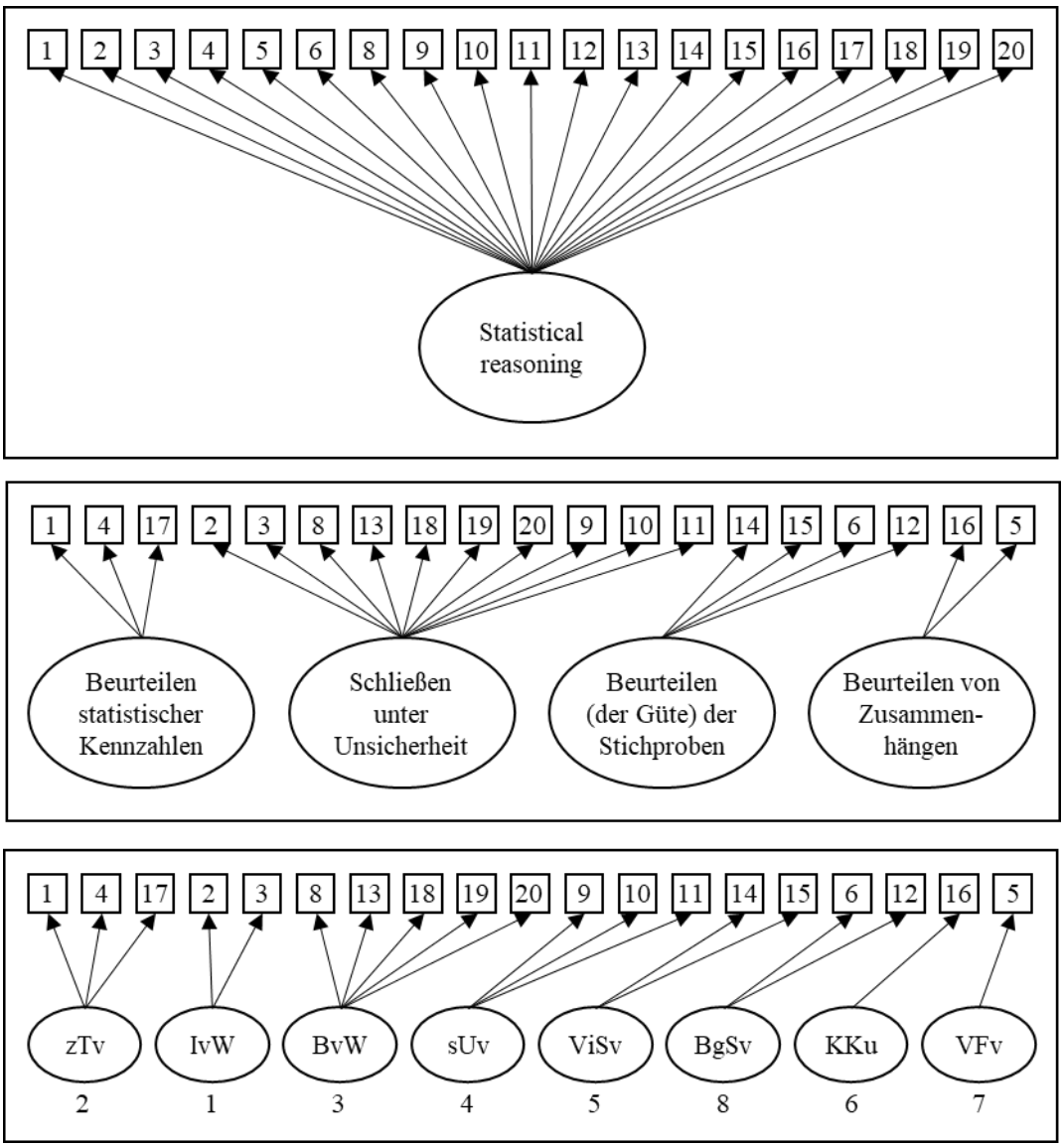


Abbildung 3. Darstellung der möglichen Struktur des SRA. Die obere Darstellung zeigt die ein-Faktor-Struktur, die mittlere Darstellung zeigt die vier-Faktor Struktur mit den Typen des statistical reasoning als Faktoren, die untere Darstellung zeigt die acht-Faktor Struktur mit den Subskalen, die Fertigkeiten (skills) des statistical reasoning darstellen, als Faktoren mit zTv = Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird, IvW = Wahrscheinlichkeiten korrekt interpretieren, BvW = Wahrscheinlichkeiten korrekt berechnen, sUv = (stochastische) Unabhängigkeit verstehen, ViSv = Variablilität innerhalb von Stichproben verstehen, BgSv = Bedeutung großer Stichproben verstehen, KKu = Zwischen Korrelation und Kausalität unterscheiden und VFv = Vierfeldertafeln korrekt interpretieren, zusätzlich wird die Nummerierung der Subskalen aus dem Original (Garfield, 2003) angezeigt.

Drittens entspräche auch eine achtdimensionale Struktur den Erwartungen an das Instrument. Hierbei lässt sich das statistical reasoning definieren als acht einzelne Fertigkeiten im Umgang mit Statistik, repräsentiert durch die acht Subskalen des SRA:

1) Wahrscheinlichkeiten korrekt interpretieren
2) Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird
3) Wahrscheinlichkeiten korrekt berechnen
4) (stochastische) Unabhängigkeit verstehen
5) Variabilität innerhalb von Stichproben verstehen
6) Zwischen Korrelation und Kausalität unterscheiden
7) Vierfeldertafeln korrekt interpretieren
8) Bedeutung großer Stichproben verstehen

Empirisch sollte sich diese Struktur in höheren (positiven) Korrelationen zwischen Items derselben Subskala als zwischen Items verschiedener Subskalen widerspiegeln und/oder in der besten Wiedergabe der Iteminterkorrelationsmatrix durch acht Faktoren (gegenüber Lösungen mit weniger oder mehr als acht Faktoren). Auch hier werden – aus Gründen der wohlwollenden Interpretation – keine zusätzlichen Annahmen über die Struktur der Zusammenhänge zwischen den Faktoren getroffen.

Versteht man die Darstellungen der Autoren dahingehend, dass die durch das SRA erfassten Merkmale Fähigkeiten im Sinne eines eigenschaftsorientierten Ansatzes darstellen, so müsste sich eine der oben dargestellten Strukturen sowohl bei Personen ohne als auch mit expliziter Unterweisung in statistischen Inhalten zeigen. Wenn die Autoren dagegen davon ausgehen, dass sich statistical reasoning als strukturierte Fähigkeit oder Fertigkeit erst durch entsprechendes Lernen bildet, sollten sich die angenommenen Strukturen in den Daten entsprechend nach einer expliziten Unterweisung zeigen.

Damit lassen sich für die erste Fragestellung folgende Hypothesen ableiten:

1a) Explorative Faktoranalysen sollten eine ein-, vier- oder acht-faktorielle Lösung nahelegen.

1b) Die Ladungen der Items auf die Faktoren sollten überwiegend der erwarteten Ladungsstruktur entsprechen (Abbildung 3, S. 94).

1c) In konfirmatorischen Faktorenanalysen sollten die ein-, vier- oder achtfaktoriellen Modelle mit entsprechender Itemzuordnung (Abbildung 3, S. 94) bessere Güte aufweisen, als ein konkurrierendes Modell.

Daneben sollten deskriptive Item- und Skaleneigenschaften (Trennschärfen, Iteminterkorrelationen sowie die interne Konsistenz) mit einer ein-, vier- oder achtdimensionalen Datenstruktur übereinstimmen.

4.2. Zeigt das SRA Sensitivität für Lehrinterventionen?

Zunächst soll kurz begründet werden, weshalb die Fragestellung nach der Sensitivität des Instruments für Interventionen gesondert aufgeführt wird. Im Sinne der Konstruktvalidierung eines Persönlichkeitsmerkmals durch nomologische Netzwerke (Cronbach & Meehl, 1955) ließe sich die Sensitivität als ein Aspekt der Konstruktvalidität verstehen und müsste demnach als Teilfragestellung dieser untersucht werden. Auch wenn das Instrument also eine Veränderung in den Messwerten nach einer Lehrintervention in erwarteter Richtung abbilden würde, müsste bei erwartungskonträrer Struktur der Messwerte die Annahme des hypothetischen Konstrukts zumindest in Frage gestellt werden. Es wäre zwar eine Veränderung beobachtet worden, aber aus Perspektive des Konstrukts könnte keine (oder eine stark eingeschränkte) Aussage darüber erfolgen, was sich verändert hat (z. B. ein anderes Persönlichkeitsmerkmal oder gänzlich andere Variablen). Das Vorliegen einer nicht-beliebigen Struktur in den Messwerten eines Instruments erfordert wiederum das Vorliegen von Zusammenhängen zwischen den Messwerten der einzelnen Items. Diese Anforderung, d. h. das Vorliegen von (erwartungskonformen) Zusammenhängen und einer (erwartungskonformen) Struktur, muss also von konstruktorientierten Tests, d. h. Tests, die beanspruchen, eine konsistente Fähigkeit, Einstellung, Persönlichkeitseigenschaft usw. zu messen, erfüllt werden. Das gilt jedoch nicht in demselben Ausmaß für kriteriumsorientierte Tests. Für Inventare, die z. B. explizit Lehrziele überprüfen sollen, lässt sich eine solche Anforderung nicht konsequent begründen. Zum einen kann in einem lehrzielvaliden Test zumindest theoretisch der Fall eintreten, dass alle untersuchten Personen identische Messwerte aufweisen, etwa wenn alle Aufgaben von allen Untersuchungsteilnehmern gelöst werden. In diesem Fall ist es nicht möglich, die faktorielle Struktur der Daten zu bestimmen. Zum anderen können sich Lehrziele

auf einzelne, nicht notwendigerweise systematisch zusammenhängende Lerninhalte beziehen. Das kann beispielsweise dann der Fall sein, wenn jeweils nur eine Aufgabe für eine bestimmte Kompetenz in den Test aufgenommen wird. In der Folge müssten die Messwerte keine bestimmte Struktur in ihren Zusammenhängen aufweisen. Damit kann ein valider lehrzielorientierter bzw. kriteriumsorientierter Test eine fehlende oder mehrdeutige, nicht interpretierbare faktorielle Struktur zeigen, während ein valider konstruktorientierter Test eine eindeutige Messwertstruktur im Sinne des angenommenen Konstrukts aufweisen muss.

Ein spezifisch für lehrzielvalide Tests sinnvolles eigenständiges Kriterium scheint dagegen die Forderung nach einer erwartungsgemäßen Veränderung in den Messwerten durch Lehre in den gemessenen Inhalten zu sein. Diese Forderung lässt sich zum einen (definitions-) analytisch begründen. Wenn die Inhalte, die in dem Test erfasst werden, (adäquat) gelehrt wurden, dann kann vernünftigerweise erwartet werden, dass die Testergebnisse höher ausfallen im Vergleich zu dem Fall, in dem diese Inhalte nicht gelehrt wurden. In seiner stärksten Form kann der dann-Teilsatz auch die Erwartung enthalten, dass alle Personen alle Aufgaben zu den gelehrten Inhalten lösen. Realistischerweise wird diese starke Formulierung hier jedoch zugunsten der schwächeren „mehr“-Relation (mehr Personen lösen Aufgaben, deren Inhalte behandelt worden sind als Aufgaben, deren Inhalte nicht behandelt worden sind) aufgegeben. Es darf jedoch nicht vergessen werden, dass letztendlich die starke Form das Ziel der Lehrbestrebungen sein muss, insbesondere in den methodischen Fächern des wissenschaftlichen Studiums. Schließlich werden auch andere Kriterien, wie die Fahrzulassung, nicht lediglich an einer größer-Relation festgemacht. Die zweite Begründung für die Forderung einer erwartungsgemäßen Änderung der Messung bei Vorliegen vs. Nichtvorliegen von Lehre ist normativer Art. In Modulbeschreibungen werden zu erreichende Ziele der Lehrbemühungen, d. h. Lehrziele formuliert, die definitionsgemäß als grundsätzlich erreichbar angenommen werden müssen.

Bezogen auf das SRA lässt sich dieses Kriterium durch drei Strategien überprüfen. Zum einen sollten Personen, die in den durch das Instrument erfassten Inhalten explizit instruiert wurden, höhere Werte erzielen, als Personen, die nicht in den entsprechenden Inhalten instruiert worden sind. Zum anderen sollten Personen nach der expliziten Unterweisung in den durch das Instrument erfassten Inhalten

höhere Werte aufweisen als vor der Unterweisung (Tabelle 4). Schließlich sollten Aufgaben, deren Inhalte in der Lehre behandelt wurden, nach der Lehrintervention insgesamt leichter, d. h. von mehr Personen, gelöst werden als vor der Lehrintervention.

Es kann außerdem danach gefragt werden, ob eine intensivere Unterweisung in den Inhalten zu einer größeren Veränderung in den korrekt gelösten Aufgaben des SRA führen sollte. Diese letzte Annahme ist jedoch an eine weitere, nicht ohne Weiteres haltbare Annahme gebunden: nur wenn angenommen werden kann, dass der Zuwachs in der statistischen Kompetenz in einer spezifischen Beziehung zur Intensität und/oder Umfang und/oder Dauer der Lehre steht, kann von Unterschieden im Ausmaß der Veränderung ausgegangen werden. Diese spezifische Beziehung kann theoretisch linearer Natur sein. Wenn die statistische Kompetenz (praktisch bedeutsam) linear mit Intensität und/oder Umfang und/oder Dauer der Lehre ansteigt, ist damit zu rechnen, dass Personen mit einer intensiveren Unterweisung eine größere Veränderung in den Messwerten aufweisen werden, als Personen mit einer weniger intensiven Unterweisung. Die Beziehung zwischen dem Ausmaß der statistischen Kompetenz und dem Ausmaß/der Intensität der Lehre kann theoretisch jedoch auch nicht-linearer Art sein. Um die Differenz einer Veränderung innerhalb eines bestimmten Zeitrahmens in den Messwerten in diesem Fall entdecken zu können, muss dieser Zeitrahmen entsprechend passend gewählt werden: eine (positive) quadratische Beziehung etwa ist im Fall niedriger Prädiktorwerte nicht von einem fehlenden Anstieg im Kriterium zu unterscheiden. Die Art der Beziehung zwischen dem Ausmaß, der Dauer oder der Intensität der Unterweisung und der statistischen Kompetenz ist, soweit der Autorin bekannt, bis heute allerdings kaum Gegenstand empirischer Forschung gewesen und so kann nicht ausgeschlossen werden, dass statistische Kompetenz in ihrem Zuwachs über die Zeit z. B. Einsichtsproblemen ähnelt (siehe Knoblich und Öllinger [2006] für einen Überblick), die über die Zeit einen klar nicht-linearen Verlauf aufweisen. Selbstverständlich gelten die zuletzt dargestellten Einschränkungen auch für die Untersuchung der längsschnittlichen Veränderung und des querschnittlichen Vergleichs der statistischen Kompetenz generell.

Tabelle 4

SRA Items und erwartete Unterschiede in Schwierigkeiten und zentraler Tendenz für Personengruppen mit unterschiedlichen Lehrplänen (auf nächster Seite fortgesetzt).

Item	Ziel der Aufgabe	im Lehrplan?	erwartete Unterschiede
1	Mittelwert ohne Ausreißer bestimmen	nicht explizit	1 = 2 = 3 = 4
2	Wahrscheinlichkeit in absolute Häufigkeit übersetzen	Stichproben 3 und 4	1, 2 < 3, 4
3	Wahrscheinlichkeit in relative Häufigkeiten in Prozent übersetzen	Stichproben 3 und 4	1, 2 < 3, 4
4	Zentrale Tendenz mit Ausreißer bestimmen	nicht explizit	1 = 2 = 3 = 4
5	Absolute Häufigkeiten in relative Häufigkeiten übersetzen	Stichproben 1, 3 und 4	1, 3, 4 > 2
6	Erkennen, dass $N = 30$ zu klein ist für Schlussfolgerungen	nicht explizit	1 = 2 = 3 = 4
7	Einschränkungen in Stichprobenauswahl und durch Selbstauskunft erkennen	nicht explizit	1 = 2 = 3 = 4
8	Gleichheit in relativen Häufigkeiten als Gleichheit in Wahrscheinlichkeit erkennen	Stichproben 3 und 4	1, 2 < 3, 4
9	Erkennen, dass die Wahrscheinlichkeiten der Reihenfolgen unabhängiger Ereignisse gleich sind	Stichproben 3 und 4	1, 2 < 3, 4
10	Erkennen, dass die relative Häufigkeit für jede Reihenfolge unabhängiger Ereignisse gleich wäre und dass die Wahrscheinlichkeit für jede Reihenfolge gleich ist	Stichproben 3 und 4	1, 2 < 3, 4
11	Erkennen, dass die Wahrscheinlichkeiten der Reihenfolgen unabhängiger Ereignisse gleich sind	Stichproben 3 und 4	1, 2 < 3, 4
12	Statistische Information statt anekdotischer Evidenz als Entscheidungsgrundlage nutzen	nicht explizit	1 = 2 = 3 = 4
13	Schwarze Seite als Ergebnis für alle sechs Würfe als am wahrscheinlichsten erkennen	Stichproben 3 und 4	1, 2 < 3, 4

Tabelle 4 *(fortgesetzt)*

Item	Ziel der Aufgabe	im Lehrplan?	erwartete Unterschiede
14	Erkennen, dass ein Anteil an neugeborenen Mädchen von 80% im Krankenhaus mit geringerer Anzahl von Geburten wahrscheinlicher ist	Stichprobe 4 (implizit)	1 = 2 = 3 <= 4
15	Erkennen, dass es keinen Unterschied zwischen den Gruppen gibt, wenn die Streuung größer ist, als die mittlere Differenz	Stichproben 1 und 4	1, 4 > 2, 3
16	Erkennen, dass die Aussage, Fernsehen schade der schulischen Leistung nicht aus einer Zusammenhangsangabe hervorgeht	Stichproben 1, 3 und 4	1, 3, 4 > 2
17	Summe der Kinder insgesamt aus der Durchschnittsangabe und dem N ableiten; Mittelwert nicht mit Modus oder Median verwechseln	nicht explizit, Stichprobe 2 (implizit)	1 =< 2 >= 3 = 4
18	Ergebnis von zwei verschiedenen Augenanzahlen gegenüber dem Ergebnis mit zwei gleichen Augenzahlen als wahrscheinlicher erkennen	Stichproben 3 und 4	1, 2 < 3, 4
19	Ergebnis von drei verschiedenen Augenanzahlen gegenüber Ergebnissen mit zwei oder mehr gleichen Augenzahlen als wahrscheinlicher erkennen	Stichproben 3 und 4	1, 2 < 3, 4
20	Ergebnis von drei gleichen Augenanzahlen gegenüber Ergebnissen mit zwei oder mehr ungleichen Augenzahlen als unwahrscheinlicher erkennen	Stichproben 3 und 4	1, 2 < 3, 4

Anmerkungen. Stichproben mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc.

Nichtsdestotrotz greifen hier zusätzlich die weiter oben dargestellten Begründungen, so dass die folgenden Hypothesen im vorliegenden Kontext als sinnvoll betrachtet werden (siehe auch Tabelle 4, S. 99):

2a) Personen, die an einer Lehrintervention teilgenommen haben, sollten durchschnittlich mehr Aufgaben zu den Inhalten, die Gegenstand der Lehrintervention waren, lösen als Personen, die an einer solchen Lehrintervention nicht teilgenommen haben.

2b) Personen sollten im Schnitt nach der Lehrintervention mehr Aufgaben zu den Inhalten, die Gegenstand der Lehrintervention waren, lösen als vor der Lehrintervention.

2c) Aufgaben zu Inhalten, die Gegenstand der Lehrinterventionen waren, sollten sich in ihren Schwierigkeitsunterschieden vor und nach der Lehrintervention stärker unterscheiden als Aufgaben, die nicht Gegenstand der Lehrintervention waren. Dabei sollten die Aufgaben zu Inhalten, die in der Lehre behandelt wurden, nach der Lehrintervention leichter gelöst werden als vor der Intervention.

Über das Ausmaß der Effekte in beiden Hypothesen sollen keine einschränkenden Annahmen getroffen werden.

4.3. Lassen sich plausible Zusammenhänge beobachten?

Die dritte Fragestellung der Arbeit dient der Analyse konvergenter und diskriminanter Konstruktvalidität des SRA als Messinstrument des statistical reasoning im Sinne eines Persönlichkeitsmerkmals, eines Lehrziels einer Statistikveranstaltung und eines kognitiven Prozesses.

Wenn statistical reasoning als Persönlichkeitsmerkmal durch das SRA erfasst wird, sollten positive Zusammenhänge zwischen dem SRA und mathematischen Fertigkeiten sowie allgemeinen kognitiven Maßen beobachtet werden. Die allgemeinen kognitiven Maße können hierbei zum einen (als grundsätzlich trainierbar angenommene) Fähigkeiten des deduktiven Schließens und (schwer trainierbare) Intelligenzkomponenten, wie sie in figuralen Reihen-Fortsetzungsaufgaben erfasst werden, darstellen. Unterschiede in den Höhen der Zusammenhänge können Hinweise über die Beschaffenheit des statistical reasoning geben. Hängt die Punktsumme des SRA deutlich höher mit mathematischen

Fertigkeiten als mit den allgemeinen kognitiven Maßen zusammen, kann das als Beleg für die Erfassung von *statistical* reasoning gewertet werden, da viele statistische Konzepte eine mathematische Grundlage haben bzw. das Verständnis mathematischer Beziehungen voraussetzen. Korreliert die Punktsumme des SRA dagegen stärker mit allgemeinen kognitiven Maßen als mit mathematischen Fertigkeiten, stellt das einen Hinweis auf die Abbildung von statstical *reasoning* durch das SRA dar. Das könnte als Beleg für die Interpretation des statistical reasoning, wie es durch das SRA gemessen wird, im Sinne eines Persönlichkeitsmerkmals gewertet werden. Einen weiteren Beleg für die diskriminante Validität würden Zusammenhangsunterschiede zwischen den kognitiven Maßen (Mathematikfertigkeiten und allgemeine kognitive Variablen) im Vergleich zu nicht-kognitiven Maßen (etwa Einstellungen zu Statistik) und der Punktsumme des SRA liefern. Hierbei sollten die Zusammenhänge zu den kognitiven Maßen höher ausfallen als die Zusammenhänge zu den nicht-kognitiven Maßen und damit inkrementell zur Erklärung der Unterschiede in den erreichten SRA-Punkten beitragen (siehe Abbildung 4 für einen Überblick).

Eine inkrementelle Varianzaufklärung in der Punktsumme des SRA durch eins der allgemeinen kognitiven Maße kann auch als Beleg der konvergenten Validität des SRA zur Erfassung des statistical reasoning als kognitiven Prozess gewertet werden (Abbildung 4).

Schließlich wäre ein hoher negativer Zusammenhang zwischen der Punktsumme des SRA und einem anderen akademischen Leistungsmaß, wie z. B. der Modulnote, ein Hinweis auf konvergente Validität des SRA im Sinne der Erfassung eines Lehrziels.

Zusammenfassend werden folgende Erwartungen bezüglich der Beziehungen zwischen der Punktsumme des SRA und einem weiteren akademischen Leistungsmaßen, allgemeinen kognitiven Maßen, mathematischen Fertigkeiten und Einstellungen zu Statistik abgeleitet:

3a) Mathematische Fertigkeiten sollten eine mindestens schwache inkrementelle Varianzaufklärung in der SRA Punktsumme über die Einstellungen zu Statistik, Figurenreihen-Fortsetzung und deduktives Schließen hinaus leisten. (Konvergente und diskriminante Validität des SRA zur Messung von *statistical* reasoning als Persönlichkeitsmerkmal)

3b) Deduktives Schließen oder Figurenreihen-Fortsetzung sollten eine mindestens schwache inkrementelle Varianzaufklärung in der SRA Punktsumme über die Einstellungen zu Statistik und mathematische Fertigkeiten hinaus leisten. (Konvergente und diskriminante Validität des SRA zur Messung von statistical *reasoning* als Persönlichkeitsmerkmal sowie zur Messung des statistical reasoning als kognitiven Prozess)

3c) Es sollte sich ein starker negativer Zusammenhang zwischen der Punktsumme des SRA und der Modulnote zeigen. (Konvergente Validität des SRA zur Messung von statistical reasoning als Lehrziel einer Statistikveranstaltung)

Schließlich können Zusammenhänge zwischen verschiedenen Einstellungen zu Statistik und der erreichten Punktsumme im SRA exploriert werden. Hierfür werden die Zusammenhänge nullter Ordnung zwischen der Anzahl gelöster SRA-Aufgaben und Einschätzungen bezüglich 1) der eigenen affektiven Einstellung zur Statistik, 2) der eigenen bzw. der erforderlichen kognitiven Kompetenz für das Lernen von Statistik, 3) dem grundsätzlichen Wert von Statistik, 4) der Schwierigkeit des Fachs Statistik, 5) das eigene Interesse an Statistik sowie 6) dem geplanten bzw. erfolgtem Aufwand für die Statistikveranstaltung berichtet.

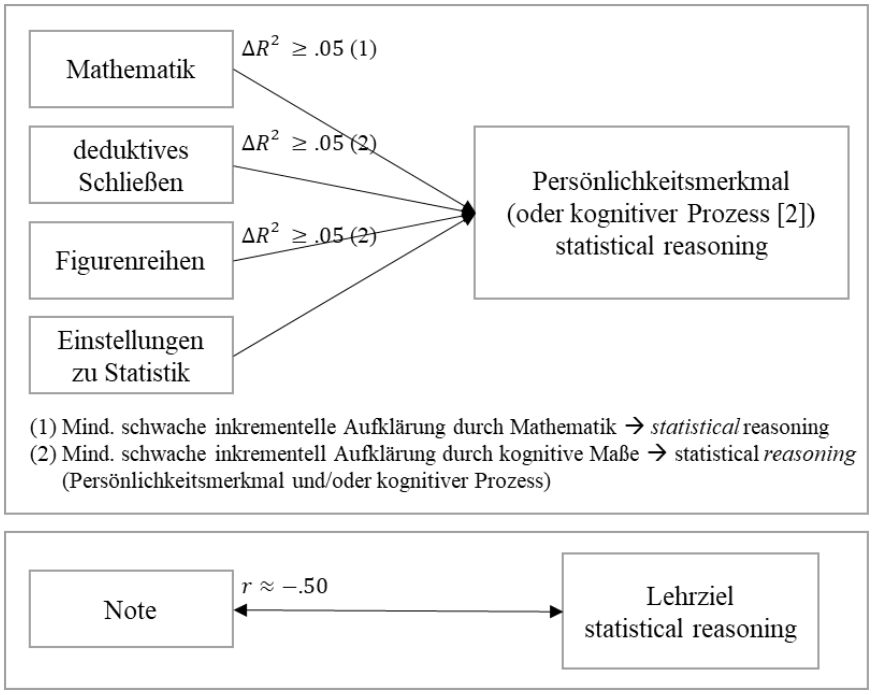


Abbildung 4. Erwartete Beziehungen zwischen der Punktsumme im SRA und kognitiven Maßen und einem anderen akademischen Leistungsmaß.

5. Methode

Als geeignete Populationen wurden für die Untersuchung einerseits Studierende der Psychologie und andererseits Studierende der Sonderpädagogik als Lehramt und Masterstudierende der Rehabilitationswissenschaften angesehen. Die Fächer sind sich in vielen Aspekten inhaltlich recht nahe, so spielt für beide Fächer etwa die Erhebung, Auswertung und Interpretation von Messungen psychologischer Konstrukte eine große Rolle. Andererseits unterscheiden sich die beiden Fächergruppen recht deutlich im Umfang und dem Detailgrad der gelehrten Statistikinhalte (siehe Abschnitt 1). Damit bietet sich die Möglichkeit, diskriminante Validitätsaspekte der Lehrerfolge in Statistik zu untersuchen.

Das Vorgehen bei der Beantwortung der Fragestellungen I und III wird korrelativ angelegt. Die Analysen der internen Struktur (Fragestellung I) beziehen sich mehrheitlich auf acht Stichproben: Studierende der Psychologie im ersten Semester, Studierende der Psychologie im zweiten Semester, Studierende der Sonderpädagogik und Studierende im Masterstudium der Rehabilitationswissenschaften jeweils zu Beginn und am Ende des Semesters. Auch wenn es sich bei den Messungen zu Beginn und am Ende des Semesters teilweise um abhängige Messwerte (max. zwei Drittel) handelt, scheint es gerechtfertigt, die Stichproben einzeln zu analysieren und damit Auskunft über die Stabilität der beobachteten Strukturen im Sinne der Replizierbarkeit zu erhalten. Bei der Beantwortung der Fragestellung I werden die Informationen aus der Messwiederholung daher versuchsplanerisch und datenanalytisch nicht berücksichtigt. Für die Analyse der Korrelationen und der Varianzaufklärung (Fragestellung III) zwischen dem SRA und den kognitiven sowie Einstellungsmaßen werden ebenfalls lediglich querschnittliche Informationen verwendet. Lediglich für die Kontrolle der Leistung im SRA zu Beginn des Semesters bei der Bestimmung der Semipartialdetermination der Prädiktoren wird die Information über den Erhebungszeitpunkt des SRA (zu Beginn des Semesters vs. am Ende des Semesters) berücksichtigt.

Die Untersuchung der Fragestellung II wurde dagegen explanativ angelegt (Döring & Bortz, 2016). Ein explanativer Forschungsansatz ermöglicht eine Aussage über differenzielle Effekte von als ursächlich vermuteten Manipulationsvariablen. Die in der vorliegenden Studie genutzte unabhängige

Manipulationsvariable liegt in der Form von Lehrinterventionen unterschiedlichen Umfangs (zwei Semesterwochenstunden [SWS] vs. vier SWS) und unterschiedlicher Inhalte (siehe Abschnitt 5.3) vor. Eine zweite unabhängige Variable liegt in Form des Messzeitpunkts (zu Beginn vs. am Ende des Semesters) vor. Damit kann die vorliegende Untersuchung auch als Interventionsstudie klassifiziert werden. Da eine randomisierte Zuweisung von Versuchspersonen zu den Untersuchungsbedingungen nicht vorgenommen werden kann, handelt es sich hierbei um eine quasi-experimentelle Anordnung. Es sollte daher bereits an dieser Stelle angeführt werden, dass die untersuchten Populationen sich über die Unterschiede in der Lehrintervention hinaus in anderen Merkmalen systematisch unterscheiden könnten.

5.1. Untersuchungsdesign

Die Untersuchung hatte damit insgesamt den Charakter einer quasi-experimentellen Anordnung mit Messwiederholung (je eine Vorher- und Nachhermessung, Tabelle 5). Natürlich vorgefundene Gruppen von Teilnehmenden, die Interventionen unterschiedlichen Umfangs und unterschiedlicher Art erfuhren, wurden zu Beginn und am Ende der Intervention untersucht. Wie es häufig der Fall ist in nicht-randomisierten Versuchsplänen, konnte auch hier eine ausgeglichene Verteilung der Untersuchungsteilnehmenden über die Gruppen und die beiden Messzeitpunkte nicht gewährleistet werden. So sind die Studierendenkohorten an privaten Hochschulen sowie in Masterstudiengängen der Sache gemäß im Umfang kleiner als diejenigen in Bachelorstudiengängen an öffentlichen Hochschulen.

5.2. Stichprobe

Die insgesamt $n = 502$ Untersuchungsteilnehmer zu mindestens einem Messzeitpunkt wurden an einer öffentlichen Universität und einer privaten Hochschule, beide in Nordrhein-Westfalen, rekrutiert. Der hauptsächliche Grund für die Auswahl der Erhebungskontexte und der Untersuchungsteilnehmer bestand in deren einfachen Verfügbarkeit für die Autorin. Bei der Stichprobengenerierung handelt es sich daher um eine ad hoc Auswahl mit den damit verbundenen Einschränkungen der Generalisierung.

Die untersuchten Personen an der öffentlichen Universität waren Studierende der humanwissenschaftlichen Fakultät, die laut der jeweiligen Prüfungsordnung ein

drei- (im Bachelor) bzw. zweiteiliges (im Master) Modul zu Forschungsmethoden im Umfang von jeweils insgesamt vier SWS absolvieren mussten. Von diesen vier SWS entfielen für beide akademischen Grade zwei SWS in Form eines Seminars auf statistische Inhalte. Die meisten Studierenden der öffentlichen Universität, die an der Untersuchung teilnahmen, befanden sich in einem Bachelorstudiengang ($n = 338$), zum größten Teil strebten diese Personen das Lehramt mit sonderpädagogischem Förderschwerpunkt (Bachelor of Arts) an ($n = 257$, siehe Tabelle 6 für eine detaillierte Übersicht). Da dieser Studiengang auch Mathematik als mögliches Fach beinhaltet, wurde bei den Studierenden erfragt, ob sie in ihrem Studium bereits an einer Mathematikveranstaltung teilgenommen haben oder zum Zeitpunkt der Studie eine Mathematikveranstaltung besuchen, um bei der Datenanalyse für mathematische Inhalte im Studium kontrollieren zu können. Für die meisten Bachelor-Studierenden ($n = 251$, 88%) war die Teilnahme an der Statistikveranstaltung zu Beginn der Studie die bisher erste.

Tabelle 5

Anzahl der Studierenden, die das SRA-Instrument bearbeiteten nach Stichproben und Zeitpunkten getrennt.

	Messwiederholungsfaktor:		
Gruppenfaktor:	Beginn des Semesters (*n*)	Ende des Semesters (*n*)	*n* beide Zeitpunkte
Statistikumfang			
Kleiner			
Stichprobe 1 (2 SWS, Bachelor)	277	163	110
Stichprobe 2 (2 SWS, Master)	34	32	15
Größer			
Stichprobe 3 (4 SWS, Bachelor)	31	29	19
Stichprobe 4 (4(+4) SWS, Bachelor)	51	43	33
Insgesamt	393	267	177

Anmerkungen. SWS: Semesterwochenstunde/n.

Die an der privaten Hochschule rekrutierten Studierenden ($n = 112$) waren zum Zeitpunkt der Erhebungen in einen rein psychologischen Studiengang (Bachelor of Science) eingeschrieben und mussten laut Prüfungsordnung insgesamt zwei Module mit jeweils vier SWS zu rein statistischen Inhalten absolvieren. Die beiden Module verteilen sich dabei auf zwei Semester, so dass die untersuchten Studierenden zum ersten Messzeitpunkt entweder an ihrem ersten Statistik-Modul teilnahmen ($n = 42$) oder bereits die Veranstaltungen im Umfang von vier SWS des ersten Statistikmoduls absolviert hatten ($n = 70$) (siehe Tabelle 6 für eine detaillierte Übersicht). Die Studierenden, die zum Zeitpunkt der Erhebung am zweiten Statistikmodul teilnahmen, nahmen laut Studienverlaufsplan zusätzlich an einer vier SWS Veranstaltung zu Forschungsmethoden und programmgestützter Datenauswertung teil. Ein/e Studierende/r nahm an der Statistikveranstaltung zum wiederholten Mal teil. Für alle Studierenden wurde außerdem noch die letzte erzielte Statistiknote erhoben. Diese Angabe konnte mehrheitlich nur von den Studierenden beantwortet werden, die bereits ein Statistikmodul laut Studienverlaufsplan absolviert hatten. Nicht überraschend ist das extrem ungleiche Geschlechterverhältnis in allen Studiengängen und das relativ homogene Alter der Studierenden an der privaten Hochschule.

Da die Anzahl und der Umfang der eingesetzten Variablen zu einer Testzeit von mehr als einer Stunde für eine Erhebung in allen Stichproben führte, wurden demografische Angaben aus Ökonomiegründen nur zu Beginn der Studie erhoben. Die dargestellten demografischen Daten fehlen daher für insgesamt $n = 90$ Studierende aus dem zweiten Messzeitpunkt.

Tabelle 6

Demografische Angaben der Studierenden (auf nächster Seite fortgesetzt).

	Stichprobe 1 *n* = 285	Stichprobe 2 *n* = 34	Stichprobe 3 *n* = 31	Stichprobe 4 *n* = 51
Alter				
Min; *Max*	18; 37	23; 37	18; 26	18; 26
M (SD)	21.11 (2.65)	25.81 (2.82)	20.27 (1.68)	20.5 (1.76)
Keine Angabe (%)	24 (8%)	2 (6%)	1 (3%)	1 (2%)
Geschlecht				
n weiblich (%)	230 (86%)	29 (85%)	26 (84%)	37 (73%)
Keine Angabe (%)	25 (9%)	2 (6%)	0 (0%)	0 (0%)
Umfang SWS Statistik	2	2	4	4(+4)
Studiengang				
BSc Psychologie			31	51
BA LA SPF	257			
BA andere	1			
MA Rehabilitations-wissenschaften		30		
MA andere		3		
Keine Angabe/fehlend	27	1		
Ausbildung/ Studium bisher				
n abgeschlossen	47 (16%)	30 (88%)	3 (10%)	2 (4%)
n angefangen	15 (5%)	1 (3%)	10 (32%)	14 (27%)
n Besuch einer Mathematik-Veranstaltung (%) [a]	117 (44%)	15 (44%)	/	/
n Besuch der Statistik-Veranstaltung bisher (%) [b]			0 (0%) [c]	1 (2%) [d]
Klausur zu der Statistik-Veranstaltung nicht bestanden [b]				2 (5%) [d]

Tabelle 6 *(fortgesetzt)*

	Stichprobe 1 *n* = 285	Stichprobe 2 *n* = 34	Stichprobe 3 *n* = 31	Stichprobe 4 *n* = 51
Letzte Note in Statistik				
1,0 bis 1,3	1	2	0	7
1,7 bis 2,3	4	3	0	8
2,7 bis 3,3	5	2	0	14
3,7 bis 4,0	0	0	0	5
5	5	3	0	11
n Note in Statistik (%)	15 (5%)	10 (29%)	0 (0%)	45 (88%)
***n* Angaben fehlen zu Z 2**	**53**	**17**	**10**	**10**

Anmerkungen. Bis auf die Anzahl der zum zweiten Zeitpunkt fehlenden Angaben der neu hinzugekommenen Studierenden (**fett** gesetzt) beziehen sich alle Angaben auf den ersten Messzeitpunkt. BA LA SPF: Bachelor of Arts mit Studienprofil Lehramt mit sonderpädagogischem Förderschwerpunkt, Z 2: Messzeitpunkt am Ende des Semesters. [a] vor und in dem aktuellen Semester, [b] die Angabe erfolgte am Ende der Intervention, [c] N = 29, [d] N = 43.

5.3. Intervention

Die Statistiklehre während des Semesters stellte die Intervention dar. Diese unterschied sich sowohl im Umfang als auch in den Inhalten für die Studierendenstichproben.

5.3.1. Stichprobe 1 (2 SWS, B.A.).

Die Inhalte des Teilmoduls für Statistik im Bachelorstudium an der öffentlichen Universität waren die Themen (jeweils im Umfang von ca. zwei Unterrichtsstunden, soweit nicht anders vermerkt):

- Allgemeines zum Forschungsprozess, Arten von Variablen, Arten von Hypothesen, Population und Stichprobe (ca. vier Unterrichtsstunden)
- Skalenniveaus, Abgrenzung deskriptive und inferenzielle Statistik, Abgrenzung uni- und bivariate Statistik, Häufigkeitsverteilung,

Abgrenzung Lage und Streuungsmaße; Lagemaße: Modus, Median, Mittelwert und deren Berechnung

- Streuungsmaße: Spannweite, Perzentile und Prozentränge, Varianz, Standardabweichung und deren Berechnung
- z-Werte und Prozentränge und deren Berechnung
- χ^2-Statistik, Kontingenzkoeffizient, Produkt-Moment-Korrelation
- Einfache und multiple lineare Regression
- Stichprobenkennwerte und Populationsparameter, Punktschätzer, Standardfehler, Stichprobenkennwerteverteilung, Konfidenzintervall
- Signifikanztest und Testhypothesen, t-Test für unabhängige und abhängige Stichproben
- SPSS: Dateneingabe, Variableneingabe und Modifikation, Syntax, Auswahl von Daten, Outputs zu deskriptiven und inferenziellen Analysen
- Diskussion von Ergebnissen

Zu allen Sitzungen und Themen gab es Übungen, die von den Studierenden in der Kontaktzeit im Plenum und Übungen, die außerhalb der Kontaktzeit freiwillig bearbeitet werden konnten. Zu allen Übungen wurden Lösungen über eine e-learning-Plattform zur Verfügung gestellt.

Die Lehre erfolgte in Studierendengruppen von ungefähr 30 Personen durch insgesamt sechs unterschiedliche Lehrende.

5.3.2. Stichprobe 2 (2 SWS, M.A.).

Die Inhalte des Teilmoduls für Statistik im Masterstudium an der öffentlichen Universität waren die Themen (jeweils im Umfang von ca. zwei Unterrichtsstunden, soweit nicht anders vermerkt):

- Messtheorie, zentrale Tendenz und Häufigkeitstabellen (ca. vier Unterrichtsstunden)
- Quantile
- Mittelwert, Varianz und Standardabweichung
- Kovarianz
- Korrelation
- Lineare Regression (ca. vier Unterrichtsstunden)

- Stichprobenkennwerte und Populationsparameter, Punktschätzer, Standardfehler, Stichprobenkennwerteverteilung, Konfidenzintervall, Hypothesentest
- Multivariate Verfahren im Überblick

5.3.3. Stichprobe 3 (4 SWS, B.Sc.).

Die Inhalte des ersten Statistikmoduls (mit drei SWS Seminar und einer SWS Übung) im Bachelorstudium an der privaten Hochschule waren die Themen (soweit nicht anders vermerkt jeweils im Umfang von ca. drei Unterrichtsstunden Seminarzeit und einer Unterrichtsstunde Übungszeit):

- Skalenniveaus
- Datenmatrix, primäre Häufigkeitsverteilung, absolute und relative Häufigkeiten, Modus, kumulierte Häufigkeiten, Median, Interquartilsbereich, sekundäre Häufigkeitsverteilung, Histogramm
- Mittelwert, Varianz, Standardabweichung, Histogramme für Gruppen, Boxplots für Gruppen, Verteilungsformen
- Phi-Koeffizient, Yules Q, Kovarianz, Produkt-Moment-Korrelation, punktbiseriale Korrelation, standardisierte Mittelwertedifferenz, Spearman-Korrelationskoeffizient
- einfache lineare Regression
- Partielle Korrelation, mutltiple lineare Regression, Moderation, Mediation, hierarchische Regression (ca. sechs Unterrichtsstunden, zwei Unterrichtsstunden Übung)
- Varianzanalyse
- Laplace Wahrscheinlichkeit, Ereignisse und Ereignisarten, Axiome der Wahrscheinlichkeit, Rechenregeln für Wahrscheinlichkeiten, Gesetz der großen Zahl, Zufallsvariablen und Wahrscheinlichkeitsfunktionen, bedingte Wahrscheinlichkeiten, stochastische Unabhängigkeit, Bayes-Theorem (ca. sechs Unterrichtsstunden mit Schwerpunkt auf Bayes, zwei Unterrichtsstunden Übung)
- Population und Stichprobe, stetige Verteilungen mit Schwerpunkt auf Standardnormalverteilung, Standardfehler, Quantile der Standardnormalverteilung (ca. drei Unterrichtsstunden Seminarzeit)

Im Seminar wurden während der Kontaktzeit zusätzliche Übungen bearbeitet. Der Schwerpunkt der Übungen im Seminar und in der Übungsstunde lag auf der Interpretation statistischer Kennzahlen im Kontext der Aufgaben (z. B. Bedeutung des statistischen Ergebnisses für die eingangs formulierte Fragestellung in der Aufgabe). Darüber hinaus hatten Studierende die Aufgabe, ausgewählte Kapitel aus einem Lehrbuch der Statistik als Prüfungsliteratur vor- und nachzubereiten (dies wurde nicht kontrolliert). Laut subjektiver Einschätzung der Lehrperson, wurde die Literatur von etwa der Hälfte der Studierenden bearbeitet.

5.3.4. Stichprobe 4 (4(+4) SWS, B.Sc.).

Studierende dieser Stichprobe (Psychologiestudierende im zweiten Semester an der privaten Hochschule) erfuhren im letzten Semester vor der Datenerhebung dieselbe Lehrintervention, wie die Stichprobe 3. Im laufenden Semester der Datenerhebung waren die Inhalte des zweiten Statistikmoduls (mit drei SWS Seminar und einer SWS Übung) die Themen (soweit nicht anders vermerkt jeweils im Umfang von ca. drei Unterrichtsstunden Seminarzeit und einer Unterrichtsstunde Übungszeit):

- Wiederholung Population und Stichprobe, stetige Verteilungen mit Schwerpunkt auf Standardnormalverteilung, Standardfehler, Quantile der Standardnormalverteilung, z-Werte
- Wiederholung Bernoulli-Theorem und starkes Gesetz der großen Zahlen, Populationsparameter vs. Stichprobenkennwerte, systematischer und unsystematischer Stichprobenfehler, Schätzer der Populationsstreuung, Standardfehler, Stichprobenkennwerteverteilung, zentraler Grenzwertsatz, Hypothesentest, bedingte Wahrscheinlichkeit p, Überblick über Ansätze von Fisher und Neyman und Pearson (ca. acht Unterrichtsstunden und zwei Unterrichtsstunden Übung)
- Kriterien der Parameterschätzungen, Konfidenzintervall
- Überblick über Verteilungen (χ^2, t, F) und Hypothesentests für unterschiedliche Stichprobenkennwerte, t-Test für unabhängige und abhängige Stichproben
- F-Test für das Prüfen von Varianzunterschieden, Kolmogorov-Smirnov-Test

- Varianzanalyse
- Hypothesentests für das Prüfen von Korrelationen
- χ^2-Tests für das Prüfen von Häufigkeitsunterschieden (darunter Vierfeldertafel), nicht-parametrische Verfahren im Überblick, Replikationsproblem
- Überblick über multivariate Verfahren

Zusätzlich nahmen die Studierenden dieser Stichprobe an einer Lehrveranstaltung zu Forschungsmethoden und der computergestützten Datenanalyse im Umfang von vier SWS teil. Die Lehre für alle Studierenden des psychologischen Bachelorstudiums an der privaten Hochschule erfolgte durch eine Lehrperson.

5.4. Untersuchungsdurchführung [23]

Die Datenerhebung erfolgte in einer paper-pencil-Form am Anfang und am Ende des Sommersemesters 2017[24]. In allen Stichproben wurden in der ersten Veranstaltungssitzung des Semesters die demografischen Variablen, das SRA sowie die Einstellung zur Statistik erhoben. In den Stichproben 3 und 4 wurden Studierende zusätzlich gebeten, Items zum Konstrukt *need for cognition*[25] zu beantworten; es folgte in der zweiten Vorlesungswoche die Erhebung von deduktivem Schließen und der Mathematik für diese beiden Stichproben. Auf Basis der Ergebnisse in den beiden Stichproben wurden anschließend zehn Mathematikitems für die erste Erhebung in den Stichproben 1 und 2 und acht Items zum deduktiven Schließen für die zweite Erhebung in den Stichproben 1 und 2 ausgewählt. In der vorletzten Vorlesungswoche des Semesters wurde die zweite Erhebung des SRA in den Stichproben 1, 3 und 4 durchgeführt. Zudem wurden die Studierenden der Stichproben 2 (in der letzten Semesterwoche) und 1 gebeten, figurale Reihenfortsetzungsaufgaben zu bearbeiten. In den Stichproben 2, 3 und 4 erfolgten alle Erhebungen während der entsprechenden Statistikveranstaltung. Aus organisatorischen Gründen wurde die Erhebung in der Stichprobe 1 in der zentralen

[23] An dieser Stelle möchte ich noch einmal meinen herzlichsten Dank an die Lehrenden aussprechen, die es mir ermöglicht haben, die – sehr zeitintensive – Datenerhebung durchzuführen.
[24] Dabei ist zu beachten, dass das Semester an der privaten Hochschule etwa einen Monat früher startete als das Semester an der öffentlichen Universität.
[25] Das Konstrukt wurde außerhalb der vorliegenden Fragestellungen erhoben.

Vorlesung des forschungsmethodischen Teilmoduls durchgeführt. Den Studierenden des psychologischen Studiengangs wurde die Teilnahme durch einen schriftlichen Nachweis über erbrachte Versuchspersonenstunden nach jeder Erhebung bescheinigt. Zusätzlich erhielten die Studierenden die Aussicht auf eine weitere Versuchspersonenstunde, wenn sie auch an der letzten Erhebung im Semester teilnehmen würden. Die Studierenden der Stichprobe 1 wurden durch die Bekanntgabe einer Klausurfrage nach der Erhebung zur Teilnahme am zweiten Messzeitpunkt motiviert.

Der zeitliche Umfang der beiden Erhebungen zu Beginn des Semesters für die Stichproben 3 und 4 betrug etwa 45 und 60 Minuten und die Erhebung am Ende des Semesters etwa 45 Minuten. Für die Stichproben 1 und 2 nahmen beide Erhebungen je etwa 60 bis 75 Minuten in Anspruch.

5.5. Variablen und Messinstrumente

Insgesamt wurden zusätzlich zu dem SRA und den demografischen Angaben weitere kognitive Leistungsmessungen sowie Einstellungsmessungen vorgenommen. Im Folgenden sollen die Instrumente kurz beschrieben werden.

5.5.1. SRA.

Das SRA wurde durch eine Übersetzerin in enger Absprache und zahlreichen Iterationen mit der Autorin aus dem Englischen ins Deutsche übertragen. Das übliche Vorgehen der Ziel- und Rückübersetzung durch jeweils bilinguale Personen oder Muttersprachler der Zielsprache wurde hier nicht als sinnvoll erachtet, da die Übertragung der Aufgaben eine besondere Sensibilität für statistische Fachbegriffe und Aufgabenstellungen erfordert. Anders, als etwa bei der Beschreibung emotionaler Zustände oder Einstellungen, kann hier die Konnotation von Begriffen, Idiomen etc. zudem als von eher unwesentlicher Bedeutung angenommen werden. Bei der Übertragung ins Deutsche wurden nur notwendige Anpassungen des Aufgabenkontextes für deutsche Verhältnisse vorgenommen, dies waren: Änderung des fiktiven Stadtnamens *Springfield* zu *Neustadt* (Item 3), Änderung der Anzahl der *malls* von 80 auf die Anzahl der *Einkaufszentren* von 20 (hierfür wurde das Verhältnis zur Gesamtbevölkerung des jeweiligen Landes in Millionen herangezogen, d. h. $80/320 \approx 20/80 \approx 0.25$) und der Anzahl der befragten

Jugendlichen von 2500 auf 550 in entsprechender Weise (Item 7); Änderung der Währungsangaben von Dollar auf Euro (Item 7, 8, 12); Änderung des Nachnamens von *Caldwell* auf *Müller*, Änderung der Automarken von *Buick* und *Oldsmobile* zu *Opel* und *Fiat*, Änderung des *Consumer Report* zu *Stiftung Warentest*, Änderung des Begriffs *rear end* bezogen auf einen Teil des Autos auf *Auspuff* (nach Absprache mit mehreren Autobesitzern und einer intensiven Recherche; hier ist im Englischen zwar vermutlich die Stoßstange gemeint, jedoch sind Probleme mit der Stoßstange am Auto in Deutschland nicht auffällig häufig) (Item 12). Das gesamte Instrument findet sich im Anhang in Tabelle 26.

Ein Item wurde dann als richtig bewertet und mit der Vergabe eines Punkts kodiert, wenn bei Einfachwahlaufgaben nur die richtige und keine (weitere) falsche Antwortoption angekreuzt wurde. Insbesondere bei den Items 5 und 17 (beide im Einfachwahlformat) wurden häufig mehrere Antwortmöglichkeiten gewählt. Da die Konstruktion und die psychometrischen Eigenschaften des SRA in der englischen Originalversion bereits ausführlich beschrieben worden sind und die Analyse der Güte der deutschen Version Hauptgegenstand der vorliegenden Studie ist, wird an dieser Stelle auf eine weitere Beschreibung verzichtet und auf die Abschnitte 3.1 (S. 58) und 6.1 (S. 121) verwiesen.

5.5.2. Kognitive Leistungsmessungen.

Für die Untersuchung der Fragestellung nach konvergenten und diskriminanten Zusammenhängen zwischen der im SRA erzielten Punktesumme und weiteren Merkmalen wurden weitere Variablen erhoben.

Mathematik.

In einigen Studien wurden als Proxy für mathematische Fertigkeiten oder Fähigkeiten die letzte Note im Fach Mathematik (z. B. in der Schule, Olani et al., 2011), die derzeitige Leistung in der Mathematik im Studium (z. B. Tempelaar et al., 2006) oder die mathematischen Anforderungen des Studiengangs (Garfield, 2003) herangezogen. Der Nachteil dieser Arten der Erfassung der mathematischen Leistungen ist hierbei zum einen, dass die Schulnote nicht nur von Fertigkeiten oder Fähigkeiten in Mathematik abhängt (z. B. Beteiligung im Unterricht) und Inhalte der Mathematik im Studium nicht notwendigerweise für Statistikveranstaltungen relevant sind. Aus diesem Grund wurden für die vorliegende Studie freigegebene

Aufgaben aus dem OECD PISA bis zur Erhebung im Jahr 2012 (mit Schwerpunkt auf mathematische Bildung 15-jähriger Schülerinnen und Schüler, Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens [BIFIE], n.d.) und der Trends in International Mathematics and Science Study für die 7./8. Klasse (TIMSS, Baumert et al., 1998) ausgewählt. Die Vorauswahl erfolgte durch zwei Lehrpersonen (darunter die Autorin) und zwei studentische Hilfskräfte, die an der Statistiklehre beteiligt waren. Die vier Personen beurteilten die Items auf Relevanz für Inhalte der Statistik. Insgesamt 31 Aufgaben wurden vorausgewählt (die Übersicht aller eingesetzten Aufgaben findet sich in Tabelle 27 im Anhang). Darunter sind zehn Aufgaben, die von allen vier Beurteilenden als hoch relevant eingeschätzt wurden und weitere 15 Aufgaben, bei denen sich drei der Beurteilenden über die hohe Relevanz einig waren. Fünf Aufgaben fanden zwei Beurteilende relevant und ein Item wurde von der Autorin in Eigenverantwortung ausgewählt. Die 31 Aufgaben wurden in der zweiten Erhebung in den Stichproben 3 und 4 (4 SWS, B.Sc. und 4(+4) SWS, B.Sc.) vorgelegt und anschließend nach dem Kriterium der Schwierigkeit und weiteren relevanten Aspekten (z. B. Auswertungseffizienz) auf 10 Items reduziert (Übersicht der Schwierigkeitsangaben aus der Originalquelle und aus den vier Stichproben finden sich im Anhang in Tabelle 31). In den Erhebungen in den Stichproben 3 und 4 wurde das schwierigste Item von 29% und das leichteste von 97% der Personen gelöst. In den Stichproben 1 und 2 lösten 36% der Personen das schwierigste Item und 96% der Personen das leichteste Item (für detailliertere Angaben siehe Tabelle 31 im Anhang). Der Grund für die Reduktion der Items lag vor allem in der begrenzten Erhebungszeit in den Stichproben 1 und 2. Da es sich hier um spezifische einzelne Aufgaben von lehrziel- bzw. kriteriumsorientierten Tests handelt, ist eine Angabe von typischen Skaleneigenschaften, wie der internen Konsistenz nicht entscheidend und wird daher nicht berichtet (siehe auch Abschnitt „Statistical reasoning als Lehrziel.“). Im Schnitt lösten die Studierenden aller Stichproben $M = 6.12$ bis $M = 7.22$ ($SD = 1.89$ bis $SD = 2.23$) der zehn Aufgaben (siehe für einen detaillierten Überblick Tabelle 34 und Abbildung 16 sowie Abbildung 17 im Anhang).

Deduktives Schließen.

Obwohl die Berücksichtigung verschiedener Formen des Schließens, z. B. in Form der Intelligenzmessung in der Forschung zu kognitiven Leistungen, darunter

akademischen Leistungsmaßen, sehr verbreitet ist, finden sich nur wenige Studien außerhalb der Forschung zu deduktivem Schließen (d. h. außerhalb der Allgemeinen Psychologie), die explizit deduktives Schließen berücksichtigen. So werden in der Denkforschung spezifische Schlussformen, wie der modus tollens z. B. in Form der Wason Aufgabe untersucht (Wason, 1960), in einen Alltagskontext eingekleidete, nicht auf spezifische Schlussformen bezogene Aufgaben im Rahmen allgemeiner Diagnostik vorgegeben (Denksporttest von Lienert, 1964) oder in einen spezifischen Kontext gesetzte Schlussformen im Rahmen von Eignungsdiagnostik getestet (z. B. Böhme, 2010). Alle der Autorin bekannten Zugänge kleiden Schlussformen jedoch entweder in einen inhaltlichen oder einen symbolisch-abstrakten Kontext (siehe Überblick bei Byrne, Evans & Newstead, 1993). Da der Kontext jedoch eine bedeutende, manchmal auch interferierende Rolle beim deduktiven Schlussfolgern spielt (z. B. Johnson-Laird, Legrenzi & Legrenzi, 1972, Byrne, 1989) und hier nur die Fertigkeit, korrekte deduktive Schlüsse in einem kontextfreien Rahmen ziehen zu können, von Interesse ist, wurden mehrere Items neu konstruiert. Zu diesem Zweck wurden verschiedene Kombinationen von Prämissen als Aufgabenstämme und richtige sowie falsche Konklusionen als Antwortoptionen erstellt. Dabei hatten acht Aufgabenstämme Prämissen deduktiver konditionaler Natur (z. B.: *wenn p, dann q. p*), zwölf Aufgabenstämme hatten Prämissen induktiver wahrscheinlichkeitsbezogener konditionaler Natur (z. B.: *wenn p, dann sehr wahrscheinlich q. p*), drei Aufgabenstämme enthielten jeweils eine Prämisse mit Quantoren (z. B.: *alle p sind q*) und weitere zwei Aufgabenstämme enthielten je zwei Prämissen mit Quantoren (z. B.: *alle p sind q. alle q sind s*). Die Inhaltsneutralität auf der einen Seite und die möglichst informelle Darstellung auf der anderen wurde durch die Verwendung von Pseudowörtern angestrebt (Bsp.: „Wenn es schnuckt [*p*], dann glackst es [*q*]. Es hat geschnuckt. [*p*]“). Die Verwendung von Pseudowörtern sollte einen möglichst natürlich-sprachlichen Kontext schaffen und damit die für die Struktur der Aufgabe irrelevante Abstraktionsnotwendigkeit reduzieren. Die zugrundeliegende Annahme bei diesem Vorgehen war, dass es für Personen einfacher ist, eine mentale Repräsentation von „glacksen“ zu bilden im Verglich zu „A“ (d. h. „glacksen“ ist konkreter als „A“). Gleichzeitig sollten durch dieses Vorgehen aber Nachteile von natürlicher Sprache (etwa Interferenzeffekte durch Wissen, Temporalität oder Assoziationen mit

bestimmten Begriffen) reduziert werden. Vorläufige und tendenzielle Bestätigung für die Überlegungen finden sich in vergleichsweise höheren Anteilen korrekter Lösungen des modus tollens (50% bis 77% mit $M = 66\%$ in den vorliegenden Stichproben gegenüber den bei Knauff [2006, S. 178] berichteten Werten von 41% bis 75% mit $M = 55\%$). Die insgesamt 25 Items (siehe Tabelle 28 im Anhang) wurden den Stichproben 3 und 4 in der zweiten Semesterwoche gemeinsam mit den Mathematikaufgaben und den Skalen zu Einstellungen zu Statistik vorgelegt. Um Ermüdungs- und Reihenfolgeeffekte bestimmen zu können, wurden die Blöcke mit Mathematikaufgaben und Aufgaben zum deduktiven Schließen (sowie die für die vorliegende Fragestellung nicht relevante Variation in den Instruktionen zur Bearbeitung der Aufgaben zum deduktiven Schließen) zufällig angeordnet. Es ließen sich keine solchen Effekte beobachten. Nach der Erhebung wurden acht Items nach der aus Sicht der Autorin zu priorisierenden inhaltlichen Bedeutung ausgewählt. Die im Rahmen der Mathematik und auch der Statistik wohl wichtigsten Schlussformen sind der modus ponens (etwa für direkte Beweise und Begründungen) und der modus tollens (für indirekte Beweise und in der Statistik in seiner wahrscheinlichkeitsbasierten Form zentral für das Konzept des Signifikanztestens). Beide Schlussformen haben auch entsprechende Fehlschlussformen, *affirming the consequent* für den modus ponens und *denying the antecedent* für den modus tollens. Aus diesem Grund wurden die acht Items mit den entsprechenden Aufgabenstämmen für die Erhebungen in den Stichproben 1 und 2 ausgewählt. Da deduktives Schließen in der Regel als kognitiver Prozess verstanden wird, werden auch für diese Items übliche Maße interner Konsistenz (Cronbach α, McDonald ω) nicht als nicht geeignet beurteilt. Für die Analysen zur Fragestellung III wurde die Anzahl der korrekt gelösten Aufgaben verwendet. Im Schnitt lösten die Studierenden aller Stichproben $M = 4.58$ bis $M = 5.21$ ($SD = 1.03$ bis $SD = 1.34$) der acht Items (siehe für einen detaillierten Überblick Tabelle 35 und Abbildung 18 sowie Abbildung 19 im Anhang). In einer studentischen Projektarbeit betreut durch die Autorin (Gebel & Sawatzky, 2018) fand sich ein Beleg für konvergente Validität der Items in Form eines hohen Zusammenhangs zwischen der Gesamtzahl korrekt gelöster Items des deduktiven Schließens und der Grundintelligenztest Skala 2 des Culture Fair Intelligence Tests (CFT 20-R, Weiß, 2006) ($r = .50$) sowie eines

moderaten bis hohen Zusammenhangs mit der Anzahl gelöster Mathematikaufgaben ($r = .38$).

Figurale Reihenfortsetzungsaufgaben.

Ein weiteres allgemeines kognitives Maß wurde in Form von figuralen Reihenfortsetzungsaufgaben erfasst. Hierfür wurden zehn Aufgaben aus dem Mensa Online-IQ-Test (Mensa The Netherlands, 2002, mit freundlicher Genehmigung) ausgewählt (siehe Tabelle 29 im Anhang). Hinweise auf konvergente Konstruktvalidität liegen für die Items in Form von moderaten bis hohen Korrelationen jeweils zu der Grundintelligenztest Skala 2 ($r = .43$, CFT 20-R, Weiß, 2006) und der Anzahl gelöster Mathematikaufgaben ($r = .49$) vor (Gebel & Sawatzky, 2018). Die Aufgaben wurden den Stichproben 1 (2 SWS, B.A.) und 2 (2 SWS, M.A.) vorgelegt. Im Schnitt lösten die Studierenden etwas mehr als die Hälfte der Items ($M = 5.67$, $SD = 1.82$ in Stichprobe 1 und $M = 5.93$, $SD = 1.84$ in Stichprobe 2). Im Anhang finden sich weitere univariate Angaben in Tabelle 36 und in Abbildung 20.

5.5.3. Einstellungen zu Statistik.

Neben kognitiven Maßen wurden verschiedene Einstellungen zu Statistik vor allem zur Bestimmung diskriminanter Validität bei allen Personen sowohl zu Beginn als auch am Ende des Semesters erhoben. Hierfür wurden die sechs Skalen des Survey of Attitudes Towards Statistics (SATS-36, Schau, et. al., 1995, Schau, 2003 und Schau, 2017, persönliche Korrespondenz, Validierung in deutscher Sprache durch Zimprich, 2012) genutzt. Die Skalen erfassen Einschätzungen bezüglich der eigenen affektiven Einstellung zur Statistik („Affekt“, Beispielitem der insgesamt sechs Items: „Ich werde Statistik mögen“ bzw. „Ich mag Statistik“), der eigenen bzw. der erforderlichen kognitiven Kompetenz für das Lernen von Statistik („Kompetenz“, Beispielitem der insgesamt sechs Items: „Ich werde in dieser Statistikveranstaltung keine Ahnung haben, worum es geht“, negativ formuliert), dem grundsätzlichen Wert von Statistik („Wert“, Beispielitem der insgesamt neun Items: „Ich werde in meinem Beruf keine Anwendungsmöglichkeiten für Statistik haben“, negativ formuliert), der Schwierigkeit des Fachs Statistik („Schwierigkeit“, Beispielitem der insgesamt sieben Items: „Statistik ist ein kompliziertes Fach“), das eigene Interesse an Statistik („Interesse“, Beispielitem der insgesamt vier Items: „Ich

bin daran interessiert, statistische Informationen zu verstehen“) sowie dem geplanten bzw. erfolgtem Aufwand für die Statistikveranstaltung („Anstrengung“, Beispielitem der insgesamt vier Items: „Ich habe vor, in meiner Statistikveranstaltung hart zu arbeiten“ bzw. „Ich habe in meiner Statistikveranstaltung hart gearbeitet“). Alle 36 Aussagen wurden auf einer siebenstufigen Ratingskala hinsichtlich der Zustimmung eingeschätzt (1 = *stimme gar nicht zu*, 4 = *weder/noch*, 7 = *stimme voll und ganz zu*). Bis auf die Skala „Schwierigkeit“ ($\alpha = .38$ bis $\alpha = .73$) zeigten alle Skalen in allen Stichproben und zu beiden Zeitpunkten in der vorliegenden Untersuchung zufriedenstellende oder hohe interne Konsistenzen in Form des Cronbach α ($\alpha = .70$ bis $\alpha = .93$). Für detaillierte Angaben zu den deskriptiven Statistiken und Cronbach α Angaben zu den Skalen aus allen Stichproben zu beiden Messzeitpunkten siehe Tabelle 37 sowie Abbildung 21, Abbildung 22, Abbildung 23 und Abbildung 24 im Anhang.

6. Ergebnisse

Im Folgenden werden die Ergebnisse der an den Fragestellungen orientierten Datenanalysen dargestellt. Zunächst soll die Analyse der internen Struktur des SRA beschrieben werden. Dieser Abschnitt dient in erster Linie der Beantwortung der ersten Fragestellung: Lassen sich Hinweise auf eine einheitliche Konstruktstruktur des statistical reasoning in den vorliegenden Daten auffinden? Nachfolgend werden empirische Antworten auf die zweite Fragestellung präsentiert: Zeigt sich ein Effekt der Lehrintervention in den Messwerten des SRA? Schließlich werden die Ergebnisse der Zusammenhangs- und Regressionsanalysen des SRA mit ausgewählten relevanten Variablen dargestellt. Die Auswertung der Daten und Darstellung der Ergebnisse erfolgte hauptsächlich mit dem Programm R (Version 3.6.1, R Core Team, 2019) und den Paketen *tidyverse* (Version 1.2.1, Wickham, 2017), *haven* (Version 2.1.1, Wickham & Miller, 2019), *psych* (Version 1.8.12, Revelle, 2018), *ggcorrplot* (Version 0.1.3, Kassambara, 2019), *cowplot* (Version 1.0.0; Wilke, 2019), *papaja* (Version 0.1.0.9842, Aust & Barth, 2018), *knitr* (Version 1.24; Xie, 2019), *rmarkdown* (Version 1.15; Allaire et al., 2019), *lavaan* (Beta-Version 0.6-4, Rosseel, 2012) sowie mit dem IBM SPSS Statistics Version 25 (c) Programm.

Übereinstimmend mit den aktuellen Empfehlungen der American Statistical Association (ASA) wird auf den Gebrauch des statistischen Signifikanzbegriffs (d. h. Formulierungen, wie „statistisch signifikant“ oder Markierung von *p*-Werten mit „***“) bei den Darstellungen verzichtet (Wasserstein, Schirm & Lazar, 2019). An einigen Stellen (z. B. bei konfirmatorischen Faktorenanalysen, Mittelwertdifferenzen, Varianzanalysen) werden jedoch *p*-Werte angeboten.

In die Datenanalysen von Skalen (Fragestellungen II und III) gingen Skalenwerte ein, die nicht mehr als 30% fehlender Werte in den Items, die in die Skalen eingehen, aufwiesen.

6.1. Fragestellung 1: Interne Struktur des SRA

Die Ergebnisse zur Analyse der internen Struktur des SRA gliedern sich in zwei Teile. Im ersten Teil werden die deskriptiven Itemeigenschaften und Beziehungen der einzelnen Items zueinander sowie zu den laut Beschreibung der Autoren angenommenen Skalen des SRA berichtet. Im zweiten Teil folgen

Faktorenanalysen zur explorativen und konfirmatorischen Untersuchung der Zusammenhänge zwischen den Items.

6.1.1. Itemeigenschaften des SRA.

Auf den folgenden Seiten werden die Schwierigkeiten, Trennschärfen, Iteminterkorrelationen sowie die internen Konsistenzen der Items bzw. Skalen beschrieben. Zuvor sollen noch die fehlenden Werte im gesamten SRA nach Stichproben getrennt berichtet werden. Insgesamt ist der Anteil der vollständig bearbeiteten Aufgaben des SRA mit 75% bis 97% recht hoch (siehe Tabelle 7 für eine detaillierte Übersicht). Die Verteilung der fehlenden Werte über die verschiedenen Stichproben der Studierenden zeigt keine offensichtlichen Auffälligkeiten. Die Qualität der Daten im Sinne des Ausmaßes von fehlenden Werten scheint daher für eine Auswertung geeignet zu sein.

Tabelle 7

Anzahl und Prozent (in Klammern) der fehlenden Werte im SRA zum ersten und zweiten Messzeitpunkt im Stichprobenvergleich.

	Erster Messzeitpunkt			Zweiter Messzeitpunkt		
S	keine	max. 30%	mehr als 30%	keine	max. 30%	mehr als 30%
1	238 (86%)	32 (12%)	7 (3%)	153 (94%)	9 (6%)	1 (1%)
2	30 (88%)	4 (12%)	/	28 (88%)	/	4 (12%)
3	31 (100%)	/	/	29 (100%)	/	/
4	49 (96%)	2 (4%)	/	41 (95%)	2 (5%)	/

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc.

Schwierigkeiten.

In Tabelle 8 finden sich die Schwierigkeiten der einzelnen Items jeweils für den ersten und zweiten Messzeitpunkt getrennt nach den unterschiedlichen Stichproben der Studierenden. Die Schwierigkeiten wurden als Anteil der korrekten Antworten an allen gegebenen Antworten für ein Item ermittelt. Da die Durchsicht der fehlenden Werte keine Hinweise auf gehäufte, nicht in Angriff genommene

Items lieferte, wurde auf eine Inangriffnahmekorrektur verzichtet. Auch eine Ratekorrektur fand nicht statt. Zu begründen ist dies zum einen darin, dass die Items des SRA zum Teil im Mehrfachantwortformat erstellt sind und sich in der Anzahl der Distraktoren teilweise stark unterscheiden, so dass die Berücksichtigung der Distraktoren eine Gewichtung bewirken würde. Zum anderen gibt es aus der Beobachtung der Studierenden bei der Durchführung und Gesprächen mit den Studierenden nach der Durchführung durch die Autorin keine Anhaltspunkte für auffälliges Raten. Schließlich weisen die meisten Aufgaben mindestens drei, oft auch vier oder mehr Distraktoren auf, so dass die Ratewahrscheinlichkeit hier in ihrer Bedeutung für das Antwortverhalten vernachlässigt werden kann. Für einen detaillierten Überblick der Optionen bei der Ermittlung von Schwierigkeiten siehe z. B. Lienert & Raatz (1998).

Da die univariaten deskriptiven Statistiken (Mittelwert, Varianz und Standardabweichung, Median und der Interquartilsbereich, Minimum und Maximum) unmittelbar aus den Schwierigkeiten erkennbar sind, werden diese für die Items des SRA nicht gesondert aufgeführt. Der Mittelwert für dichotome mit „1“ und „0“ kodierte Items ergibt sich aus der Anzahl der Personen, die das Item mit „1“ beantwortet haben relativ zu der Gesamtzahl der Antworten und ist damit mit der unkorrigierten Schwierigkeit identisch. Die Varianz dichotomer Items bestimmt sich aus dem Produkt der Schwierigkeit (Anteil der Personen, die das Item mit „1“ beantwortet haben) und der „Leichtigkeit“ (Anteil der Personen, die das Item mit „0“ beantwortet haben) und erreicht ihren maximal möglichen Wert für eine Schwierigkeit von $P = .5$ (bzw. 50%). Je leichter oder schwieriger ein Item ist, d. h. je mehr die Schwierigkeit von 50% abweicht, desto kleiner sind in der Folge die Varianz und die Standardabweichung. Der Median (sowie die untere und obere Interquartilsgrenzen) liegen bei Schwierigkeiten unter 50% (bzw. unter 25% und 75%) bei „0“ und bei Schwierigkeiten über 50% (bzw. über 25% und 75%) bei „1“. Entsprechend liegt das Minimum bei „0“ und das Maximum bei „1“ sofern die Schwierigkeiten einen Wert ungleich 0 (bzw. 0%) oder 1 (bzw. 100%) betragen.

Zusätzlich ist in der Tabelle 8 die Differenz der Schwierigkeit abgetragen, die für die zweite Fragestellung der Arbeit eine Rolle spielt; auf diesen Aspekt wird an entsprechender Stelle eingegangen (siehe 6.2.3). Für eine bessere Übersicht der Schwierigkeiten sind der Median und der Interquartilsbereich in einem Boxplot in

Tabelle 8

Schwierigkeiten (Anteil der Personen, die das Item richtig gelöst haben in %) der SRA-Items für beide Zeitpunkte sowie die Differenz der Schwierigkeiten zwischen dem Beginn und dem Ende des Semesters für alle Stichproben.

S	Z	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	63.9	89.4	34.1	81.5	33.1	83.5	12.5	89.5	86	23.7	71.6	60.8	29.7	25.3	14.6	73	36.4	11.7	20	26.4
		(263)	(263)	(267)	(265)	(266)	(266)	(265)	(267)	(264)	(266)	(264)	(265)	(266)	(265)	(261)	(263)	(261)	(266)	(265)	(265)
	2	50.3	88.2	32.9	78.3	33.8	64.2	27.2	95.1	89.5	29.0	78.1	69.8	22.8	34.6	8.1	70.0	42.9	10.6	13.8	18.8
		(161)	(161)	(161)	(161)	(160)	(162)	(162)	(162)	(162)	(162)	(160)	(162)	(162)	(162)	(161)	(160)	(161)	(160)	(160)	(160)
	D	**-13.6**	**-1.2**	**-1.2**	**-3.2**	0.7	**-19.3**	14.7	5.5	3.5	5.3	6.5	9	**-6.9**	9.3	**-6.5**	**-3**	6.5	**-1**	**-6.2**	**-7.7**
2	1	52.9	94.1	29.4	82.4	33.3	70.6	18.2	85.3	88.2	14.7	82.4	70.6	18.2	8.8	12.9	88.2	52.9	17.6	14.7	23.5
		(34)	(34)	(34)	(34)	(33)	(34)	(33)	(34)	(34)	(34)	(34)	(34)	(33)	(34)	(31)	(34)	(34)	(34)	(34)	(34)
	2	46.4	92.9	21.4	78.6	21.4	53.6	10.7	85.7	89.3	28.6	82.1	53.6	10.7	3.6	14.3	85.7	60.7	10.7	7.1	14.3
		(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)
	D	**-6.5**	**-1.3**	**-8**	**-3.8**	**-11.9**	**-17**	**-7.5**	0.4	1.1	13.9	**-0.2**	**-17**	**-7.5**	**-5.3**	1.4	**-2.5**	7.8	**-6.9**	**-7.6**	**-9.2**
3	1	41.9	71	22.6	83.9	41.9	54.8	16.1	83.9	71	19.4	54.8	54.8	32.3	32.3	3.2	61.3	41.9	9.7	19.4	38.7
		(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)
	2	62.1	89.7	41.4	58.6	20.7	79.3	6.9	96.6	96.6	24.1	82.8	44.8	17.2	31	0	82.8	34.5	17.2	27.6	41.4
		(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)
	D	20.1	18.7	18.8	**-25.3**	**-21.2**	24.5	**-9.2**	12.7	25.6	4.8	27.9	**-10**	**-15**	**-1.2**	**-3.2**	21.5	**-7.5**	7.6	8.2	2.7
4	1	52.9	88.2	37.3	80.4	43.1	86.3	11.8	90.2	92.2	29.4	86.3	70.6	21.6	31.4	10.2	70.6	45.1	11.8	11.8	21.6
		(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(49)	(51)	(51)	(51)	(51)	(51)
	2	60.5	90.7	39.5	72.1	25.6	58.1	18.6	97.6	97.7	30.2	92.9	67.4	18.6	27.9	7	67.4	44.2	9.5	7.1	11.9
		(43)	(43)	(43)	(43)	(43)	(43)	(43)	(42)	(43)	(43)	(42)	(43)	(43)	(43)	(43)	(43)	(43)	(42)	(42)	(42)
	D	7.5	2.5	2.3	**-8.3**	**-17.6**	**-28.1**	6.8	7.4	5.5	0.8	6.6	**-3.1**	**-3**	**-3.5**	**-3.2**	**-3.1**	**-0.9**	**-2.2**	**-4.6**	**-9.7**

Anmerkungen. *n* jeweils in Klammern angegeben. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., Z: Zeitpunkt mit 1 = zu Beginn der Intervention, 2 = am Ende der Intervention und D = Differenz der Schwierigkeiten wobei positive Werte eine Erhöhung des Schwierigkeitsindex anzeigen (d. h. mehr Studierende lösen das Item anteilig am Ende des Semesters; negative Werte sind **fett** gesetzt).

Abbildung 5 visualisiert. Aus der Darstellung wird ersichtlich, dass die Items des SRA in fast allen Stichproben eine tendenziell bimodale Verteilung der Schwierigkeiten aufweisen: die Items sind überwiegend eher leicht oder eher schwer. Lediglich für die Stichprobe 3 zu Beginn des Semesters lässt sich eine in etwa normale Verteilung der Schwierigkeiten beobachten. Zudem zeigen die Verteilungen für alle Stichproben zu beiden Messzeitpunkten, dass (mindestens) 25% der Items von teilweise weit weniger als 50% der Studierenden gelöst werden. So liegt die obere Grenze des ersten Quartils zu beiden Zeitpunkten bei allen Stichproben bei maximal ca. 30%. Das bedeutet, dass 25% der Items von maximal ca. 30% der Studierenden korrekt beantwortet wurden.

Insgesamt können die Aufgaben des SRA daher zwar insgesamt als mittelschwer gelten, jedoch werden einzelne Items von anteilig sehr vielen oder aber sehr wenigen Studierenden gelöst.

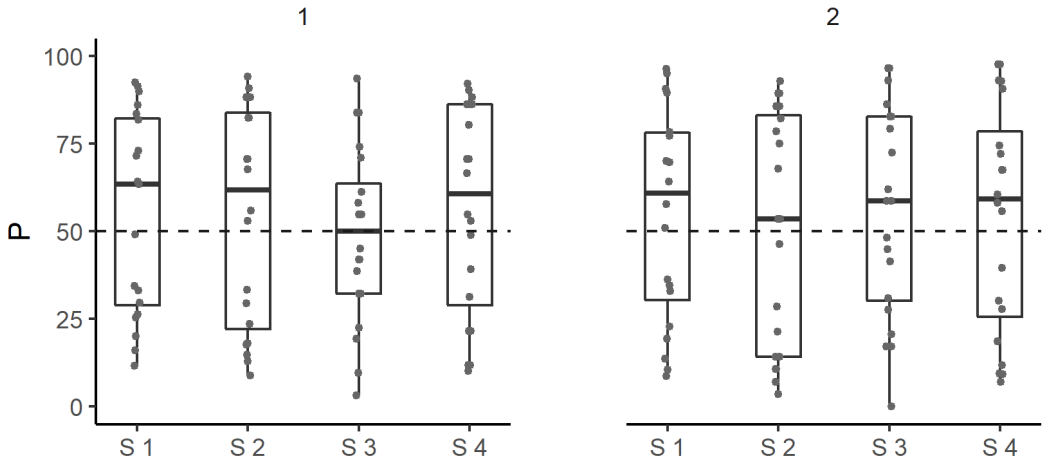


Abbildung 5. Boxplots der Verteilung der Schwierigkeiten in den Stichproben S 1 (2 SWS, B.A.), S 2 (2 SWS, M.A.), S 3 (4 SWS, B.Sc.) und S 4 (4(+4) SWS, B.Sc.) zu Beginn (1) und am Ende (2) des Semesters mit Referenz zu einer Schwierigkeit von 50% (gestrichelte Linie).

Korrigierte Trennschärfen.

Wie in der Fragestellung I erläutert, lassen sich über die Struktur des SRA verschiedene Erwartungen formulieren. Für die Struktur der Trennschärfekoeffizienten bedeuten diese Erwartungen für den Fall der eindimensionalen Struktur (mindestens mehrheitlich) positive Zusammenhänge der einzelnen Items mit der Gesamtskala. Im Fall einer vierdimensionalen Struktur sollten die Zusammenhänge der den Typen von statistical reasoning zugeordneten Items mit den Typenskalen mehrheitlich positiv sein. Wenn dem Instrument eine achtdimensionale Struktur zugrunde liegt, sollten den Subskalen zugeordnete Items mit der zugehörigen Subskala positiv zusammenhängen. Da die jeweiligen Skalen (Gesamtskala, Typenskalen sowie Subskalen) auch das jeweilige Item mit enthalten, würden über Produkt-Moment-Korrelationen operationalisierte Zusammenhänge die tatsächlichen Korrelationen überschätzen (siehe dazu z. B. Lienert & Raatz, 1998). Daher werden hier um das jeweilige Item bereinigte Korrelationen (auch als *part-whole*-korrigiert bezeichnet) berichtet (berechnet mit der Funktion *alpha* aus dem Paket *psych*, Revelle, 2018). Der dabei genutzte Koeffizient ist die Produkt-Moment-Korrelation.[26]

Die Korrelogramme (erstellt mit dem Paket *ggcorrplot*, Kassambara, 2019) in Abbildung 6 zeigen die um das Item bereinigten Item-Skala-Korrelationen im jeweiligen Ausmaß für den ersten bzw. zweiten Messzeitpunkt und die vier Stichproben der Studierenden im Vergleich. Insgesamt scheinen lediglich die Items 18, 19 und 20 sowie 9, 10 und 11 recht konsistent über Stichproben und Messzeitpunkte hohe Korrelationen zu der jeweils entsprechenden Subskala („Wahrscheinlichkeiten korrekt berechnen“ bzw. „(stochastische) Unabhängigkeit verstehen“), nicht aber der Typenskala oder der Gesamtskala aufweisen. Besonders auffällig sind hier die relativ zahlreichen negativen Korrelationen der Items zu den Subskalen, Typen oder der Gesamtskala. Negative Korrelationen zeigen an, dass Personen, die das entsprechende Item lösten, insgesamt einen geringeren Wert in der

[26] Die Produkt-Moment-Korrelation ist mathematisch äquivalent zu der punktbiserialen Korrelation, die in manchen Lehrbüchern für die Berechnung von Zusammenhängen zwischen (natürlich) dichotomen und metrischen Variablen empfohlen wird (z. B. Bortz, 2005, Eid et al., 2017 oder Lienert und Raatz, 1998, in den ersten beiden Quellen findet sich auch die Erläuterung der Äquivalenz). Die Verwendung der punktbiserialen Korrelation hat vermutlich eher historisch-pragmatische Gründe, da diese in der Berechnung ohne Computerprogramme wesentlich einfacher erfolgt, als die Produkt-Moment-Korrelation.

Subskala, dem Typ des statistical reasoning bzw. der Gesamtskala erzielten, als Personen, die das Item nicht lösten.

Abbildung 6. Korrigierte Trennschärfen (Produkt-Moment-Korrelationen) für die Items des SRA zu den jeweiligen Subskalen, den Typen des statistical reasoning und der Gesamtskala correct statistcal reasoning für die Stichproben S 1 (2 SWS, B.A.), S 2 (2 SWS, M.A.), S 3 (4 SWS, B.Sc.) und S 4 (4(+4) SWS, B.Sc.) zu Beginn (obere Grafik) und am Ende (untere Grafik) des Semesters. Farblich umrandete Bereiche (rot für Typen des statistical reasoning, blau für Subskalen) sollten positive Korrelationen enthalten. Item 15 in S 3 am Ende des Semesters wurde von keiner Person gelöst, in der Folge können keine Korrelationen mit dem Item ermittelt werden, die jeweiligen Stichprobengrößen können der Tabelle 9 (Gesamtskala), der Tabelle 10 (Typ) sowie der Tabelle 11 (Subskala) entnommen werden.

Die empirisch ermittelten Trennschärfen stimmen damit nur im Fall der Subskala „(stochastische) Unabhängigkeit verstehen“ (Items 9, 10, und 11) mit den zu erwartenden Trennschärfen überein. Für die Subskala „Wahrscheinlichkeiten korrekt berechnen“ zeigt sich ein erwartetes Ergebnismuster lediglich für drei der fünf Items (18, 19 und 20). Insgesamt werden die Trennschärfen damit als nicht erwartungsgemäß eingeschätzt.

Iteminterkorrelationen.

Auch Iteminterkorrelationen können, wie die Trennschärfen, deskriptiv auf Übereinstimmung mit den vermuteten Datenstrukturen inspiziert werden. Für dichotome Variablen, wie im vorliegenden Fall, können für diesen Zweck, wie auch für metrische Variablen, grundsätzlich Produkt-Moment-Korrelationen genutzt werden.[27] Der Produkt-Moment-Koeffizient ist jedoch in seiner maximalen Höhe abhängig von Randverteilungen der zu korrelierenden Variablen. So erreicht der Produkt-Moment-Korrelationskoeffizient einen maximal möglichen Wert von lediglich $r_{XY} = \phi_{XY} \approx .50$ und einen minimalen Wert von $r_{XY} = \phi_{XY} \approx -.15$ bei Schwierigkeiten von $P_X = .80$ und $P_Y = .90$ (Ekström, 2011). Wie bereits weiter oben gezeigt, weisen die Items in den vorliegenden Daten zum Teil extreme Randverteilungen auf und so kann erwartet werden, dass die Produkt-Moment-Korrelationskoeffizienten in ihren Höhen davon beeinflusst werden. Bortz (2005) führt als eine Lösungsmöglichkeit für diesen Fall eine Korrektur des Produkt-Moment- bzw. ϕ-Koeffizienten um den maximal möglichen Wert des Koeffizienten bei gegebenen Randverteilungen auf. Allerdings weist er auch darauf hin, dass dieses Vorgehen kritisiert werden kann, da Produkt-Moment-Korrelationen in anderen Fällen von nicht identischen Randverteilungen nicht entsprechend angepasst werden (S. 230). Alternativ bieten z. B. Lienert & Raatz (1998) an, bei extremen Randverteilungen auf den Einsatz der Koeffizienten der Produkt-Moment-Familie zu verzichten und stattdessen den Q-Koeffizienten von Yule (1900) zu nutzen. Da die Produkt-Moment-Korrelation und der Q-Koeffizient allerdings unterschiedliche Informationen zur Verfügung stellen, sollen hier beide Kennzahlen berichtet werden. Während die Produkt-Moment-Korrelation eine Aussage über die Vorhersagekraft eines Items für ein anderes erlaubt (hier: wenn ich das Antwortverhalten im

[27] zur Äquivalenz der Produkt-Moment-Korrelation und des ϕ-Koeffizienten siehe z. B. Bortz (2005) oder Eid et al. (2017).

Allgemeinen von Personen auf Item *X* kenne, wie gut kann ich das Antwortverhalten im Allgemeinen für Item *Y* vorhersagen?), beantwortet der Yule Q-Koeffizient die Frage, inwiefern die Antwortoption eines Items im Speziellen von der speziellen Antwort auf ein anderes Item abhängt (hier: Wie viele Personen lösen das Item *Y* anteilig, wenn sie auch das Item *X* gelöst haben?[28]). Ein Produkt-Moment-Korrelationskoeffizient von $r = |\pm 1|$ würde bedeuten, dass aus der Kenntnis der Ausprägung eines Items die Ausprägung des anderen Items fehlerfrei vorhergesagt werden kann. Ein Q-Koeffizient von $Q = 1$ dagegen sagt lediglich aus, dass alle Personen, die Item 1 gelöst haben, auch das Item 2 gelöst haben[29], hier kann es durchaus sein, dass es Personen gibt, die Item 1 zwar nicht gelöst haben, jedoch das Item 2. Eine fehlerfreie Vorhersage ist daher nur aus einer Antwortoption für eine Antwortoption der beiden Items möglich. Da bei der vorliegenden Fragestellung damit zu rechnen ist, dass die Schwierigkeiten (auch theoretisch zu erwartend) unterschiedlich sind, scheint es daher sinnvoll, die Q-Koeffizienten zumindest ergänzend zu berichten.

Abbildung 7 und Abbildung 8 zeigen die Iteminterkorrelationen in Form von Korrelogrammen. Aus Platzgründen werden hier nur für zwei ausgewählte Stichproben (1 und 3) die Iteminterkorrelationen für den zweiten Messzeitpunkt (Ende des Semesters) aufgeführt. Im Anhang (Abbildung 26 bis Abbildung 31) finden sich die Korrelogramme für den ersten Messzeitpunkt sowie die anderen beiden Stichproben und beide Messzeitpunkte. Zunächst zeigen sich, wie aufgrund der deutlich ungleichen Randverteilungen erwartet, teilweise extreme Unterschiede in der Höhe der Produkt-Moment- und der Q-Koeffizienten. Ein Blick auf die Schwierigkeiten der betroffenen Items bestätigt die sehr hohen (z. B. Items 8 und 9) oder sehr niedrigen (z. B. Items 13 und 18) Schwierigkeiten, die sehr unausgewogene Randverteilungen anzeigen. Die Produkt-Moment-Korrelationen fallen wie angesichts der Schwierigkeiten erwartet, überwiegend eher klein aus. In Tabelle 11 sind die höchsten und niedrigsten durchschnittlichen sowie die Mediane

[28] etwas vereinfacht, tatsächlich handelt es sich um den Anteil der Differenz zwischen Konkordanzen und Diskonkordanzen im Vergleich zur Summe der Konkordanzen und Diskonkordanzen:

$$\frac{n_{11} \cdot n_{22} - n_{12} \cdot n_{21}}{n_{11} \cdot n_{22} + n_{12} \cdot n_{21}}$$

mit n_{11} und n_{22}: Anzahl der Häufigkeiten von gleichen Ausprägungen auf beiden Variablen und n_{12} und n_{21}: Anzahl der Häufigkeiten von ungleichen Ausprägungen auf beiden Variablen.

[29] natürlich gilt dies ganz allgemein für beliebige dichotome Ausprägungen und nicht nur für Lösungsverhalten oder „richtig“ und „falsch“ Angaben.

der Iteminterkorrelationen (Produkt-Moment-Koeffizienten) der Subskalen-Items für alle Stichproben pro Messzeitpunkt angegeben. Für alle Subskalen mit Ausnahme der Subskala „(stochastische) Unabhängigkeit verstehen“ lässt sich in mindestens einer Stichprobe zu mindestens einem Messzeitpunkt eine negative durchschnittliche oder Median-Korrelation beobachten. Lediglich für die Subskala „Bedeutung großer Stichproben verstehen“ zum zweiten Messzeitpunkt liegt die niedrigste durchschnittliche und Median-Korrelation nicht im negativen Bereich bei $r = .01$. Für die Subskala „(stochastische) Unabhängigkeit verstehen“ lassen sich konsistent über alle Stichproben und beide Messzeitpunkte im Durchschnitt und im Median mindestens kleine positive Effekte der Produkt-Moment-Zusammenhänge beobachten ($.13 < M_r$ bzw. $Md_r < .26$).

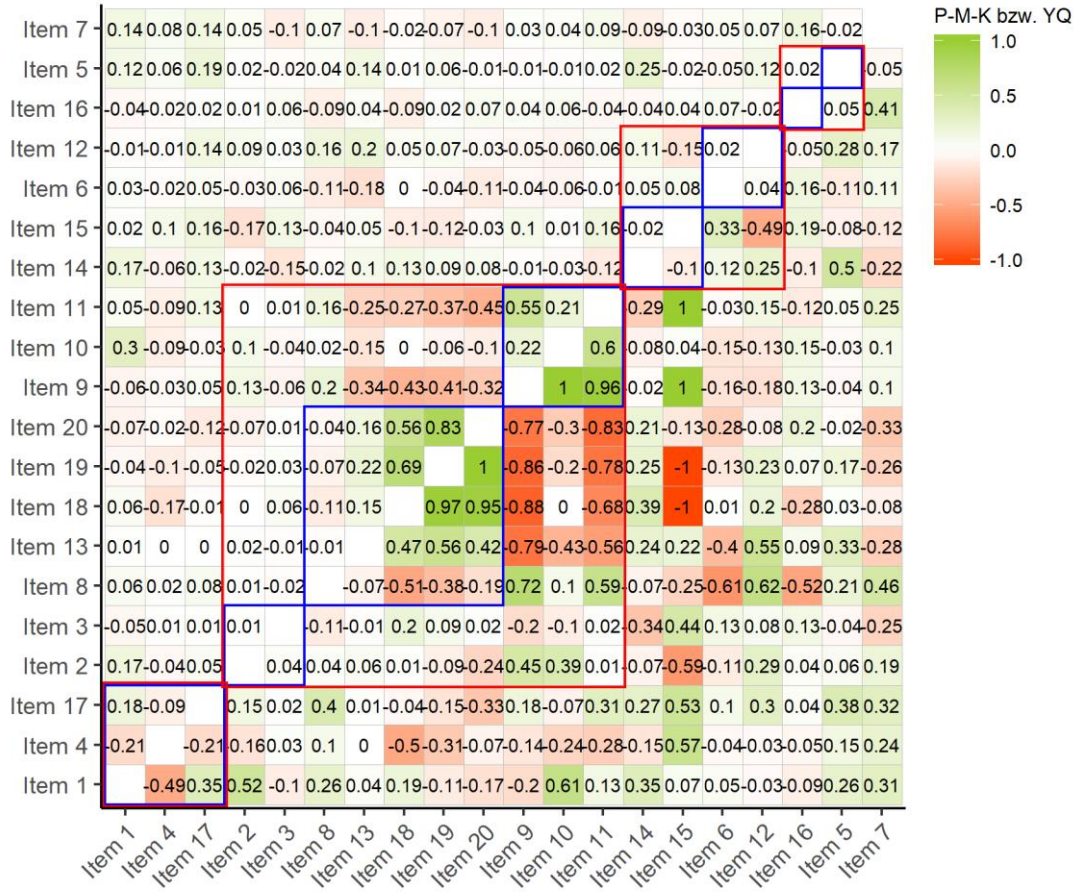


Abbildung 7. Iteminterkorrelationen für Stichprobe 1 (2 SWS, B.A.) mit Produkt-Moment-Korrelationskoeffizienten oberhalb der Diagonalen und Yules Q Koeffizienten unterhalb der Diagonalen am Ende des Semesters. Farblich umrandete Felder markieren zusammengehörige Items nach Subskalen (blau) und Typen des statistical reasoning (rot).

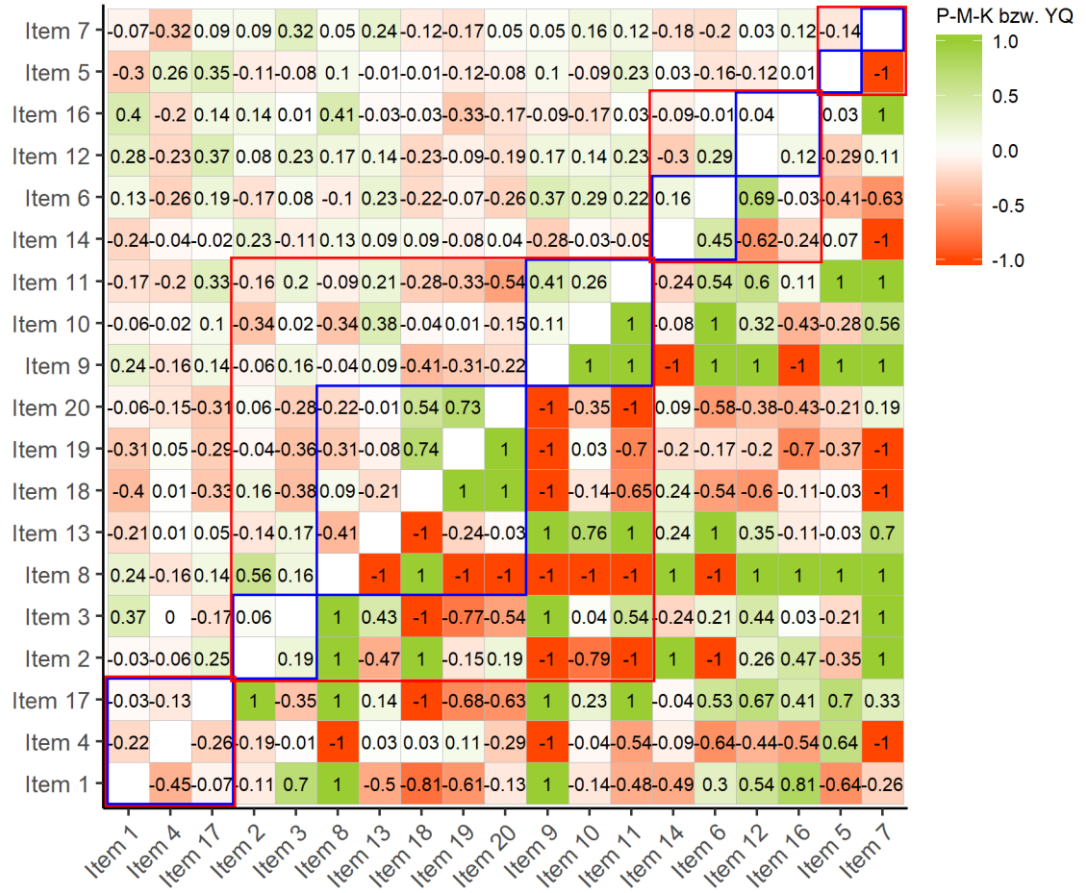


Abbildung 8. Iteminterkorrelationen für Stichprobe 3 (4 SWS, B.Sc.) mit Produkt-Moment-Korrelationskoeffizienten oberhalb der Diagonalen und Yules Q Koeffizienten unterhalb der Diagonalen am Ende des Semesters. Farblich umrandete Felder markieren zusammengehörige Items nach Subskalen (blau) und Typen des statistical reasoning (rot), Item 15 wurde von keiner Person gelöst, in der Folge können keine Korrelationen mit dem Item ermittelt werden.

Die Q-Koeffizienten zeigen über die Stichproben hinweg ebenfalls ein eher hypothesenkonträres Bild, viele der Koeffizienten sind negativ. In einigen Fällen lässt sich ein Q-Koeffizient von $Q = -1$ beobachten. Inhaltlich bedeutet dies, dass keine Person beide Items gelöst hat, jedoch das eine Item löste und das andere nicht und umgekehrt. Dies ist etwa für die Items 15 und 18 aus der Stichprobe 1 im zweiten Messzeitpunkt der Fall (Abbildung 7). Für die beiden Items gab es $n = 129$ Personen, die weder das Item 15 noch das Item 18 lösten. Insgesamt $n = 17$ Studierende lösten das Item 18 und $n = 13$ Studierende lösten das Item 15. Wenn man nun annimmt, dass beide Items eine konsistente Fähigkeit oder Fertigkeit

erfassen, würde man erwarten, dass (zumindest) ein Teil der Studierenden, die eines der beiden (als sehr schwer einzuordnenden) Items lösen, auch das andere Item korrekt beantwortet. Dies ist hier (und in allen anderen Fällen, in denen der Q-Koeffizient $Q = -1$ beträgt) jedoch nicht der Fall. Alle Studierenden, die das Item 15 korrekt beantwortet hatten, beantworteten das Item 18 falsch und alle Studierenden, die das Item 18 korrekt lösten, lösten das Item 15 nicht. Über alle Stichproben und beide Zeitpunkte kann man erkennen, dass diese, wenn man von dem Stichprobenfehlereinfluss absieht, unplausiblen Werte auch in den Interkorrelationen von Items, die einem Typ des statistical reasoning oder einer Subskala zugeordnet werden, zu finden sind. Konsistent über alle Stichproben hinweg zeigen sich positive und zum großen Teil hohe Korrelationen (sowohl in Produkt-Moment- als auch in Q-Koeffizienten) lediglich für die Items 18, 19 und 20 (diese sind Teil der Subskala „Wahrscheinlichkeiten korrekt berechnen“) sowie, in etwas schwächerem Ausmaß, für die Items 9, 10 und 11 (Subskala „[stochastische] Unabhängigkeit verstehen“).

Insgesamt ergibt daher auch die Inspektion der Iteminterkorrelationen ein Bild, das nur sehr eingeschränkt den in Fragestellung I formulierten Erwartungen entspricht.

Interne Konsistenz.

Cronbachs α (Cronbach, 1951) ist das am häufigsten eingesetzte Maß für die Bestimmung der internen Konsistenz (vgl. z. B. Döring & Bortz, 2016). Das Maß kann herangezogen werden, um die Rolle eines übergreifenden Faktors in den Varianzen der einer Skala zugeordneten Items zu quantifizieren (Cronbach, 1951). Hohe positive Werte weisen hierbei auf das Vorliegen eines zugrundeliegenden Faktors hin, negative Werte sind Hinweise auf negative Zusammenhänge zwischen den Items (wenn sichergestellt ist, dass die Items im Sinne des Konstrukts gleichgerichtet kodiert sind). In der Tabelle 9 sind die Cronbach α für die Gesamtskala statistical reasoning abgetragen (erstellt mit der Funktion *alpha* des Pakets *psych*, Revelle, 2019). Insgesamt zeigen sich hier, übereinstimmend mit den Iteminterkorrelationen, über die verschiedenen Stichproben und die beiden Messzeitpunkte hinweg niedrige und stellenweise negative Konsistenzen. Auch bei Weglassen einzelner Items aus der Skala bleiben die Werte sehr niedrig. Wenn die Items nach den Typen des statistical reasoning gruppiert werden, verbessert sich die

Güte der Konsistenz nicht: außer für den Typ „Beurteilen statistischer Kennzahlen“ zum ersten Messzeitpunkt ergibt sich in mindestens einer der Stichproben zu beiden Messzeitpunkten ein negatives Cronbach α (Tabelle 10). Auch hier zeigt sich eine nur geringe Erhöhung der Konsistenzen, wenn ein Item aus der Typenskala entfernt wird. In Übereinstimmung mit den Iteminterkorrelationen zeigt sich für die internen Konsistenzen der Subskalen lediglich für „Berechnen von Wahrscheinlichkeiten“ (Items 8, 13, 18, 19 und 20) und „stochastische Unabhängigkeit“ (Items 9, 10 und 11) mindestens in einer Stichprobe zu beiden Zeitpunkten ein befriedigendes oder hohes Cronbach α (Tabelle 11).

Zusammenfassend liefern die durchgeführten Analysen der Beziehungen zwischen den Items wenig Bestätigung der postulierten ein-, vier- oder achtdimensionalen Struktur.

Tabelle 9

*Cronbac*h α, die höchsten Cronbach α be*i Weglassen der einzelnen Items der Skala sowie die durchschnittlichen und der Median der Korrelationen zwischen den Items der Gesamtskala des SRA zu Beginn und nach Ende der Intervention.*

		n					*r*	
S	Z	*Min*	*Max*	*k*	α	α_{max}	*M*	*Md*
1	1	263	275	19	.37	.40	.03	.02
	2	160	163	19	.24	.30	.01	.01
2	1	31	34	19	**-.80**	**-.45**	-.02	-.03
	2	28	32	19	.03	.19	-.01	-.02
3	1	31	31	19	.27	.34	.01	.00
	2	29	29	18	**-.24**	.00	-.01	-.03
4	1	49	51	19	.44	.49	.04	.04
	2	42	43	19	.42	.50	.03	.03

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., Z: Zeitpunkt mit 1 = zu Beginn der Intervention, 2 = am Ende der Intervention, *n*: kleinstes (*Min*) und größtes (*Max*) für die paarweisen Korrelationen, *k* = Anzahl der Items, α: Cronbachs α (negative Werte sind **fett** gesetzt), α_{max}: höchstes Cronbach α beim Weglassen einzelner Items (negative Werte sind **fett** gesetzt), *r*: Mittelwert (*M*) und Median (*Md*) der Iteminterkorrelationen der gesamten Skala.

Tabelle 10

Niedrigste und höchste Cronbach α, höchste Cronbach α bei Weglassen der einzelnen Items der jeweiligen Itemgruppe sowie durchschnittliche und Median der Korrelationen zwischen den Items für die nach Typen gruppierten Items des SRA zu Beginn und nach Ende der Intervention über alle Stichproben der Studierenden.

									r			
		n			α		α_{max}		M		Md	
Typ	Z	*Min*	*Max*	*k*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*
Kenn-	1	31	272	3	**-.04**	.20	.11	.36	-.04	.07	-.16	.05
zahlen	2	29	162	3	**-.52**	.10	**-.06**	.53	-.13	.03	-.18	.03
Stich-	1	31	271	4	**-.40**	.30	.13	.38	-.08	.06	-.08	.14
proben	2	28	163	3	**-.06**	.09	.13	.44	-.06	.05	-.14	.16
Unsicher-	1	31	275	10	**-.53**	.32	**-.10**	.41	-.04	.06	-.06	.03
heit	2	28	163	10	**-.05**	.25	.13	.39	-.01	.03	-.04	.01
Zusammen-	1	31	270	2	**-.30**	.12	.00	.07	-.13	.07	-.13	.07
hänge	2	29	161	2	**-.84**	.04	.00	.09	-.30	.02	-.30	.02

Anmerkungen. Z: Zeitpunkt mit 1 = zu Beginn der Intervention, 2 = am Ende der Intervention, Kennzahlen: „Beurteilen statistischer Kennzahlen“, Stichproben: „Beurteilen (der Güte) der Stichproben“, Unsicherheit: „Schließen unter Unsicherheit“, Zusammenhänge: „Beurteilen von Zusammenhängen“, *n*: kleinstes (*Min*) und größtes (*Max*) für die paarweisen Korrelationen, *k* = Anzahl der Items, α: kleinstes (*Min*) und größtes (*Max*) Cronbach α innerhalb der vier Stichproben von Studierenden (negative Werte sind **fett** gesetzt), α_{max}: kleinstes (*Min*) und größtes (*Max*) höchstes Cronbach α beim Weglassen einzelner Items (negative Werte sind **fett** gesetzt), *r*: kleinster (*Min*) und größter (*Max*) Mittelwert (*M*) und Median (*Md*) der Iteminterkorrelationen der Skala.

Tabelle 11

Niedrigste und höchste Cronbach α, höchste Cronbach α bei Weglassen der einzelnen Items der jeweiligen Itemgruppe sowie durchschnittliche und Median der Korrelationen zwischen den Items für die nach Typen gruppierten Items des SRA zu Beginn und nach Ende der Intervention über alle Stichproben der Studierenden.

									r			
		n			α		α_{max}		M		Md	
Subskala	Z	*Min*	*Max*	*k*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*
zentrale	1	31	272	3	**-.04**	.20	.11	.36	-.04	.07	-.16	.05
Tendenz	2	29	162	3	**-.52**	.10	**-.06**	.53	-.13	.03	-.18	.03
Berechnen v.	1	31	275	5	**-.28**	.59	.50	.66	-.06	.22	-.02	.15
Wahrsch.	2	28	163	5	.44	.67	.71	.80	.09	.29	-.05	.22
Interpretieren	1	31	274	2	**-.07**	.23	.00	.16	-.03	.16	-.03	.16
v. Wahrsch.	2	29	162	2	**-.18**	.09	.00	.06	-.09	.06	-.09	.06
Stichproben-	1	31	270	2	**-.66**	.36	.01	.22	-.26	.22	-.26	.22
größe	2	28	163	2	.03	.44	.01	.29	.01	.29	.01	.29
Unab-	1	31	272	3	.40	.60	.46	.82	.21	.32	.13	.27
hängigkeit	2	28	163	3	.38	.65	.49	.84	.26	.42	.19	.29
Variabilität	1	31	271	2	**-.30**	.33	.01	.26	-.13	.26	-.13	.26
	2	29	163	2	**-.14**	.06	.00	.03	-.08	.03	-.08	.03

Anmerkungen. zentrale Tendez: „Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird“, Berechnen v. Wahrsch.: „Wahrscheinlichkeiten korrekt berechnen“, Interpretieren v. Wahrsch.: „Wahrscheinlichkeiten korrekt interpretieren“, Stichprobengröße: „Bedeutung großer Stichproben verstehen“, Unabhängigkeit: „(stochastische) Unabhängigkeit verstehen“, Variabilität: „Variablilität innerhalb von Stichproben verstehen“, die Subskalen „zwischen Korrelation und Kausalität unterscheiden“ und „Vierfeldertafeln korrekt interpretieren“ enthalten jeweils nur ein Item, Z: Zeitpunkt mit 1 = zu Beginn der Intervention, 2 = am Ende der Intervention, *n*: kleinstes (*Min*) und größtes (*Max*) für die paarweisen Korrelationen, *k* = Anzahl der Items, α: kleinstes (*Min*) und größtes (*Max*) Cronbach α innerhalb der vier Stichproben von Studierenden (negative Werte sind **fett** gesetzt), α_{max}: kleinstes (*Min*) und größtes (*Max*) höchstes Cronbach α beim Weglassen einzelner Items, *r*: kleinster (*Min*) und größter (*Max*) Mittelwert (*M*) und Median (*Md*) der Iteminterkorrelationen der Skala. Umrandet sind Cronbach α-Werte ab $\alpha \gtrsim .70$.

6.1.2. Faktorielle Struktur des SRA.

Besonders bei eher niedrigen und/oder uneinheitlichen Zusammenhängen zwischen der relativ hohen Anzahl von Items können Faktorenanalysen zu besser interpretierbaren Ergebnissen führen als etwa die Inspektion von Iteminterkorrelationen (Bortz, 2005). Im Folgenden soll das Instrument daher mit Hilfe von explorativen und konfirmatorischen Faktorenanalysen untersucht werden.

Um die Frage nach der faktoriellen Struktur beantworten zu können, muss zunächst konkretisiert werden, welche Art der Datenstruktur im vorliegenden Fall von Interesse ist. Die Zusammenhänge zwischen den Items stellen die für eine konstruktorientierte Klärung der Frage nach der Anzahl der Faktoren am besten geeignete Strukturart dar. Die typischerweise hierfür eingesetzten Verfahren sind die Faktorenanalyse und Hauptfaktorenanalyse (die letzte stellt eine modifizierte Form der Hauptkomponentenanalyse dar, siehe dazu etwa Rietz, 1996, Fabrigar, Wegener, McCallum und Strahan, 1999 oder Brown, 2015). Die Hauptfaktorenanalyse und Faktorenanalyse unterscheiden sich vor allem darin, dass für Faktorenanalysen eine Reihe von Güteindizes angegeben werden können, die für die Hauptfaktorenanalysen nicht berichtet werden können (siehe auch hierzu z. B. Rietz, 1996, Brown, 2015 oder Eid et al. 2017). Für eine zuverlässige Schätzung von Faktoren und Faktorladungen durch die Faktorenanalyse (anhand der Maximum Likelihood Schätzung) müssen jedoch verschiedene Voraussetzungen der Daten vorliegen, die zentrale von den denen die multivariate Normalverteilung ist (Rietz, 1996, Eid et al., 2017, Brown, 2015). Bei den vorliegenden Daten ist diese Voraussetzung jedoch nicht erfüllt, da dichotome Indikatoren vorliegen. Alternativ wurde von Christoffersson (1975) und Muthén (1978) eine Schätzung nach dem Kriterium der nach Mittelwerten und Varianzen gewichteten kleinsten Quadrate (Weighted Least Squares Mean and Variance adjusted, WLSMV) vorgeschlagen, die zu Ergebnissen führt, die im einfaktoriellen Fall in Kennwerte der ein- oder zwei-Parameter logistischen Modelle (1PL bzw. 2PL) der Item Response Theorie überführt werden können (Brown, 2015). Eine weitere Möglichkeit der Schätzung von Faktoren bei nicht multivariat normalverteilten Indikatoren stellt das Weighted Least Squares (WLS) Verfahren dar (Eid et al., 2017).

Alternativ kann jedoch auch von Interesse sein, die Streuung in den Items auf möglichst sparsame Weise zu reduzieren, ohne zwischen Konstrukt und

einzigartigem Varianzanteil einer Variablen (uniqueness) und Messfehler zu unterscheiden. Da die Items auch als (Lehrziel-) Kriterien verstanden werden können und zudem die dimensionsanalytische Reduktion der Items anhand der Zusammenhänge einerseits und der Streuung der einzelnen Items andererseits unterschiedliche Ergebnisse in der Anzahl der Faktoren bzw. Komponenten hervorbringen können (siehe z. B. Rietz, 1996), wird auch eine Hauptkomponentenanalyse als ein mögliches sinnvolles Verfahren für die vorliegende Fragestellung angesehen.

Wie im Abschnitt zu Iteminterkorrelationen bereits erörtert wurde, kann es eine wichtige Rolle spielen, welches Zusammenhangsmaß für die Analyse von dichotomen Variablen gewählt wird. Da Items mit ähnlicheren oder gleichen Randverteilungen bei demselben Ausmaß des Zusammenhangs höhere Produkt-Moment-Korrelationen aufweisen, können sogenannte Schwierigkeitsfaktoren (Guilford, 1941) auftreten. Ein Lösungsvorschlag hierfür ist die Nutzung von Yule Q (z. B. Allerbeck, 1972) oder aber der tetrachorischen Koeffizienten (z. B. Christoffersson, 1975, Guilford, 1941). Tetrachorische Koeffizienten als Spezialfall polychorischer Korrelationen behandeln dichotome Variablen im Sinne von Schwellenwerten. Hierbei wird angenommen, dass ein zugrundeliegendes kontinuierliches Merkmal eine bestimmte Schwelle erreichen muss, damit die Ausprägung der manifesten Variablen den Wert 1 annehmen kann. In dem hier vorliegenden Fall kann daher auch für die Nutzung von tetrachorischen Korrelationen argumentiert werden: die zugrundeliegende (kontinuierlich verteilte) spezifische Fähigkeit oder Fertigkeit muss ein bestimmtes Ausmaß erreichen, damit das entsprechende Item gelöst werden kann.

Da zwischen den einzelnen Subskalen und Typen des statistical thinking sowohl positive als auch negative Korrelationen erwartet werden können (siehe Abschnitt 4.1), wird als Rotationsverfahren die oblique Oblimin-Rotation verwendet. Dieses Rotationsverfahren lässt Korrelationen beliebiger Höhe zwischen den Faktoren bzw. Komponenten zu und kann zu einer besseren Interpretation der Ladungsmatrizen führen.

Für die Analysen der faktoriellen Struktur der Daten werden daher Hauptkomponenten- und Faktorenanalysen jeweils mit Nutzung von Produkt-Moment-Korrelationsmatrizen und tetrachorischen Korrelationsmatrizen mit einer

WLS- (in explorativen Faktorenanalysen) bzw. WLSMV- (in konfirmatorischen Faktorenanalysen) Schätzung und zugelassenen Korrelationen zwischen den Komponenten bzw. Faktoren herangezogen. Die explorativen Hauptkomponenten- bzw. Faktorenanalysen wurden mit dem Paket *psych* (Revelle, 2018) und die konfirmatorischen Faktorenanalysen mit dem Paket *lavaan* (Rosseel, 2012) im Programm R (R Core Team, 2019) berechnet. Für die explorativen Hauptkomponentenanalysen wurde die Funktion *principal*, für die explorativen Faktorenanalysen die Funktion *fa* und für die konfirmatorischen Faktorenanalysen die Funktion *cfa* genutzt.

Exploration der Anzahl von Faktoren.

Laut Hypothese 1a) stellt sich zunächst die Frage nach der Anzahl der Faktoren, die die empirische Datenstruktur am besten wiedergeben. Für die Bestimmung der Anzahl der Faktoren wurden die gängigen Kriterien für Hauptkomponentenanalysen herangezogen (z. B. Brown, 2015, Eid et al., 2017). Dies sind im Einzelnen der Scree-Test (Cattell, 1966, zitiert nach Eid et al., 2017), das Kaiser-Guttman Kriterium (oder Kaiser-Kriterium oder Guttman-Kaiser Kriterium, Guttman, 1954 und Kaiser, 1960, zitiert nach Fabrigar et al., 1999) und die Parallelanalyse (Horn, 1965). Ergebnisse der Parallelanalysen sowie die entsprechenden Screeplots finden sich im Anhang in Abbildung 32 bis Abbildung 35.

Für die explorativen Faktorenanalysen wurden absolute Abweichungsindizes (χ^2-Statistik, Root Mean Sqare Error of Approximation, RMSEA und die Summe der Residuenquadrate im Verhältnis zu den Freiheitsgraden, RMS, Revelle, 2018) sowie komparative Indizes (Comparative Fit Index, CFI und Tucker Lewis Non-Normed Index, TLI), die sich im vorliegenden Fall alle auf das Nullmodell der vollständigen Unabhängigkeit beziehen, herangezogen. Die bei den explorativen Faktorenanalysen vorgegebene Faktorenanzahl wurde von einem bis neun Faktoren variiert. Die Festlegung der maximalen Anzahl auf neun Faktoren erfolgte willkürlich als die höchste Anzahl der theoretisch sinnvoll zu erwartenden Faktoren plus eins. Eine Übersicht der Indizes je Messzeitpunkt für jede der vier Stichproben findet sich in Abbildung 36 bis Abbildung 39 und in Tabelle 41 im Anhang. Die Ergebnisse der verschiedenen Kriterien für die Bestimmung der optimalen Anzahl der Faktoren bzw. Komponenten wurden kombiniert und auf dieser Basis die

mehrheitlich als optimal zu betrachtende Anzahl der Faktoren bzw. Komponenten gewählt. Das konkrete Vorgehen bestand darin, dass zunächst die häufigste ermittelte Anzahl der Faktoren bzw. Komponenten identifiziert wurde. So ergaben das Scree- und das Kaiser-Guttman-Kriterium für die Stichprobe 1 zu Beginn des Semesters acht Komponenten bei der Hauptkomponentenanalyse auf Basis der Produkt-Moment-Korrelationsmatrix (Abbildung 38). In dem Fall, wenn sich keine mehrheitliche Anzahl an Faktoren bzw. Komponenten anhand der Kriterien identifizieren ließ und eine der ermittelten Anzahlen im Sinne der Hypothesen war (ein/e, vier oder acht Faktoren bzw. Komponenten), wurde diese als ermittelte Faktoren- bzw. Komponentenanzahl gewählt. In der Stichprobe 2 zu Beginn des Semesters zeigten das Scree-, das Kaiser-Guttman und das Parallelanalysenkriterium etwa unterschiedliche Komponentenanzahlen an (drei bzw. sechs Komponenten, sieben Komponenten und eine Komponente, siehe Abbildung 38). In diesem Fall wurde das Ergebnis im Sinne einer ein-Komponenten-Struktur interpretiert. Lag für die ermittelte Anzahl der Faktoren bzw. Komponenten laut einzelner Kriterien kein Modalwert vor und waren mehrere (oder keine) ermittelte Anzahlen hypothesenkonform, so wurde sowohl im hypothesenkonformen als auch im hypothesenkonträren Fall die sparsamste Lösung gewählt. In der Stichprobe 4 am Ende des Semesters, z. B., deuteten die unterschiedlichen Kriterien auf eine vier-, acht- und ein-Komponentenlösung im Rahmen der Hauptkomponentenanalyse anhand der Produkt-Moment-Korrelationsmatrix hin. In diesem Fall wurde die sparsamste Lösung (eine Komponente) gewählt. In der Stichprobe 3 am Ende des Semesters deuteten die Kriterien auf eine fünf-, sieben- und zwei-Komponentenlösung. Auch hierbei wurde die sparsamste Lösung (zwei Komponenten) gewählt (Abbildung 38). Die so gewählten Anzahlen der Komponenten bzw. Faktoren sind zusammengefasst in Tabelle 12 dargestellt.

Bei Übereinstimmung der empirischen Datenstruktur mit der in Abschnitt 4.1 abgeleiteten erwarteten Struktur der Daten sollten die explorativen Komponenten- bzw. Faktorenanalysen mehrheitlich auf eine ein-, vier- oder achtdimensionale Lösung hinweisen. Aus den Ergebnissen in Tabelle 12 geht hervor, dass die Hauptkomponentenanalysen der Produkt-Moment-Korrelationsmatrizen lediglich in einer Stichprobe und zu einem Messzeitpunk (Stichprobe 1 zu Beginn des Semesters) nach mehr als einem Kriterium auf eine hypothesenkonforme

Datenstruktur hinweisen (acht Komponenten). Faktorenanalysen der Produkt-Moment-Korrelationsmatrizen ergeben ebenfalls nur in einer Stichprobe zu einem

Tabelle 12

Übersicht der ermittelten Anzahl der Faktoren in Abhängigkeit von der genutzten Korrelation und des genutzten Verfahrens.

		PMK				TCK			
		PC		FA		PC		FA	
	n	K	Eind.	F	Eind.	K	Eind.	F	Eind.
Z 1									
S 1	277	8	2/3	(3)	2/8	8	2/3	(8)	2/8
S 2	34	(1)	1/3	(3)	4/8	(3)	1/3	(1);(4)	1/8
S 3	31	3	2/3	(4)	5/8	7	2/3	(5)	2/8
S 4	51	(8)	1/3	3	6/8	4	2/3	(6)	2/8
Z 2									
S 1	163	(4);(8)	1/3	(3)	2/8	(4)	1/3	(7)	2/8
S 2	32	(4)	1/3	/	/	4	2/3	(4)	2/8
S 3[a]	29	(2)	1/3	(4)	3/8	6	2/3	(1)	2/8
S 4	43	(1);(4);(8)	1/3	(2)	3/8	7	2/3	(3)	2/8

Anmerkungen. Z 1: zu Beginn des Semesters, Z 2: am Ende des Semesters, S 1 = 2 SWS, B.A., S 2 = 2 SWS, M.A., S 3 = 4 SWS, B.Sc., S 4 = 4(+4) SWS, B.Sc., PMK: Produkt-Moment-Korrelationsmatrix, TCK: Tetrachorische Korrelationsmatrix, PC: Hauptkomponentenanalyse, FA: Faktorenanalyse, K: Anzahl der ermittelten Komponenten, F: Anzahl der ermittelten Faktoren, Eind.: Eindeutigkeit der ermittelten Anzahl (dargestellt ist die ermittelte Anzahl über die genutzten Kriterien), für die Hauptkomponentenanalysen wurden das Screeplot-Kriterium, das Kaiser-Guttman-Kriterium und das Ergebnis der Parallelanalyse (insgesamt 3 Kriterien) genutzt, für die Faktorenanalyse wurden das Screeplot-Kriterium, das Kaiser-Guttman-Kriterium, das Ergebnis der Parallelanalyse, der Comparative-Fit-Index, der Tucker-Lewis-Non-Normed-Index, der Root Mean Square Error of Approximation, die Summe der Residuenquadrate im Verhältnis zu den Freiheitsgraden sowie der p-Wert der χ^2-Statistik, Werte in Klammern zeigen uneindeutige Anzahlen an und/oder deutliche Mängel in einigen der Kriterien (siehe Text).

[a] Item 15 wurde in dieser Stichprobe von keiner Person gelöst und wurde daher aus den Analysen ausgeschlossen.

Messzeitpunkt eine erwartete Faktorenanzahl nach mehr als der Hälfte der Kriterien (Stichprobe 3 zu Beginn des Semesters). Die Hauptkomponentenanalysen der tetrachorischen Korrelationsmatrizen ergeben in drei Stichproben-Messzeitpunkt-Kombinationen eine erwartungsgemäße Komponentenanzahl (acht in Stichprobe 1 zu Beginn des Semesters und jeweils vier in Stichprobe 4 zu Beginn des Semesters und Stichprobe zwei am Ende des Semesters). Die Faktorenanalysen der tetrachorischen Korrelationsmatrizen ergeben für alle Stichproben und Zeitpunkte hinsichtlich aller Kriterien eine nur sehr schlechte Modellpassung, so dass eine eindeutige Interpretation hinsichtlich der optimalen Faktorenanzahl schwierig erscheint. Werden die Ergebnisse der Screeplot-, Kaiser-Guttman- sowie Parallelanalysenkriterien herangezogen, so zeigen sich für drei Stichproben-Messzeitpunkt-Kombinationen erwartungskonforme Anzahlen an Faktoren (acht Faktoren in Stichprobe 1 zu Beginn des Semesters, vier Faktoren in Stichprobe 2 am Ende des Semesters sowie ein Faktor für Stichprobe 3 am Ende des Semesters) laut mindestens zwei der Kriterien.

Konsistent kommen die verschiedenen Verfahren (Hauptkomponenten- und Faktorenanalysen anhand der unterschiedlichen Korrelationsmatrizen) lediglich in Stichprobe 1 zu Beginn des Semesters und in Stichprobe 2 am Ende des Semesters zu einer erwartungskonformen Dimensionsanzahl. Insgesamt finden sich daher, auch bei Nutzung verschiedener faktorenanalytischer Techniken, eher schwache Hinweise für die theoretisch erwartete Anzahl der Komponenten bzw. Faktoren. Darüber hinaus ist festzuhalten, dass die explorativ ermittelte Anzahl der Komponenten bzw. Faktoren über die verschiedenen Substichproben die faktorenanalytischen Techniken als uneinheitlich eingeschätzt werden kann.

Schließlich ist zu beachten, dass bei der Durchführung der meisten Analysen zum Teil gravierende statistisch-mathematische Probleme auftraten. Diese betrafen zum einen die Simulation und das Resampling der Datenstrukturen (zur Simulierung der Eigenwerte) im Rahmen der Parallelanalysen und deuten damit auf uneinheitliche und widersprüchliche Muster der Streuungen der bzw. Zusammenhänge zwischen den Items hin. Zum anderen traten Ultra-Heywood-Fälle (negative Varianzen oder Korrelationen mit Werten größer als $|\pm 1|$), nicht-positiv-definite Matrizen, leere Zellen bei tetrachorischen Korrelationsmatrizen als Grundlage der Schätzung sowie Konvergenzprobleme innerhalb der

Faktorenanalysen auf (detaillierte Übersicht im Anhang, S. 296). Lediglich für die Stichprobe 1 und nur bei der Analyse der Produkt-Moment-Korrelationsmatrix mittels beider Verfahren (Hauptkomponenten- und Faktorenanalyse) zu beiden Messzeitpunkten zeigten sich keine problematischen Auffälligkeiten.

Exploration der Ladungsmuster.

Laut Hypothese 1b) stellt sich anschließend die Frage nach dem Ladungsmuster der Items auf die ermittelten Komponenten bzw. Faktoren. Zwar gibt es wenige Hinweise auf die hypothesenkonforme Anzahl von Komponenten/Faktoren, es ist jedoch denkbar, dass sich zumindest zum Teil relativ klare und erwartungsgemäße Komponenten bzw. Faktoren in den Daten wiederfinden. Tabelle 13 und Tabelle 14 zeigen die Ladungsmatrizen der einzelnen Items auf die oblique rotierten Komponenten bzw. Faktoren für die Stichprobe 1 zu Beginn und am Ende des Semesters. Die Ladungsmatrizen für die Stichproben 2 bis 4 finden sich im Anhang in Tabelle 42, Tabelle 43 und Tabelle 44.

In Stichprobe 1 zeigt sich konsistent über alle faktoriellen Analysen und für beide Zeitpunkte eine Komponente bzw. ein Faktor, auf den die Items 18, 19 und 20 hoch positiv laden. Ebenso laden die Items 9, 10 und 11 auf eine weitere Komponente bzw. einen weiteren Faktor zu Beginn des Semesters in den beiden achtdimensionalen Lösungen. Besondere Aufmerksamkeit verdienen die ähnlichen Ladungsmuster für die beiden Itemgruppen in den Dimensionsanalysen anhand der verschiedenen Zusammenhangsmaße trotz der berichteten statistisch-mathematischen Probleme bei der Analyse der tetrachorischen Korrelationsmatrizen. Die Ladungen der beiden Itemgruppen scheinen daher robust zu sein. Items 9 und 11 laden am Ende des Semesters relativ konsistent deutlich negativ auf den ersten, durch Items 18, 19 und 20 markierten Faktor (bzw. die Komponente). Dieses Muster findet sich auch in der dreifaktoriellen Lösung zu Beginn des Semesters. Für die anderen Items lässt sich kein eindeutiges Ladungsmuster beobachten. Items 1, 12 und 15 etwa laden zu Beginn des Semesters auf eine gemeinsame Komponente, jedoch lässt sich das Ergebnis am Ende des Semesters in den Daten nicht mehr beobachten. Items 17, 13 und 14 markieren in der Hauptkomponentenanalyse ebenfalls eine Komponente zu Beginn, nicht jedoch am Ende des Semesters.

Tabelle 13

Übersicht der Ladungen auf Komponenten bzw. Faktoren für Stichprobe 1 zu Beginn des Semesters.

	PCA								FA (TCK)								FA (PMK)		
	K1	K2	K3	K4	K5	K6	K7	K8	F1	F2	F3	F4	F5	F6	F7	F8	F1	F2	F3
1	-.05	-.09	.54	.13	.21	.21	-.01	-.38	-.03	-.05	.38	.13	.10	.10	.03	-.17	-.03	.03	.2
4	-.10	-.02	.06	0	.6	.04	.27	.23	-.09	-.03	.05	.39	.02	0	.13	.10	-.05	.03	.12
17	.03	.20	-.1	.74	.16	.04	-.04	.06	0	.21	0	.15	.42	.01	-.03	.07	-.13	-.04	.22
2	.05	.04	.07	.12	.06	-.04	-.07	.82	.12	.01	0	.08	.14	-.02	-.05	.4	.06	-.04	.05
3	.17	.09	.05	.08	.03	.78	.04	-.13	.22	.08	.06	-.01	.06	.39	.06	-.06	.14	.08	-.03
8	.04	-.02	.05	-.02	.74	-.02	-.13	-.04	.04	-.04	.07	.48	-.02	-.04	-.09	-.01	0	-.07	.11
13	.09	-.44	-.08	.40	.20	-.09	.06	.11	.10	-.33	-.01	.16	.22	-.04	.06	.05	.22	.08	.03
18	.65	-.07	.04	-.17	.04	.17	.15	.14	.65	-.05	-.01	.03	-.08	.07	.08	.06	.55	.08	-.05
19	.91	.02	.02	.03	0	.06	-.03	.01	.78	.01	.03	0	.03	.01	-.02	.01	.83	-.05	.08
20	.86	-.07	-.06	.04	-.03	-.04	-.07	-.03	.72	-.07	-.03	-.03	.02	-.03	-.05	-.02	.79	-.03	-.01
9	-.04	.69	-.2	.17	.17	-.06	-.04	-.12	-.11	.57	-.09	.11	.06	-.04	-.05	-.10	-.38	-.16	.17
10	.05	.61	.21	-.01	-.13	.03	.24	-.01	.06	.62	.08	-.05	.03	-.01	.08	-.02	-.23	0	.16
11	-.12	.77	.05	-.03	-.06	.05	-.07	.17	-.13	.62	.01	-.04	-.01	.02	-.07	.09	-.42	-.16	.15
14	-.03	-.13	.18	.70	-.29	.03	.13	.07	-.04	-.06	.18	-.15	.43	.02	.12	.05	0	.10	.16
15	-.17	-.05	-.05	.10	-.20	.39	-.56	.05	-.17	-.04	-.05	-.15	.06	.22	-.24	.05	-.03	-.04	-.78
6	.03	.17	.2	.21	.04	-.52	-.08	-.35	-.02	.21	.17	.05	.11	-.27	-.06	-.16	-.17	-.12	.22
12	.16	.05	.68	.01	-.14	-.21	.01	.10	.19	.08	.41	-.06	.05	-.13	.03	.04	.06	-.03	.23
16	-.12	-.02	-.07	.10	-.08	.13	.78	-.06	-.13	-.01	-.03	-.04	.08	.07	.41	-.04	-.01	.97	.02
5	-.16	.03	.67	-.02	.16	.09	-.07	.11	-.14	.03	.42	.14	.03	.03	0	.05	-.12	-.02	.22

Anmerkungen. PC: Hauptkomponentenanalyse, FA: Faktorenanalyse, TCK: Tetrachorische Korrelationsmatrix als Grundlage, PMK: Produkt-Moment-Korrelationsmatrix als Grundlage, Anteil aufgeklärter Varianz beträgt 20% (FA PMK) bis 60% (PCA).

Tabelle 14

Übersicht der Ladungen auf Komponenten bzw. Faktoren für Stichprobe 1 am Ende des Semesters.

	PCA								FA (TCK)							FA (PMK)		
	K1	K2	K3	K4	K5	K6	K7	K8	F1	F2	F3	F4	F5	F6	F7	F1	F2	F3
1	-.06	-.11	.81	-.02	.13	.06	.00	-.11	.07	.43	.27	.00	-.07	.07	-.10	-.02	-.02	.23
4	-.28	-.23	-.45	-.08	.07	.11	.26	.13	-.24	-.36	.00	.08	.09	-.06	.21	-.06	-.02	-.14
17	-.04	.07	.24	.45	.18	.37	-.19	.00	-.09	.08	.31	.25	-.03	.24	-.09	-.09	.01	.12
2	-.16	-.11	.40	.21	-.17	-.49	.09	.22	-.12	.34	.00	-.32	-.01	.21	.08	-.05	.01	.11
3	.07	-.10	.03	.30	-.64	.29	-.12	.12	.15	.01	-.30	.25	-.08	.28	.04	.00	.00	-.82
8	.09	.45	-.12	.42	.01	.01	.32	-.22	-.18	-.03	.06	-.03	.44	.16	-.17	-.13	-.08	.10
13	.01	-.61	.05	.26	.03	.12	.40	.04	.37	-.14	.18	.05	.14	.21	.18	.31	.00	.03
18	.79	-.07	.17	.05	-.04	-.01	-.10	-.14	.80	.10	.03	-.02	-.05	.04	-.10	.70	-.09	.03
19	.92	-.04	-.01	.06	.01	-.03	.03	.06	.80	.02	.01	-.05	.03	.03	.02	.90	.01	.04
20	.89	-.02	-.11	-.08	.04	.03	.11	.10	.77	-.02	-.03	.04	.08	-.09	.06	.81	.06	-.01
9	-.23	.74	-.03	.03	.07	.01	.09	.16	-.57	.26	-.03	.04	.16	-.07	-.06	-.57	.07	.12
10	.08	.34	.58	-.25	-.05	-.03	.20	.18	-.07	.57	-.03	.00	.08	-.10	-.03	-.15	.07	.16
11	-.23	.64	.12	.17	-.09	.18	.02	-.05	-.49	.21	-.04	.19	.14	.06	-.16	-.55	-.02	.08
14	.10	-.03	.10	.11	.73	.03	-.17	.00	.11	.02	.52	-.02	-.06	-.05	-.04	.14	-.03	.26
15	-.05	.01	.04	-.11	-.09	.84	.03	.05	-.12	-.03	.02	.57	-.05	-.03	.05	-.13	.06	-.15
6	-.07	-.01	-.05	.05	.03	.01	-.83	.09	-.17	-.05	.04	.06	-.50	.04	-.06	-.05	.06	-.08
12	.02	.01	-.09	.77	.01	-.24	-.02	.00	.00	-.10	.14	-.16	.15	.42	-.03	.06	-.05	.09
16	.07	.05	-.04	.00	.00	.01	-.07	.89	-.03	.21	-.03	.10	-.16	.02	.34	.00	1.00	.00
5	-.03	-.09	.07	.32	.52	.15	.15	.23	-.02	.01	.45	.06	.10	.12	.11	.04	.01	.16

Anmerkungen. PC: Hauptkomponentenanalyse, FA: Faktorenanalyse, TCK: Tetrachorische Korrelationsmatrix als Grundlage, PMK: Produkt-Moment-Korrelationsmatrix als Grundlage, Anteil aufgeklärter Varianz beträgt 25% (FA PMK) bis 65% (PCA).

In der Stichprobe 2 (Tabelle 42) zeigen sich für die Items 18, 19 und 20 ebenfalls konsistent hohe Ladungen auf einer Komponente/einen Faktor. Hier lädt zusätzlich das Item 8 mehrheitlich hoch mit entgegengesetztem Vorzeichen auf derselben Komponente/demselben Faktor. Dieses Ergebnis ist hypothesenkonträr, da das Item 8 derselben Subskala zugeordnet ist, wie die Items 18, 19 und 20. Für die Items 9, 10 und 11 lassen sich auch hier mehrheitlich Ladungen mit entgegengesetztem Vorzeichen auf die Dimension beobachten, der Items 18, 19 und 20 angehören. Für die anderen Items lässt sich kein eindeutiges hypothesenkonformes Muster erkennen.

Für die Stichprobe 3 (Tabelle 43) lässt sich ebenfalls ein überwiegend einheitliches Ladungsmuster für die Items 18, 19 und 20 in konsistenter Weise beobachten: die Itemgruppe lädt moderat bis hoch auf eine Dimension. Für den ersten Messzeitpunkt ist darüber hinaus zu beobachten, dass auch Item 13 in ähnlicher Höhe und mit demselben Vorzeichen auf die entsprechende Komponente/den Faktor lädt. Item 9 und 11 laden auf diese Komponente/diesen Faktor konsistent über alle Verfahren und beide Messzeitpunkte hoch mit entgegengesetztem Vorzeichen. Die Items 6 und 12 laden konsistent positiv auf eine Komponente/einen Faktor. Diese Dimension lässt sich inhaltlich jedoch nicht im Sinne der Hypothesen interpretieren, da zu Beginn des Semesters und am Ende des Semesters verschiedene weitere Items auf diese laden (Item 5 zu Beginn und z. B. Items 9, 10 und 11 am Ende des Semesters) trotz der gleichen Dimensionsanzahl und des gleichen Verfahrens (Faktorenanalysen anhand Produkt-Moment-Korrelationsmatrizen). Hierbei muss allerdings beachtet werden, dass Item 15 aus der Analyse am Ende des Semesters wegen fehlender Streuung (das Item wurde von keiner Person gelöst) ausgeschlossen werden musste und in der Folge die Zusammenhangsstrukturen nicht ohne Einschränkungen verglichen werden können.

Die Ladungen in der Stichprobe 4 (Tabelle 44) sind für die Items 18, 19 und 20, wie auch in den anderen Stichproben, hoch und im Vorzeichen identisch einer Komponente/einem Faktor klar zuzuordnen. Eine zweite Komponente/ein zweiter Faktor lässt sich eingeschränkt durch die moderaten bis hohen Ladungen der Items 9 und 11 erkennen. Auf diese Komponente/diesen Faktor laden in überwiegend uneinheitlicher Weise auch andere Items.

Insgesamt lassen sich lediglich die Items 18, 19 und 20 zum einen sowie in stärker eingeschränktem Maß die Items 9 und 11 zum anderen über alle Stichproben hinweg relativ eindeutig einer Dimension zuordnen. Items 18, 19 und 20 stellen einen Teil der Subskala „Wahrscheinlichkeiten korrekt berechnen“ und Items 9 und 11 einen Teil der Subskala „(stochastische) Unabhängigkeit verstehen“ dar. Items 8 und 13, die ebenfalls der Subskala „Wahrscheinlichkeiten korrekt berechnen“ zugeordnet sind, laden inkonsistent und teilweise mit entgegengesetztem Vorzeichen auf die durch Items 18, 19 und 20 markierte Dimension. Item 10, das mit den Items 9 und 11 zur Subskala „(stochastische) Unabhängigkeit verstehen“ gehört, lädt dagegen mehrheitlich nur sehr schwach (nahe 0) auf die durch Items 9 und 11 markierte Dimension. Damit zeigen sich trotz der statistisch-mathematischen Herausforderungen bei der Dimensionsanalyse robuste Ergebnisse für die Items 18, 19 und 20 sowie 9 und 10, nicht jedoch für die anderen Items des SRA. Dieses Ergebnis kann als nur zu einem geringen Teil übereinstimmend mit der Hypothese 1b) gewertet werden.

Konfirmatorische Faktorenanalysen.

Für die konfirmatorische Prüfung der Datenstruktur laut Hypothese 1c) wurde zunächst ein Nullmodell auf Basis der Ergebnisse aus den explorativen Faktorenanalysen formuliert. Hier hatten sich jeweils drei bzw. zwei der Items als einer Dimension zugehörig erwiesen (Items 18, 19 und 20 sowie Items 9 und 11). Nach einer inhaltlichen Analyse wurden die ähnlichen Aufgabenkontexte der Items als Erklärung für die relativ engen Zusammengehörigkeit vermutet. Die Items 18, 19 und 20 enthalten das Werfen von Würfeln, die Items 9 und 11 sowie 10 enthalten Münzwürfe im Aufgabenstamm. Dementsprechend wurde ein Nullmodell mit insgesamt drei Faktoren formuliert: einem „Würfelaufgaben“-Faktor, dem Items 18, 19 und 20, einem „Münzwurfaufgaben“-Faktor, dem Items 9, 10 und 11 und einem „Restitems“-Faktor, dem alle übrigen Items zugeordnet wurden. Die konfirmatorischen Faktorenanalysen wurden durchgeführt mit der Funktion *cfa* aus dem Paket *lavaan* (Rosseel, 2012) unter Verwendung des WLSMV-Schätzers.

Das Nullmodell wurde dann den laut Hypothesen zu erwartenden Modellen gegenübergestellt. Dies waren das einfaktorielle, vierfaktorielle sowie achtfaktorielle Modell mit einer entsprechenden Itemzuordnung (siehe Abbildung 3). In Tabelle 15 und Tabelle 16 sind die Ergebnisse der konfirmatorischen Faktorenanalysen

abgetragen. Wie aus den Übersichten erkennbar ist, sind die Modelle lediglich in sieben von 16 Fällen (Beginn des Semesters) bzw. vier von 16 Fällen (Ende des Semesters) konvergiert. Die leeren Zeilen der Tabellen zeigen an, dass für die entsprechenden Stichproben keine mathematische Lösung für die Strukturgleichungen gefunden werden konnte.

Tabelle 15

Ergebnisse der konfirmatorischen Faktorenanalysen für den ersten Messzeitpunkt (Beginn des Semesters).

S	Modell	χ^2	df	p	CFI	TLI	RMSEA	N
1	Nullmodell	178.43	149	.05	0.92	0.91	0.03	239
	1 Faktor	259.44	152	< .001	0.70	0.66	0.05	
	4 Faktoren							
	8 Faktoren							
2	Nullmodell							31
	1 Faktor	169.60	152	0.16	0.85	0.83	0.06	
	4 Faktoren							
	8 Faktoren	133.48	126	0.31	0.94	0.91	0.04	
3	Nullmodell							31
	1 Faktor	173.09	152	0.12	0.79	0.77	0.07	
	4 Faktoren							
	8 Faktoren							
4	Nullmodell	189.99	149	0.01	0.72	0.68	0.08	49
	1 Faktor	205.68	152	0.002	0.64	0.59	0.09	
	4 Faktoren							
	8 Faktoren							

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., CFI: Comparative-Fit-Index, TLI: Tucker-Lewis-Non-Normed-Index, RMSEA: Root Mean Square Error of Approximation, Modelle mit fehlenden Angaben sind nicht konvergiert.

Zu Beginn des Semesters konnte ein laut Hypothesen erwartetes Modell mit dem Nullmodell für zwei Stichproben verglichen werden. Beide Vergleiche zeigen eine leichte Überlegenheit des Nullmodells in allen Gütekriterien. Für die Stichproben 2 und 3 lassen sich moderate bis hohe Güten für die

hypothesenkonformen Modelle berichten, allerdings treten bei der Gleichungslösung Probleme, wie nicht positiv-definite Varianz-Kovarianz-Matrizen, negative Eigenwerte und negative geschätzte Varianzen auf. Diese lassen sich als Hinweise auf eine mangelnde empirische Identifizierbarkeit des Modells in den Daten interpretieren.

Tabelle 16

Ergebnisse der konfirmatorischen Faktorenanalysen für den zweiten Messzeitpunkt (Ende des Semesters).

Z	S	M	χ^2	df	p	CFI	TLI	RMSEA	N
2	1	Nullmodell							153
		1 Faktor	170.57	152	.14	0.92	0.91	0.03	
		4 Faktoren							
		8 Faktoren							
	2	Nullmodell							34
		1 Faktor							
		4 Faktoren							
		8 Faktoren							
	3 [a]	Nullmodell	103.34	132	.97	1.00	1.34	< 0.01	29
		1 Faktor	107.81	135	.96	1.00	1.32	< 0.01	
		4 Faktoren							
		8 Faktoren							
	4	Nullmodell	118.64	149	.97	1.00	2.19	< 0.01	41
		1 Faktor							
		4 Faktoren							
		8 Faktoren							

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., CFI: Comparative-Fit-Index, TLI: Tucker-Lewis-Non-Normed-Index, RMSEA: Root Mean Square Error of Approximation, Modelle mit fehlenden Angaben sind nicht konvergiert.

[a] Item 15 wurde in dieser Stichprobe von keiner Person gelöst und wurde daher aus den Analysen ausgeschlossen.

Am Ende des Semesters konnte lediglich ein Modellvergleich zwischen dem formulierten Nullmodell und den erwarteten Modellen vorgenommen werden (Stichprobe 3). Hierbei zeigt sich, dass sich die Modelle in ihrer Güte praktisch nicht unterscheiden. Für die anderen Stichproben kann wegen nicht konvergierter Lösungen keine vergleichende Beurteilung vorgenommen werden. Zu berücksichtigen ist hierbei außerdem für die Analysen der Stichproben 3 und 4, dass statistisch-mathematische Probleme (negative geschätzte Varianzen bzw. Kovarianzen und Eigenwerte) die Aussagekraft der Ergebnisse deutlich reduzieren.

Insgesamt bleibt damit die Frage nach der Passungsgüte der hypothetisierten Modelle gegenüber einem rivalisierenden Nullmodell zwar größtenteils offen. Auf der Basis der durch die berichtete Analyse ermittelten Ergebnisse kann allerdings auch kein Beleg für die bessere Passung der ein- vier- oder achtdimensionalen Modelle laut der Darstellung in Abbildung 3 erbracht werden.

6.2. Fragestellung 2: Effekte der Lehrinterventionen

Für die Analyse der Effekte der Lehrinterventionen in den vier untersuchten Stichproben werden im Folgenden Unterschiede im Lösungsverhalten in allen sowie ausgewählten Items zwischen den Stichproben am Ende des Semesters und die durchschnittlichen Veränderungen im Lösungsverhalten im Lauf des Semesters bei ausgewählten Stichproben betrachtet. Zusätzlich werden Differenzen der Schwierigkeiten nach Itemgruppen verglichen. Zuvor sollen jedoch allgemeine Ergebnisse für die Lehrinterventionen berichtet werden.

Tabelle 17 zeigt die zentralen Tendenzen, Streuungsangaben und Differenzen für die im SRA erreichte Gesamtpunktzahl aller untersuchten Studierenden zu Beginn und am Ende des Semesters. In die Auswertung gingen damit alle zu jeweils einem Messzeitpunkt vorliegenden Messwerte ein, nicht nur die Messwerte von Personen, die an beiden Messzeitpunkten teilgenommen hatten. Aus der Darstellung wird deutlich, dass die durchschnittlich (und im Median) gelöste Anzahl der Items in keiner der Stichproben und zu keinem Zeitpunkt über 50% (d. h. zehn der 20 Items) hinausgeht. Weiterhin ist die geringe Streuung der Gesamtpunktzahl auffällig. Diese geht in keiner der Stichproben zu keinem Zeitpunkt über zweieinhalb Punkte von 20 möglichen (12.5%) hinaus. Übereinstimmend damit zeigt auch der Variationskoeffizient einen nur geringen

Tabelle 17

Uni- und bivariate Statistiken für die Gesamtpunktsumme im SRA nach Stichproben und Zeitpunkten.

S	Z	M	SD	Md	IQB		M_n	M_x	n	v	V	e	d_{min}	d_{max}	t	p [a]
1	1	9.58	2.37	10	8	11	2	16	267	1.19	25%	-0.05	-0.02	-0.02	-0.22	0.59
	2	9.53	2.17	9	8	11	4	16	162		23%					
2	1	9.56	1.37	9.5	9	10	6	12	34	1.52	14%	-0.84	-0.50	-0.61	-2.53	0.99
	2	8.71	1.70	9	8	9	5	13	28		19%					
3	1	8.55	2.32	8	7	9.5	4	13	31	2.01	27%	1.00	0.43	0.61	1.70	0.05
	2	9.55	1.64	10	8	10	7	14	29		17%					
4	1	9.92	2.42	10	8.5	11	5	16	51	1.05	24%	-0.53	-0.22	-0.22	-1.10	0.86
	2	9.40	2.36	9	8	11	5	17	43		25%					

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., Z: Zeitpunkt mit 1 = zu Beginn des Semesters, 2 = am Ende des Semesters, M_n = Minimum, M_x = Maximum, v = Varianzverhältnis, V = Variationskoeffizient; e = Differenz der Punktesumme zwischen Ende und Beginn des Semesters, d_{min} und d_{max} wurden anhand jeweils einer der beiden Standardabweichung der Stichprobe berechnet, [a] p = p-Wert für die Fläche zur rechten Seite des t-Werts auf der t-Verteilung. Es handelt sich um eine Mischung aus unabhängigen und abhängigen Daten (siehe Text).

Anteil der Standardabweichung an der mittleren Punktsumme im SRA mit maximal ca. 30% (Stichprobe 3 zu Beginn des Semesters). Der Vergleich der Streuungen zeigt ein ungefähr gleiches Verhältnis der Varianzen zu den beiden Messzeitpunkten für alle Stichproben mit Ausnahme der Stichprobe 3. Bei den Psychologiestudierenden im ersten Semester lässt sich eine deutliche Verringerung in der Unterschiedlichkeit des Lösungsverhaltens am Ende des Semesters feststellen. Überraschenderweise lässt sich für drei der vier Stichproben ein Abfall der durchschnittlich gelösten Itemanzahl im SRA beobachten: im Schnitt lösten alle Personen, die am Ende des Semesters an der Erhebung teilnahmen weniger SRA-Aufgaben als alle Personen, die zu Beginn des Semesters an der Erhebung teilnahmen. Der negative Unterschied zuungunsten des Semesterendes ist in zwei der Stichproben als kleiner (Stichprobe 4) bzw. moderater (Stichprobe 2) Effekt zu beurteilen. Da in die Bestimmung der Differenzen sowohl unabhängige Messwerte

(Personen, die entweder zu Beginn [$n = 159$ für Stichprobe 1] oder am Ende des Semesters [$n = 54$ für Stichprobe 1] an der Erhebung teilnahmen) als auch abhängige Messwerte (Personen, die sowohl zu Beginn als auch am Ende des Semesters an den Erhebungen teilnahmen, $n = 108$ für Stichprobe 1) eingingen, sollten die ermittelten standardisierten Effektgrößen und t- sowie p-Werte nur mit Vorsicht interpretiert werden.

6.2.1. Gruppenunterschiede.

Die Hypothese 2a) bildet die Frage nach Unterschieden im Lösungsverhalten zwischen den Studierendenstichproben in Abhängigkeit von der Behandlung spezifischer SRA-Inhalte ab. Für die Prüfung dieser Hypothese wurden zunächst zwei SRA-Itemgruppen gebildet. Eine Gruppe enthielt Items, für die zweifelsfrei entschieden werden konnte, dass deren Inhalte in einer oder mehreren Stichproben in der Statistiklehre behandelt worden sind (Items 2, 3, 8, 9, 10, 11, 13, 18, 19, 20, siehe Tabelle 4). Diese Items bildeten Inhalte ab, die bei den Studierenden der Psychologie (Stichproben 3 und 4) behandelt wurden. Eine zweite Gruppe enthielt lediglich Items, für die zweifelsfrei entschieden werden konnte, dass deren Inhalte in keiner der Stichproben explizit behandelt wurden (Items 1, 4, 6, 7, 12, siehe Tabelle 4). Da die Anzahl der Items unterschiedlich ist und die Punktesummen für die beiden Itemgruppen nicht unmittelbar verglichen werden können, wurden die Punktesummen in Prozent der maximal möglichen (d. h. X Punkte von 10 und Y Punkte von 5 möglichen) umgerechnet. Es sollte hierbei beachtet werden, dass die Gruppe der Items, deren Inhalte nicht explizit Gegenstand der Statistiklehre waren, lediglich halb so viele Items umfasst, wie die Itemgruppe, deren Inhalte nicht gelehrt wurden. Damit findet bei Vergleichen eine indirekte doppelte Gewichtung der Items, deren Inhalte nicht behandelt worden sind, statt. In Tabelle 18 und Abbildung 9 sind die uni- und bivariaten Statistiken sowie Wertepaare (Abbildung 9) für die Prozent gelöster Aufgaben der beiden Itemgruppen je Stichprobe abgetragen. Es wird hierbei ersichtlich, dass in drei der vier Stichproben prozentual mehr Items korrekt gelöst wurden, deren Inhalte in keiner der Stichproben behandelt worden sind. Während ein solcher Unterschied für die Studierenden der Sonderpädagogik und der Masterstudierenden mit der Hypothese 2a) vereinbar ist, erscheint dieser für die Stichprobe der Psychologiestudierenden im zweiten Semester kontraintuitiv. Es

könnte plausiblerweise erwartet werden, dass Studierende im Schnitt mehr Items korrekt lösen, deren Inhalte innerhalb der Lehrveranstaltung behandelt worden sind als Items, deren Inhalte nicht Gegenstand der Lehrveranstaltung waren. Lediglich bei einer der beiden Stichproben, bei denen ein Unterschied zugunsten der Items, deren Inhalte behandelt worden sind (Stichproben 3 und 4) erwartet werden kann, zeigt sich der Unterschied, wie erwartet: Studierende der Psychologie im ersten Semester lösen im Schnitt etwas mehr Items (3% bzw. $d_{diff} = 0.14$), deren Inhalte behandelt wurden. Weiterhin fällt auf, dass sich die Streuungen im Lösungsverhalten zwischen den Items, deren Inhalte in keiner Stichprobe behandelt wurden und Items, deren Inhalte nur bei Psychologiestudierenden behandelt worden sind, für alle Stichproben in großem Ausmaß unterscheiden (Varianzverhältnis zwischen $2.29 < \nu < 3.9$). Während ein solcher Unterschied für die Psychologiestudierenden intuitiv plausibel scheint, da eine größere Streuung im Lösungsverhalten von Inhalten, die nicht behandelt wurden, Unterschiede z. B. im Vorwissen reflektieren könnte, ist der Unterschied in den Varianzen bei den Sonderpädagogik- und Masterstudierenden damit nicht erklärbar. Es lässt sich in diesem Zusammenhang ferner festhalten, dass die Streuung der prozentual gelösten Aufgaben für die Itemgruppe „Inhalte in Stichproben 3 und 4 behandelt“ in allen vier Studierendenstichproben als sehr klein bewertet werden kann ($11.68\% < SD < 12.59\%$). Schließlich lässt sich kein nennenswerter linearer Zusammenhang zwischen den prozentual gelösten Aufgaben der beiden Itemgruppen bei den Studierenden beobachten ($r = .00$ für Stichprobe 1, $r = .09$ für Stichprobe 2, $r = -.01$ für Stichprobe 3 und $r = .09$ für Stichprobe 4, siehe auch Abbildung 9). Dies ist vor dem Hintergrund der nur äußerst eingeschränkten Zusammenhänge zwischen den Items (siehe 6.1.1 Iteminterkorrelationen.) nicht überraschend und mit den Erwartungen im Sinne der Lehrzielvalidität vereinbar.

Tabelle 18

Uni- und bivariate Statistiken für die Prozent gelöster Aufgaben in Items, deren Inhalte in der Lehre behandelt vs. nicht behandelt wurden nach Stichproben am Ende des Semesters.

S	BI	M	SD	n	v	V	e_{diff}	SD_{diff}	d_{diff}	n_{diff}	t	p [a]
1	1	47.75	12.59	160	2.87	26%	-9.88	24.83	-0.4	160	-5.02	< .001
	2	57.79	21.32	163		37%						
2	1	44.29	11.68	28	3.9	26%	-4.29	23.48	-0.18	28	-0.95	.35
	2	45.16	23.08	31		51%						
3	1	53.45	12.03	29	2.29	23%	3.1	21.89	0.14	29	0.75	.23
	2	50.34	18.22	29		36%						
4	1	49.76	11.79	42	3.27	24%	-5.48	23.6	-0.23	42	-1.49	.93
	2	55.35	21.31	43		38%						

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., BI: Behandlung der Inhalte mit 1 = in Stichproben 3 und 4 behandelt, 2 = in keiner Stichprobe behandelt, v = Varianzverhältnis, V = Variationskoeffizient; e_{diff} = mittlere Differenz zwischen Itemgruppen, deren Inhalte behandelt vs. nicht behandelt wurden über Personen, SD_{diff} = Standardabweichung der mittleren Differenzen, d_{diff} = standardisierte mittlere Differenz zwischen Itemgruppen, deren Inhalte behandelt vs. nicht behandelt wurden über Personen. Informationen über Quantile finden sich in Abbildung 9

[a] p = p-Wert für das obere Ende der t-Verteilung.

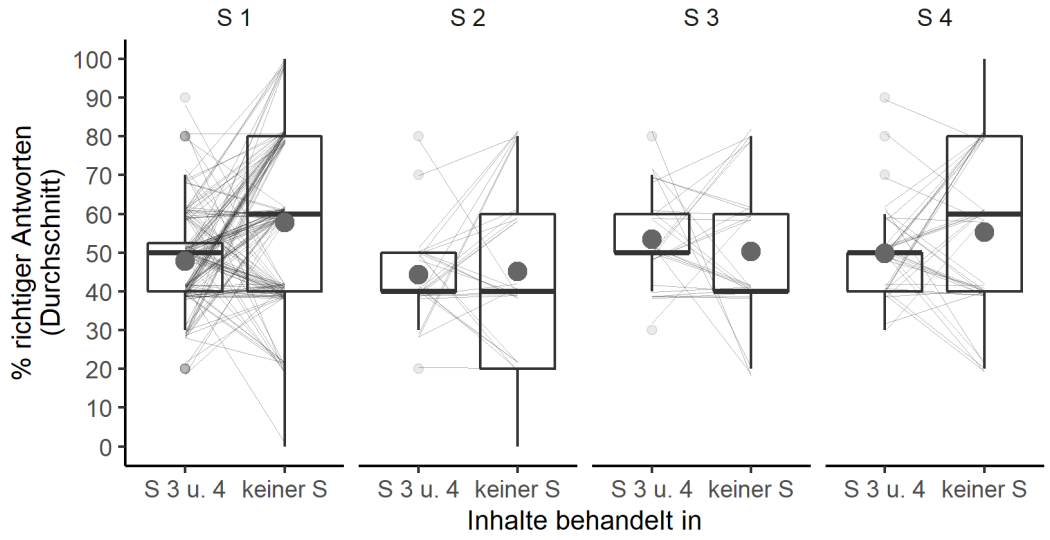


Abbildung 9. Boxplots, Mittelwerte (graue große Punkte) und verbundene Wertepaare (hellgraue dünne Linien) für die Prozent gelöster Aufgaben in Items, deren Inhalte in der Lehre behandelt (S 3 u. 4) vs. nicht behandelt (keiner S) wurden nach Stichproben (S 1 bis S 4) am Ende des Semesters.

Ein t-Test für unabhängige Stichproben des Unterschieds im Lösungsausmaß (prozentual gelöste Anzahl der Items, deren Inhalte nur bei den Psychologiestudierenden behandelt wurden) zwischen den Psychologiestudierenden einerseits und den Sonderpädagogik- und Masterstudierenden andererseits ergab $t(257) = -2.35$, $p = -.02$. Dabei beträgt der Unterschied $e = -4.03\%$ gelöster Aufgaben zugunsten der Psychologiestudierenden und stellt damit einen eher kleinen Effekt dar ($d = 0.33$, Cohen, 1988). In Abbildung 10 A ist der Effekt als dunkelgraue Linie abgetragen. Zur Gegenüberstellung des Unterschieds in Items, deren Inhalte behandelt wurden, dient die hellgraue gestrichelte Linie, die den Unterschied der Studierenden in Items, deren Inhalte nicht behandelt wurden darstellt. Auch hierfür findet sich ein kleiner Unterschied zwischen den beiden Stichprobengruppen ($e = -2.44\%$, $t[264] = -0.82$, zweiseitige $p = -.41$, $d \approx 0.12$) und die beiden Effektgrößen unterscheiden sich um eine Effektgröße kleinen Ausmaßes ($\Delta d \approx 0.21$). Während ein Unterschied in den Items zu behandelten Inhalten inhaltslogischen Erwartungen entspricht, ist der Unterschied zwischen Items, die in keiner der Stichproben behandelt wurden, erwartungskonträr. Es lässt sich jedoch zumindest ein erwartungsgemäßer kleiner Unterschied in den Mittelwertdifferenzen beobachten.

Werden Mediane statt Mittelwerte betrachtet (Abbildung 10 B), nivellieren sich die in den Mittelwerten beobachteten Unterschiede zwischen den Studierendengruppen. In beiden Studierendengruppen lässt sich im Median ein Lösungsvorteil des gleichen Ausmaßes für die Items beobachten, deren Inhalte in keiner Stichprobe behandelt wurden. Beide Studierendengruppen lösten im Median 10% mehr Items, deren Inhalte nicht Gegenstand der Lehre waren. Inhaltslogischen Überlegungen gemäß sollte jedoch ein Vorteil für die in den Stichproben 3 und 4 behandelte Inhalte in den Stichproben 3 und 4 beobachtet werden.

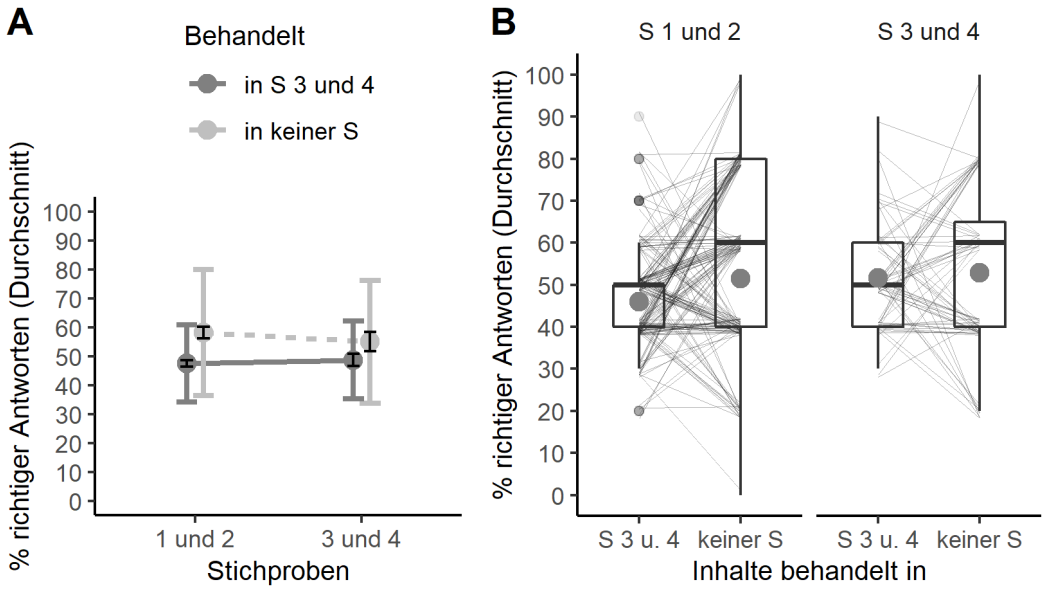


Abbildung 10. A: Interaktionsdiagramm für die Prozent gelöster Aufgaben in Items, deren Inhalte in der Lehre der Psychologiestudierenden behandelt (in S 3 und 4) vs. in keiner Stichprobe behandelt wurden für Sonderpädagogik- und Masterstudierende (Stichproben 1 und 2) und Psychologiestudierende (Stichproben 3 und 4), dicke schattierte Fehlerbalken stellen Standardabweichungen und dünne schwarze Fehlerbalken den einfachen Standardfehler dar. B: Boxplots, Mittelwerte (graue große Punkte) und verbundene Wertepaare (hellgraue dünne Linien) für die Prozent gelöster Aufgaben in Items, deren Inhalte in der Lehre der Psychologiestudierenden behandelt (S 3 u. 4) vs. in keiner Stichprobe behandelt wurden für Sonderpädagogik- und Masterstudierende (Stichproben 1 und 2) und Psychologiestudierende (Stichproben 3 und 4) am Ende des Semesters.

Da durch eine Gruppierung der Stichproben beim Vergleich der prozentual gelösten Aufgaben zu bei Psychologiestudierenden behandelten Inhalten mögliche differenzielle Unterschiede zwischen den einzelnen Stichproben (Psychologiestudierende am Ende des ersten Semesters vs. Psychologiestudierende am Ende des zweiten Semesters oder Sonderpädagogikstudierende vs. Masterstudierende) verdeckt werden, wurde zusätzlich eine einfaktorielle Varianzanalyse berechnet. Diese ergab mit $F(3; 255) = 2.99$, $p = .03$ eine kleine bis moderate partielle Varianzaufklärung von 3% ($\eta^2 = .03$, $f = .18$, Cohen, 1988, S. 283, 284-285). Die Unterschiede zwischen den Stichproben reflektierten hierbei die Erwartungen: Psychologiestudierende am Ende des ersten Semesters zeigten einen moderaten ($d = 0.46$) Vorteil gegenüber den Sonderpädagogikstudierenden und einen großen Vorteil ($d = 0.74$) gegenüber den Masterstudierenden. Der Vorteil für die Psychologiestudierenden am Ende des zweiten Semesters lag in geringerem Ausmaß vor ($d = 0.16$ gegenüber Sonderpädagogikstudierenden und $d = 0.44$ gegenüber Masterstudierenden). Da die Fehlervarianzen innerhalb der vier Stichproben als ähnlich beurteilt werden können (Levene-Statistik $F[3; 255] = 0.8$), erfolgte die Ermittlung der standardisierten Mittelwertdifferenzen anhand der Fehlerstandardabweichung ($SD_{Fehler} = 12.31\%$). Abbildung 11 zeigt die Unterschiede zwischen den Stichproben im Anteil gelöster Items, deren Inhalte bei Psychologiestudierenden behandelt wurden (dunkelgraue Linie) im Vergleich zum Anteil gelöster Items, deren Inhalte in keiner der Stichproben explizit behandelt worden sind (hellgraue gestrichelte Linie). Aus der Darstellung wird deutlich, dass die Ergebnismuster zweier Stichproben mit den inhaltslogischen Erwartungen übereinstimmen: Masterstudierende unterscheiden sich nicht im Anteil gelöster Items in beiden Itemgruppen. Da die Inhalte, die in beiden Itemgruppen erfasst werden, nicht Gegenstand der Lehre bei Masterstudierenden waren, ist dieses Ergebnis erwartungsgemäß. Studierende der Psychologie am Ende des ersten Semesters lösten anteilig mehr Items, deren Inhalte in dieser Stichprobe explizit behandelt worden sind. Auch dieses Ergebnis ist erwartungskonform. Für die beiden verbleibenden Stichproben wären die beobachteten Ergebnismuster in entsprechender Weise nach inhaltslogischen Überlegungen zu erwarten gewesen. So hätte das Ergebnismuster der Masterstudierenden auch bei Studierenden der Sonderpädagogik und das Ergebnismuster der Psychologiestudierenden am Ende des

ersten Semesters auch bei Psychologiestudierenden am Ende des zweiten Semesters beobachtet werden sollen. Die Betrachtung der Mediane der einzelnen Stichproben (Abbildung 9) bestätigt und verstärkt das anhand der Mittelwertvergleichen gewonnene Bild. Die Mediane der Masterstudierenden für die beiden Itemgruppen sind gleich, während sich die Mediane der Itemgruppen für Psychologiestudierende am Ende des Semesters um 10% gelöster Aufgaben zugunsten der behandelten Inhalte unterscheiden. Bei den Studierenden der Sonderpädagogik und der Psychologie am Ende des zweiten Semesters lässt sich umgekehrt ein Vorteil von 10% gelösten Aufgaben mehr zu Inhalten, die in keiner Stichprobe behandelten wurden beobachten.

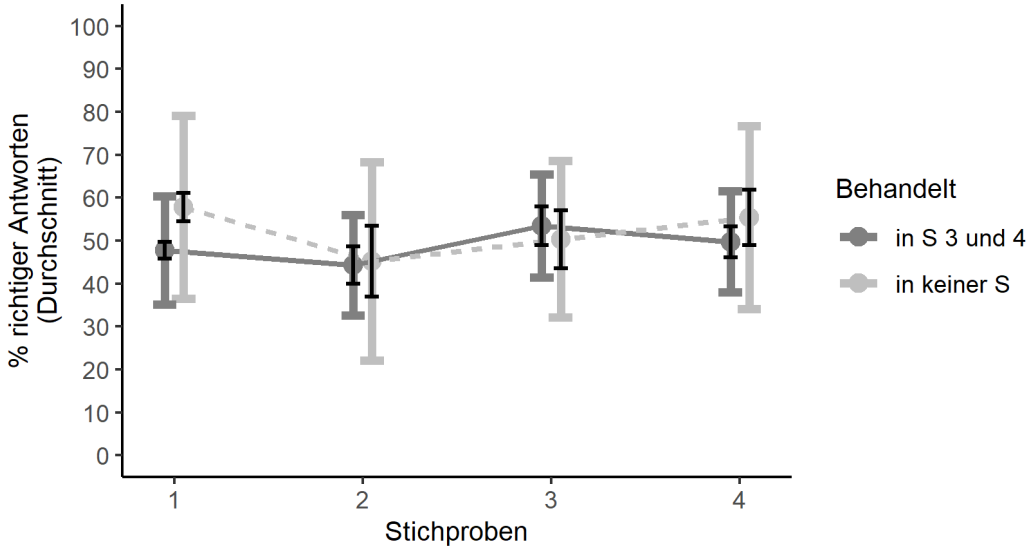


Abbildung 11. Interaktionsdiagramm für die Prozent gelöster Aufgaben in Items, deren Inhalte in der Lehre der Psychologiestudierenden behandelt (in S 3 und 4) vs. in keiner Stichprobe behandelt wurden für die untersuchten Stichproben (1: Sonderpädagogikstudierende, 2: Masterstudierende, 3: Psychologiestudierende im ersten Semester und 4: Psychologiestudierende im zweiten Semester), dicke schattierte Fehlerbalken stellen Standardabweichungen und dünne schwarze Fehlerbalken den einfachen Standardfehler dar.

Insgesamt kann die Hypothese 2a) damit teilweise bestätigt werden. Zwar lassen sich erwartungsgemäße durchschnittliche Unterschiede zugunsten der gelehrten Inhalte bei den Psychologiestudierenden beobachten, jedoch ist der Unterschied lediglich bei Psychologiestudierenden am Ende des ersten Semesters

deutlich ausgeprägt. Einschränkend muss zudem beachtet werden, dass der Vergleich der Mediane lediglich für die Psychologiestudierenden im ersten Semester ein hypothesenkonformes Ergebnis liefert. Weitere Einschränkungen liegen dadurch vor, dass sich bei der Betrachtung der Items zu in keiner Stichprobe behandelten Inhalten die Stichproben lediglich in zwei Fällen wie erwartet unterscheiden bzw. ähneln.

6.2.2. Veränderungen.

Bevor die Ergebnisse für Hypothese 2b) berichtet werden, laut der Personen nach der Lehrintervention im Schnitt mehr Aufgaben zu in der Lehre behandelten Inhalten lösen sollten als vor der Lehrintervention, sollen zunächst Veränderungen in allen SRA-Items aufgeführt werden. Aus den Angaben in Tabelle 19 wird deutlich, dass lediglich in zwei Stichproben ein Zuwachs in der Anzahl insgesamt gelöster SRA-Items beobachtet werden konnte. Dabei handelt es sich um einen sehr geringen Effekt bei den Studierenden der Sonderpädagogik (Stichprobe 1) und einen geringen Effekt bei den Studierenden der Psychologie im ersten Semester (Stichprobe 3) im Umfang rund eines halben gelösten Items. Während in der Stichprobe 4 (Studierende der Psychologie im zweiten Semester) keine nennenswerte Veränderung in der Anzahl gelöster Items beobachtet werden konnte, lösten Studierende im Master of Arts (Stichprobe 2) annähernd ein Item (0.69) weniger am Ende im Vergleich zum Beginn des Semesters. Da mittlere Differenzen ein verzerrtes Bild der tatsächlichen Veränderungen liefern können, wenn der Zusammenhang zwischen den Wertepaaren negativ ist, wurden Korrelationen zwischen den Messwertpaaren ermittelt. Diese sind für alle Stichproben moderat bis hoch positiv ($.36 < r < .59$). Damit können negative Zusammenhänge zwischen den Messwertreihen als Erklärung für beobachtete geringe mittlere Differenzen ausgeschlossen werden. Die Streuungen der Messwertdifferenzen in den vier Stichproben ist mit $1.84 < SD_{diff} < 2.24$ und einem Varianzverhältnis von maximal $\nu = 1.5$ als ähnlich zu beurteilen. Auffälligkeiten in Streuungsunterschieden zeigen sich jedoch zum einen für die Masterstudierenden im Vergleich zu den anderen drei Stichproben zu Beginn des Semesters und für die Psychologiestudierenden im ersten Semester zwischen dem Beginn und dem Ende des Semesters. Die Streuung in der Anzahl gelöster Aufgaben zu Beginn des

Semesters bei Masterstudierenden ist im Vergleich zu den anderen Studierenden als deutlich (um den Faktor $\nu \approx 2.2$) kleiner einzuschätzen. Die Streuung in der Anzahl gelöster Aufgaben am Ende des Semesters im Vergleich zu der Streuung zu Beginn des Semesters ist bei den Studierenden der Psychologie im ersten Semester um den Faktor $\nu \approx 2.3$ wesentlich geringer.

Tabelle 19

Veränderungen in der Lösung aller SRA-Items zwischen Beginn und Ende des Semesters nach Stichproben.

S	Z	M	SD	n	ν	V	e_{diff}	SD_{diff}	d_{diff}	r_{diff}	t	p [a]
1	1	9.53	2.28	108	1.08	24%	0.23	2.24	0.1	.50	1.07	.29
	2	9.76	2.19	108		22%						
2	1	9.69	1.49	13	1.57	15%	-0.69	1.84	-0.38	.42	-1.3	.78
	2	9	1.87	13		21%						
3	1	8.74	2.23	19	2.26	26%	0.53	2.2	0.24	.36	1.02	.16
	2	9.26	1.48	19		16%						
4	1	9.88	2.53	33	1.28	26%	-0.06	2.18	-0.03	.59	-0.16	.56
	2	9.82	2.24	33		23%						

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., Z: Zeitpunkt mit 1 = zu Beginn des Semesters, 2 = am Ende des Semesters, ν = Varianzverhältnis, V = Variationskoeffizient, e_{diff} = mittlere Differenz der gelösten Items über den Verlauf des Semesters (positive Vorzeichen zeigen einen Zuwachs an gelösten Aufgaben am Ende des Semesters an), SD_{diff} = Standardabweichung der mittleren Differenzen, d_{diff} = standardisierte mittlere Differenz der gelösten Items über den Verlauf des Semesters (positive Vorzeichen indizieren einen Zuwachs an gelösten Aufgaben), r_{diff} = Korrelation zwischen den Messwertpaaren (Anzahl gelöster Aufgaben zu Beginn und am Ende des Semesters). Informationen über Quantile finden sich in Abbildung 40 im Anhang.

[a] p: p-Wert für das obere Ende der t-Verteilung.

Für die Prüfung der Hypothese 2b) wurden Varianzanalysen mit Messwiederholungen herangezogen. Hierbei zeigte die Varianzanalyse für die beiden Stichprobengruppen, in denen die Inhalte der Items (Items 2, 3, 8, 9, 10, 11, 13, 18, 19, 20, siehe Tabelle 4) behandelt (Psychologiestudierende) bzw. nicht

behandelt wurden (Sonderpädagogikstudierende und Masterstudierende) zu den zwei Messzeitpunkten (zu Beginn und am Ende des Semesters) eine sehr geringe Varianzaufklärung durch den Interaktionsterm der Stichprobengruppenzugehörigkeit und des Messzeitpunkts mit R^2 bzw. η^2 von .04% ($F[1; 171] = 0.68$, $p = .41$). Die Varianzanalyse mit dem vierstufigen Faktor „Stichprobe“ (Psychologiestudierende im ersten Semester vs. Psychologiestudierende im zweiten Semester vs. Sonderpädagogikstudierende vs. Masterstudierende) zeigte eine etwas höhere Varianzaufklärung von 0.8%, die jedoch ebenfalls als schwacher Effekt zu beurteilen ist ($f = 0.1$, Cohen, 1988, S. 283). Die Interaktionseffekte der Varianzanalyse sind in Abbildung 12 B abgetragen.

Für eine differenzielle Analyse der Veränderungen in den Stichproben wurden zusätzlich t-Tests für abhängige Stichproben bzw. Wertepaare durchgeführt. Wie bei den Darstellungen im vorangegangenen Abschnitt wurden auch hier aus Gründen der leichteren Interpretation und besseren Vergleichbarkeit Prozentwerte für die Anzahl der korrekt gelösten Aufgaben (X von 10) angegeben. Für Psychologiestudierende im ersten Semester ließ sich eine Veränderung von $e_{diff} = 4.74\%$ ($SD_{diff} = 18.1\%$, $M_{Beginn} = 47.9\%$, $SD_{Beginn} = 12.84\%$, $M_{Ende} = 52.63\%$, $SD_{Ende} = 12.28\%$) beobachten ($t[18] = 1.14$, $p = .14$, $r = -.03$). Diese Veränderung ist mit einer standardisierten mittleren Differenz von $d_{diff} = 0.26$ zwar als gering aber im Sinne der Hypothesen zu beurteilen. Bei Psychologiestudierenden im zweiten Semester lässt sich das Ergebnis nicht replizieren. Hier wurde eine Veränderung von lediglich $e_{diff} = 1.21\%$ ($SD_{diff} = 16.16\%$, $M_{Beginn} = 49.09\%$, $SD_{Beginn} = 14\%$, $M_{Ende} = 50.03\%$, $SD_{Ende} = 12.87\%$) beobachtet ($t[32] = 0.43$, $p = .33$). Diese Veränderung ist mit $d_{diff} = 0.08$ zwar numerisch als im Sinne der Hypothesen, jedoch sehr nahe null zu bewerten. Anhand der Ergebnisse der Varianzanalysen und t-Tests lässt sich Hypothese 2b) daher nur eingeschränkt bestätigen.

Im Sinne diskriminanter Ergebnisse in Abhängigkeit davon, ob Iteminhalte innerhalb der Lehre behandelt worden sind, liegen weitere Einschränkungen durch minimale Zuwächse bei Sonderpädagogik- ($d_{diff} = 0.02$) und Masterstudierenden ($d_{diff} = 0.05$) für Inhalte vor, die lediglich bei Psychologiestudierenden behandelt wurden. Diese Ergebnisse können als denen aus der Stichprobe der Psychologiestudierenden des zweiten Semesters ähnlich angesehen werden. Damit

liegt eine – relativ schwache – Bestätigung für Hypothese 2b) lediglich in Form der Ergebnisse aus der Stichprobe 3 vor.

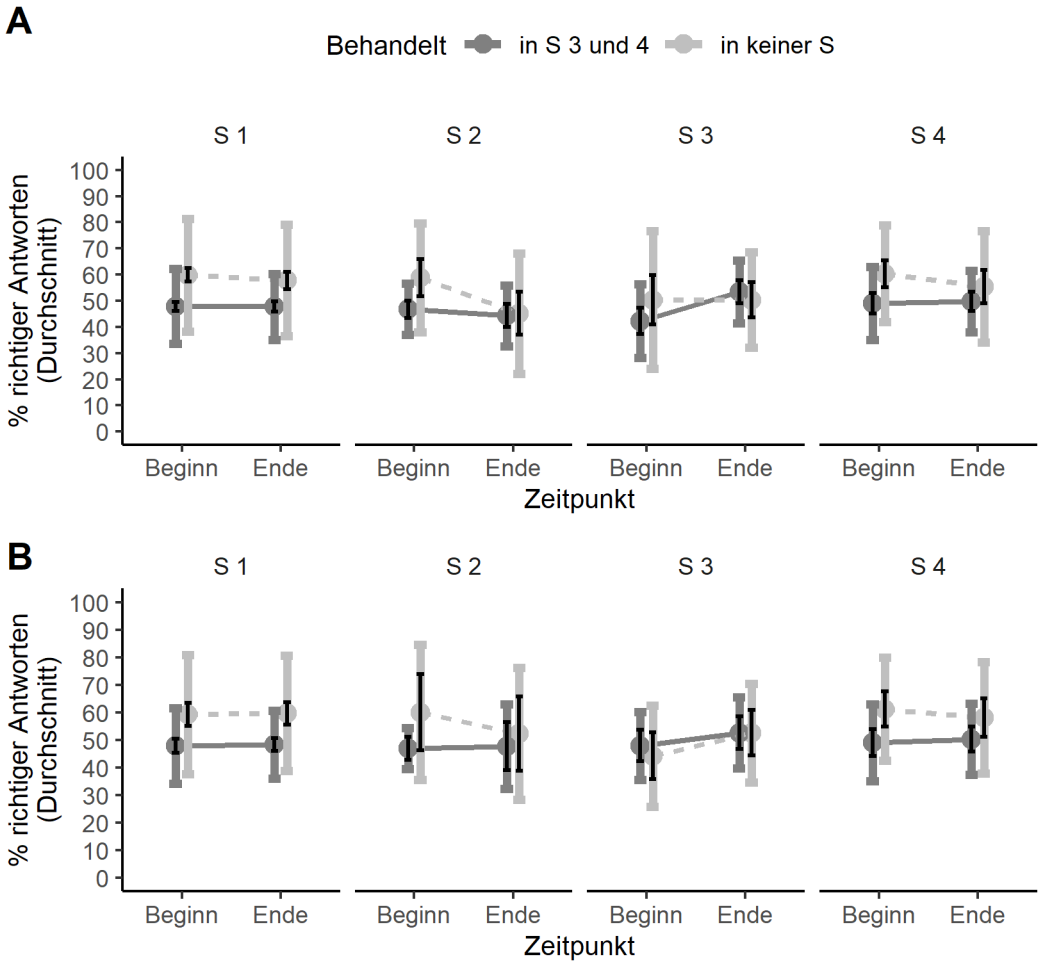


Abbildung 12. Durchschnitte prozentual gelöster Aufgaben nach Itemgruppen (in Stichproben 3 und 4 behandelt vs. in keiner der Stichproben behandelt) für jede der untersuchten Stichproben (S 1: Studierende der Sonderpädagogik, S 2: Masterstudierende, S 3: Psychologiestudierende im ersten Semesters und S 4: Psychologiestudierende im zweiten Semester) zu Beginn und am Ende des Semesters. A: Werte aller Studierenden wurden berücksichtigt ($n \approx 270$ bzw. $n \approx 160$ für S 1 zu Beginn bzw. am Ende des Semesters, $n \approx 30$ für S 2 und S 3 zu beiden Zeitpunkten und $n = 51$ bzw. $n \approx 40$ für S 4 zu Beginn bzw. am Ende des Semesters), B: Nur gepaarte Werte wurden berücksichtigt ($n = 107$ für S 1, $n = 13$ für S 2, $n = 19$ für S 3 und $n = 33$ für S 4).

Explorativ kann darüber hinaus ein (schwacher) Selektionseffekt in der Stichprobe der Psychologiestudierenden im ersten Semester im Verlauf des Semesters vermutet werden (Abbildung 12 A vs. B). Während für die anderen Stichproben kaum Unterschiede im Ausmaß und Unterschied der gelösten Items in Abhängigkeit vom Messzeitpunkt und behandelten und nicht behandelten Inhalten für alle erhobenen Studierenden und Studierende, die an beiden Messzeitpunkten teilnahmen beobachtet werden können, unterscheiden sich die Werte für Psychologiestudierende leicht. Es ist daher möglich, dass sich in dieser Stichprobe Studierende, die an beiden Messzeitpunkten an der Erhebung (und möglicherweise an allen Lehrveranstaltungssitzungen) teilnahmen von Studierenden, die entweder zur ersten oder zur zweiten Erhebung anwesend waren (und möglicherweise nicht konsistent an den Lehrveranstaltungssitzungen teilnahmen) systematisch im Vorwissen unterscheiden.

6.2.3. Veränderungen und Unterschiede der Schwierigkeiten.

Zur Prüfung der Hypothese 2c) werden die Unterschiede in den Schwierigkeiten und den Schwierigkeitsdifferenzen zwischen den beiden Messzeitpunkten für die beiden Itemgruppen zu behandelten bzw. nicht behandelten Inhalten und die Stichproben analysiert. Auch hierfür wurden die Schwierigkeitsdifferenzen für die Items 2, 3, 8, 9, 10, 11, 13, 18, 19, 20 zum einen und die Items 1, 4, 6, 7, 12 (siehe Tabelle 4) zum anderen gruppiert. Da lediglich zehn Schwierigkeitsdifferenzen miteinander verglichen werden, wird auf die Darstellung von statistischen Kennzahlen, wie Mittelwerten, Standardabweichungen, unstandardisierten und standardisierten Differenzen verzichtet. Aus Abbildung 13 (untere Grafik) und Tabelle 8 wird ersichtlich, dass sich die Verteilung der Schwierigkeitsdifferenzen lediglich für die Stichprobe 3 deutlich von der Verteilung in den anderen Stichproben unterscheidet. Bis auf ein Item zu den Inhalten, die bei Studierenden der Psychologie behandelten wurden, waren für die Psychologiestudierenden im ersten Semester alle Items etwas (Schwierigkeitsdifferenz von 2.67%) oder merklich (Schwierigkeitsdifferenz von 27.92%) leichter zu lösen. Dasselbe Muster wäre auch bei Psychologiestudierenden im zweiten Semester zu erwarten gewesen. Hier zeigt sich jedoch ein ähnliches Ergebnismuster, wie bei den Sonderpädagogikstudierenden. Die Schwierigkeiten

verändern sich insgesamt nicht in nennenswerter Weise: die Differenzen liegen zwischen -7.67% und 6.53% für die Sonderpädagogikstudierenden und zwischen -9.66% und 7.42% für die Psychologiestudierenden im zweiten Semester. Zum Vergleich sind in Abbildung 13 (untere Grafik) die Verteilungen der Items abgetragen, deren Inhalte in keiner der Stichproben behandelt wurden. Bis auf die Verteilungen bei den Masterstudierenden unterscheiden sich die Verteilungen der beiden Itemgruppen relativ deutlich in ihrer Breite.

Ergänzend zu der visuellen Inspektion wurden für den Vergleich der Schwierigkeitsdifferenzen nach Itemgruppen asymptotische Fisher-Pitman-Permutationstests durchgeführt. Das Verfahren ergab für die beiden Stichproben 1 und 4 (Sonderpädagogikstudierende und Psychologiestudierende im zweiten Semester) recht hohe Wahrscheinlichkeiten ($p = .66$, $z = 0.44$ bzw. $p = .27$, $z = 1.1$) für die beobachteten Summenunterschiede der Schwierigkeitsdifferenzen zwischen den Gruppen, wenn die einzelnen Schwierigkeitsdifferenzen den beiden Gruppen per Zufall zugeordnet würden. Geringere Wahrscheinlichkeiten ergeben sich für die Stichproben 2 und 3 ($p = .06$, $z = 1.89$ bzw. $p = .21$, $z = 1.26$ und $p = .10$, $z = 1.63$ ohne Ausreißer). Zu berücksichtigen ist beim Fisher-Pitman-Test allerdings, dass hier Summen verglichen werden, d. h. im vorliegenden Fall die Summe der Schwierigkeitsdifferenzen für eine Itemgruppe. Damit stellt sich die Frage, inwiefern sich eine solche Angabe dafür eignet, um die Veränderung in den Schwierigkeiten angemessen wiederzugeben.

Ungeachtet der Verfahrensweise, d. h. einer visuellen Inspektion der Verteilungen oder der Bestimmung von bedingten Wahrscheinlichkeiten in Form von p-Werten, liefern lediglich die Daten aus der Stichprobe der Psychologiestudierenden im ersten Semester Belege für die Hypothese 2c). Die Hypothese lässt sich daher lediglich in eingeschränktem Ausmaß bestätigen.

Die Verteilungen der Schwierigkeiten am Ende des Semesters (Abbildung 13) unterscheiden sich darüber hinaus für die Studierenden der verschiedenen Stichproben eher geringfügig. Die Mediane für die Schwierigkeiten der Stichproben 3 und 4 (Psychologiestudierende) liegen mit $Md = 41.4$ und $Md = 34.9$ nur ca. 4% bis 16% über den Medianen der Stichproben 1 und 2 (Sonderpädagogikstudierende und Masterstudierende). Eine visuelle Inspektion ergibt nur für die Stichprobe der Psychologiestudierenden im ersten Semester (Stichprobe 3) einen erkennbaren

Unterschied der Schwierigkeitsverteilung insgesamt im Vergleich zu den drei übrigen Stichproben. Auch dieses Ergebnismuster lässt sich nur teilweise mit der Erwartung vereinbaren, dass Items leichter gelöst werden, wenn Studierende in den Inhalten unterrichtet worden sind.

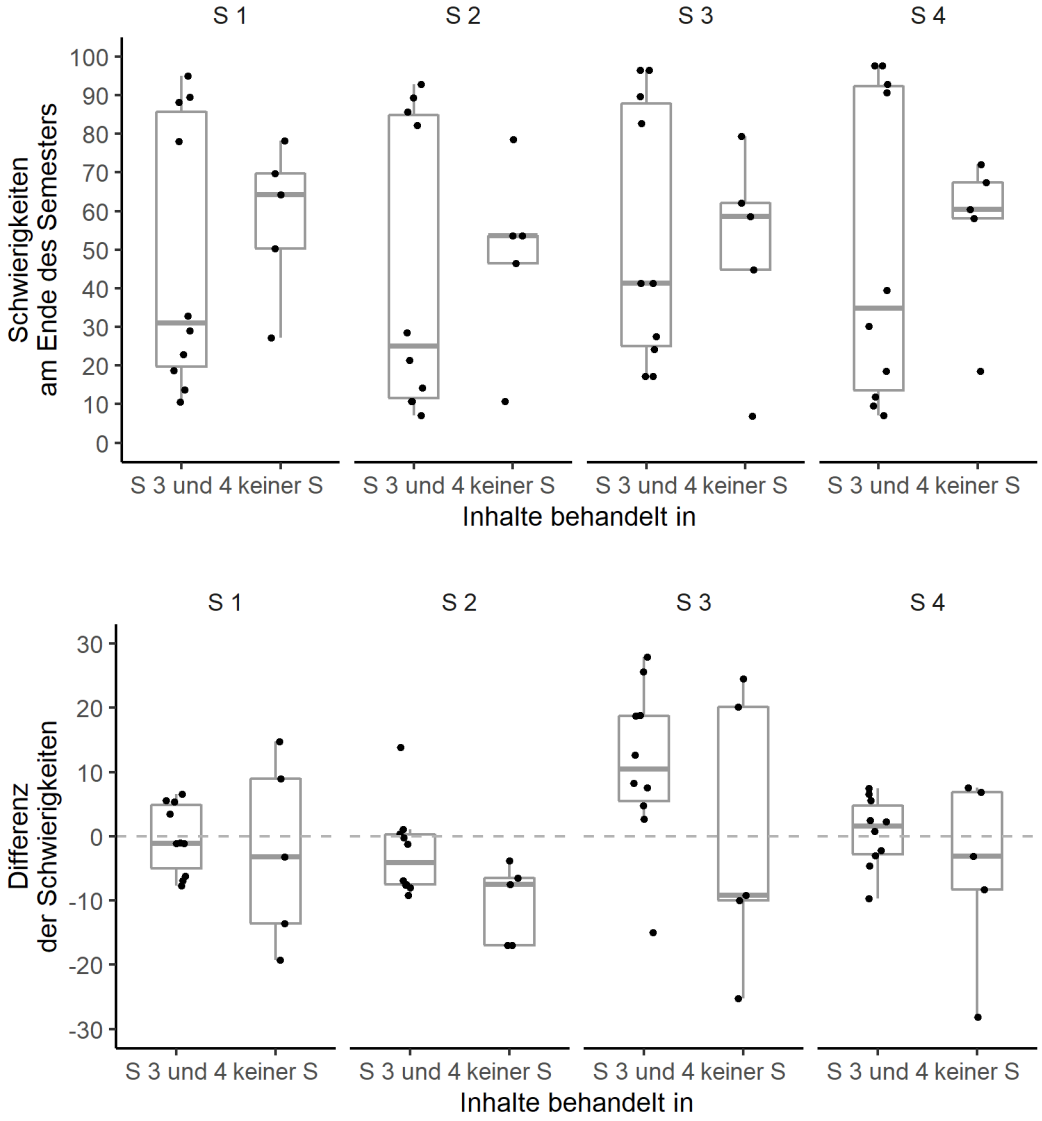


Abbildung 13. Übersicht der Schwierigkeiten am Ende des Semesters (obere Grafik) und Schwierigkeitsdifferenzen (untere Grafik, positive Differenzen zeigen am Ende des Semesters leichtere Items an).

6.3. Fragestellung 3: Zusammenhänge zu anderen Merkmalen

Um Aufschlüsse über die konvergente und diskriminante Validität des SRA im Sinne eines Persönlichkeitsmerkmals, eines kognitiven Prozesses und eines Lehrziels zu erhalten, wurden korrelative und regressive Beziehungsmuster zwischen der Anzahl der im SRA gelösten Items („SRA“) und den Einstellungen zu Statistik, der Anzahl korrekt gelöster Mathematikaufgaben („Mathematik“), der Anzahl korrekt gelöster Aufgaben zu deduktivem Schließen („deduktives Schließen“), der Anzahl korrekt vervollständigter Figurenreihenaufgaben („Figurenreihen“) und der Note für ein statistisches Modul analysiert. Deskriptive univariate Statistiken und Verteilungsübersichten für die Prädiktoren sind in Tabelle 34 und Abbildung 16 und Abbildung 17 (Mathematik), Tabelle 35 und Abbildung 18 und Abbildung 19 (deduktives Schließen), Tabelle 36 und Abbildung 20 (Figurenreihen), Tabelle 37 und Abbildung 21 bis Abbildung 24 (Einstellungen zu Statistik) im Anhang sowie für Noten in Tabelle 6 aufgeführt. Eine Übersicht über die Verteilung der Punktsummen im SRA findet sich im Anhang (Abbildung 25).

Da die Freiheitsgrade für die Stichproben 2 und 3 (Masterstudierende und Psychologiestudierende im ersten Semester) sehr gering sind ($n = 13$ und $n = 19$) werden die Analysen auf die Stichproben 1 (Sonderpädagogikstudierende mit 2 SWS Statistiklehreumfang) und 4 (Psychologiestudierende im zweiten Semester mit 4+4 SWS Statistiklehreumfang) beschränkt. In beiden Stichproben werden für Mathematik und deduktives Schließen die kurzen Itemsammlungen (Summen von zehn bzw. acht Items) für eine bessere Vergleichbarkeit der Ergebnisse zwischen den beiden Stichproben verwendet. In die Analysen gingen auch hier lediglich die Summenwerte ein, die maximal 30% nicht bearbeitete Items (fehlende Werte) aufwiesen. Eine Übersicht der Mittelwerte, Standardabweichungen und Korrelationen nullter Ordnung zwischen allen aufgeführten Variablen findet sich in Tabelle 21 (Stichprobe 1, Sonderpädagogikstudierende, S. 171) und Tabelle 22 (Stichprobe 4, Psychologiestudierende im zweiten Semester, S. 176).

6.3.1. Inkrementelle Varianzaufklärung.

Für die Beantwortung der Hypothesen 3a) und 3b) über die mindestens schwache inkrementelle Varianzaufklärung durch kognitive Maße wurden mehrere multiple Regressionsmodelle mit dem Kriterium „Anzahl der im SRA gelösten Items am Ende des Semesters“ (SRA [t2]) bestimmt und verglichen. Hierbei wurde für jede der beiden Stichproben ein Basismodell mit den Subskalen zu den Einstellungen zu Statistik (SATS-Subskalen: Affekt, kognitive Kompetenz, Wert, Schwierigkeit, Interesse und Anstrengung) einerseits zu Beginn des Semesters (t1) und andererseits am Ende des Semesters (t2) als Prädiktoren formuliert. Für die Bestimmung der inkrementellen Varianzaufklärung von Mathematik gingen in das Basismodell zusätzlich deduktives Schließen für die Stichprobe 4 und deduktives Schließen sowie Figurenreihen für die Stichprobe 1 als Prädiktoren ein. Für die Bestimmung der inkrementellen Varianzaufklärung von deduktivem Schließen enthielt das Basismodell die Einstellungen zu Statistik (t1 oder t2) sowie Mathematik für Stichprobe 4 und zusätzlich Figurenreihen für die Stichprobe 1 als Prädiktoren. Für die Bestimmung der inkrementellen Varianzaufklärung der Figurenreihenaufgaben in Stichprobe 1 enthielt das Basismodell die Prädiktoren Einstellung zu Statistik (t1 oder t2), Mathematik und deduktives Schließen. Zusätzlich wurden die aufgeführten Basismodelle hinsichtlich der Punktsumme im SRA zu Beginn des Semesters (SRA [t1]) variiert. Die jeweiligen Basismodelle wurden dann mit einem Modell verglichen, das neben den Prädiktoren des Basismodells den relevanten Prädiktor enthielt. Tabelle 20 sind die jeweiligen inkrementellen Varianzaufklärungsanteile der insgesamt 20 Regressionsmodellvergleiche zu entnehmen.

Wie aus Tabelle 20 hervorgeht, lässt sich lediglich eine stabile und im Effekt annähernd erwartete Varianzaufklärung durch Figurenreihen beobachten ($\Delta R^2 = .04$ bzw. $\Delta R^2 = .036$ vs. $\Delta R^2_{erwartet} \geq .05$). Damit lassen sich Unterschiede in der Anzahl gelöster SRA-Aufgaben am Ende des Semesters auch über die Anzahl gelöster SRA-Aufgaben zu Beginn des Semesters (und den Einstellungen zu Statistik am Ende des Semesters sowie Mathematik und deduktives Schließen) hinaus durch die Anzahl korrekt vervollständigter Figurenreihen vorhersagen. Werden im Basismodell sowohl Einstellungen zu Beginn als auch Einstellungen am Ende sowie die anderen kognitiven Prädiktoren berücksichtigt, sinkt der

Tabelle 20

Anteil inkrementell aufgeklärter Varianz (ΔR^2) in der Anzahl gelöster SRA-Items am Ende des Semesters durch verschiedene Prädiktoren über Einstellungen zu Statistik in Abhängigkeit vom Messzeitpunkt und der Anzahl gelöster SRA-Items zum ersten Messzeitpunkt hinaus.

AV: SRA (t2)	Einstellungen (t1)		Einstellungen (t2)	
	ohne	mit	ohne	mit
Stichprobe und Prädiktoren	SRA (t1)	SRA (t1)	SRA (t1)	SRA (t1)
1 ($n = 103$)				
Mathematik	.03	.02	.02	.004
Figurenreihen	.03	.02	.04	.036
deduktives Schließen	.04	.01	.04	.01
4 ($n = 29$)				
Mathematik	.064	.04	.065	.01
deduktives Schließen	.03	.02	.04	.001

Anmerkungen. SRA (t2): Anzahl der gelösten SRA-Items am Ende des Semesters, Einstellungen (t1): Einstellungen zu Statistik zu Beginn des Semesters im Modell berücksichtigt, Einstellungen (t2): Einstellungen zu Statistik am Ende des Semesters im Modell berücksichtigt, ohne SRA (t1): Anzahl der gelösten SRA-Items zu Beginn des Semesters im Modell nicht berücksichtigt, mit SRA (t1): Anzahl der gelösten SRA-Items zu Beginn des Semesters im Modell berücksichtigt.

Varianzaufklärungsanteil durch Figurenreihen auf rund 2% ($\Delta R^2 = .025$). Der Effekt entspricht einer Semipartialkorrelation von $r_{spart} \approx .16$ und kann noch als klein, oder in Cohens Worten als „just barely escaping triviality“ bewertet werden (Cohen, 1988, S. 413)[30]. Für die anderen untersuchten Prädiktoren lässt sich – wenn der „Startwert“ in der Anzahl gelöster SRA-Aufgaben berücksichtigt wird (SRA [t1]) – keine nennenswerte inkrementelle Varianzaufklärung beobachten. Dieses Ergebnis spricht gegen die Gültigkeit der Hypothese 3a), laut der das SRA mathematiknähere Inhalte gegenüber allgemeinkognitiven Inhalten erfasst. Mit Hypothese 3b) sind die Ergebnisse insofern vereinbar, als dass eins der beiden allgemeinen kognitiven Maße, die Anzahl korrekt vervollständigter Figurenreihen,

[30] Hierbei sind zwar einige Details, wie die Anzahl der genutzten Prädiktoren zu beachten (siehe Cohen, 1988, Kap. 9), jedoch können Korrelationskoeffizienten beliebiger Art (d. h. nullter Ordnung, semipartialer oder partieller Art) als annähernd ihren Effektgrößen entsprechend verglichen werden ($r \approx \pm.10$, $r \approx \pm.30$, $r \gtrsim |\pm.50|$).

in einer Stichprobe (Sonderpädagogikstudierende) eine schwache (allerdings noch schwächere als erwartete) inkrementelle Varianzaufklärung zeigt. Damit liegt in begrenztem Ausmaß konvergente und diskriminante Validität des SRA als Erfassung allgemeinkognitionsnaher Personenmerkmale bzw. allgemeinkognitiver Prozesse vor.

Wird die Anzahl der gelösten Aufgaben im SRA am Ende des Semesters ohne die Berücksichtigung der zu Beginn des Semesters gelösten SRA-Aufgabensumme betrachtet, zeigt sich für Figurenreihen in Stichprobe 1, für deduktives Schließen für beide Stichproben und für Mathematik in Stichprobe 4 ein Effekt annähernd der erwarteten Größe ($\Delta R^2 = .03$ bzw. $\Delta R^2 = .065$ vs. $\Delta R^2_{erwartet} \geq .05$). Eine solche Betrachtung kann sinnvoll sein, wenn von Interesse ist, welche Fertigkeiten einen Beitrag zur Steigerung der Lösung statistischer Aufgaben ungeachtet des statistischen Vorwissens leisten. Wie in der Diskussion erörtert wird, wäre eine solche Betrachtung allerdings erst dann von Relevanz, wenn die SRA-Aufgaben ein sinnvolles Kriterium der statistischen Kompetenz darstellen. Zu beachten ist bei der Interpretation der Ergebnisse, dass lediglich lineare Beziehungen zwischen den Prädiktoren und dem Kriterium untersucht wurden.

6.3.2. Zusammenhang mit akademischem Leistungsmaß.

Für die Beantwortung der Hypothese 3c) zum Zusammenhang zwischen der Modulnote für die vergangene Statistikveranstaltung und der SRA-Punktesumme wurden der Produkt-Moment-Korrelationskoeffizient und Polynomterme innerhalb einer multiplen linearen Regression zur Bestimmung nicht-linearer Beziehungen genutzt. In die Analyse gingen lediglich Psychologiestudierende im zweiten Semester ein. Der lineare Zusammenhang erklärte die Beziehung zwischen der Note für das erste Statistikmodul und der wenige Monate später erzielten SRA-Punktsumme mit $r = -.31$ und $R^2 = .09$ ($n = 45$) relativ zur Anzahl genutzter Polynomterme am besten (siehe Abbildung 14 A). Die Hinzunahme von quadratischen, kubischen Termen sowie eines Terms vierten Grades zeigte mit 2% bzw. 3% nur äußerst geringfügig mehr aufgeklärte Varianz. Damit scheint die Beziehung zwischen der Note im ersten Statistikmodul und der SRA-Punktezahl zu Beginn des zweiten Semesters am besten durch einen linearen Zusammenhang moderaten Ausmaßes repräsentiert zu werden. Der Zusammenhang zwischen der

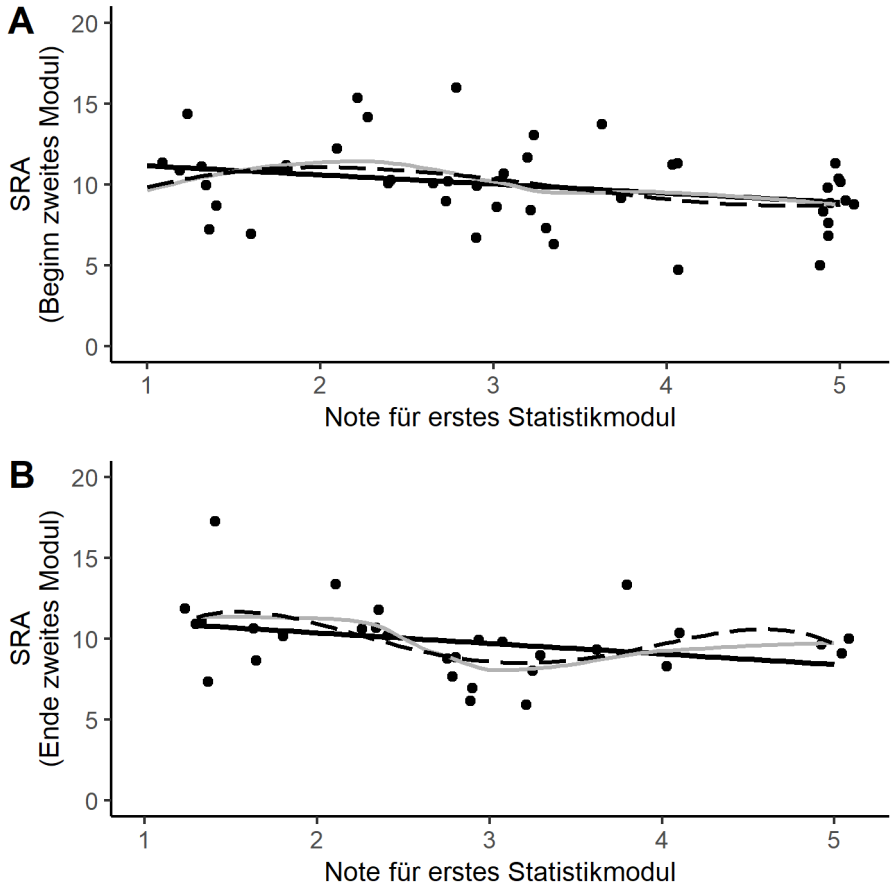


Abbildung 14. Streudiagramm für die Anzahl der gelösten SRA-Items zu Beginn (A) und am Ende (B) des zweiten Statistikmoduls und die im ersten Statistikmodul erzielte Note bei Psychologiestudierenden im zweiten Semester mit linearer (schwarze durchgezogene Linie), Loess-geschätzter (graue durchgezogene Linie) und der Polynomfunktion mit bester Anpassung an Loess (kubisch in A, vierten Grades in B, schwarze gestrichelte Linie).

Note im ersten Statistikmodul und der SRA-Punktezahl am Ende des darauffolgenden Semesters wird dagegen wesentlich besser durch eine kurvilineare Beziehung beschrieben. Während der lineare Zusammenhang in diesem Fall dem oben berichteten sehr ähnelt ($r = -.33$ und $R^2 = .11$, $n = 30$), zeigt die Berücksichtigung eines quadratischen Terms 9% (Steigerung mittleren Ausmaßes) und die eines Terms vierten Grades weitere 5% (Steigerung geringen Ausmaßes) mehr im Kriterium aufgeklärte Varianz (siehe auch Abbildung 14 B). Unter Berücksichtigung des Sparsamkeitsprinzips für statistische Modelle kann hier daher eine (mindestens) quadratische Beziehung beobachtet werden und insgesamt eine recht hohe Varianzaufklärung von 20%.

Die Hypothese 3c) lässt sich damit als mindestens in Teilen bestätigt ansehen. Der Zusammenhang zwischen der Note in einem ersten Statistikmodul hängt im mittleren Ausmaß mit der erzielten SRA-Punktesumme zu Beginn des zweiten Statistikmoduls und im eher hohen Ausmaß mit der erzielten SRA-Punktesumme am Ende des zweiten Statistikmodul zusammen.

6.3.3. Exploration ausgewählter Zusammenhänge.

In Tabelle 21 und Tabelle 22 sind die Zusammenhänge in Form von Produkt-Moment-Korrelationen zwischen den untersuchten Variablen für die Stichproben der Sonderpädagogikstudierenden und der Psychologiestudierenden aus dem zweiten Semesters dargestellt. Aufgrund der sehr geringen Stichproben der Masterstudierenden ($n = 13$ Personen) und der Psychologiestudierenden im dritten Semester ($n = 16$ bis $n = 19$ Personen) werden die Ergebnisse für diese beiden Stichproben an dieser Stelle nicht berichtet (eine Übersicht der Korrelationen findet sich hierfür im Anhang in Tabelle 45 und Tabelle 46).

Im Folgenden sollen die Zusammenhänge zwischen den verschiedenen Einstellungsfacetten zu Statistik und den SRA-Punktesummen beleuchtet werden. Bei der Interpretation der Zusammenhänge sollte berücksichtigt werden, dass die Erfassung einiger Einstellungsfacetten in drei Skalen durch zwei unterschiedliche Formulierungsformen erfolgte. Sonderpädagogikstudierende gaben für die Skalen zur affektiven Einstellung zu Statistik („Affekt“), die selbsteingeschätzte eigene kognitive Kompetenz für das Erlernen von Statistik („Kompetenz“) sowie den Aufwand für die Lehrveranstaltung („Anstrengung“) bei der Erhebung zu Beginn des Semesters die für das kommende Semester erwartete Einschätzungen an (Bsp.: „Ich werde Statistik mögen“, „Ich werde in der Statistikveranstaltung keine Ahnung haben, worum es geht“ usw.). Am Ende des Semesters erfolgte eine zum Teil retrospektive Einschätzung (Bsp.: „Ich mag Statistik“, „Ich hatte keine Ahnung, worum es in der Statistikveranstaltung geht“). Damit ist es möglich, bei den Studierenden der Sonderpädagogik für die drei Skalen einen Zusammenhang zwischen den durch Studierende vorgenommenen Prognosen mit der tatsächlichen (selbsteingeschätzten) Entwicklung vorzunehmen. Die Psychologiestudierenden im zweiten Semester schätzten mittels der Vergangenheitsversion der drei Skalen das

Tabelle 21

Mittelwerte, Standardabweichungen und Produkt-Moment-Korrelationen für Skalen bzw. Punktesummen aus Stichprobe 1 (2 SWS, B.A.).

		M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	SRA (1)	9.55	2.38																
2	SRA (2)	9.53	2.17	**.51**															
3	dedukt. Schließen	5.21	1.34	**.36**	.29														
4	Figurenreihen	5.67	1.82	.22	.29	.32													
5	Mathematik	7.22	1.89	.33	.32	.32	.20												
6	Affekt (1)	3.72	1.17	.17	.23	.15	.19	**.37**											
7	Kompetenz (1)	4.31	1.16	.23	.27	.19	.16	**.40**	**.83**										
8	Wert (1)	4.54	0.97	.04	.17	.04	.11	.14	**.51**	**.49**									
9	Schwierigkeit (1)	4.59	0.82	-.12	-.15	-.06	-.11	-.25	**-.71**	**-.70**	**-.43**								
10	Interesse (1)	4.34	1.25	-.02	.21	-.04	.08	.03	**.44**	**.42**	**.63**	-.34							
11	Anstrengung (1)	5.25	1.10	-.03	.05	-.15	-.18	-.07	.06	.09	.22	.07	**.35**						
12	Affekt (2)	3.84	1.18	.28	.18	.17	.25	.20	**.68**	**.69**	**.40**	**-.61**	.29	.04					
13	Kompetenz (2)	4.63	1.09	.31	.30	.22	.27	.21	**.66**	**.72**	**.37**	**-.61**	**.36**	.10	**.76**				
14	Wert (2)	4.30	1.09	.09	.14	.05	.17	.07	**.38**	**.35**	**.53**	-.30	**.41**	.10	**.43**	**.41**			
15	Schwierigkeit (2)	4.60	0.88	-.24	-.11	-.11	-.15	.03	**-.56**	**-.56**	**-.37**	**.65**	-.28	.06	**-.62**	**-.64**	-.29		
16	Interesse (2)	3.90	1.39	.08	.16	-.03	.09	.05	.29	.29	**.38**	-.17	**.46**	.14	**.37**	**.36**	**.75**	-.25	
17	Anstrengung (2)	4.75	1.30	-.16	-.09	-.17	-.19	-.07	-.24	-.15	-.01	.25	.05	.29	-.09	-.05	.19	.28	.20

Anmerkungen. N beträgt 104 bis 108 (nicht unterlegt) und 259 bis 270 (grau unterlegt), moderate bis hohe Effekte (ab ≈ ±.40) sind **fett** gesetzt, dedukt. Schließen: deduktives Schließen, der Messzeitpunkt ist in Klammern angegeben (1: zu Beginn des Semesters, 2: am Ende des Semesters).

erste zurückliegende Statistikmodul (Erhebung zu Beginn des zweiten Semesters) und das zweite zurückliegende Statistikmodul (Erhebung am Ende des zweiten Semesters) ein. Die Korrelationen zwischen den Skalen zu den beiden Messzeitpunkten zeigen in dieser Stichprobe die Konsistenz der Einschätzungen bzw. des Verhaltens über zwei inhaltlich unterschiedliche Statistikmodule an.

Die Zusammenhänge zwischen dem erwarteten und dem tatsächlichen Affekt sowie der erwarteten und tatsächlichen selbsteingeschätzten Kompetenz bei Sonderpädagogikstudierenden waren mit $r = .68$ und $r = .72$ hoch. Für die geplante und tatsächliche Anstrengung in dem Modul war die Korrelation mit $r = .29$ moderat und damit wesentlich kleiner (großer Effekt mit Cohen $q = .56$, Lenhard & Lenhard, 2016). Die Zusammenhänge zwischen den beiden Messzeitpunkten für die Einschätzung des Werts von Statistik („Wert“) mit $r = .53$, der wahrgenommenen Schwierigkeit von Statistik („Schwierigkeit“) mit $r = .65$ sowie dem eigenen Interesse an Statistik („Interesse“) mit $r = .46$ lassen sich als hoch einschätzen.

Bei Psychologiestudierenden lassen sich für den Affekt, die Kompetenz und die erbrachte Anstrengung für beide Statistikmodule hohe lineare Zusammenhänge beobachten ($r = .70$, $r = .68$ und $r = .72$). Auch der eingeschätzte Wert von Statistik zeigte eine recht hohe Stabilität über das zweite Semester ($r = .73$). Deutlich geringere Stabilität liegt für die wahrgenommene Schwierigkeit von Statistik ($r = .30$) sowie das Interesse an Statistik ($r = .35$) vor (Cohen $q \approx .50$, großer Effekt, Lenhard & Lenhard, 2016). Die beiden letztgenannten Korrelationen könnten damit die Unterschiede in den Inhalten des zweiten Statistikmoduls (Inferenzstatistik) gegenüber dem ersten (deskriptive Statistik und Wahrscheinlichkeitsrechnung) anzeigen. Hierbei sollte allerdings beachtet werden, dass die eingeschränkte interne Konsistenz mit Cronbachs $\alpha = .58$ und $\alpha = .55$ für die Skala „Schwierigkeit“ ebenfalls einen Grund für die geringeren beobachteten Korrelationen darstellen kann.

Für den Affekt zeigen sich für die beiden Stichproben leicht unterschiedliche Zusammenhänge zu den SRA-Punktsummen zu Beginn und am Ende des Semesters. Während bei Sonderpädagogikstudierenden für den Zusammenhang zwischen dem Affekt und der im SRA erzielten Punktzahl unabhängig vom Messzeitpunkt kleine bis höchstens moderate Effekte zu beobachten sind ($.17 < r < .28$), zeigen sich für

Psychologiestudierende im zweiten Semester höhere Zusammenhänge. So korreliert der Affekt zu Beginn des Semesters in tendenziell hohem Maß mit der SRA-Punktsumme zu Beginn ($r = .42$) und in mindestens moderatem Maß am Ende des Semesters ($r = .35$). Der Affekt am Ende des Semesters zeigt ebenfalls eine relativ enge Beziehung zu der SRA-Punktesumme am Ende des Semesters ($r = .40$). Damit erweist sich die affektive Einstellung zu Statistik in der Stichprobe der Psychologiestudierenden im zweiten Semester als etwas bedeutsamer (Cohen $q \approx .10$, kleiner Effekt, Lenhard & Lenhard, 2016) zur Vorhersage der SRA-Punktesumme im Vergleich zur Rolle des Affekts bei den Sonderpädagogikstudierenden.

Für die selbsteingeschätzte Kompetenz zeigt sich ein ähnliches Bild. Bei Sonderpädagogikstudierenden liegen auch hier höchstens moderate Effekte unabhängig vom Messzeitpunkt vor ($.23 < r < .31$). Für die Studierenden der Psychologie lässt sich zwischen der selbsteingeschätzten Kompetenz und der SRA-Punktesumme am Ende des Semesters eine hohe Korrelation ($r = .51$) beobachten, nicht aber für Kompetenz am Anfang des Semesters und der SRA-Punktesumme zu Beginn oder am Ende des Semesters ($r = .22$ und $r = .26$). Die wahrgenommene kognitive Kompetenz scheint daher vor allem am Ende des Moduls eine hohe Vorhersagekraft für die erzielte SRA-Punktsumme bei Psychologiestudierenden im zweiten Semester zu haben.

Die Zusammenhänge zwischen dem wahrgenommenen Wert von Statistik zu Beginn und am Ende des Semesters mit der erzielten SRA-Punktsumme zu Beginn und am Ende des Semesters sind für beide Stichproben insgesamt sehr klein bis höchstens moderat ($.04 < r < .29$). Auch hierbei fallen die Produkt-Moment-Korrelationen für die Stichprobe der Psychologiestudierenden etwas höher aus ($.16 < r < .29$) im Vergleich zu den Studierenden der Sonderpädagogik ($.04 < r < .17$). Eine Erklärung für die (geringen, Cohen $q \approx .10$, Lenhard & Lenhard, 2016) Unterschiede liegt möglicherweise darin, dass die wichtige Rolle der Statistik für das Studium bei Psychologiestudierenden in allen inhaltlichen Modulen unterstrichen wird. Eine Bestätigung für eine solche Deutung liegt etwa in den höheren Mittelwerten für den Wert der Statistik bei Psychologiestudierenden im Vergleich zu Sonderpädagogikstudierenden vor ($M_\Psi = 5.00$, $SD_\Psi = 0.81$ zu

Beginn und $M_{\Psi} = 5.11$, $SD_{\Psi} = 0.87$ am Ende vs. $M_{SoP} = 4.54$, $SD_{SoP} = 0.97$ zu Beginn und $M_{SoP} = 4.30$, $SD_{SoP} = 1.09$ am Ende des Semesters).

Die zu Beginn des Semesters eingeschätzte Schwierigkeit von Statistik hängt für beide Stichproben in eher geringem Maß negativ mit den erzielten SRA-Punktsummen zu Beginn und am Ende des Semesters zusammen ($-.18 < r < -.12$). Eine höhere wahrgenommene Schwierigkeit geht mit tendenziell eher kleineren SRA-Punktsummen einher. Für die am Ende des Semesters eingeschätzte Schwierigkeit zeigen sich ebenfalls eher geringe Effekte ($|.09| < r < |.24|$), jedoch ist eine Richtungsänderung in den Zusammenhängen bei den Psychologiestudierenden zu beobachten. Tendenziell gehen hier mit einer höher eingeschätzten Schwierigkeit von Statistik mehr erzielte Punkte im SRA sowohl zu Beginn ($r = .12$) als auch am Ende des Semesters ($r = .09$) einher. Bei Konstanthalten der selbsteingeschätzten kognitiven Kompetenz am Ende des Semesters lässt sich darüber hinaus ein Suppressionseffekt beobachten: der Zusammenhang verstärkt sich auf $r = .25$ (Zusammenhang zwischen Schwierigkeit am Ende des Semesters und der erzielten SRA-Punktsumme zu Beginn des Semesters) und $r = .36$ (Zusammenhang zwischen Schwierigkeit am Ende des Semesters und der erzielten SRA-Punktsumme am Ende des Semesters). Der Suppressionseffekt lässt sich auch bei Studierenden der Sonderpädagogik für den Zusammenhang zwischen der Schwierigkeit am Ende des Semesters und der am Ende des Semesters erzielten SRA-Punktsumme beobachten: Die Korrelation ändert sich von $r = -.11$ auf $r = .16$ bei Konstanthalten der am Ende des Semesters rückblickend selbsteingeschätzten Kompetenz. Ungeachtet des wahrgenommenen Erfolgs bei der Bewältigung der Lehrinhalte („Kompetenz“) zeigt eine höher eingeschätzte Schwierigkeit der Statistik eine kleine (bei Sonderpädagogikstudierenden) bis moderate (bei Psychologiestudierenden) positive Beziehung zu der gelösten SRA-Punktesumme am Ende des Semesters.

Für das Interesse an Statistik lassen sich mehrheitlich nur äußerst geringe Zusammenhänge beobachten ($-.02 < r < .08$). Lediglich bei Studierenden der Sonderpädagogik liegen höhere Zusammenhänge für das angegebene Interesse zu Beginn des Semesters ($r = .21$) bzw. am Ende des Semesters ($r = .16$) und der am Ende des Semesters erzielten SRA-Punktesumme vor. Insgesamt scheint die lineare

Beziehung zwischen dem Interesse an Statistik und der SRA-Punktesumme daher sehr schwach bis schwach zu sein.

Ein interessantes Ergebnismuster lässt sich für die selbsteingeschätzte rückblickende Anstrengung für die Lehrveranstaltungen bei Psychologiestudierenden im zweiten Semester beobachten. Die Produkt-Moment-Korrelationen bewegen sich hier zwar im geringen bis höchstens moderaten Bereich ($-.28 < r < -.15$), sind allerdings durchweg negativ. Eine höhere Anstrengung im Verlauf des Semesters geht damit tendenziell mit einer geringeren SRA-Punktsumme bei Psychologiestudierenden einher. Hierbei supprimiert, ähnlich, wie bei eingeschätzter Schwierigkeit von Statistik, die am Ende des Semesters berichtete kognitive Kompetenz den Zusammenhang zwischen der erfolgten Anstrengung im Verlauf des Semesters und der am Ende des Semesters erzielten SRA-Punktsumme ($r = -.24$ vs. $r = -.33$ bei Kontrolle der Kompetenz, Cohen $q \approx .10$, Lenhard & Lenhard, 2016, Cohen, 1988). Die Zusammenhänge bei Studierenden der Sonderpädagogik sind dagegen überwiegend sehr gering bis gering ($-.16 < r < .05$) und ändern sich bei Kontrolle der kognitiven Kompetenz nicht in auffälligem Ausmaß ($-.19 < r < .01$).

Bei Konstanthalten der erhobenen kognitiven Variablen (Anzahl gelöster Mathematikaufgaben, Anzahl gelöster Aufgaben zu deduktivem Schließen und bei Studierenden der Sonderpädagogik Anzahl der richtig fortgesetzten Figurreihen) zeigen sich mindestens moderate Zusammenhänge für Studierende der Psychologie nur für Affekt und Kompetenz am Ende des Semesters mit der am Ende des Semesters erzielten SRA-Punktsumme ($r = .36$ und $r = .37$). Für Studierende der Sonderpädagogik lassen sich für Affekt und Kompetenz am Ende des Semesters mit der zu Beginn des Semesters erzielten SRA-Punktesumme moderate Effekte beobachten ($r = .26$ und $r = .27$). Damit scheinen einzelne Einstellungen zu Statistik auch einen inkrementellen Wert bei der Vorhersage von SRA-Punktsummen aufzuweisen.

Tabelle 22

Mittelwerte, Standardabweichungen und Produkt-Moment-Korrelationen für Skalen bzw. Punktesummen aus Stichprobe 4 (4(+4) SWS, B.Sc.).

		M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	SRA (1)	9.92	2.42																	
2	SRA (2)	9.40	2.36	**.59**																
3	ded. Schl. (kurz)	4.67	1.16	.20	.24															
4	Figurenreihen																			
5	Mathematik (kurz)	7.21	2.16	**.56**	**.52**	**.44**														
6	Affekt (1)	4.57	1.28	**.42**	**.35**	.17		**.47**												
7	Kompetenz (1)	5.04	1.06	.22	.26	.03		**.35**	**.78**											
8	Wert (1)	5.00	0.81	.28	.26	.30		**.65**	**.36**	**.39**										
9	Schwierigkeit (1)	4.38	0.72	-.18	-.16	-.30		-.33	**-.60**	**-.62**	-.03									
10	Interesse (1)	5.01	1.07	.01	.08	.31		.25	**.48**	.34	**.64**	-.02								
11	Anstrengung (1)	5.26	1.21	-.28	-.15	-.32		-.28	.23	.30	-.01	-.03	.17							
12	Affekt (2)	3.97	1.40	.23	**.40**	-.01		.16	**.70**	**.57**	.14	**-.39**	.27	.16						
13	Kompetenz (2)	4.78	0.97	.28	**.51**	.28		**.35**	**.60**	**.68**	.25	**-.51**	.24	.11	**.73**					
14	Wert (2)	5.11	0.87	.16	.29	**.55**		**.55**	.28	.31	**.73**	-.20	**.54**	-.10	.22	**.36**				
15	Schwierigkeit (2)	4.37	0.69	.12	.09	-.23		-.02	-.18	-.23	-.01	.30	-.21	-.02	-.31	**-.38**	-.19			
16	Interesse (2)	4.80	1.04	.05	.08	.26		.19	.08	.09	**.47**	.06	**.35**	.23	.19	.25	**.69**	-.25		
17	Anstrengung (2)	5.13	1.20	-.17	-.24	-.27		-.28	.08	.20	.06	.14	.09	**.72**	.17	.09	-.03	-.16	**.40**	
18	Note ($n = 26$)	2.68	1.09	-.29	-.33	**-.51**		**-.46**	**-.60**	**-.64**	**-.43**	**.55**	-.31	-.26	**-.37**	**-.51**	**-.68**	**.46**	**-.56**	-.26

Anmerkungen. *N* beträgt 22 (nicht unterlegt, außer bei Note) und 29 (grau unterlegt), moderate bis hohe Effekte (ab $\approx \pm.40$) sind **fett** gesetzt, dedukt. Schl.: deduktives Schließen, der Messzeitpunkt ist in Klammern angegeben (1: zu Beginn des Semesters, 2: am Ende des Semesters).

7. Diskussion

Die Ergebnisse der Untersuchungen sollen im Folgenden zunächst im Sinne der Fragestellungen zusammengefasst und interpretiert werden. Dabei finden auch Einschränkungen in der Aussagekraft der Ergebnisse Berücksichtigung. Vor dem Hintergrund der bisherigen Ergebnisse werden dann die beiden Zielfragen der Arbeit vorläufig beantwortet: „Eignet sich das SRA, um den Beherrschungsstand in Statistik bei Studierenden zu erfassen?“ und „Eignet sich das SRA, um die Veränderung im Beherrschungsstand durch Lehre zu erfassen?“. Im Anschluss soll auf allgemeine Einschränkungen der vorliegenden Arbeit sowie der bisherigen Forschungsbemühungen insgesamt hingewiesen werden.

7.1. Zusammenfassung und kritische Interpretation der Ergebnisse

Die vorliegende Arbeit setzte sich zum Ziel, drei in der empirischen Forschung zu den Eigenschaften des SRA noch offene Fragestellungen zu beantworten. Die erste dieser Fragestellungen betrifft die interne Struktur des SRA. So ist empirisch noch nicht erforscht worden, inwiefern das SRA eine mit den Darstellungen der Autoren (Garfield & Chance, 2000, Garfield, 2003) vereinbare ein-, vier- oder achtdimensionale Struktur im Sinne des correct statistical reasoning aufweist. Die zweite Fragestellung beinhaltet die Frage nach der Sensitivität des SRA für Lehrinterventionen. Es liegen hierzu noch keine Untersuchungen über Unterschiede in der Anzahl gelöster Aufgaben im SRA für Lernendengruppen in Abhängigkeit von der Behandlung (vs. Nichtbehandlung) bestimmter statistischer Inhalte vor. Die dritte und letzte Fragestellung betrifft die empirische Verwandtschaft des statistical reasoning, wie es mit dem SRA erfasst wird, mit anderen kognitiven und nicht-kognitiven Konstrukten. Insbesondere die Nähe und Abgrenzung zu mathematischen und allgemeinkognitiven Maßen ist hierbei von Bedeutung.

7.1.1. Fragestellung 1: Interne Struktur des SRA

Die Hypothesen zur ersten Fragestellung nach der internen Struktur des SRA 1a) bis 1c) können durch die vorliegenden Daten nicht bestätigt werden. Zwar ergeben die Analysen mit einzelnen Verfahrenskombinationen (z. B. Anwendung

des Kaiser-Guttman-Kriteriums bei der Hauptkomponentenanalyse anhand der Produkt-Moment-Korrelationsmatrix bei der Stichprobe 1 zu Beginn und am Ende des Semesters) eine der Hypothese 1a) entsprechende Faktorenanzahl, keine der Stichprobenanalysen ergibt jedoch eine einheitliche oder eindeutige Faktoren- bzw. Komponentenanzahl. Läge den Daten eine ein-, vier- oder achtdimensionale Struktur zugrunde, sollten sich zumindest mehrheitlich hypothesenkonforme Ergebnismuster in der ermittelten Dimensionsanzahl über die acht Stichproben zeigen. Damit kann zunächst nicht davon ausgegangen werden, dass das SRA ein ein-, vier- oder achtdimensionales Konstrukt im Sinne der Hypothese 1a) einheitlich abbildet.

Auch die erwarteten Ladungsmuster laut Hypothese 1b) lassen sich in den Daten nur in äußerst eingeschränktem Maß erkennen. Lediglich zwei Itemgruppen (Items 18, 19 und 20 sowie Items 9 und 11), d. h. fünf der insgesamt 20 Items laden überwiegend einheitlich über alle Stichproben und beide Zeitpunkte hinweg auf einem bzw. zwei Faktoren. Dieses Ergebnis ist zwar hypothesenkonform und kann als konvergenter Validitätsbeleg verstanden werden, jedoch laden andere, laut Erwartung demselben Faktor zugeordnete Items (8 und 13 sowie 10) nicht oder nur in inkonsistenter Weise auf den gemeinsamen Faktor. Eine Erklärung für dieses Muster, d. h. die recht konsistenten Ladungen der o. g. Items auf jeweils einen gemeinsamen Faktor, könnte darin bestehen, dass die Itemgruppen jeweils abhängige Items darstellen. Items 19 und 20 teilen sich den Aufgabenstamm (siehe Tabelle 26) und unterscheiden sich lediglich durch die Aufgabenstellung „welches […] ist am wahrscheinlichsten?“ und „welches […] ist am unwahrscheinlichsten?“. Item 18 unterscheidet sich von den Items 19 und 20 dadurch, dass statt der Ergebnisse von drei Würfelwürfen die Ergebnisse von zwei Würfelwürfen vorgegeben werden. Damit weisen die drei Items ein sehr ähnliches Aufgabenformat auf und zwei der Items sind inhaltlich voneinander abhängig. Ähnliches gilt für die Items 9 und 11: sie teilen sich den Aufgabenstamm und unterscheiden sich, ebenso, wie die Items 19 und 20 lediglich durch die gegensätzlich formulierte Aufgabenstellung. Damit sind die beiden Items inhaltlich ebenfalls voneinander abhängig. Die hohen Iteminterkorrelationen und in der Folge die hohen Ladungen auf einen gemeinsamen Faktor könnten daher auch ein statistisches Artefakt in der Form eines Methodenfaktors darstellen. Um diese Erklärung ausschließen zu können, müssten in einer oder mehreren neuen Stichproben die in den Items

erfassten Konzepte in unterschiedlichen Itemformen oder Inhaltskontexten vorgegeben werden. Insgesamt wird die Hypothese 1b) damit als nicht bestätigt beurteilt.

Nicht überraschenderweise bestätigt sich das Bild bei den konfirmatorischen faktoriellen Analysen der empirischen Datenstruktur laut Hypothese 1c). In den beiden Stichproben, in denen eine Konvergenz für die Schätzung der Modellparameter für das Nullmodell und eines der hypothetisierten Modelle erreicht werden konnte, zeigt das Nullmodell mit jeweils einem „Würfelaufgaben"-, einem „Münzwurfaufgaben"- und einem „Restaufgaben"-Faktor bessere Modellpassung im Vergleich zu dem postulierten Modell. Auch wenn die Ergebnisse der Prüfung der Hypothese 1c) mit äußerster Vorsicht betrachtet werden sollten, da lediglich zwei Vergleiche aus 24 theoretisch möglichen (Vergleich des Nullmodells mit dem 1-, 4- und 8-Faktor-Modell in insgesamt acht Stichproben) vorgenommen werden konnten, kann für die Hypothese 1c) kein Beleg erbracht werden.

Insgesamt deuten zusätzlich die statistisch-mathematischen Probleme bei der Durchführung der Parallelanalysen (z. B. singuläre Datenstrukturen aufgrund leerer Kreuztabellenzellen), der explorativen Faktoren- und Hauptkomponentenanalysen (negative Eigenwerte, nicht positiv definite Matrizen, Varianzen gleich null, Ultra-Heywood-Fälle usw.) sowie der konfirmatorischen Faktorenanalysen (fehlende Konvergenz) in den meisten Stichproben auf eine defizitäre faktorielle Struktur der vorliegenden Daten. Da die Probleme auch in Analysen der Stichprobe 1 auftreten, können die teilweise sehr geringen Stichprobenumfänge (z. B. in Stichproben 2 und 3) als Erklärung hierbei nachgeordnet werden.

Bei den vorgenommenen explorativen faktorenanalytischen Verfahren wurde für die Schätzung der Parameter das WLS-Vorgehen gewählt. Es kann kritisch angemerkt werden, dass alternative Verfahren, wie der WLSMV-Schätzer, der in dieser Arbeit bei den konfirmatorischen Verfahren eingesetzt wurde, auch für die explorativen Faktorenanalysen genutzt werden sollte. Auch können prinzipiell noch zahlreiche weitere Analysetechniken bei den Datenstrukturanalysen zum Einsatz kommen. So könnten Prokrustes-Zielrotationen zusätzliche Auskunft über die Modellpassung liefern, konfirmatorische Faktorenanalysen anhand der Produkt-Moment-Korrelationsmatrizen durchgeführt oder andere Schätzer eingesetzt werden. Inwiefern ein solches Vorgehen noch als hypothesenprüfend beurteilt werden kann,

ist allerdings unklar („if you torture the data long enough it will confess“, Coase, zitiert nach Good, 1983, S. 289). Daneben sollten alle genutzten Verfahren, die sich mathematisch zum Teil nur geringfügig unterscheiden, zu zumindest ähnlichen Ergebnissen führen, oder mit den Worten von Cronbach (1970):

> The variation between [results] is confusing, just as it confuses the beginning student of geography to find different maps picturing Greenland in different ways. These differences are of little concern to the nonspecialist; *the important thing is that all maps agree that there is such a large island in the North Atlantic*. (S. 310, Hervorhebung hinzugefügt)

Daher liefern nennenswerte Unterschiede in den Ergebnissen der Analysen (etwa in der ermittelten Anzahl von Faktoren) Hinweise darauf, dass in Cronbachs Worten, die Existenz der „großen Insel im Nordatlantik“ mindestens bezweifelt werden kann. Die Ergebnisse aller durchgeführten Analysen deuten zudem auf eine empirische Zusammengehörigkeit der Items 18, 19 und 20 sowie, in eingeschränkterem Maß, der Items 9 und 11. Dieser Umstand führt vor Augen, dass stabile und robuste Zusammenhänge sowohl in Stichproben sehr unterschiedlicher Größen als auch bei Verwendung unterschiedlicher Zusammenhangskoeffizienten sowie Schätz- und Extraktionsmethoden beobachtet werden können. Es kann daraus geschlossen werden, dass die übrigen Items des SRA ein solches stabiles Zusammenhangsmuster möglicherweise grundsätzlich nicht aufweisen. Ungeachtet der in der vorliegenden Arbeit formulierten Fragestellungen kann es natürlich von großem Interesse sein, die Struktur einer Itemsammlung, wie sie im SRA vorliegt, in größerem Detail und aus verschiedenen (auch mathematisch-statistischen) Gesichtspunkten zu betrachten. Dies ist jedoch nicht das primäre Ziel der vorliegenden Untersuchungen.

Die Ergebnisse der faktorenanalytischen Untersuchungen decken sich mit den ermittelten Itemeigenschaften. So liegen auch in den Mustern der Trennschärfen, der Iteminterkorrelationen und der internen Konsistenzen kaum Belege für die postulierte ein-, vier- oder achtdimensionale Struktur vor. Alle Analysen liefern jedoch übereinstimmend Hinweise darauf, dass die Itemgruppen 18, 19 und 20 sowie 9 und 11 in konsistenter Weise zusammenhängen.

Bei der angebotenen Interpretation der Daten sollten einige einschränkende Bedingungen beachtet werden. So können die Stichprobengrößen insgesamt als klein gelten, insbesondere bei der Verwendung von Schätzern, die keine multivariate

Normalverteilung voraussetzen (Eid et al., 2017). Eine weitere Einschränkung betrifft die relativ zahlreichen geringen und hohen Schwierigkeiten. So liegen die Schwierigkeiten in den meisten Stichproben für mindestens fünf Items unter 30% bzw. über 75% (ebenfalls mind. 5 Items). Dieser Umstand führt dazu, dass Zusammenhänge zwischen den Items nicht angemessen bestimmt werden können. Zwar gibt es Verfahren, wie den Yules Q Koeffizienten oder die tetrachorische Korrelation, die die eingeschränkte Streuung in den Variablen mathematisch beachten. Jedoch bleibt das Problem ein Folgendes: wenn die meisten (oder wenigsten) Personen ein Merkmal in einer bestimmten Ausprägung aufweisen, dann steigt die Wahrscheinlichkeit dafür, dass es sich um bloßen Zufall handelt, wenn die wenigen Personen mit der anderen Ausprägung ein bestimmtes anderes Merkmal zeigen oder nicht zeigen. Mit anderen Worten: wenn von den wenigen Personen, die Item 18 gelöst haben, keine das Item 15 löst, das ebenfalls nur von sehr wenigen Personen gelöst wurde, dann kann das durchaus ein Zufallsergebnis darstellen. Um das Problem zu lösen, müssen Populationen unterschiedlichen Kompetenzgrades mit (sehr) großen Stichprobenumfängen getestet werden. Dies war in der vorliegenden Untersuchung nicht der Fall.

Es kann weiterhin der Einsatz von Cronbachs α als Angabe für die interne Konsistenz kritisiert werden. Die Kennzahl setzt τ-äquivalente Indikatoren voraus und liefert daher lediglich für eindimensionale Skalen mit Items, die in gleichem Ausmaß auf die latente Variable laden, angemessen hohe Schätzungen der internen Konsistenz. In der vorliegenden Arbeit wird jedoch für jede der untersuchten Skalen (d. h. die Gesamtskala statistical reasoning, die Typen des statistical reasoning und die Subskalen des statistical reasoning) analysiert, inwiefern es sich um jeweils eindimensionale homogene Konstrukte handelt. In dem Fall etwa, wenn statistical reasoning, wie es durch das SRA erfasst wird, eine achtdimensionale Struktur aufwiese, wäre dies erkennbar in wesentlich höheren Cronbach α-Werten für Subskalen im Unterschied zu den Cronbach α-Werten für die Typen oder die Gesamtskala des correct statistical reasoning. Da auch die gegenüber der τ-Äquivalenz schwächere Anforderung der τ-Kongenerität durch das McDonald ω keine nennenswerten Verbesserungen der internen Konsistenz erbrachte, wurde auf

die Darstellung der weniger eindeutig interpretierbaren Kennzahl[31] verzichtet. Aus den genannten Gründen scheint die Verwendung von Cronbachs α in dem vorliegenden Zusammenhang daher gerechtfertigt.

Wie oben bereits erwähnt muss außerdem bei der Interpretation aller faktorenanalytisch ermittelten Ergebnisse (d. h. Parallelanalysen, Screeplots, Ladungstabellen) beachtet werden, dass aufgrund zahlreicher statistisch-mathematischer Probleme bei der Schätzung der Parameter die Aussagekraft der ermittelten Faktoren- bzw. Komponentenanzahl sowie der Ladungsmuster äußerste Vorsicht geboten ist. Dieser Umstand ist jedoch selbst ein Hinweis auf eine inkonsistente Struktur der Daten.

Zuletzt sollte beachtet werden, dass die in der vorliegenden Arbeit genutzten dimensionsanalytischen Verfahren eine Einfachstruktur in den Daten annehmen. Hierarchische empirische Beziehungen (etwa in Form von sogenannten „Methodenfaktoren“ durch ähnliche Itemformate oder Itemkontexte) könnten daher zu einer eingeschränkten Aussagekraft der Ergebnisse führen.

7.1.2. Fragestellung 2: Effekte der Lehrintervention

Die Hypothesen zur zweiten Fragestellung nach Gruppenunterschieden und Veränderungen der erzielten Punkte in Items zu in der Lehre behandelten Inhalten sowie der Schwierigkeiten dieser Items 2a) bis 2c) können in eingeschränktem Maß bestätigt werden. Die Ergebnisse der Varianzanalyse und der Mittelwertvergleiche zwischen den vier Stichproben am Ende des Semesters zeigen für die in den Stichproben 3 und 4 behandelten Inhalte hypothesenkonforme geringe Unterschiede in den erzielten Punkten zugunsten der Stichproben 3 und 4. Studierende der Psychologie erreichen – erwartungsgemäß – in den behandelten Inhalten am Ende des Semesters im Schnitt 2% bis 9% höhere Lösungen in den entsprechenden Items im Vergleich zu Studierenden der Sonderpädagogik, bei denen dieselben Inhalte nicht behandelt wurden. Die Effekte der Unterschiede sind mit einer Varianzaufklärung von ca. 3% recht gering und mit standardisierten

[31] Das McDonald ω lässt sich nicht ganz einfach interpretieren, weil es auf der Basis von explorativen Strukturgleichungsmodellen ermittelt wird, in denen bei besserer Passung einzelne Items weggelassen werden. Aus Sicht der Autorin handelt es sich bei der Kennzahl daher eher um eine explorative Möglichkeit der Untersuchung interner Konsistenz bzw. Homogenität einer bereits theoretisch konfirmierten Skala.

Mittelwertdifferenzen zwischen $0.16 < d < 0.74$ geringen bis moderat-hohen Ausmaßes. Die normativ-praktische Relevanz der Unterschiede lässt sich damit als eher unbedeutend einschätzen. Für die Lehre über ein bzw. zwei Hochschulsemester wären sicherlich höhere Effekte zu wünschen. Ein Vergleich der Lösungen von Items zu behandelten Inhalten mit den Lösungen von Items zu nicht behandelten Inhalten bei Studierenden der Psychologie zeigt zudem für Studierende im zweiten Semester (Stichprobe 4) höhere Lösungen von Items zu nicht-behandelten Inhalten (wenn auch in geringem Ausmaß mit $d = .23$). Dieses Ergebnis ist überraschend und zunächst nicht erwartungsgemäß, da das zweite Semester dieser Studierendenstichprobe einen sehr großen Schwerpunkt auf Wahrscheinlichkeiten beinhaltet und die Items mehrheitlich Inhalte zur Wahrscheinlichkeit abbilden. Unter Beachtung dieser Einschränkungen können die Ergebnisse als vorläufiger jedoch eingeschränkter Beleg für die Hypothese 2a) betrachtet werden.

Die Hypothese 2b) lässt sich ebenfalls nur in eingeschränktem Maß bestätigen. Es lässt sich zwar eine Steigerung von ca. 5% bzw. 1% für die Studierenden der Psychologie nach dem ersten bzw. zweiten Semester Statistiklehre beobachten, allerdings sind die Effekte sehr gering ($d = .26$ und $d = .08$). Da auch für die Studierenden der Sonderpädagogik und die Studierenden im Master of Arts kaum Zuwächse in der Lösung derselben Items beobachtet werden können (0.3%, $d = .02$ und ca. 1%, $d = .05$) ist sowohl die konvergente als auch die diskriminante Validität des Ergebnisses eher gering.

Auch für die Hypothese 2c) laut der Items behandelter Inhalte für Studierende, die in den Inhalten instruiert wurden, leichter zu lösen sein sollten, liegen lediglich eingeschränkte Belege vor. Allein bei Psychologiestudierenden im ersten Semester lassen sich relativ klare Steigerungen in der Lösungshäufigkeit der entsprechenden Items beobachten. Die Differenzen selbst bewegen sich jedoch, mit Ausnahme eines Items (das am Ende des Semesters seltener gelöst wird, als zu Beginn des Semesters), im Bereich von ca. 3% bis 28% und sind damit als höchstens moderat in ihrer praktischen Bedeutsamkeit zu beurteilen. In der Stichprobe der Psychologiestudierenden im zweiten Semester lassen sich keine hypothesenkonformen Unterschiede der Schwierigkeitsdifferenzen beobachten.

Insgesamt fallen die Ergebnisse zur Analyse der Sensitivität des SRA für Lehre ähnlich zu denen von Gundlach et al. (2015) aus. Die Autoren berichten einen

Anstieg von 4% gelöster Items nach einer traditionellen Lehrveranstaltung. Die Schwierigkeiten der SRA-Items veränderten sich in der vorliegenden Untersuchung in stärker erwartungsgemäßem Maß verglichen mit denen von Olani et al. (2011) berichteten. Während die Autoren lediglich bei sieben von 15 Items Anstiege der Schwierigkeit (d. h. Leichtigkeit) beobachten konnten, sind neun von zehn Items in der vorliegenden Arbeit durch Psychologiestudierende nach dem ersten Semester häufiger gelöst worden. Mit einer durchschnittlichen Lösung von ca. 53% und 50% der Items zu behandelten Inhalten lassen sich die Ergebnisse der vorliegenden Studie als etwas geringer einordnen, als die von Garfield (2003, 57% und 61%) und Gundlach et al. (2015, 54%, 58% und 60%) berichteten. Die relative Ähnlichkeit der empirischen Ergebnisse insgesamt lässt sich auch als Beleg einer annähernd äquivalenten Übertragung des SRA ins Deutsche werten.

Hinsichtlich der Sensitivität des SRA für Lehrinterventionen können die Ergebnisse insgesamt als numerisch mindestens teilweise im Sinne der Hypothesen bewertet werden. Aus praktischen Gesichtspunkten kann das SRA jedoch nicht als sensitiv für Lehrinterventionen gelten.

Für die erfolgten Analysen wurden Itemgruppen betrachtet und es kann sicher von Interesse sein, die Lösungshäufigkeiten für einzelne Items über die Stichproben zu vergleichen. Auch weitere Follow-Up-Erhebungen könnten ein vollständigeres Bild über den Verlauf des Lösungsverhaltens liefern. Weiterhin könnten gezielte Modifikationen der Lehre daraufhin untersucht werden, ob z. B. eine eher technisch orientierte Lehre dazu führt, dass bestimmte Items des SRA von mehr Studierenden gelöst werden. Bevor jedoch solche detaillierteren empirischen Antworten auf die Frage nach der Lehrzielvalidität der Items formuliert werden, scheint es sinnvoll, zunächst die Items im Hinblick ihrer Eignung als Kriterium zu beleuchten. Einige Gedanken und Vorschläge hierfür werden im weiteren Verlauf dargelegt (Abschnitte 7.2.1 und 8).

Zu beachten ist bei der Interpretation der Ergebnisse zudem, dass die Vergabe eines Punkts für die korrekte Lösung eines Items „streng“ vorgenommen wurde: wenn Studierende neben der korrekten Antwortoption zusätzlich eine nicht zutreffende Antwortoption bei Items im Einfachwahlformat wählten, wurde die Antwort als „nicht korrekt“ mit der Vergabe von 0 Punkten bewertet. Es ist daher möglich, dass sich die Ergebnismuster bei einer wohlwollenderen Bewertung

ändern. Die Wahl einer nicht-korrekten Antwortoption wird jedoch in der vorliegenden Arbeit als Hinweis auf noch bestehende Unklarheiten bzw. Defizite im Verständnis gedeutet und damit als (noch) Nicht-Beherrschung des betreffenden Inhalts interpretiert.

Die sicherlich größte Einschränkung der Aussagekraft der Ergebnisse ist die lose Verbindung zwischen der Lehre in den Statistikveranstaltungen der Studierenden der Psychologie (Stichproben 3 und 4) und den Iteminhalten. In der Folge kann nicht entschieden werden, inwiefern die geringen und teilweise nicht hypothesenkonformen Effekte auf Lehrdefizite oder Defizite im Erhebungsinstrument zurückzuführen sind. Dass spezifische Iteminhalte in den Veranstaltungen nicht explizit aufgenommen wurden, ist in erster Linie der Tatsache geschuldet, dass die Studierenden im Rahmen der Lehre auf Modulprüfungen vorbereitet werden sollen. Das führt zu gewissen Einschränkungen in den Entscheidungsfreiräumen von Lehrenden über die zu lehrenden Inhalte. Somit lagen für die Umsetzung der Validitätsprüfung praktische Grenzen vor. Auch hier jedoch sei noch einmal auf die Relevanz und Angemessenheit der SRA-Items als Kriterium grundsätzlich verwiesen: stellen die Items des SRA ein normativ-rational sinnvolles Lehrziel dar, so ist es sicher von hoher Bedeutung, Veränderungen in der Lehre so lange vorzunehmen, bis Studierende eine angemessene Lösungshäufigkeit in den entsprechenden Items erreichen. Wenn jedoch andere Lehrziele als die im SRA abgebildeten eine höhere Relevanz haben, dann wäre es vermutlich von größerem Nutzen, die Bemühungen in der Lehre auf diese Ziele auszurichten.

Eine weitere Einschränkung betrifft mögliche systematische Stichprobenfehler. Es ist vorstellbar, dass Studierende, die sich in den behandelten (modulprüfungsrelevanten) Inhalten besonders sicher oder aber besonders unsicher fühlen, an den letzten Sitzungen der Statistikveranstaltung nicht mehr teilnehmen. Dies hätte zur Folge, dass die statistischen Stichprobenkennwerte die tatsächlichen Populationsparameter in systematischer Weise unter- oder überschätzen.

Schließlich sollte bei der Interpretation der Ergebnisse die zum Teil sehr geringe Anzahl von Wertepaaren beachtet werden. Während etwas mehr als 100 Studierende der Sonderpädagogik ($n = 107$ bzw. $n = 108$) sowohl zu Beginn als auch am Ende des Semesters befragt werden konnten, liegen für die anderen Stichproben der Studierenden lediglich 13 (Studierende im Master of Arts), 19

(Studierende der Psychologie im ersten Semester) und 33 (Studierende der Psychologie im zweiten Semester) vollständige Messwertpaare vor.

7.1.3. Fragestellung 3: Zusammenhänge zu anderen Merkmalen

Für die Hypothesen 3a) bis 3c) zur Fragestellung nach Zusammenhängen zwischen den im SRA erreichten Punkten und kognitiven Maßen sowie einem anderen akademischen Leistungsmaß können ebenfalls in eingeschränktem Maß Belege erbracht werden. Mathematische Fertigkeiten zeigen in linearen Regressionen eine sehr geringe bis geringe inkrementelle Varianzaufklärung ($.004 < \Delta R^2 < .065$). Hierbei zeigen sich Unterschiede in den Effekten für die Varianzaufklärung der am Ende des Semesters erzielten SRA-Punktesumme bei der vorherigen Berücksichtigung der zu Beginn des Semesters erzielten SRA Punktesumme in Abhängigkeit von der Berücksichtigung des Zeitpunkts der Einstellungserfassung (schwacher Interaktionseffekt). So reduziert sich der Anteil der durch mathematische Fertigkeiten aufgeklärten Varianz bei Berücksichtigung der SRA-Punktesumme zu Beginn des Semesters deutlicher für Einstellungen zu Statistik am Ende des Semesters (gegenüber Einstellungen zu Statistik zu Beginn des Semesters). Dieses Ergebnis spricht dafür, dass bei Personen mit gleichen Einstellungen zu Statistik am Ende (nicht aber zu Beginn) des Semesters und gleichen Leistungen in den Statistikitems zu Beginn des Semesters sowie gleichen Leistungen in den Figurenreihenaufgaben und Aufgaben zum deduktiven Schließen das Ausmaß mathematischer Fertigkeiten keinen (bzw. einen äußerst geringen) Beitrag zur Vorhersage der Leistungen in Statistik am Ende des Semesters leistet. Eine mögliche Erklärung für dieses Ergebnismuster liegt darin, dass die Einstellungen zu Statistik am Ende des Semesters stärker von den Leistungen in Statistik zu Beginn des Semesters abhängen, als von den Leistungen in Statistik am Ende des Semesters. Diese Erklärung findet Bestätigung in den Korrelationen nullter Ordnung in der Stichprobe der Sonderpädagogik-Studierenden (Tabelle 21). Es ist vorstellbar, dass die Kompetenz, besonders zu Beginn des Semesters, einen Einfluss auf die Einstellungen zu Statistik nimmt, darunter insbesondere die wahrgenommene eigene und als notwendig eingeschätzte Kompetenz in Statistik sowie die affektive Einstellung zu Statistik. Eine solche Erklärung hat natürlich lediglich hypothetischen Charakter und lässt sich im Rahmen dieser Arbeit nicht überprüfen.

Ein sehr ähnliches Ergebnismuster zeigt sich für das deduktive Schließen. Auch hier variiert die inkrementelle Varianzaufklärung ($.001 < \Delta R^2 < .04$) in Abhängigkeit von der Berücksichtigung der SRA Punktesumme zu Beginn des Semesters in Kombination mit der Berücksichtigung des Zeitpunkts der Einstellungserfassung. Werden Einstellungen zu Statistik am Ende des Semesters, mathematische Fertigkeiten, die Anzahl der gelösten Figurenreihen-Aufgaben und der SRA-Punktesumme zu Beginn des Semesters kontrolliert, reduziert sich die Varianzaufklärung für das deduktive Schließen auf ungefähr null.

Für die Figurenreihen-Items zeigt sich hingegen keine nennenswerte Reduktion in der inkrementellen Varianzaufklärung von $.02 < \Delta R^2 < .04$ bei gleichzeitiger Kontrolle der anderen Variablen. Dieses Ergebnis, übereinstimmend mit dem von Martin et al. (2017) berichteten Befund, kann als Hinweis darauf gewertet werden, dass die SRA-Items tatsächlich einen „reasoning"-Aspekt erfassen. Es sollte hierbei allerdings berücksichtigt werden, dass der Effekt der inkrementellen Varianzaufklärung als gering einzuschätzen ist.

Damit liegen eingeschränkte Belege für die Hypothese 3a) vor, dass die Items des SRA zur Messung des statistical reasonings im Sinne eines statistischen Kompetenz-Konstrukts (*statistical* reasoning) herangezogen werden können. Betrachtet man die Leistungen der Studierenden im SRA ausschließlich am Ende des Semesters, kann eine geringe inkrementelle Varianzaufklärung beobachtet werden und somit ein schwacher Hinweis auf konvergente und diskriminante Validität. Werden jedoch die Leistungen von Studierenden im SRA zu Beginn des Semesters berücksichtigt, können keine Belege für die Hypothese gefunden werden.

Die Hypothese 3b) lässt sich, wenn auch in schwachem Ausmaß, bestätigen. Für die Figurenreihenaufgaben, wie sie in sprachfreien Intelligenztests verwendet werden, lässt sich eine geringe, jedoch stabile inkrementelle Varianzaufklärung beobachten. Damit kann vermutet werden, dass das SRA einen allgemeinen kognitiven Prozess oder ein allgemein-kognitives Konstrukt (statistical *reasoning*) erfasst.

Für die Hypothese 3c) lassen sich engere Zusammenhänge zwischen der SRA-Punktzahl und einem akademischen Leistungsmaß beobachten als die von Tempelaar et al. (2006) berichteten. Für Studierende der Psychologie im zweiten Semester (Stichprobe 4) liegt jeweils eine moderate negative Korrelation zwischen

der Statistiknote für das erste Semester und der SRA Punktesumme zu Beginn ($r = -.31$, $n = 45$) und am Ende des darauffolgenden Semesters ($r = -.33$, $n = 30$) vor. Das Ergebnis liefert einen Hinweis auf konvergente Validität des SRA zur Messung von Lehrzielen in einer Statistikveranstaltung, auch wenn der Zusammenhang niedriger ausfällt, als in der Hypothesenformulierung gefordert. Eine mögliche Erklärung für die geringere Korrelation kann darin liegen, dass die Erhebung des SRA zwei Monate nach der Prüfungsleistung stattfand und höher ausgefallen wäre, wenn die beiden Maße zeitgleich erhoben worden wäre. Auf der anderen Seite bleibt die Korrelation auch nach fünf Monaten stabil, so dass ein Vergessen von Inhalten entweder nur in dem Zeitraum zwischen der Prüfung und der Erhebung vorlag oder aber als Hinweis auf die Messung eines stabilen kognitiven Merkmals gedeutet werden kann.

Einige der bereits genannten Einschränkungen gelten auch für die Interpretation der Ergebnisse zu den Hypothesen 3a) bis 3c). Diese betreffen in erster Linie die möglicherweise systematisch verzerrten Stichproben. Da die Erhebung der in den Regressionsanalysen genutzten Variablen bei den Psychologiestudierenden an drei und bei den Sonderpädagogikstudierenden an zwei Lehrveranstaltungssitzungen erfolgte und die Erhebungen einen recht hohen kognitiven und zeitlichen Aufwand darstellten, kann vermutlich mit Varianzeinschränkungen in den Variablen gerechnet werden. Dies hat zum einen zur Folge, dass die Ergebnisse sich nur auf eine spezielle Population verallgemeinern lassen und zum anderen, dass die Zusammenhänge in ihren Höhen unterschätzt werden.

Darüber hinaus muss auch mit Messfehlern insbesondere bei der Verwendung der Variablen zur Erfassung kognitiver Maße gerechnet werden. Zwar liegen einige Belege für die Validität der genutzten Verfahren, wie den Figurenreihenaufgaben sowie den Aufgaben zum deduktiven Schließen vor, jedoch handelt es sich hier nicht um Instrumente, für die langjährige empirische Erfahrungen gesammelt werden konnten. Auch die Aspekte der Reliabilität hätten eine Unterschätzung der tatsächlichen Zusammenhänge zwischen den Variablen zur Folge.

Es kann zudem kritisch angemerkt werden, dass die Prüfung der Voraussetzungen der linearen Regression nicht gesondert aufgeführt wurde. Zwar

wurden alle Variablen auf ihre Verteilungen hin inspiziert und auch die Linearität, wo möglich, auf den besten Modellfit untersucht, jedoch wurde auf eine gesonderte statistische Regressionsdiagnostik verzichtet. Bei mathematisch präzisen (und als vollständig angestrebten) Modellformulierungen und einer repräsentativen Stichprobe ist die Inspektion des Kreuzvalidierungsfehlers, die Sicherstellung oder Quantifizierung der Messfehlerfreiheit der unabhängigen Variablen, die Bestimmung des Ausmaßes der Multikollinearität und der Homoskedastizität von hoher Bedeutung für die Beurteilung des postulierten Modells. In der vorliegenden Fragestellung sind jedoch eher partielle Korrelationskoeffizienten sowie der Semipartialdeterminationskoeffizient (ΔR^2) von Interesse und weniger ein gesamtes Vorhersagemodell des SRA-Punktewerts. Aus diesem Grund erscheint es zulässig auf eine ausführliche Regressionsdiagnostik zu verzichten.

Die Ergebnisse der Korrelations- und Regressionsanalysen liefern Hinweise auf eine ganze Reihe sehr interessanter Beziehungen (Suppressions-, Redundanz- und Interaktionseffekte) zwischen den unterschiedlichen Variablen. Analysen solcher Beziehungsmuster könnten Antworten geben auf Fragen, wie „verändert sich die Einstellung in Abhängigkeit vom Vorwissen?“, „ist die Einstellung nur dann wichtig für die Leistung in Statistik, wenn ein bestimmtes Maß an Vorwissen besteht?“, „Weshalb weist die wahrgenommene Schwierigkeit von Statistik eine positive Beziehung zu der Leistung in Statistik bei gleicher eingeschätzter kognitiver Kompetenz auf?“ und viele andere mehr. Alle diese Fragen lassen sich jedoch erst dann sinnvoll und zielführend beantworten, wenn die Messung statistischer Leistung oder Kompetenz in adäquater Weise vorgenommen werden kann.

7.1.4. Eignung des SRA für die Messung statistischer Kompetenz

Vor dem Hintergrund der Ergebnisse der vorliegenden Untersuchungen und der bisherigen Forschung lassen sich kaum Belege dafür erbringen, dass das SRA statistical reasoning im Sinne eines ein-, vier- oder achtdimensionalen Konstrukts erfasst. Die Ergebnisse von Martin et al. (2017) sowie die der vorgelegten Korrelations- und Regressionsanalysen legen zudem nahe, dass die SRA-Messwerte einen nennenswerten Varianzanteil mit allgemein-kognitiven und mathematischen Variablen gemeinsam haben. Inkrementell für die Varianzaufklärung im SRA-Wert zeigen sich bei Martin et al. (2017) sowie in der vorliegenden Untersuchung vor

allem intelligenznahe Variablen. Inwiefern ein anhand des SRA gewonnener Testscore statistische Kompetenz im Sinne eines Konstrukts anzeigt, ist damit insgesamt unklar.

Hinsichtlich der Validität des SRA zur Erfassung des statistical reasoning im Sinne eines Lehrziels erweist sich die Befundlage ebenfalls als nicht eindeutig, auch hier können jedoch einige Einschränkungen der Validität identifiziert werden. Zwar lässt sich ein Zusammenhang zwischen einem akademischen Leistungsmaß und der SRA-Punktsumme beobachten, allerdings sind die konvergenten und diskriminanten Effekte für die erfolgreiche Lösung der SRA-Items zwischen verschiedenen Studierendengruppen insgesamt gering. Schließlich stellt der fehlende Bezug auf eine Aufgabengrundmenge eine grundlegende Einschränkung der Inhaltsrepräsentativität des Instruments dar.

Auf der Basis der hier vorgelegten Ergebnisse sowie der Ergebnisse aus Martin et al. (2017), Kunzi et al. (2018) kann vermutet werden, dass Items des SRA möglicherweise spezifische kognitive Prozesse, wie z. B. Typ-2-Prozesse im Sinne der Zwei-System-Theorien erfassen. Welche kognitiven Prozesse dies sind, kann jedoch auf Basis der vorliegenden Informationen nicht beantwortet werden. Darüber hinaus bleibt auch hier offen, inwiefern diese Prozesse im Sinne von statistical reasoning interpretiert werden können.

Damit muss zumindest in Frage gestellt werden, ob mit Hilfe des SRA das Konzept „statistical reasoning“ in operationale Begriffe überführt wird. In der Folge kann nicht davon ausgegangen werden, dass das Konzept „statistical reasoning“ sich eignet, um die Beherrschung statistischer Inhalte in analytische (d. h. theoretisch-inhaltliche) Begriffe zu überführen. Das SRA wird daher als nicht optimal geeignet beurteilt, um die Beherrschung statistischer Inhalte bei Studierenden und die Veränderungen der Beherrschung statistischer Inhalte durch Hochschullehre bei Studierenden zu erfassen.

7.2. Einschränkungen der Forschungsbemühungen allgemein

Sowohl für die vorliegende Arbeit als auch für die bisherigen Forschungsbemühungen zur Erfassung von statistischen Kompetenzen lassen sich zudem Einschränkungen hinsichtlich drei allgemeiner Aspekte identifizieren. Diese betreffen die explizite Formulierung der Zielsetzung, die Güte der theoretischen Grundlagen und die Güte der methodischen Umsetzung bei der Entwicklung der Erhebungsverfahren.

7.2.1. Zielsetzung der Messung.

Hinsichtlich der Zielsetzung stellt sich als Erstes die Frage nach dem Einsatzzweck und der angestrebten Aussagekraft von durch Erhebungsverfahren erzielten Scores. Für die Konstruktion von Messinstrumenten statistischer Kompetenz ergeben sich unterschiedliche Konsequenzen, ob etwa eine Aussage über die Beherrschung von grundlegenden mathematisch-statistischen Konzepten oder aber datenanalytisch-praktischen Verfahren getroffen werden soll. Beispiele hierfür wären die Erfassung des Lösungsverhaltens bei Aufgaben zur mathematischen Begründung von Schätzerkriterien (Erwartungstreue, Konsistenz, Effizienz und Suffizienz) gegenüber Aufgaben zum Vergleich verschiedener zentraler Tendenzen (Modus, Median und Mittelwert). Während im ersten Fall mathematische Begründungen in der Form von Beweisen, wie sie im Mathematikstudium typisch sind notwendig wären, müssten im zweiten Fall vermutlich verschiedene Verteilungen vor dem Hintergrund inhaltlicher Fragestellungen vorgegeben werden. Dass es hierbei auf die genaue Gestaltung der Aufgaben ankommt, zeigen etwa die unterschiedlichen Schwierigkeiten (sowie die negativen oder sehr geringen positiven Korrelationen) über alle acht Stichproben hinweg für die beiden Items 1 und 4 des SRA (siehe Tabelle 26). Beide Items beziehen sich auf das Konzept der zentralen Tendenz und sind oberflächlich betrachtet sehr ähnlich gestaltet. Jedoch wird das Item 4 in sieben von acht Stichproben von ca. 20%-30% mehr Studierenden gelöst, als das Item 1. Dies könnte z. B. auf eine alltagsnähere Formulierung im Item 4 („typische Anzahl“ vs. „so akkurat wie möglich […] tatsächliches Gewicht“ in Item 1) zurückzuführen sein. Für die Interpretation des Test- oder Itemwerts einer Person stellt sich dann die Frage, worüber der oder die Testanwendende eine Auskunft erhält.

Eine weitere die Zielsetzung von Messverfahren betreffende Frage bezieht sich auf den Kontext, über den eine Aussage gemacht werden soll. So werden sicherlich unterschiedliche Items (oder zumindest Antwortoptionen) benötigt werden, um datenanalytisch-praktische Fertigkeiten z. B. im alltäglichen Kontext (etwa nach der Forderung von Wallman, 1993 oder Watson & Callingham, 2003) und im Kontext einer spezifischen empirischen Wissenschaftsdisziplin zu erfassen. Als Beispiel soll auch hierfür das Item 1 aus dem SRA angeführt werden. Im alltäglichen Kontext ist die Entfernung der deutlich abweichenden Werte zweckdienlich (etwa, um ein tragfähiges Regal für den Gegenstand zu konzipieren). Darüber hinaus kann meist recht klar entschieden werden, dass oder ob die Werte (fast) ausschließlich Messfehler darstellen. Demgegenüber ist der Umgang mit Ausreißerwerten in der wissenschaftlichen Psychologie wesentlich komplexer: hier ist zum einen oft nicht zu entscheiden, inwiefern sehr hohe Werte (ausschließlich) Messfehler darstellen und zum anderen ist der unmittelbare Zweck der Messung (etwa Konzeption von Interventionen) nicht klar oder nicht hilfreich für die Entscheidung nach der Berücksichtigung von deutlich abweichenden Werten. Schließlich greift für die Gestaltung von statistischen Modellen auch die Definition des Erwartungswerts, der trotz Ausreißer- und Extremwerten (solange diese unsystematischer Natur sind), die im Sinne des Kleinste-Quadrate-Kriteriums genaueste Angabe liefert. Diese Besonderheiten führen dazu, dass die Wahl der Antwortoption „Die 9 Zahlen aufaddieren und sie durch 9 teilen“ in bestimmten Kontexten zumindest weniger eindeutig als Fehlverständnis klassifiziert werden kann.

Ein letzter, für eine klare Zielsetzung einer Messung relevanter Aspekt, der hier beleuchtet werden soll, betrifft den formativen vs. summativen Verwendungszweck der Testscores. Soll eine Erhebung Auskunft über Ansatzpunkte für Interventionen liefern, so stellen sich andere Anforderung an Testitems, als wenn lediglich der status quo in einem bestimmten Merkmal von Interesse ist. Bezogen auf das Item 1 des SRA reicht die Information darüber, welche Antwortoption eine Lernende oder ein Lernender gewählt hat aus, um beurteilen zu können, ob die Person die Aufgabe korrekt gelöst hat bzw. welcher Art die spezifische nicht-korrekte Antwort ist. Eine Information darüber, an welcher konkreten Stelle eine Lehrintervention im Falle einer nicht-korrekten Antwort

ansetzen sollte, lässt sich daraus jedoch nicht ableiten. Die Gestaltung des Items (bzw. des Instruments) erlaubt damit keine Aussage über die Gründe einer nicht-korrekten Beurteilung. Ein ähnlicher Einwand (oder Vorwurf) wurde an den Forschungsansatz zu Heuristiken und Urteilsverzerrungen (z. B. Gigerenzer, 1996) herangetragen (siehe auch Abschnitt 2.4.1) und betrifft den zweiten weiter oben aufgeführten Punkt der Güte theoretischer Grundlagen von Forschungsbemühungen.

7.2.2. Theoretische Grundlage der Messung.

Kurt Lewin begründete die Relevanz theoretischer Grundlagen für die Bearbeitung praktischer Fragestellungen in seiner Schriftensammlung von 1951 treffend mit den Worten

> The greatest handicap of applied psychology has been the fact that, without proper theoretical help, it had to follow the costly, inefficient, and limited method of trial and error. […] close cooperation between theoretical and applied psychology […] can be accomplished […] if the theorist does not look toward applied problems with highbrow aversion […], and if the applied psychologist realizes that there is nothing so practical as a good theory. (S. 169)

Auch wenn Lewin in diesem Zusammenhang die Aversion von theoretisch arbeitenden Wissenschaftlern praktischen Fragestellungen gegenüber anmerkt, ist für die vorliegende Arbeit vor allem die an vielen Stellen defizitäre Berücksichtigung der Theorie durch angewandte Wissenschaftler relevant. Ein großer Teil der Forschung zum Lernen und Lehren von Statistik, ebenso wie ein großer Teil der Forschung zu Urteilsverzerrungen beim Schließen unter Unsicherheit ist durch „the costly, inefficient, and limited method of trial and error“ (Lewin, S. 169) gekennzeichnet. So liegen verschiedene Studien zu Interventionsansätzen für die Förderung statistischer Kompetenzen vor (siehe z. B. die *Special Issue on Statistics and the Undergraduate Curriculum* in der Zeitschrift *The American Statistician*, Überbick der Literatur bis 1996 von Becker [1996] und Artikel in der Zeitschrift *Statistics Education Research Journal*), jedoch prüfen nur wenige Arbeiten spezifische theoretisch begründete Aspekte (z. B. Gigerenzer & Hoffrage, 1995 oder Penna, Agus, Peró-Cebollero, Guàrdia-Olmos & Pessa, 2014 zu

Repräsentationsformaten von Informationen). Keine der Autorin bekannte Studie bezieht jedoch explizit breiter angelegte theoretische Ansätze, wie etwa die Cognitive Load Theory (Sweller, van Merrienboer & Paas, 1998) oder die Theorie mentaler Modelle (Johnson-Laird, 2012) mit ein. In der Folge erlauben die Ergebnisse der Forschungsbemühungen nur wenig Generalisierung über die in den Studien selbst untersuchten Interventionen hinaus. Hat sich etwa ein bestimmtes Instruktionsmittel als hilfreich bei der Lehre eines statistischen Inhalts gezeigt, so stellt sich die Frage, welcher Aspekt des Lehrmittels hierfür kritisch gewesen ist und ob auch für den Einsatz in anderen statistischen Inhalten ähnliche Effekte zu erwarten sind. Die Identifikation von kritisch-wirksamen Interventionsansätzen oder -techniken sowie der interventionssensitiven Inhalte stellen gleichermaßen das Ziel der Entwicklung von Theorien dar.

Bei der Konzipierung von Erhebungsinstrumenten der kritischen Zielvariablen muss daher die theoretische Grundlage bewusst berücksichtigt werden, wenn eine Aussage über die Gültigkeit der Theorie möglich sein soll. Die empirische Prüfung der Theorie anhand des Erhebungsinstruments erlaubt in der Folge eine Aussage über die Bewährung der Theorie, die wiederum genauere und allgemeinere Aussagen über zu erwartende empirische Effekte macht. Ein solches Vorgehen hat zur Folge, dass sich theoretische und empirische Analysen gegenseitig korrigieren und ergänzen und somit sowohl zu (rein theoretischem) Erkenntnisfortschritt, als auch zum größeren praktischen Nutzen führen. Für formative Messziele, wie sie für die Lehre allgemein vermutet werden können, ist eine sinnvolle theoretische Grundlage daher von hoher Relevanz.

Um Ansatzpunkte für die Verbesserung der Lehre statistischer Inhalte zu identifizieren, werden sich nicht alle theoretischen Grundlagen gleichermaßen eignen. Ansätze, die den Einfluss von als stabil angenommenen Persönlichkeitsmerkmalen, wie etwa Persönlichkeitseigenschaften, Intelligenz oder (Leistungs-) Motivation zur Erklärung von Kompetenzunterschieden in Statistik nutzen, sind hier möglicherweise von nachgeordnetem Interesse. So ist zwar anzunehmen, dass etwa fluide Intelligenz eine Rolle beim Erwerb statistischer Kompetenz spielen wird. Wird jedoch fluide Intelligenz als nur in eingeschränktem Maß trainierbar angenommen, so liefert das Wissen über den Einfluss dieses Merkmals nur wenige Ansatzpunkte für Veränderungen in Lehrinterventionen.

Ähnliche Folgen für mangelnde Auskunft über Interventionsansätze hat die Nutzung der Kompetenzbegriffe im engeren Sinne (z. B. Weinert, 2014, OECD, 2003). Erreichen Personen nicht den geforderten Ausprägungsgrad der Kompetenz, bleibt zunächst unklar, welche Merkmale der Person – oder der Lehre – dafür verantwortlich sind und in der Folge ist es zunächst nicht möglich, auf spezifische Defizite ausgerichtete Interventionen abzuleiten.

Angemessener scheinen für die Auseinandersetzung mit statistischer Kompetenz und deren Förderung im Rahmen der Lehre kognitionspsychologische Ansätze zu sein. Ein entscheidender Vorteil der kognitionspsychologischen Ansätze für die Erfassung von statistischer Kompetenz besteht darin, dass Messergebnisse ein differenziertes Bild über den Beherrschungsstand von gelehrten Inhalten liefern können. Enthält etwa ein Messinstrument Items zur Erfassung der mentalen Verfügbarkeit der Merkmale eines Begriffs (z. B. „Sensibilität für Ausreißer“ als Merkmal des arithmetischen Mittels und „Plausibilität oder Informationsgehalt“ als Merkmale für Ausreißer) und zusätzlich Items zur Erfassung der Verfügbarkeit von Relationen zwischen Begriffen (z. B. die Relationen in der Proposition „unplausible Ausreißer beeinträchtigen die Interpretation des arithmetischen Mittels stärker als die des Medians oder des Modalwerts“) sowie Items zur Erfassung von Schemata (z. B. „Wenn in einer Datenreihe unplausible Ausreißer vorkommen, sollte zusätzlich ein arithmetisches Mittel ohne Ausreißer als zentrale Tendenz angegeben werden“), so gibt es die Möglichkeit, Defizite in der Beherrschung eines Inhalts genauer zu bestimmen. Diese differenzierende Feststellung von Lehr- bzw. Lernbedarfen kann dann im Rahmen formativer Lehrevaluationsbemühungen genutzt werden. Ein weiterer Vorteil der kognitionspsychologischen Ansätze besteht darin, dass sowohl Interventionsansätze als auch Operationalisierungen der Statistikkompetenz, d. h. der Interventionswirkung, unmittelbar in das nomologische Netzwerk der zum Teil hoch informativen Modellvorschläge (z. B. Theorie mentaler Modelle, Johnson-Laird, 2012) eingebettet und empirischen Prüfungen unterzogen werden können. Messinstrumente, die Statistikkompetenz in den Begriffen der Kognitionswissenschaften erfassen, würden damit auch einen Beitrag zur Erforschung von Lehr- und Lernprozessen leisten.

7.2.3. Methodische Umsetzung der Messung.

Der dritte und letzte Aspekt, der den dargestellten Ansätzen zur Operationalisierung der Statistikkompetenz gemeinsam ist, betrifft das methodische Vorgehen bei der Konstruktion der Messinstrumente. Wie in den Teilen 2. und 3. der vorliegenden Arbeit dargestellt, ziehen bestimmte Entscheidungen hinsichtlich der Zielsetzung und theoretischen Bedeutung der Erhebungen bestimmte methodische Besonderheiten nach sich. Für die in der vorliegenden Arbeit vorgestellten Ansätze, die Messungen statistischer Kompetenz gebrauchen, können mindestens zum Teil Lehrziele als Messgegenstand vermutet werden. So erheben Watson und Callingham (2003) statistical literacy im Rahmen des Schulunterrichts und Garfield (2003), Park (2012), Ziegler (2015), delMas et al. (2007), Sabbag & Zieffler (2015), Sabbag et al. (2018), Lane-Getaz (2007) und Allen (2006) statistical reasoning oder andere statistische Kompetenzen innerhalb der Hochschullehre. Zur Messung von Lehrzielen, d. h. der Leistung in einem curricular verankerten Inhalt in Bezug auf ein zu erreichendes objektives Ziel, muss jedoch, wie im Verlauf der Arbeit dargestellt, insbesondere der Aspekt der Inhaltsvalidität oder Inhaltsrepräsentativität berücksichtigt werden. Osburn (1968) beschreibt die grundlegende Herausforderung hierbei folgendermaßen:

> Few measurement specialists would quarrel with the premise that the fundamental objective of achievement testing is generalization. Yet the fact is that current procedures for the construction of achievement tests do not provide an unambiguous basis for generalization to a well defined universe of content. At worst, achievement tests consist of arbitrary collections of items thrown together in a haphazard manner. At best, such tests consist of items judged by subject matter experts to be relevant to and representative of some incompletely defined universe of content. *In neither case can it be said that there is an unambiguous basis for generalization*. This is *because the method of generating items and the criteria for the inclusion of items in the test cannot be stated in operational terms* (S. 95, Hervorhebung hinzugefügt)

Das grundlegende Problem besteht damit in der nicht hinreichend begründbaren – jedoch angestrebten – Verallgemeinerung von den gemessenen Items auf die Gesamtheit dieser Items. Hat eine Person daher etwa das Item 1 im SRA zur

Berücksichtigung von Ausreißern gelöst, kann nicht von der Kompetenz einer Person zum Umgang mit Ausreißern gesprochen werden. Dies ist lediglich dann möglich, wenn das Item repräsentativ für eine Grundmenge von Items zum Umgang von Ausreißern ist. In keinem der vorgestellten Verfahren wird die Aufgabengrundmenge oder das universe of content in „operational terms" beschrieben, wie dies lehrziel- oder kriteriumsorientierte Skalen (auch criterion-referenced genannt) voraussetzen. Tatsächlich definieren die im Rahmen dieser Arbeit vorgestellten Erhebungsinstrumente teilweise relativ breit und eher ungenau gefasste Verhaltensziele (behavioral objectives). Damit sind die Lehrziele allerdings so ungenau angegeben, dass für ein jedes Lehrziel viele unterschiedliche Arten von Items formuliert werden können und damit keine adäquate Definition des Kriteriums (Lehrziels) zur Verfügung steht. Für das Item 1 (sowie die Items 4 und 17) des SRA kann etwa angenommen werden, dass es Auskunft geben soll über „knowing which [measures of centre, spread, and position] are best to use under different conditions, and how they do or do not represent a data set" (Garfield, 2003, S. 25). Für diese Beschreibung des Verhaltensziels lassen sich unterschiedliche Klassen von Items generieren, z. B. Items mit grafischen Inhalten „Entscheiden Sie für die folgende [schiefe] Verteilung, welche Zahl sie am besten repräsentiert" oder Items zu Variablen unterschiedlicher Informationsdichte (Skalenniveaus) und viele mehr. In der Folge ist die Aussage über die Fertigkeit einer Person, „understands how to select an appropriate average" (Subskala „Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird") nicht zulässig. Auf diese Gefahr hat Popham (1974) mit den Worten „[…] a criterion-referenced test that doesn't spell out its criterion satisfactorily might just as well be a *cloud-referenced* test for all the good it will do" (S. 615, Hervorhebung im Original) in seinem Artikel „An approaching peril: cloud-referenced tests" hingewiesen.

Besondere Gefahren ergeben sich im Rahmen der Test- bzw. Skalenentwicklung zudem mit den verfügbaren statistischen Auswertungsverfahren, allen voran Methoden zur Modellierung latenter Variablen (Item-Response-Modelle

oder Strukturgleichungsmodelle allgemein). Auch hier trifft Osburn (1968) den – aus Sicht der Autorin – Kern des Problems:

> Once the concept of a latent variable is introduced, attention to generalization over a defined universe of content is diverted to the measurement of the latent hypothetical continuum. […] Furthermore, the answer to the question, 'What is the test measuring?', can be couched in terms of inferred relationships between observables and the hypothetical continuum. Somewhere along the line we usually give the latent variable a name such as vocabulary, number facts, etc. and *we* usually *do not hesitate to endow the test scores with all surplus meanings that our label implies*. *Thus, by taking an arbitrary collection of items and referring them to a latent variable with a name, we create the illusion of generalization*. Statistical analysis of test data is, of course, very useful. But *no amount of item analysis or factor analysis can provide a firm basis for generalization to a universe of content*. (S. 96, Hervorhebungen hinzugefügt).

Auch wenn Osburn sich in seinen Äußerungen auf kriteriums- oder lehrzielorientierte Tests bezieht, gilt seine Warnung auch für konstruktorientierte Verfahren.

Die Messintention, ihre theoretische Einbettung sowie das methodische Vorgehen bei der Konzeption des Messinstruments stellen damit ein wechselseitig abhängiges Gebilde dar, das nur dann einen wissenschaftlichen Erkenntnismehrwert produziert, wenn die drei Komponenten aufeinander abgestimmt sind. Für die zukünftigen Forschungsbemühungen zur Statistikkompetenz könnten die einzelnen Komponenten noch etwas passgenauer aufeinander eingestellt werden. Einige Vorschläge, wie dies gestaltet werden könnte, finden sich im nächsten und schließenden Teil der Arbeit.

8. Empfehlungen und Ausblick

> *We need an instrument to assess where we are currently with respect to [...] the statistical thinking ability of [students]. Without some measure of [...] statistical literacy/thinking, we have no objective way to assess the effects of any changes in curriculum [or teaching methods]. The development of such instruments and execution of trials will be costly and time-consuming [...]. However without [it], we are then left to argue about the curriculum and statistical literacy/thinking based on [...] anecdotes [...] and gut feeling.*
>
> MacKay, 2016, S. 193

Das zentrale Anliegen der vorliegenden Arbeit bestand darin, zu beurteilen, inwiefern das SRA zur Messung der Beherrschung von Statistik im Rahmen der Hochschullehre in der Psychologie und der Sonderpädagogik herangezogen werden kann. Nach der theoretischen und empirischen Analyse des Instruments kann derzeit nicht davon ausgegangen werden, dass sich das Instrument gut eignet, um den Beherrschungsstand und Änderungen im Beherrschungsstand durch Hochschullehre in den beiden Fächern zu erfassen. Die Frage danach, wie die Beherrschung statistischer Inhalte während der Hochschulausbildung erfasst werden kann, ist jedoch von zentraler Bedeutung sowohl vor dem Hintergrund der Anforderungen an eine Akademikerin oder einen Akademiker als auch für die Beurteilung und Modifikation der Lehre in empirischen Hochschulfächern.

Eine zufriedenstellende Antwort auf diese Frage sollte Antworten auf drei Teilfragestellungen beinhalten. Zunächst muss geklärt werden, was unter „Beherrschung" verstanden werden kann. Dann muss begründet werden, welcher Beherrschungsgrad vorliegen soll, um als ausreichend im Sinne einer Mindestanforderung an Absolventen der Psychologie oder Sonderpädagogik zu gelten. Schließlich muss entschieden werden, welche Inhalte und Konzepte der Statistik beherrscht werden müssen, um die Mindestanforderung für den jeweiligen berufsqualifizierenden Abschluss zu erfüllen. Wenn zumindest vorläufig befriedigende Antworten auf diese Teilfragen formuliert sind, können Kriterien abgeleitet werden, die bei der Bewertung oder Konstruktion eines

Erhebungsinstruments, das für die Messung des Beherrschungsstands von statistischen Methoden geeignet ist, berücksichtigt werden müssen.

8.1. Mögliche Kriterien für ein Erhebungsinstrument statistischer Kompetenz

„Beherrschung“ oder „Kompetenz“ stellen zumindest derzeit keine eindeutig definierten fachpsychologischen Begriffe[32] dar (siehe Weinert, 2001) und können daher nur in einem alltagsgebräuchlichen Sinn verstanden werden. Die Nutzung dieser Begriffe soll deutlich machen, dass die für das Erfüllen der Forderung des HRG (Studierende sollen „zu wissenschaftlicher […] Arbeit und zu verantwortungsvollem Handeln […] befähigt” sein, [§ 7 HRG, 1999]) notwendigen Konstrukte[33] erst noch bestimmt werden müssen. Bei der Auseinandersetzung damit, was Studierende an kognitiven Fähigkeiten und/oder Fertigkeiten im Umgang mit statistischen Methoden aufweisen müssen, kommen, wie oben bereits erwähnt, unterschiedliche Konzepte in Betracht. Neben den in der vorliegenden Arbeit vorgestellten Konzepten aus der Forschung zu Statistikkompetenz stellt die pädagogische Psychologie Konstrukte, wie deklaratives, prozedurales, domänenspezifisches oder domänenübergreifendes Wissen oder Schemata zur Verfügung. Den Kognitionswissenschaften kann man sich zusätzlich zu den Begriffen, die auch in der pädagogischen Psychologie verwendet werden, weitere Begriffe, wie propositionale Wissensrepräsentation von Relationen, Konzepte, Begriffe oder semantische Netzwerke entleihen. In einem ersten Schritt sollten daher die Vor- und Nachteile identifiziert werden, die mit der Nutzung der Konstrukte oder Konstruktkombinationen für eine Messung der Beherrschung statistischer Methoden einhergehen. Hierbei ist es notwendig, normative und empirische Antwortentwürfe zunächst getrennt zu betrachten. Normativ muss untersucht und festgesetzt werden, was genau es bedeuten soll, etwas zu beherrschen und damit die Messintention zu konkretisieren: Verstehen wir unter „Beherrschen“ das Kennen von Fakten, etwa welche Koeffizienten es gibt, um Zusammenhänge zu bestimmen?

[32] wohl aber im Rahmen des lehrzielorientierten Testens, in dem Kompetenz definiert wird als Lösungswahrscheinlichkeit einer Aufgabengrundmenge (siehe hierzu Klauer, 1987)

[33] Für die Abschnitte 8.1 und 8.2 soll mit „Konstrukt“ ein beliebiges analytisch expliziertes, d. h. theoretisch begründetes und empirisch prüfbares Konzept gemeint sein. Es sind hier ausdrücklich nicht nur Persönlichkeitsmerkmale (z. B. Fähigkeiten, Dispositionen, Fertigkeiten), sondern auch kognitive Prozesse gemeint.

Ist es die Fertigkeit, Koeffizienten ohne Zuhilfenahme von Computerprogrammen bestimmen zu können, etwa das Berechnen des Yules Q-Koeffizienten per Hand? Ist es das (kritische) Interpretierenkönnen, wie z. B. erkennen können, dass ein Produkt-Moment-Korrelationskoeffizient für die Quantifizierung einer quadratischen Beziehung (unangemessen) genutzt worden ist? Auf der anderen Seite muss theoriegebunden empirisch untersucht werden, welche psychologischen Merkmale und Prozesse für den Erwerb der normativ gesetzten Fähigkeiten oder Fertigkeiten notwendig sind (etwa domänenübergreifende präpositionale Wissensrepräsentation von Relationen zwischen Schemata).

Auch die Frage danach, welches Ausmaß der Beherrschung vorliegen muss, um etwa der Anforderung des HRG (1999) gerecht zu werden, sollte normativ und empirisch geklärt werden. Normativ sollte festgesetzt werden, was ein ausreichender Beherrschungsgrad von Statistik ist. Ist es beispielsweise ausreichend, wenn Studierende die Ergebnisse einer multiplen linearen Regression in 50%[34] oder 90% der gestellten Aufgaben erfolgreich verstehen und kritisch bewerten? Gehört es dazu, die Aussage eines Signifikanztests in mindestens 80% oder 10% der gestellten Aufgaben verstehen und kritisch bewerten zu können? Wie verhält es sich mit dem Verstehen und kritischen Beleuchten des Ergebnisses einer messwiederholten multivariaten Varianzanalyse oder den Voraussetzungen und Einschränkungen dynamischer Wachstumsmodelle? Empirisch sollte dann untersucht werden, in welchem Ausmaß bzw. in welcher Qualität die jeweiligen ermittelten psychologischen Merkmale und Prozesse vorliegen müssen, damit die normativ gesetzten Kriterien (z. B. Identifizierenkönnen von Einschränkungen der Aussagekraft einer multiplen linearen Regression) erfüllt sind. Auch wenn die Auseinandersetzung mit dem als notwendig bestimmten Beherrschungsgrad in erster Linie der Zielerreichungsprüfung dient, so kann es auch für die Untersuchung von Instruktionserfolg in der Hochschullehre sinnvoll sein, Beherrschungsgrade im Sinne normativ (oder als praktisch bedeutsam) gesetzter Mindesteffekte zu nutzen.

Schließlich ist eine Eingrenzung von statistischen Inhalten oder Konzepten, die von Studierenden in einem bestimmten Grad hinsichtlich bestimmter psychologischer Merkmale und Prozesse beherrscht werden müssen, unabdingbar.

[34] wie es etwa derzeit in den üblichen Modulprüfungen der Fall ist (Bestehensgrenze von 50% oder 60%).

Es muss daher eine sinnvolle Eingrenzung auf die für die jeweiligen Wissenschaften Psychologie und Sonderpädagogik als Lehramt wichtigsten Inhalte oder Konzepte vorgenommen werden. Auch hier sollten theoretisch-rationale Überlegungen mit empirischen Ergebnissen Hand in Hand gehen. Theoretisch-rational muss bestimmt werden, welche statistischen Konzepte eine notwendige Basis bilden für ein relevantes statistisches Zielkonzept (z. B. das Konzept des arithmetischen Mittels als Grundlage aller linearen Verfahren) und welche statistischen Inhalte für Datenanalysen (im Allgemeinen [etwa die Bedeutung eines *p*-Wertes] und spezifisch im Hinblick auf den empirischen Umgang mit Variablen des jeweiligen Fachs [etwa unterschiedliche Korrelationskoeffizienten]) unabdingbar sind. Empirisch kann zum einen untersucht werden, welche statistischen Methoden aus der Perspektive von Professorinnen und Professoren im jeweiligen Fachgebiet notwendigerweise beherrscht werden müssen, um den wichtigen wissenschaftlichen Debatten folgen zu können. Zum anderen muss auch empirisch geprüft werden, welche statistischen Inhalte und Konzepte und in jeweils welchem Ausmaß benötigt werden, um die von Fachdozierenden bzw. -professorinnen und -professoren vorgeschlagenen Methoden beherrschen zu können. So kann etwa das Ergebnis einer (linearen) multiplen Regression vermutlich nur dann im Hinblick auf die Aussagekraft kritisch bewertet werden, wenn die Konzepte des Durchschnitts und der Korrelationen nullter Ordnung sowie Partial- und Teilpartialkorrelationen verstanden und verinnerlicht worden sind. Dies ist jedoch letzten Endes eine empirisch zu klärende Hypothese. Eine solche Frage müsste daher sowohl normativ-rational als auch empirisch beantwortet werden.

8.2. Vorschlag für ein Forschungsprogramm

Die drei umrissenen Teilfragestellungen können natürlich nicht (vollständig) voneinander isoliert beantwortet werden. In Abbildung 15 wird daher ein Vorschlag für ein Forschungsprogramm präsentiert, das die drei Teilfragestellungen in ein gemeinsames Rahmenkonzept integriert. Die Grundstruktur des Forschungsprogramms kann als Regressionsmodell verstanden werden: ein bestimmtes Konstrukt, z. B. deklaratives domänenspezifisches Wissen, oder ein Bündel verschiedener Konstrukte wird als Prädiktor bzw. Gruppe von Prädiktoren für die Schätzung des Kriteriums, beispielsweise das Erkennen von

Einschränkungen einer statistischen Auswertungsprozedur in einem Artikel, herangezogen. Damit kann die Eignung der als Prädiktoren genutzten Konstrukte für die Messung statistischer Beherrschung im Sinne der Kriteriumsvalidität beurteilt werden. Bevor dieser Schritt jedoch erfolgen kann, müssen sowohl das Kriterium als auch der Prädiktor (bzw. die Prädiktoren) näher spezifiziert werden.

Das Kriterium (bzw. die Kriterien) sollte dabei sowohl nach normativen und rationalen als auch nach empirischen Gesichtspunkten ausgewählt werden. So könnten zum einen Forderungen des HRG (1999) und notwendige Bedingungen für Wissenschaftlichkeit für die Auswahl berücksichtigt werden (❶). Zum anderen sollten aber auch Einschätzungen von Experten (Professorinnen und Professoren aller Fachbereiche des betreffenden Studiengangs) herangezogen werden (❷). Dies könnte etwa in der Form einer Befragung umgesetzt werden, in der Professorinnen und Professoren angeben, woran sich die Beherrschung statistischer Inhalte durch Studierende ihrer Meinung nach äußert. Schließlich sollte eine Kriteriumsvalidierung des gewählten Kriteriums zumindest angestrebt werden, z. B. in Form von Studien zum Zusammenhang des Kriteriums mit der Qualität der wissenschaftlichen und beruflichen Leistungen von Akademikern[35] (❷). Die gleiche Strategie kann auch für die Bestimmung des Kriteriumgrades („wie viel bzw. wie sicher?") angewendet werden. Ein solcher Auswahlprozess könnte dann ergeben, dass etwa das Erkennen von Einschränkungen statistischer Verfahren in einem Artikel ein robustes Kriterium statistischer Beherrschung darstellt.

Grundsätzlich ähnlich kann auch bei der Auswahl des Prädiktors (oder der Prädiktoren) verfahren werden. Auf Basis rationaler oder normativer Gesichtspunkte kann ein Konstrukt vorausgewählt werden. So könnte deklaratives domänenspezifisches Wissen auf der Basis des aktuellen Literaturstands als Prädiktor für das Erkennenkönnen von Einschränkungen statistischer Verfahren in einem Artikel in Frage kommen (❶). Die Auswahl dieses Konstrukts kann sodann empirisch geprüft werden. Für eine empirische Prüfung müssen im ersten Schritt sowohl statistische Inhalte als auch formale Gestaltungsmerkmale der Testitems bestimmt werden (①). Die relevanten statistischen Inhalte, auf die sich das

[35] Der Autorin ist bewusst, dass die Wahl von „Erfolgsindikatoren" für wissenschaftliche Arbeit selbst mit vielen Herausforderungen und Problemen verbunden ist.

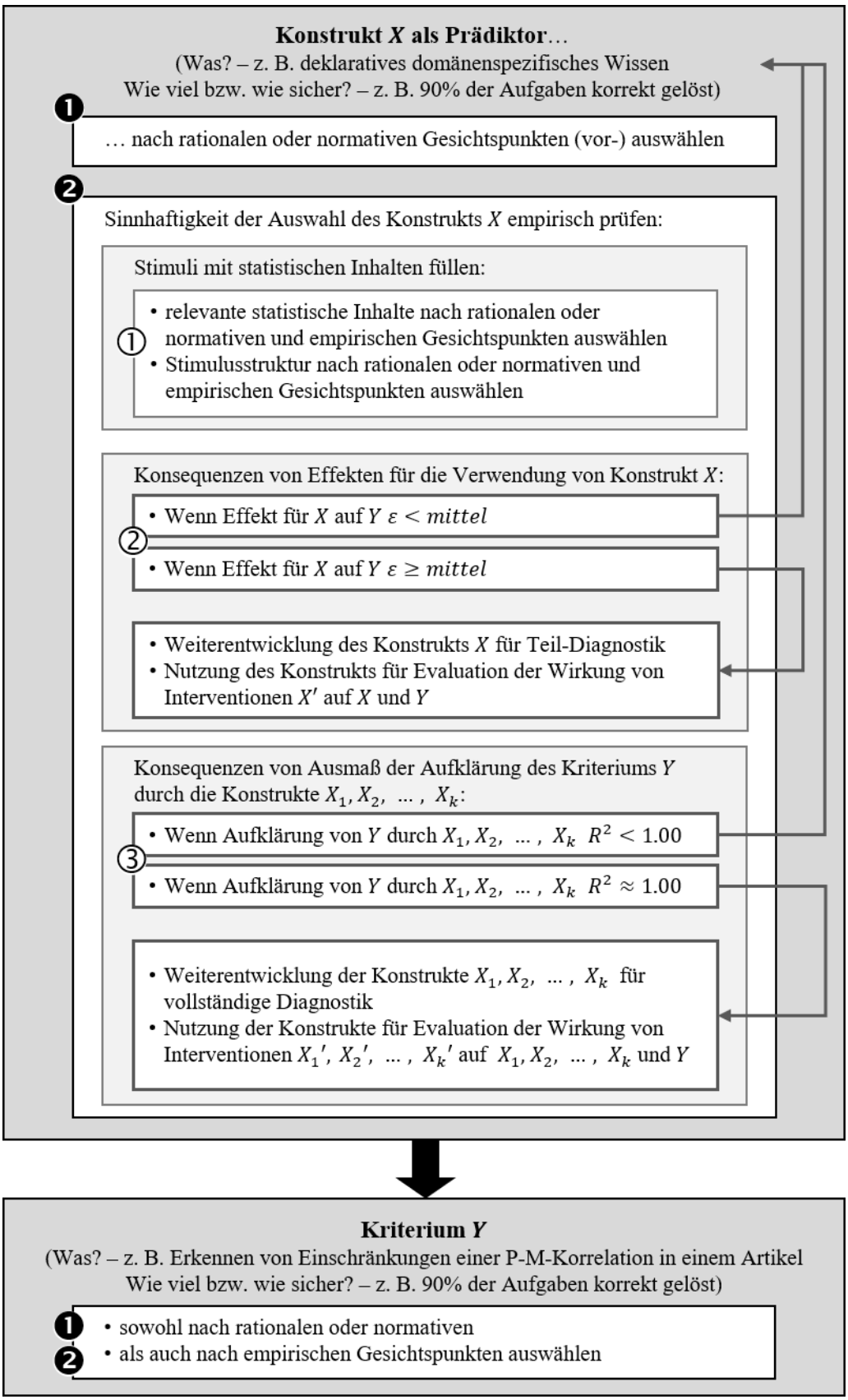


Abbildung 15. Vorschlag eines Forschungsprogramms zur Messung von Statistikkompetenz (siehe Text für Erläuterungen).

Konstrukt deklaratives domänenspezifisches Wissen beziehen soll, sollten dabei sowohl normativ bzw. rational als auch empirisch begründet werden. So könnte sich etwa nach normativ-rationaler Analyse ergeben, dass deklaratives domänenspezifisches Wissen über das Konzept der Kovarianz und des arithmetischen Mittels notwendig sein müsste, um Einschränkungen der üblichen statistischen Verfahren in einem Artikel (lineare Verfahren) erkennen zu können. Eine empirische Analyse, wiederum in Form einer Expertenbefragung (hier Professorinnen und Professoren der Statistik und Fachdidaktik), könnte die Überlegungen bestätigen oder ergänzen. Auch die Gestaltung der Testitems sollte sich nach Ergebnissen von rationalen (siehe etwa Tversky, 1964, zur Anzahl von Antwortoptionen bei Multiple-Choice Items, oder die Formulierung von operationalen Begriffen für die Itemgenerierung, z. B. Hively et al., 1968) und normativen als auch empirischen Analysen (etwa Vergleich der Testwerte für Multiple-Choice vs. offenes Antwortformat-Items) richten. Wenn Iteminhalte und Itemformat ausgewählt worden sind, kann man nun die Eignung des ausgewählten Konstrukts für die Erklärung des Kriteriums empirisch prüfen (②). An dieser Stelle kann eine erste Schleife derart formuliert werden, dass wenn die Erklärung des Kriteriums durch das ausgewählte Konstrukt eine festgelegte Effektstärke unterschreitet, die Verwendung dieses Konstrukts zunächst abgebrochen wird und ein neues Konstrukt (zunächst nach raional-normativen Gesichtspunkten) ausgewählt wird (❶). Trägt das vorausgewählte Konstrukt nennenswert zur Vorhersage des Kriteriums bei, kann dieses Konstrukt für diagnostische Fragestellungen weiterverwendet werden. Diese Fragestellungen können sich dabei zum einen auf die Feststellung des Beherrschungsgrades von Statistik richten. Zum anderen können Interventionen, die zum Ziel haben, das Konstrukt bei Studierenden zu erhöhen, hier etwa Instruktionen, die das deklarative domänenspezifische Wissen über Kovarianz und des arithmetischen Mittels erhöhen sollen, entwickelt und in ihrem Erfolg überprüft werden. Eine zweite Schleife (③) kann im Anschluss hieran formuliert werden: wenn die Varianz im Kriterium durch den Prädiktor nur unvollständig aufgeklärt wird, beginnt die Suche nach einem ergänzenden relevanten Konstrukt (❶). Wird die Varianz im Kriterium statistisch vollständig erklärt, können alle als Prädiktoren genutzten Konstrukte für eine (nahezu) vollständige Diagnostik des Beherrschungsgrades und der Evaluation von

Lehrinterventionen genutzt werden (dies ist selbstverständlich eine rein theoretische Überlegung, da in absehbarer Zukunft nicht davon ausgegangen werden kann, dass ein solcher Fall in den Human- oder Sozialwissenschaften eintritt). Dieser gesamte Prozess müsste nun auf alle identifizierten Kriterien des Regressionsmodells angewendet werden.

Das präsentierte Forschungsprogramm kann natürlich nicht innerhalb einer Qualifikationsarbeit oder weniger Projekte umgesetzt werden. Es kann jedoch als Orientierungshilfe für die sinnvolle Ableitung von Anforderungen an ein gründlich untersuchtes und gut begründetes Messinstrument statistischer Beherrschung dienen. Gerade vor dem Hintergrund der herausragenden Bedeutung von Forschungsmethoden scheint es unentbehrlich, durch eine systematische, *wissenschaftliche* Vorgehensweise die Qualität der forschungsmethodischen Hochschulausbildung zu sichern.

9. Literaturverzeichnis

Agnoli, F. (1991). Development of judgmental heuristics and logical reasoning: Training counteracts the representativeness heuristic. *Cognitive Development*, *6*(2), 195–217.

Agnoli, F., & Krantz, D. H. (1989). Suppressing natural heuristics by formal instruction: The case of the conjunction fallacy. *Cognitive Psychology*, *21*(4), 515–550.

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K. Atkins, A. et al. (2019). *rmarkdown: dynamic documents for R. R package version 1.15*. URL https://rmarkdown.rstudio.com.

Allen, K. (2006). *The Statistics Concept Inventory: Development and analysis of a cognitive assessment instrument in statistics*. ProQuest LLC, Ann Arbor, MI. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=msn&AN=MR270 8547&site=ehost-live

Allerbeck, K. (1972). *Datenverarbeitung in der empirischen Sozialforschung. Eine Einführung für Nichtprogrammierer*. Wiesbaden: Springer Fachmedien.

American Statistical Association. (2005). *Guidelines for assessment and instruction in statistics education: College report*. Alexandria, VA: Author. Retrieved from http://www.amstat.org/education/gaise/

American Statistical Association. (2016). *Guidelines for assessment and instruction in statistics education: College report*. Alexandria, VA: Author. Retrieved from http://www.amstat.org/education/gaise/

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning & Verbal Behavior*, *22*(3), 261-295. http://dx.doi.org/10.1016/S0022-5371(83)90201-3

Andersen, H. & Hepburn, B. (2016). Scientific method. In E. N. Zalta (ed.) *The stanford encyclopedia of philosophy* (Summer 2016 Edition). Retrieved from https://plato.stanford.edu/archives/sum2016/entries/scientific-method/

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Baumert, J., Lehmann, R., Lehrke, M., Clausen, M., Hosenfeld, I. et al. (1998). *Testaufgaben Naturwissenschaften TIMSS 7./8. Klasse (Population 2)*. Materialien aus der Bildungsforschung Nr. 61. Berlin: Max-Planck-Institut für Bildungsforschung.

Becker, B. (1996). A look at the literature (and other resources) on teaching statistics. *Journal of Educational and Behavioral Statistics, 21*(1), 71-90. Retrieved from http://www.jstor.org/stable/1165256

Beyth-Marom, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal*, *7*(2), 20–39.

Biehler, R. (2001). Statistische Kompetenz von Schülerinnen und Schülern – Konzepte und Ergebnisse empirischer Studien am Beispiel des Vergleichens von statistischen Verteilungen. In M. Borovcnik, J. Engel & D. Wickmann (Hrsg.), *Anergungen zum Stochastikunterricht* (97-114). Hildesheim: Franzbecker.

Biehler, R., & Engel, J. (2015). Stochastik: Leitidee Daten und Zufall. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme, H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (pp. 221-251). Berlin: Springer.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The solo taxonomy*. New York: Academic Press.

Biggs, J. B. & Collis, K. F. (1991). Multimodal learning and the quality of intelligent behavior. In H. A. H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 57-76), Hillsdale, NJ: Erlbaum.

Bohner, G., Moskowitz, G. B. & Chaiken, S. (1995). The interplay of heuristic and systematic processing of social information. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 6, pp. 33–68). Chichester, UK: Psychology Press.

Böhme, H. F. (2010) *Entwicklung und Erprobung eines Kurztests zum Konditionalen Schlussfolgern*. Unveröffentlichte Dissertation. Universität Jena.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler* (6., vollständig überarbeitete und aktualisierte Aufl.). Heidelberg: Springer Medizin.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY, US: The Guilford Press.

Bruner, J. S. (1964). The course of cognitive growth. *American Psychologist*, *19*(1), 1–15. https://doi.org/10.1037/h0044160

Bundesinstitut Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens (n. d.) *Mathematikkompetenz. Sammlung freigegebener PISA-Aufgaben. Charakteristika, Lösungen und Bewertungsrichtlinien*. Retrieved from https://www.bifie.at/wp-content/uploads/2017/04/PISA_Aufgabensammlung_Mathematik.pdf

Byrne, R. M. J. (1989). Everyday reasoning with conditional sequences. *The Quarterly Journal of Experimental Psychology Section A*, *41*(1), 141–166. https://doi.org/10.1080/14640748908402357

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306-307. http://dx.doi.org/10.1207/s15327752jpa4803_13

Callingham, R. & Watson, J. M. (2017). The development of statistical literacy at school. *Statistics Education Research Journal,* 17(1), 181-201.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105. http://dx.doi.org/10.1037/h0046016

Case, R. (1987). The structure and process of intellectual development. *International Journal of Psychology*, *22*(5/6), 571. https://doi.org/10.1080/00207598708246796

Central Advisory Council For England (1959). *15 to 18 (The Crowther Report)*. London: HMSO. Retrieved from http://www.educationengland.org.uk/documents/crowther/crowther1959-1.html

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752-766. http://dx.doi.org/10.1037/0022-3514.39.5.752

Chalmers, A. (2013). *What is this thing called science?* (4th Ed.). Maidenhead, Berkshire: Open University Press.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*(1), 5-32. http://dx.doi.org/10.1007/BF02291477

Cobb, G. W. (1992). Teaching statistics. In L. Steen (Ed.), *Heeding the call for change: Suggestions for curricular action* (pp. 3-43). Washington: Mathematical Association of America.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801–823.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, *49*(12), 997. https://doi.org/10.1037/0003-066X.49.12.997

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Routledge, https://doi.org/10.4324/9780203771587

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. http://dx.doi.org/10.1007/BF02310555

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd Ed.). New York, NY: Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. http://dx.doi.org/10.1037/h0040957

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137-163. http://dx.doi.org/10.1111/j.2044-8317.1963.tb00206.x

de Ayala, R. J. (2009). *Methodology in the social sciences. The theory and practice of item response theory.* New York, NY, US: Guilford Press.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58.

Döring, N. (2019). *Statistik*. In M. A. Wirtz (Hrsg.) Dorsch – Lexikon der Psychologie. Retrieved from https://portal.hogrefe.com/dorsch/statistik/

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5., vollständig überarbeitete, aktualisierte und erweiterte Aufl.). Berlin: Springer.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Eckart, W. U. (2009). *Geschichte der Medizin. Fakten, Konzepte, Haltungen* (6. völlig neu bearbeitete Aufl.). Heidelberg: Springer Medizin. https://doi.org/10.1007/978-3-662-07472-5

Eichler, A. & Vogel, M. (2011). *Leitidee Daten und Zufall: Von konkreten Beispielen zur Didaktik der Stochastik* (2., aktualisierte Aufl.). Wiesbaden: Springer Fachmedien.

Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5., korrigierte Aufl.). Weinheim: Beltz.

Ekström, J. (2011). The phi-coefficient, the tetrachoric correlation coefficient, and the Pearson-Yule debate. *UCLA: Department of Statistics, UCLA*. Retrieved from https://escholarship.org/uc/item/7qp4604r

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Faust, D. & Meehl, P. E. (1992). Using scientific methods to resolve questions in the history and philosophy of science: Some illustrations. *Behavior Therapy*, 23(2), 195-211. http://dx.doi.org/10.1016/S0005-7894(05)80381-8

Feldman, R. (1998). Charity, principle of. In *The Routledge encyclopedia of philosophy*. Taylor and Francis. Retrieved 20 Sep. 2019, from https://www.rep.routledge.com/articles/thematic/charity-principle-of/v-1. doi:10.4324/9780415249126-P006-1

Feyerabend, P. (1993). *Against method* (3rd ed.). London: Verso.

Franklin, C. & Garfield, J. B. (2006). The GAISE Project: Developing statistics education guidelines for pre K-12 and college courses. In G. Burrill (Ed.),

Thinking and reasoning with data and chance: 2006 NCTM yearbook (pp. 345–375). Reston, VA: National Council of Teachers of Mathematics.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Funke, J. & Spering, M. (2006) Methoden der Denk- und Problemlöseforschung. In J. Funke (Hrsg.), *Denken und Problemlösen* (Enzyklopädie der Psychologie, Serie Kognition, Bd. 8, S. 647-744). Göttingen: Hogrefe.

Gal, I. (1995). Statistical tools and statistical literacy: the case of the average. *Teaching Statistics, 17*(3), 97-99.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1–51.

Garfield, J. (1991). Evaluating students' understanding of statistics: Development of the Statistical Reasoning Assessment. In R. G. Underhill *North American Chapter of the International Group for the Psychology of Mathematics Education, Proceedings of the 13th Annual Meeting* (pp. 307-313). Blacksburg, VA: Virginia Tech. Retrieved from https://eric.ed.gov/?q=ED352274

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education [Online], 10*(3), https://doi.org/10.1080/10691898.2002.11910676

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22–38. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2009-03900-002&site=ehost-live

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44–63. https://doi.org/10.2307/749110

Garfield, J. & Ben-Zvi, D. (2004). Statistical literacy, reasoning, and thinking: goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.) *The challenge of developing statistical literacy, reasoning and thinking* (3-16). Dordrecht: Springer Science+Business Media.

Garfield, J. & Chance, B. (2000). Assessment in statistics education: issues and challenges. *Mathematics Thinking and Learning, 2*, 99-125.

Garfield, J. B., & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review, 67*(1), 1–12.

Garfield, J., delMas, R. & Chance, B. (2003). *The Web-based ARTIST: Assessment Ressource Tool for Improving Statistical Thinking*. Paper presented in Symposium Assessment of Statstical Resoning to Enhance Educational Quality at American Educational Research Association Annual Meeting, Chicago.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883–898.

Gebel, T. & Sawatzky, A. (2018). *Was ist wichtiger für das Lernen von Statistik, Vorwissen oder IQ?* Unveröffentlichte Projektarbeit.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology, 2*(1), 83-115, https://doi.org/10.1080/14792779143000033

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103*(3), 592-596. http://dx.doi.org/10.1037/0033-295X.103.3.592

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, *8*(2), 195–204. https://doi.org/10.1177/0959354398082006

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory & Psychology*, *20*(6), 733–743. https://doi.org/10.1177/0959354310378184

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Gigerenzer, G. & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421–440. https://doi.org/10.1177/0149206314547522

Gigerenzer, G., & Regier, T. (1996). How do we tell an association from a rule? Comment on Sloman (1996). *Psychological Bulletin, 119*(1), 23-26. http://dx.doi.org/10.1037/0033-2909.119.1.23

Gigerenzer, G., Todd, P. M. & The ABC Research Group. (1999). *Evolution and cognition. Simple heuristics that make us smart.* New York, NY, US: Oxford University Press.

Gilbert, D. T. & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60*(4), 509-517. http://dx.doi.org/10.1037/0022-3514.60.4.509

Gilbert, D. T., Pelham, B. W. & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54*(5), 733-740. http://dx.doi.org/10.1037/0022-3514.54.5.733

Gilovich, T., & Griffin, D. (2002). Introduction – heuristics and biases: then and now. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 1-18). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511808098.002

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511808098

Glaser, R. (1963). Instructional technology and the measurement of learing outcomes: Some questions. *American Psychologist, 18*(8), 519-521. http://dx.doi.org/10.1037/h0049294

Good, I. (1983). The philosophy of exploratory data analysis. *Philosophy of Science, 50*(2), 283-295. Retrieved from http://www.jstor.org/stable/188015

Grice, H. P. (1991) Logic and conversation. In S. Davies (Ed.), *Pragmatics. A reader* (pp. 305-315). Oxford: Oxford University Press.

Groening, M. & Cohen, X. D. (executive producers). (2013). Calculon 2.0. In *Futurama* [Fernsehserie], Season 7. Los Angeles, CA: The Curiosity Company, Los Angeles, CA: 20th Century Fox Television.

Guilford, J. P. (1941). The difficulty of a test and its factor composition. *Psychometrika*, *6*, 67–77. https://doi.org/10.1007/BF02292175

Gundlach, E., Richards, K. A. R., Nelson, D., & Levesque-Bristol, C. (2015). A comparison of student attitudes, statistical reasoning, performance, and

perceptions for web-augmented traditional, fully online, and flipped sections of a statistical literacy class. *Journal of Statistics Education, 23*(1).

Haack, S. (1999). Defending science – within reason. *Principia: An International Journal of Epistemology 3*(2), 187-212.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R. et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Research in Education, 11*(4), 14–26. https://doi.org/10.1177/0034523717723385

Hájek, A. (2012). Interpretations of probability. In E. N. Zalta (ed.) *The Stanford encyclopedia of philosophy* (Winter 2012 Edition). Retrieved from https://plato.stanford.edu/archives/win2012/entries/probability-interpret/

Haller, H., & Kraus, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1-20.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*(1), 1-47. http://dx.doi.org/10.2307/1169908

Hammock, J. (1960). *Criterion measures: instruction vs. selection research*. Paper read at meetings of the American Psychological Association, Chicago, IL. Retrieved from https://www.learningoutcomesassessment.org/wp-content/uploads/2019/08/Hammock.pdf

Hicks, K. L., Harrison, T. L., & Engle, R. W. (2015). Wonderlic, working memory capacity, and fluid intelligence. *Intelligence, 50,* 186-195. http://dx.doi.org/10.1016/j.intell.2015.03.005

Hively, W., Patterson, H. L. & Page, S. H. (1968). A “universe-defined” system of arithmetic achievement tests. *Journal of Educational Measurement, 5*(4), 275-290.

Hochschulrahmengesetz (HRG) idF vom 19.01.1999 (BGBl. I S. 18) zuletzt geändert durch Artikel 6 Absatz 2 des Gesetzes vom 23.05.2017 (BGBl. I S. 1228).

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. http://dx.doi.org/10.1007/BF02289447

Johnson-Laird, P. N. (2012). Inference in mental models. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning.* (pp. 134–154). New York: Oxford University Press.

Johnson-Laird, P. N., Legrenzi, P., & Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology, 63*(3), 395-400. http://dx.doi.org/10.1111/j.2044-8295.1972.tb01287.x

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49-81). New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511808098.004

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267-293). New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press, 3–20.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press.

Kant, I. (1784). Beantwortung der Frage: Was ist Aufklärung? In: *Berlinische Monatsschrift, 1784, H. 12,* S. 481-494. In Deutsches Textarchiv. Retrieved from http://www.deutschestextarchiv.de/kant_aufklaerung_1784

Kassambara, A. (2019). *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'*. R package version 0.1.3. https://CRAN.R-project.org/package=ggcorrplot

Klauer, K. J. (1987). *Kriteriumsorientierte Tests. Lehrbuch der Theorie und Praxis lehrzielorientierten Messens*. Göttingen: Hogrefe.

KMK (2003). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss Beschluss vom 4.12.2003*. München: Luchterhand Verlag.

KMK (2004a). *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschlussvom15.10.2004*. Neuwied: Luchterhand Verlag.

KMK (2004b). *Bildungsstandards im Fach Mathematik für den Hauptschulabschluss Beschluss vom15.10.2004*. Neuwied: Luchterhand Verlag.

KMK (2012). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife (Beschluss der Kultusministerkonferenz vom 18.10.2012)*. Köln: Wolters Kluwer.

Knauff, M. (2006) Deduktion und logisches Denken. In J. Funke (Hrsg.), *Denken und Problemlösen* (Enzyklopädie der Psychologie, Serie Kognition, Bd. 8, S. 167-264). Göttingen: Hogrefe.

Knoblich, G. & Öllinger, M. (2006) Einsicht und Umstrukturierung beim Problemlösen. In J. Funke (Hrsg.), *Denken und Problemlösen* (Enzyklopädie der Psychologie, Serie Kognition, Bd. 8, S. 1-85). Göttingen: Hogrefe.

Krosnick, J. A., Li, F., & Lehman, D. R. (1990). Conversational conventions, order of information acquisition, and the effect of base rates and individuating information on social judgments. *Journal of Personality and Social Psychology*, *59*(6), 1140–1152. https://doi.org/10.1037/0022-3514.59.6.1140

Kuhn, T. S. (1970). The structure of scientific revolutions. In O. Neurath, R. Carnap & C. Morris (Eds.) *Foundations of the unity of science* (International encyclopedia of unified science, vol. I-II) (2nd enlarged ed.). Chicago: The University of Chicago Press.

Kunzi, S., Müller, M., Pewinsky, L., Rath, D. & Sawatzky, A. (2018). *Problemlösekompetenz – kann man statistische Konzepte auch durch Alltagsprobleme erfassen?* Unveröffentlichte Projektarbeit.

Lane-Getaz, S. J. (2007). *Development and validation of a research-based assessment: Reasoning about P-values and Statistical Significance*. *Dissertation Abstracts International Section A: Humanities and Social Sciences*. ProQuest Information & Learning.

Lane-Getaz, S. J. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal, 12*(1), 20-47.

Lecoutre, M.P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23*(6), 557-568.

Lecoutre, M.-P., Durand, J.-L., & Cordier, J. (1990). A study of two biases in probabilistic judgments: Representativeness and equiprobability. In J.-P. Caverni, J.-M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases.* (pp. 563–575). Oxford: North-Holland.

Lenhard, W. & Lenhard, A. (2016). *Berechnung von Effektstärken*. Retrieved from https://www.psychometrica.de/effektstaerke.html. Dettelbach: Psychometrica. DOI: 10.13140/RG.2.1.3478.4245

Lewin, K. (1951). *Field theory in social science: selected theoretical papers (Edited by Dorwin Cartwright)*. Oxford: Harpers. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1951-06769-000&site=ehost-live

Lienert, G.-A. (1964). *Denksporttest*. Göttingen: Hogrefe.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*(1), 37–44. https://doi.org/10.1177/0272989X0102100105

Liu, H. C. & Garfield, J. (2002). Sex differences in statistical reasoning. *Bulletin of Educational Psychology, National Taiwan Normal University, 34*(1), 123-138.

Martin, N., Hughes, J., & Fugelsang, J. (2017). The roles of experience, gender, and individual differences in statistical reasoning. *Statistics Education Research Journal*, *16*(2), 454–475.

Mayo, D. and Spanos, A. (2010). Introduction and background: part I: central goals, themes, and questions. In D. Mayo & A. Spanos (Eds.) *Error and inference: Recent exchanges on experimental reasoning, reliability and the objectivity and rationality of science*, S. 1-14. Cambridge: Cambridge University Press.

McKelvie, S. J. (1989). The Wonderlic Personnel Test: Reliability and validity in an academic setting. *Psychological Reports, 65*(1), 161–162. https://doi.org/10.2466/pr0.1989.65.1.161

Meehl, P. E. (1954/1996/2003). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press. https://doi.org/10.1037/11281-000, retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.693.6031&rep=rep1&type=pdf

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Mensa The Netherlands (2002). *Online IQ-Test* [Website]. Retrieved from https://www.mensa.de/ueber-den-iq/online-tests-raetsel/online-iq-test/

Messick, S. (1987). Validity. *ETS Research Report Series, 1987*(2). i-208, doi:10.1002/j.2330-8516.1987.tb00244.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Miller, G. A., Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40–48. https://doi.org/10.1037/0021-843X.110.1.40

Moore, D. S. (1992). Teaching statistics as a respectable subject. In F. Gordon & S. Gordon (Eds.). *Statistics for the twenty-first century*, MAA Notes 26. Washington, DC: Mathematical Association of America.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551-560. http://dx.doi.org/10.1007/BF02293813

Nachtigall, C., Kroehne, U., Funke, F., & Steyer, R. (2003). (Why) should we use SEM? Pros and cons of structural equation modeling. *MPR-Online, 8*(2), 1–22. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=pdx&AN=0165949&lang=de&site=ehost-live

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Niilinuoto, I. (2015). Scientific progress. In E. N. Zalta (ed.) *The Stanford encyclopedia of philosophy* (Summer 2015 Edition). Retrieved from https://plato.stanford.edu/archives/sum2015/entries/scientific-progress/

Nuerk, H.-C., Engel, J. & Martignon, L. (2015). Statistical literacy: Eine Basiskompetenz in der Informationsgesellschaft. *Lernen und Lernstörungen*, 4, 85-90.

Oakes, M. W. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester: Wiley.

OECD (2003). *The PISA 2003 assessment framework – mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD. Retrieved from http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/33694881.pdf

OECD (2004). *Lernen für die Welt von morgen – erste Ergebnisse von PISA 2003*. Paris: OECD. Retrieved from http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/34474315.pdf

Olani, A., Hoekstra, R., Harskamp, E., & van der Werf, G. (2011). Statistical reasoning ability, self-efficacy, and value beliefs in a university statistics course. *Electronic Journal of Research in Educational Psychology*, *9*(1), 49–72. Retrieved from http://www.investigacion-psicopedagogica.org/revista/new/english/ContadorArticulo.php?530

Open Science Collaboration, Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A. et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), cc4716. https://doi.org/10.1126/science.aac4716

Osburn, H. G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurement*, *28*(1), 95–104. https://doi.org/10.1177/001316446802800109

Park, J. (2015). *Developing and validating an instrument to measure college students' inferential reasoning in statistics: An argument-based approach to*

validation. *Dissertation Abstracts International Section A: Humanities and Social Sciences*. ProQuest Information & Learning.

Penna, M. P., Agus, M., Peró-Cebollero, M., Guàrdia-Olmos, J., & Pessa, E. (2014). The use of imagery in statistical reasoning by university undergraduate students: A preliminary study. *Quality & Quantity: International Journal of Methodology*, *48*(1), 173–187. https://doi.org/10.1007/s11135-012-9757-5

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). New York: Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, thinking and reasoning* (pp. 17–46). Dordrecht, The Netherlands: Kluwer Academic Press.

Politzer, G., & Macchi, L. (2005). The representation of the task: The case of the Lawyer-Engineer problem in probability judgement. In V. Girotto & P. N. Johnson-Laird (Eds.), *The shape of reason: Essays in honour of Paolo Legrenzi.* (pp. 119–135). New York, NY: Psychology Press.

Popham, W. J. (1974). Minimal competencies for objectives-oriented teacher education programs. *Journal of Teacher Education*, *25*(1), 68–73. https://doi.org/10.1177/002248717402500117

Poser, H. (2012). *Wissenschaftstheorie. Eine philosophische Einführung* (2., überarbeitete und erweiterte Aufl.). Stuttgart: Philipp Reclam jun.

Prinz, W., Müsseler, J. & Rieger, M. (2017). Einleitung – Psychologie als Wissenschaft. In J. Müsseler & M. Rieger (Hrsg.), *Allgemeine Psychologie* (3. Aufl.) (1-10). Berlin: Springer.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Report of the Committee of Inquiry into the Teaching of Mathematics in Schools (1982). *Mathematics counts (The Cockcroft Report)*. London: HMSO. Retrieved from

http://www.educationengland.org.uk/documents/cockcroft/cockcroft1982.html

Revelle, W. (2018). *psych: Procedures for personality and psychological research*. Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.12

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145-154. http://dx.doi.org/10.1007/s11336-008-9102-z

Rietz, C. (1996). *Faktorielle Invarianz: die inferenzstatistische Absicherung von Faktorstrukturvergleichen*. Bonn: PACE.

Romeijn, J.-W. (2017). Philosophy of statistics. In E. N. Zalta (ed.) *The Stanford encyclopedia of philosophy* (Spring 2017 Edition). Retrieved from https://plato.stanford.edu/archives/spr2017/entries/statistics/

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. URL http://www.jstatsoft.org/v48/i02/

Sá, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, *91*(3), 497-510. http://dx.doi.org/10.1037/0022-0663.91.3.497

Sabbag, A. G., & Zieffler, A. (2015). Assessing learning outcomes: An analysis of the Goals-2 instrument. *Statistics Education Research Journal*, *14*(2), 93–116.

Sabbag, A., Garfield, J., & Zieffler, A. (2018). Assessing statistical literacy and statistical reasoning: The REALI instrument. *Statistics Education Research Journal*, *17*(2), 141–160.

Sawatzky, A. & Rietz, C. (2016). *Status Quo der statistischen Ausbildung in der Heilpädagogik, L. A.* Unveröffentlichte Projektdaten.

Schaeffer, R. (1990). The ASA–NCTM quantitative literacy project: An overview. In D. Vere-Jones, S. Carlyle, & B. P. Dawkins (Eds.), *Proceedings of the Third International Conference on Teaching Statistics* (pp. 1–5). Voorburg: International Statistical Institute.

Schau, C. (2003). *Students' attitudes: The "other" important outcome in statistics education*. Paper presented at the Joint Statistical Meetings, August, San Francisco, CA.

Schau, C., Stevens, J., Dauphinee, T. L. & Del Vecchio, A. (1995). The development and validation of the Survey of Attitudes Toward Statistics. *Educational and Psychological Measurement, 55*, 868-875.

Schwarz, N., Strack, F., Hilton, D., & Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: The contextual relevance of "irrelevant" information. *Social Cognition: A Journal of Social, Personality, and Developmental Psychology*, *9*(1), 67–84.

Shamos, M. H. (1995). *The myth of scientific literacy*. Bew Brunswick, NJ: Rutgers University Press.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. doi:10.1007/s11336-008-9101-0

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99-118. http://dx.doi.org/10.2307/1884852

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. http://dx.doi.org/10.1037/0033-2909.119.1.3

Sloman, S. A. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 379-398). Cambridge: Cambridge University Press.

Snee, R. D. (1990). Statistical thinking and its contribution to total quality. *American Statistician*, *44*(2), 116. https://doi.org/10.2307/2684144

Sommer, M., Müller-Wille, S. & Reinhardt, C. (2017). *Handbuch Wissenschaftsgeschichte*. Stuttgart: J. B. Metzler.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*(2), 342-357. http://dx.doi.org/10.1037/0022-0663.89.2.342

Stanovich, K., & West, R. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645-665. doi:10.1017/S0140525X00003435

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296. http://dx.doi.org/10.1023/A:1022193728205

Tempelaar, D. T. (2004). *Statistical Reasoning Assessment: an analysis of the SRA instrument*. Paper presented at the ARTIST Roundtable Conference on Assessment in Statistics at Lawrence University, Appleton, WI.

Tempelaar, D. T., Gijselaers, W. H. & van der Loeff, S. S. (2006) Puzzles in statistical reasoning. *Journal of Statistics Education, 14*(1), doi: 10.1080/10691898.2006.11910576

Tempelaar, D. T., van der Loeff, S. S., & Gijselaers, W. H. (2007). A structural equation model analyzing the relationship of students' attitudes toward statistics, prior reasoning abilities and course performance. *Statistics Education Research Journal*, *6*(2), 78–102.

Toplak, M., West, R., & Stanovich, K. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. http://dx.doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84-98). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511809477.007

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Viswanathan, M. (1993). Measurement of individual differences in preference for numerical information. *Journal of Applied Psychology, 78*(5), 741-752. http://dx.doi.org/10.1037/0021-9010.78.5.741

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, 88, 1-8.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology, 12*, 129-140. http://dx.doi.org/10.1080/17470216008416717

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05." *The American Statistician, 73*(Supplement 1), 1–19.

Watson, J. (1997). Assessing statistical thinking using the media. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107–121). Amsterdam, The Netherlands: IOS Press.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2), 3–46.

Watson, J.M., & Callingham, R.A. (2005). Statistical literacy: From idiosyncratic to critical thinking. In G. Burrill & M. Camden (Eds.), *Curricular development in statistics education. International Association for Statistical Education (IASE) Roundtable, Lund, Sweden, 2004* (pp. 116-162). Voorburg, The Netherlands: International Statistical Institute.

Watson, J., Collis, K., Callingham, R., & Moritz, J. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation, 1*(3), 247–275. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=eoah&AN=11315957&site=ehost-live

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science & Technology, 34*(1), 1. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=tnh&AN=9212598&site=ehost-live

Weinert, F. E. (1999). *Concepts of competence*. Unveröffentlichtes Manuskript. Retrieved from

https://pdfs.semanticscholar.org/8b88/efa9dd5e0a4b605aea6e5e3b9ec640beb089.pdf

Weinert, F. E. (2001). Concept of competence: a conceptual clarification. In D. S. Rychen & L. H. Salganik (eds.), *Defining and selecting key competencies* (S. 45-65). Seattle: Hogrefe & Huber.

Weinert, F. E. (2014). *Leistungsmessung in Schulen* (3., aktualisierte Aufl.). Weinheim: Beltz und Bonn: Kultusministerkonferenz.

Weiß, R. H. (2006). *CFT 20-R. Grundintelligenzstest. Skala 2*. Göttingen: Hogrefe.

Wickham, W. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1*. https://CRAN.R-project.org/package=tidyverse

Wickham, H. & Miller, E. (2019). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.1.1. https://CRAN.R-project.org/package=haven

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review*, 67(3), 223–265.

Wilke, C. O. (2019). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.0.0. https://CRAN.R-project.org/package=cowplot

Xie Y. (2014) knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch and R. D. Peng, (Eds), *Implementing reproducible computational research*. Chapman and Hall/CRC. ISBN 978-1466561595

Xie Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman and Hall/CRC. ISBN 978-1498716963

Xie Y. (2019). *knitr: A general-purpose package for dynamic report generation in R*. R package version 1.24.

Xie Y., Allaire, J. J. & Grolemund, G. (2018). *R Markdown: the definitive guide*. Chapman and Hall/CRC. ISBN 9781138359338. URL https://bookdown.org/yihui/rmarkdown

Yule, G. (1900). On the association of attributes in statistics: With illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 194*, 257-319. Retrieved from http://www.jstor.org/stable/90759

Zieffler, A., & Catalysts for Change. (2019). *Statistical thinking: A simulation approach to uncertainty* (4.2th ed.). Minneapolis, MN: Catalyst Press. http://zief0002.github.io/statistical-thinking/

Zieffler, A., Garfield, J. & Fry, E. (2018). What is statistics education? In D. Ben-Zvi, K. Maker & J. Garfield (eds.) International handbook of research in statistics education (pp. 37-70). Cham: Springer.

Ziegler, L. A. (2015). *Reconceptualizing statistical literacy: Developing an assessment for the modern introductory statistics course*. *Dissertation Abstracts International Section A: Humanities and Social Sciences*. ProQuest Information & Learning.

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence, 1*(1), 5-31. http://dx.doi.org/10.1016/0160-2896(77)90025-3

Zimprich, D. (2012). Attitudes toward statistics among Swiss psychology students. *Swiss Journal of Psychology, 71*(3), 149-155. http://dx.doi.org/10.1024/1421-0185/a000082

10. Anhang

Vorkommenshäufigkeit der Konzepte für Beherrschung statistischer Kompetenz

Tabelle 23

Treffer für Konzepte statistischer Kompetenz in verschiedenen Datenbanken

	PsycINFO und PSYNDEX	Scopus	Web of Science
statistical numeracy	12	157 (13)	17
statistical literacy	101	1184 (293)	199
statistical thinking	113	3389 (439)	336
statistical reasoning	258	6734 (526)	406
statistical cognition	12	38 (7)	4
statistische Kompetenz	1	2 (1)	0

Anmerkungen. Suche erfolgte mit: „[Begriff]“ als *basic search* nicht weiter eingeschränkt und den Datenbanken PsycINFO und PSYNDEX mit dem EBSCOhost, in *all fields* (bzw. *article title, abstract, keyword* in Klammern) in der Scopus Datenbank und in *all databases* in *topic* im Web of Science

Zum Vergleich: „deductive reasoning“ ergibt 4,092 und „inductive reasoning“ 1159 Treffer in den Datenbanken PsycINFO und PSYNDEX

(Datum: 30.08.2019)

Korrelationen des correct statistical reasoning mit akademischen Leistungsmaßen (aus Tempelaar et al., 2006)

Tabelle 24.

Korrelationen des Gesamtwerts für die Subskalen des correct reasoning im SRA mit Hochschulleistungen in Statistik und Mathematik aus Tempelaar et al., 2006

Leistung	Statistik	Mathematik
Hausaufgaben		
1. Zeitabschnitt	−.02	−.12
2. Zeitabschnitt	−.09	−.14
3. Zeitabschnitt	−.13	−.06
Abschlusstest		
1. Zeitabschnitt	.24	.28
2. Zeitabschnitt	.06	.18
3. Zeitabschnitt	.07	.13
Kurztest		
1. Zeitabschnitt	.01	
2. Zeitabschnitt	−.01	

Anmerkungen.

Itemzuordnung im Statistical Reasoning Assessment

Tabelle 25.

Zuordnung der SRA Items zu Subskalen des correct reasoning und Subskalen der misconceptions (modifiziert nach Garfield, 2003)

correct reasoning		misconceptions	
Subskala	Items		Items
Wahrscheinlichkeiten korrekt interpretieren	2, 3		
		Orientierung am Ergebnis	2, 3, 11, 12
Bedeutung großer Stichproben verstehen	6 (mc), 12		
		Gesetz der kleinen Zahl	12, 14
		Überzeugung, dass Gruppen nur verglichen werden können, wenn sie dieselbe Stichprobengröße haben	6 (mc)
Variabilität innerhalb von Stichproben verstehen	14, 15		
		Fehlverständnis von Mittelwerten	1, 15, 17
Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird	1, 4, 17		
(stochastische) Unabhängigkeit verstehen	9, 10 (mc), 11	Repräsentativitätsheuristik	9, 10 (mc), 11
Wahrscheinlichkeit korrekt berechnen	8, 13, 18, 19, 20	Gleichwahrscheinlichkeits-verzerrung	13, 18, 19, 20
Zwischen Korrelation und Kausalität unterscheiden	16 (mc)	Gleichsetzen von Korrelation mit Kausalität	16 (mc)
		Gute Stichproben müssen einen hohen Anteil der Population darstellen	7, 16
Vierfeldertafeln korrekt interpretieren	5		

Anmerkungen. mc = Mehrfachwahlformat (multiple choice).

Eingesetzte Skalen

- Deutsche Version des Statistical Reasoning Assessment mit der Zuordnung der Antwortoptionen zu den correct reasoning skills (cr) und den misconceptions (m)
- Aufgaben zur Mathematik
- Instruktion und Aufgaben zum deduktiven Schließen
- Instruktion und figurale Reihenvervollständigungsaufgaben aus dem Mensa Online IQ-Test

Tabelle 26.

Statistical Reasoning Assessment (deutsch) mit der Zuordnung der Antwortoptionen zu den correct reasoning skills (cr) und den misconceptions (m) (auf den nächsten Seiten fortgesetzt).

1 Ein kleiner Gegenstand wurde im Sachkundeunterricht von jedem einzelnen der insgesamt neun Schüler auf der gleichen Waage gewogen. Die notierten Gewichte (in Gramm) jedes Schülers sind unten aufgeführt.

6,2 6,0 6,0 15,3 6,1 6,3 6,2 6,15 6,2

Die Schüler wollen so akkurat wie ihnen möglich das tatsächliche Gewicht des Objekts bestimmen. Welche der folgenden Methoden würden Sie ihnen zur Nutzung empfehlen?

m1a ○ a Die häufigste Zahl benutzen, welche 6,2 ist.

○ b Die 6,15 benutzen, da sie die genauste Messung ist.

m1b ○ c Die 9 Zahlen addieren und sie durch 9 teilen.

cr2 ○ d Die 15,3 rauswerfen, die anderen 8 Zahlen addieren und sie durch 8 teilen.

2 Die folgende Nachricht ist auf die Flasche eines verschreibungspflichtigen Medikaments gedruckt:

WARNUNG: Bei Anwendungen auf Hautregionen besteht eine 15%ige Wahrscheinlichkeit, dass sich ein Ausschlag entwickelt. Wenn sich ein Ausschlag entwickelt, konsultieren Sie Ihren Arzt.

Welche der folgenden ist die beste Interpretation dieser Warnung?

○ a Benutzen Sie das Medikament nicht auf der Haut, es besteht eine hohe Wahrscheinlichkeit, dass sich ein Hautausschlag entwickelt.

○ b Für die Anwendung auf der Haut verwenden Sie nur 15% der empfohlenen Dosis.

○ c Wenn sich ein Ausschlag entwickelt, wird er wahrscheinlich 15% Ihrer Haut betreffen.

cr1 ○ d Bei ungefähr 15 von 100 Leuten, die dieses Medikament benutzen, entwickelt sich ein Ausschlag.

m2 ○ e Es besteht kaum eine Wahrscheinlichkeit, bei der Nutzung des Medikaments einen Ausschlag zu bekommen.

3 Die Mitarbeiter des Meteorologischen Zentrums in Neustadt wollten die Genauigkeit ihrer Wettervorhersage bestimmen. Sie durchsuchten ihre Aufzeichnungen nach den Tagen, an denen die Vorhersage eine 70%ige Regenwahrscheinlichkeit angab. Sie verglichen diese Vorhersagen mit Meldungen, ob es an den jeweiligen Tagen wirklich geregnet hatte oder nicht.
Die Vorhersage von 70%iger Regenwahrscheinlichkeit kann als sehr genau angesehen werden, wenn es:

m2 ○ a an 95%-100% dieser Tage regnete.

m2 ○ b an 85%-94% dieser Tage regnete.

○ c an 75%-84% dieser Tage regnete.

cr1 ○ d an 65%-74% dieser Tage regnete.

○ e an 55%-64% dieser Tage regnete.

Tabelle 26 *(fortgesetzt)*

4 Eine Lehrerin will die Sitzordnung in ihrer Klasse ändern, in der Hoffnung, dass dies die Anzahl der Meldungen, die ihre Schüler machen, erhöhen wird. Zunächst entscheidet sie sich, zu schauen, wie viele Meldungen ihre Schüler in der aktuellen Sitzordnung machen. Die Aufzeichnung der Anzahl der Meldungen, die ihre 8 Schüler in einer Unterrichtsstunde gemacht haben, ist unten aufgeführt.

Schülerinitialen	**A.A.**	**R.F.**	**A.G.**	**J.G.**	**C.K.**	**N.K.**	**J.L.**	**A.W.**
Anzahl der Meldungen	0	5	2	22	3	2	1	2

Sie will diese Daten zusammenfassen, indem sie die typische Anzahl von Meldungen an diesem Tag berechnen will. Welche der folgenden Methoden würden Sie ihr zur Nutzung empfehlen?

cr2 ○ a Die häufigste Zahl benutzen, welche 2 ist.

cr2 ○ b Die 8 Zahlen addieren und sie durch 8 dividieren.

○ c Die 22 rauswerfen, die anderen 7 Zahlen addieren und sie durch 7 teilen.

○ d Die 0 rauswerfen, die anderen 7 Zahlen addieren und sie durch 7 teilen.

5 Ein neues Medikament wird getestet, um seine Wirksamkeit in der Behandlung gegen Ekzeme, eine Entzündungserscheinung der Haut, zu bestimmen. Dreißig Patienten mit Ekzemen wurden für die Teilnahme an der Studie ausgewählt. Die Patienten wurden zufällig in zwei Gruppen aufgeteilt. Zwanzig Patienten in einer Experimentalgruppe bekamen das Medikament, während zehn Patienten in einer Kontrollgruppe kein Medikament bekamen. Die Ergebnisse nach zwei Monaten sind unten aufgeführt.

	Experimentalgruppe (Medikament)	**Kontrollgruppe (kein Medikament)**
Gebessert	8	2
Keine Besserung	12	8

Basierend auf den Daten, glaube ich, dass das Medikament:

○ 1) eher wirksam war.

Wenn Sie sich für Option 1) entschieden haben, markieren Sie diejenige der unten aufgeführten Erklärungen, die Ihr Argument für Option 1) am besten beschreibt:

○ a 40% der Leute (8/20) in der Experimentalgruppe haben sich gebessert.

○ b 8 Leute haben sich in der Experimentalgruppe gebessert, während sich nur 2 in der Kontrollgruppe besserten.

○ c Die Anzahl der Leute in der Experi-mentalgruppe, die sich gebessert haben, ist nur 4 weniger, als die Anzahl derer, die sich nicht gebessert haben (12-8), während die Differenz in der Kontrollgruppe 6 ist (8-2).

○ d 40% der Patienten in der Experimentalgruppe haben sich gebessert (8/20), während sich in der Kontrollgruppe nur 20% (2/10) gebessert haben.

cr7

○ 2) im Grunde nicht wirksam war.

Wenn Sie sich für Option 2) entschieden haben, markieren Sie diejenige der unten aufgeführten Erklärungen, die Ihr Argument für Option 1) am besten beschreibt:

○ a In der Kontrollgruppe haben sich 2 Leute sogar ohne das Medikament gebessert.

○ b In der Experimentalgruppe ging es mehr Leuten nicht besser, als besser (12 vs. 8).

○ c Der Unterschied zwischen den Anzahlen derer, die sich gebessert haben und derer, die sich nicht gebessert haben, ist in beiden Gruppen ungefähr gleich (4 vs. 6).

○ d In der Experimentalgruppe besserten sich nur 40% der Patienten (8/20).

Tabelle 26 *(fortgesetzt)*

6 Unten aufgeführt sind mehrere mögliche Gründe, warum man die Ergebnisse des oben beschriebenen Experiments in Frage stellen könnte. Kreuzen Sie alle Gründe an, denen Sie zustimmen.

m8 ○ a Es ist nicht legitim, die zwei Gruppen zu vergleichen, weil es verschiedene Anzahlen von Patienten in jeder Gruppe gibt.

cr8 ○ b Die Stichprobe von 30 ist zu klein, um das Ziehen von Schlussfolgerungen zu erlauben.

○ c Die Patienten hätten nicht zufällig in die Gruppen verteilt werden dürfen, weil die härtesten Fälle vielleicht per Zufall in einer Gruppe gelandet sein könnten.

○ d Mir wurden nicht genug Informationen gegeben, wie die Ärzte entschieden, ob Patienten sich gebessert haben oder nicht. Ärzte könnten in ihrem Urteil voreingenommen sein.

○ e Ich stimme keiner dieser Aussagen zu.

7 Eine Firma für Marktforschung wurde gebeten zu bestimmen, wie viel Geld Teenager (Alter 13 – 19) für aufgezeichnete Musik (Kassetten, CDs und Schallplatten) ausgeben. Die Firma wählte bundesweit 20 Einkaufszentren zufällig aus. Jeweils ein Feldforscher stand an einem zentralen Ort im Einkaufszentrum und bat Vorbeigehende, die in einem geeigneten Alter zu sein schienen, einen Fragebogen auszufüllen. Insgesamt 550 Fragebögen wurden von Teenagern ausgefüllt. Auf Basis dieser Umfrage hat die Firma berichtet, dass der durchschnittliche Teenager in diesem Land jedes Jahr 155€ für aufgezeichnete Musik ausgibt.
Unten werden mehrere Aussagen aufgeführt, die diese Umfrage betreffen. Bitte kreuzen Sie jede Aussage an, der Sie zustimmen.

○ a Der Durchschnitt basiert auf der Schätzung der Teenager dafür, was sie ausgeben und könnte sich daher ziemlich von dem unterscheiden, was sie wirklich ausgeben.

m3 ○ b Sie hätten die Umfrage an mehr als 20 Einkaufszentren durchführen sollen, wenn sie einen Durchschnitt auf Basis von Teenagern im ganzen Land wollten.

m3 ○ c Die Stichprobe von 550 Teenagern ist zu klein, um das Treffen von Schlussfolgerungen über das ganze Land zu erlauben.

○ d Sie hätten Teenager fragen sollen, die aus Musikgeschäften herauskommen.

○ e Der Durchschnitt könnte eine dürftige Schätzung der Ausgaben von allen Teenagern sein, da Teenager nicht zufällig ausgewählt wurden, um den Fragebogen auszufüllen.

○ f Der Durchschnitt könnte eine dürftige Schätzung der Ausgaben von allen Teenagern sein, da nur Teenager in Einkaufszentren ausgewählt wurden.

○ g Einen Durchschnitt in diesem Fall auszurechnen ist unangemessen, da es eine große Variation in dem gibt, was Teenager ausgeben.

○ h Ich stimme keiner dieser Aussagen zu.

Tabelle 26 *(fortgesetzt)*

8 Zwei Behälter, mit A und B beschriftet, werden mit roten und blauen Murmeln in folgender Menge befüllt:

Behälter	Rot	Blau
A	6	4
B	60	40

Jeder Behälter wird kräftig geschüttelt. Nachdem Sie einen Behälter ausgesucht haben, greifen Sie herein und nehmen sich, ohne zu gucken, eine Murmel heraus. Wenn die Murmel blau ist, dann gewinnen Sie 50€. Welcher Behälter bietet Ihnen die höchste Wahrscheinlichkeit, eine blaue Murmel herauszunehmen?

- ○ a Behälter A (mit 6 roten und 4 blauen).
- ○ b Behälter B (mit 60 roten und 40 blauen).

cr3a ○ c Gleiche Wahrscheinlichkeit bei beiden Behältern.

9 Welche der folgenden Reihenfolgen ist am wahrscheinlichsten das Resultat vom fünfmaligen Werfen einer fairen Münze?

m5	○	a	Kopf-Kopf-Kopf-Zahl-Zahl	bzw.:	K-K-K-Z-Z
m5	○	b	Zahl-Kopf-Kopf-Zahl-Kopf	bzw.:	Z-K-K-Z-K
	○	c	Zahl-Kopf-Zahl-Zahl-Zahl	bzw.:	Z-K-Z-Z-Z
m5	○	d	Kopf-Zahl-Kopf-Zahl-Kopf	bzw.:	K-Z-K-Z-K
cr4	○	e	Alle Reihenfolgen sind gleich wahrscheinlich.		

10 Wählen Sie eine oder mehrere Erklärungen für die Antwort, die Sie in der vorherigen Aufgabe gegeben haben.

- ○ a Da die Münze fair ist, müsste man ungefähr die gleichen Anzahlen von Kopf und Zahl bekommen.
- ○ b Da Münzwürfe zufällig sind, müsste die Münze häufig abwechselnd auf Kopf und Zahl landen.
- ○ c Jede der Reihenfolgen könnte auftreten.

cr4 ○ d Wenn man die Münze immer wieder fünfmal wirft, würde jede dieser Reihenfolgen ungefähr so häufig auftreten, wie jede andere Reihenfolge.

m5 ○ e Wenn man ein paar Mal hintereinander Kopf wirft, erhöht sich die Wahrscheinlichkeit, beim nächsten Wurf Zahl zu bekommen.

cr4 ○ f Jede Reihenfolge von fünf Würfen hat exakt die gleiche Wahrscheinlichkeit aufzutreten.

11 Unten aufgeführt sind dieselben Reihenfolgen von Kopf- und Zahlwürfen, die in Aufgabe 9 aufgelistet sind. Welche der Reihenfolgen ist das am wenigsten wahrscheinliche Resultat vom fünfmaligen Werfen einer Münze?

m2 ○ a K-K-K-Z-Z

m2 ○ b Z-K-K-Z-K

m5 ○ c Z-K-Z-Z-Z

m2 ○ d K-Z-K-Z-K

cr4 ○ e Alle Reihenfolgen sind gleich unwahrscheinlich.

Tabelle 26 *(fortgesetzt)*

12 Die Müllers wollen ein neues Auto kaufen und sie haben ihre Auswahl auf einen Fiat oder einen Opel eingeschränkt. Zuerst konsultieren sie einen Beitrag aus der Stiftung Warentest, welcher die Raten der Reparaturen für verschiedene Autos verglich. Aufzeichnungen der Reparaturen an 400 Autos jeden Typs zeigten etwas weniger mechanische Probleme mit dem Fiat, als mit dem Opel.
Die Müllers haben dann mit drei Freunden gesprochen, zwei Opel-Besitzern und einem früheren Fiat-Besitzer. Beide Opel-Besitzer berichteten, dass sie ein paar mechanische Probleme haben, aber nichts Bedeutendes. Der Fiat-Besitzer, allerdings, explodierte, als sie ihn fragten, wie er sein Auto fand:

Zuerst ging die Benzineinspritzung kaputt - 250 Euro! Als nächstes bekam ich Probleme mit dem Auspuff und musste ihn ersetzen. Letztendlich entschied ich mich, es zu verkaufen, nachdem das Getriebe hin war. Ich würde niemals wieder einen Fiat kaufen.

Die Müllers wollen ein Auto kaufen, bei dem es weniger wahrscheinlich ist, dass erhebliche Reparaturen anfallen. Nach dem, was die Familie Müller jetzt weiß, welches Auto würden Sie ihnen zum Kauf empfehlen?

m4 ○ a Ich würde ihnen den Opel zum Kauf empfehlen, in erster Linie wegen des ganzen Ärgers, den ihr Freund mit dem Fiat hatte. Da sie keine ähnlichen Horrorgeschichten über den Opel gehört haben, sollten sie sich dafür entscheiden.

cr8 ○ b Ich würde ihnen den Kauf des Fiats trotz der schlechten Erfahrung ihres Freundes empfehlen. Es ist nur ein Fall, während die berichteten Informationen aus der Stiftung Warentest auf vielen Fällen basieren. Und laut der Daten ist es bei dem Fiat etwas weniger wahrscheinlich, dass erhebliche Reparaturen benötigt werden.

m2 ○ c Ich würde ihnen sagen, dass es egal ist, welches Auto sie kaufen. Sogar wenn eines der Modelle wahrscheinlicher erhebliche Reparaturen benötigen könnte, könnten sie trotzdem zufällig ein bestimmtes Auto abbekommen, das viele Reparaturen benötigen würde. Sie könnten genauso gut eine Münze werfen.

13 Fünf Seiten eines fairen Würfels werden schwarz angemalt und eine Seite wird weiß angemalt. Der Würfel wird sechsmal geworfen. Welches der folgenden Ergebnisse ist wahrscheinlicher?

cr3b ○ a Schwarze Seite oben bei fünf Würfen, weiße Seite oben beim anderen Wurf.

m2 ○ b Schwarze Seite oben bei allen sechs Würfen.

m7 ○ c Die Antwortoptionen a) und b) sind gleich wahrscheinlich.

14 Die Hälfte aller Neugeborenen sind Mädchen und die andere Hälfte sind Jungen. Krankenhaus A zeichnet einen Durchschnitt von 50 Geburten pro Tag auf. Krankenhaus B zeichnet einen Durchschnitt von 10 Geburten pro Tag auf. In welchem Krankenhaus wird die Geburt von 80% (oder mehr) weiblichen Geburten an einem bestimmten Tag wahrscheinlicher registriert?

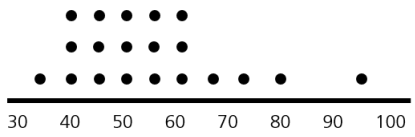
○ a Krankenhaus A (mit 50 Geburten am Tag)

cr5 ○ b Krankenhaus B (mit 10 Geburten am Tag)

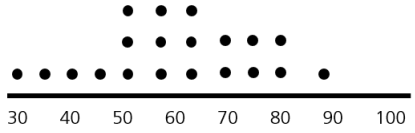
m4 ○ c Es ist bei beiden Krankenhäusern gleich wahrscheinlich.

Tabelle 26 *(fortgesetzt)*

15 Vierzig Studierende nahmen an einer Studie über die Wirkung von Schlaf auf Testergebnisse teil. Zwanzig der Studierenden meldeten sich freiwillig dafür, die ganze Nacht vor dem Test aufzubleiben und zu lernen. Die anderen zwanzig Studierenden (die Kontrollgruppe) gingen am Abend vor dem Test gegen 23:00 Uhr ins Bett. Die Testergebnisse für jede Gruppe werden in der Grafik unten aufgeführt. Jeder Punkt repräsentiert das Ergebnis einer/eines bestimmten Studierenden. Die zwei Punkte über der 80 in der unteren Grafik zeigen zum Beispiel, dass zwei Studierende in der Schlafgruppe 80 Punkte im Test erzielten.



Testergebnisse: Kein-Schlaf Gruppe



Testergebnisse: Schlaf Gruppe

Schauen Sie sich die zwei Grafiken sorgfältig an. Wählen Sie dann von den sechs möglichen Fazits, die unten aufgeführt sind, dasjenige aus, dem Sie <u>am meisten</u> zustimmen.

○ a Die Kein-Schlaf-Gruppe war besser, weil keine/r dieser Studierenden weniger als 40 Punkte erzielte und das beste Ergebnis von einer/einem Studierenden dieser Gruppe erreicht wurde.

m1c ○ b Die Kein-Schlaf-Gruppe war besser, weil der Durchschnitt hier etwas höher erscheint, als der Durchschnitt der Schlaf-Gruppe.

○ c Es gibt keinen Unterschied zwischen den zwei Gruppen, weil es eine beträchtliche Überschneidung in den Ergebnissen der beiden Gruppen gibt.

cr5 ○ d Es gibt keinen Unterschied zwischen den zwei Gruppen, weil die Differenz zwischen ihren Durchschnitten im Vergleich zum Umfang der Variationen in den Ergebnissen klein ist.

○ e Die Schlaf-Gruppe war besser, weil mehr Studierende in dieser Gruppe 80 Punkte oder mehr erzielten.

m1c ○ f Die Schlaf-Gruppe war besser, weil der Durchschnitt hier etwas höher erscheint, als der Durchschnitt der Kein-Schlaf-Gruppe.

Tabelle 26 *(fortgesetzt)*

16 500 Grundschüler führten für einen Monat täglich Buch über die Stunden, die sie für Fernsehen aufbrachten. Die durchschnittliche Anzahl der Stunden pro Woche, die für Fernsehen aufgebracht wurden, war 28. Die Forscher, die diese Studie durchführten, bekamen auch die Zeugnisse für jeden der Schüler. Sie fanden heraus, dass die Schüler, die in der Schule gut waren, weniger Zeit für Fernsehen aufbrachten, als die Schüler, die schlecht waren. Unten aufgeführt sind mehrere mögliche Aussagen hinsichtlich der Ergebnisse dieser Forschungsstudie. Bitte kreuzen Sie jede Aussage an, der Sie zustimmen.

m3 ○ a Die Stichprobe von 500 ist zu klein, um das Ziehen von Schlussfolgerungen zu erlauben.

m6 ○ b Wenn ein Schüler die Dauer der Fernsehzeit verringern würde, würde sich seine schulische Leistung verbessern.

cr6 ○ c Auch wenn Schüler, die gut waren, weniger fernsehen, bedeutet das nicht unbedingt, dass Fernsehen der schulischen Leistung schadet.

m3 ○ d Ein Monat ist keine hinreichende Zeitspanne, um zu schätzen, wie viel Zeit Schüler wirklich mit Fernsehen verbringen.

m6 ○ e Die Forschungsstudie demonstriert, dass Fernsehen eine schlechtere schulische Leistung verursacht.

○ f Ich stimme keiner dieser Aussagen zu.

17 Die Schulkommission eines kleinen Dorfes wollte die durchschnittliche Anzahl von Kindern pro Haushalt in ihrer Stadt bestimmen. Sie dividierten die Gesamtanzahl von Kindern im Dorf durch 50, die Gesamtanzahl der Haushalte. Welche der folgenden Aussagen muss richtig sein, wenn der Durchschnitt der Kinder pro Haushalt 2,2 ist?

m1d ○ a Die Hälfte der Haushalte des Dorfs hat mehr als zwei Kinder.

○ b Mehr Haushalte des Dorfs haben drei Kinder, als zwei Kinder.

cr2 ○ c Es gibt insgesamt 110 Kinder im Dorf.

○ d Es kommen 2,2 Kinder im Dorf auf jeden Erwachsenen.

m1a ○ e Die häufigste Anzahl von Kindern in einem Haushalt ist 2.

○ f Nichts von dem oben Aufgeführten.

18 Wenn zwei Würfel gleichzeitig geworfen werden, ist es möglich, dass eins der folgenden zwei Ergebnisse eintritt: *Ergebnis 1:* Eine 5 und eine 6 werden gewürfelt. *Ergebnis 2:* Eine 5 wird zweimal gewürfelt.
Wählen Sie die Antwort, der Sie am meisten zustimmen:

m7 ○ a Die Wahrscheinlichkeit für beide Ergebnisse ist gleich.

cr3b ○ b Die Wahrscheinlichkeit, *Ergebnis 1* zu erhalten ist höher.

○ c Die Wahrscheinlichkeit, *Ergebnis 2* zu erhalten ist höher.

○ d Es ist unmöglich, eine Antwort zu geben.

Tabelle 26 *(fortgesetzt)*

19			Wenn drei Würfel gleichzeitig geworfen werden, welche der folgenden Ergebnisse ist am wahrscheinlichsten?
cr3b	○	a	*Ergebnis 1*: „Eine 5, eine 3 und eine 6“.
	○	b	*Ergebnis 2*: „Dreimal eine 5“.
	○	c	*Ergebnis 3*: „Zweimal eine 5 und eine 3“.
m7	○	d	Alle drei Ergebnisse sind gleich wahrscheinlich.
20			Wenn drei Würfel gleichzeitig geworfen werden, welche der folgenden Ergebnisse ist am unwahrscheinlichsten?
	○	a	*Ergebnis 1*: „Eine 5, eine 3 und eine 6“.
cr3b	○	b	*Ergebnis 2*: „Dreimal eine 5“.
	○	c	*Ergebnis 3*: „Zweimal eine 5 und eine 3“.
m7	○	d	Alle drei Ergebnisse sind gleich unwahrscheinlich.

Anmerkungen. cr1 = Wahrscheinlichkeiten korrekt interpretieren (correctly interprets probabilities), cr2 = Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird (understands how to select an appropriate average), cr3 = Wahrscheinlichkeiten korrekt berechnen (correctly computes probability), cr3a = Wahrscheinlichkeit als Verhältnis verstehen (understands probabilities as ratios), cr3b = kombinatorisch denken (uses combinatorial reasoning), cr4 = (stochastische) Unabhängigkeit verstehen (understands independence), cr5 = Variablilität innerhalb von Stichproben verstehen (understands sampling variability), cr6 = zwischen Korrelation und Kausalität unterscheiden (distinguishes between correlation and causation), cr7 = Vierfeldertafeln korrekt interpretieren (correctly interprets two-way tables), cr8 = Bedeutung großer Stichproben verstehen (understands importance of large samples), m1 = Fehlverständnis von Mittelwerten (misconception involving averages), m1a = Mittelwert als am häufigsten vorkommender Messwert (averages are the most common number), m1b = fehlende Berücksichtigung von Ausreißern bei der Bestimmung des Durchschnitts (fails to take outliers into consideration when computing the mean), m1c = Vergleich von Gruppen anhand ihrer Mittelwerte (compares groups based on their averages), m1d = Verwechseln des Durchschnitts mit dem Median (confuses mean with median), m2 = Orientierung am Ergebnis (outcome orientation), m3 = gute Stichproben müssen einen hohen Anteil der Population darstellen (good samples have to represent a high percentage of the population), m4 = Gesetz der kleinen Zahl (law of small numbers), m5 = Repräsentativitätsheuristik (representativeness misconception), m6 = Gleichsetzen von Korrelation mit Kausalität (correlation implies causation), m7 = Gleichwahrscheinlichkeitsverzerrung (equiprobability bias), m8 = Überzeugung, dass Gruppen nur verglichen werden können, wenn sie dieselbe Stichprobengröße haben (groups can only be compared if they are the same size)

Tabelle 27

Aufgaben zur Mathematik (auf den nächsten Seiten fortgesetzt).

1 Drei Fünftel der Kinder einer Klasse sind Mädchen. Wenn 5 Mädchen und 5 Jungen dazukommen, welche der folgenden Aussagen über die Klasse ist dann wahr?

k
- ○ a In der Klasse gibt es mehr Mädchen als Jungen.
- ○ b Es gibt gleich viele Jungen wie Mädchen in der Klasse.
- ○ c In der Klasse gibt es mehr Jungen als Mädchen.
- ○ d Aufgrund dieser Informationen kann man nicht sagen, ob es mehr Mädchen oder mehr Jungen in der Klasse gibt.

2 [a] Jonas hat 5 Hüte weniger als Maria, und Clarissa hat dreimal so viele Hüte wie Jonas. Welcher der folgenden Ausdrücke steht für die Anzahl von Clarissas Hüten, wenn Maria n Hüte hat?

- ○ a $5-3n$
- ○ b $3n$
- ○ c $n-5$
- ○ d $3n-5$

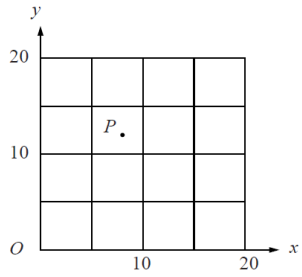
k
- ○ e $3(n-5)$

3 $P = LW$. Wenn $P = 12$ und $L = 3$, wie groß ist dann W?

- ○ a $\frac{3}{4}$
- ○ b 3

k
- ○ c 4
- ○ d 12
- ○ e 36

4 [a] Welches der folgenden Koordinatenpaare beschreibt die Koordinaten des Punktes P im Bild am besten?



k
- ○ a $(8; 12)$
- ○ b $(8; 8)$
- ○ c $(12; 8)$
- ○ d $(12; 12)$

Tabelle 27 *(fortgesetzt)*

5 [a] $\frac{3}{4}+\frac{2}{3}\cdot\frac{1}{4}=$

- ○ a $\frac{1}{8}$
- ○ b $\frac{5}{16}$
- ○ c $\frac{17}{48}$
- ○ d $\frac{5}{6}$

k ○ e $\frac{11}{12}$

6 Mit Infusionen (oder einem intravenösen Tropf) werden Patientinnen und Patienten mit Flüssigkeiten und Medikamenten versorgt.

Pflegekräfte müssen die Tropfrate D in Tropfen pro Minute für Infusionen berechnen.
Sie verwenden dazu die Formel

$$D=\frac{dv}{60n}\text{, wobei gilt:}$$

d ist der Tropffaktor gemessen in Tropfen pro Milliliter (ml)
v ist das Volumen der Infusion in ml
n ist die Anzahl Stunden, die die Infusion angeschlossen bleiben muss.

a Eine Pflegekraft möchte die Infusionsdauer verdoppeln.
Beschreiben Sie genau, wie sich D verändert, wenn n verdoppelt wird, aber d und v sich nicht ändern.

[offene Antwort: **D halbiert sich**]

b Pflegekräfte müssen auch das Volumen v der Infusion anhand der Tropfrate D berechnen.
Eine Infusion mit einer Tropfrate von 50 Tropfen pro Minute muss einem Patienten 3 Stunden lang verabreicht werden. Für diese Infusion ist der Tropffaktor 25 Tropfen pro Milliliter.
Wie groß ist das Volumen der Infusion in ml?

[offene Antwort: **360**]

Tabelle 27 *(fortgesetzt)*

7 Diese beiden Anzeigen sind in einer Zeitung erschienen in einem Land, in dem die Währungseinheit zeds ist.

GEBÄUDE A
Büroräume zu vermieten
85 - 95 Quadratmeter
475 *zeds* pro Monat
100 - 120 Quadratmeter
800 *zeds* pro Monat

GEBÄUDE B
Büroräume zu vermieten
35 - 260 Quadratmeter
90 *zeds* pro Quadratmeter
pro Jahr

Eine Firma ist daran interessiert, ein 110 Quadratmeter großes Büro in diesem Land für ein Jahr zu mieten. In welchem Bürogebäude, A oder B, sollte sie das Büro mieten, um den niedrigeren Preis zu bekommen?

k
- ○ a Gebäude A ist günstiger, die Miete hier kostet [offene Antwort: **9600** bzw. **800**] im Vergleich zur Miete von [offene Antwort: **9900** bzw. **825**] in Gebäude B.
- ○ b Gebäude B ist günstiger, die Miete hier kostet _____ im Vergleich zur Miete von _____ in Gebäude A.

8 Die Tabelle stellt einen Zusammenhang zwischen x und y dar.

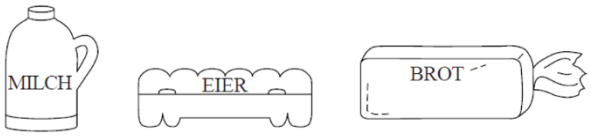
x	**y**
1	1
2	?
4	7
7	13

Wie lautet die fehlende Zahl in der Tabelle?

- ○ a 2

k
- ○ b 3
- ○ c 4
- ○ d 5
- ○ e 6

9 Peter hat 7 Euro, um Milch, Brot und Eier zu kaufen. Im Geschäft fand er die folgenden Preisangaben:



In welchem Fall wäre es sinnvoller zu schätzen, anstatt mit den Zahlen genau zu rechnen?

k
- ○ a Wenn Peter entscheiden möchte, ob 7 Euro ausreichen.
- ○ b Wenn der Verkäufer jeden einzelnen Preis in die Kasse eintippt.
- ○ c Wenn Peter gesagt wird, wieviel er bezahlen muss.
- ○ d Wenn der Verkäufer Peters Wechselgeld zählt.

Tabelle 27 *(fortgesetzt)*

10 Welcher der folgenden Ausdrücke ist gleich $m + m + m + m$, wenn m eine positive Zahl ist?

○ a $m + 4$

k ○ b $4m$

○ c m^4

○ d $4(m + 1)$

11 [a] Christian hat versucht, drei aufeinanderfolgende natürliche Zahlen zu finden, deren Summe 81 ist.
Er hat folgende Gleichung aufgeschrieben: $(n - 1) + n + (n + 1) = 81$. Wofür steht das n?

○ a Für die kleinste der drei natürlichen Zahlen.

k ○ b Für die mittlere der drei natürlichen Zahlen.

○ c Für die größte der drei natürlichen Zahlen.

○ d Für die Differenz zwischen der kleinsten und der größten der drei natürlichen Zahlen.

12 In Marks Garten gibt es 84 Reihen mit Kohl. In jeder Reihe sind 57 Kohlköpfe.
Welche der folgenden Gleichungen bietet die BESTE Möglichkeit, die Gesamtzahl der Kohlköpfe abzuschätzen?

○ a $100 \cdot 50 = 5000$

○ b $90 \cdot 60 = 5400$

k ○ c $80 \cdot 60 = 4800$

○ d $80 \cdot 50 = 4000$

13 Die Tabelle zeigt Werte von x und y, wobei x proportional zu y ist.

x	**y**
3	7
6	Q
P	35

Welches sind die Werte von P und Q?

○ a $P = 14$ und $Q = 31$

○ b $P = 10$ und $Q = 14$

○ c $P = 10$ und $Q = 31$

○ d $P = 14$ und $Q = 15$

k ○ e $P = 15$ und $Q = 14$

Tabelle 27 *(fortgesetzt)*

14 [a] Peter kauft 70 Stück einer Ware und Susi kauft 90 Stück. Jedes Stück kostet gleichviel. Alle Stücke zusammen kosten 800 Euro. Wieviel muss Susi zahlen?

[offene Antwort: **450**]

15 In welcher Zeile sind alle Brüche gleich groß?

○ a $\frac{3}{4}, \frac{6}{8}, \frac{12}{14}$

○ b $\frac{3}{5}, \frac{5}{7}, \frac{19}{15}$

k ○ c $\frac{3}{8}, \frac{6}{16}, \frac{12}{32}$

○ d $\frac{5}{10}, \frac{10}{15}, \frac{1}{2}$

16 Diese Tabelle zeigt einige Temperaturangaben, die an vier Tagen zu jeweils unterschiedlichen Zeiten gemessen wurden.

TEMPERATUREN					
	6 Uhr	9 Uhr	12 Uhr	15 Uhr	20 Uhr
Montag	15°	17°	20°	21°	19°
Dienstag	15°	15°	15°	10°	9°
Mittwoch	8°	10°	14°	13°	15°
Donnerstag	8°	11°	14°	17°	20°

Wann wurde die höchste Temperatur aufgezeichnet?

○ a Montag um 12 Uhr

k ○ b Montag um 15 Uhr

○ c Dienstag um 12 Uhr

○ d Mittwoch um 15 Uhr

17 Sie bereiten Ihr eigenes Salatdressing zu.
Hier ist ein Rezept für 100 Milliliter (ml) Dressing.

Salatöl:	60 ml
Essig:	30 ml
Sojasauce:	10 ml

Wie viele Milliliter (ml) Salatöl brauchen Sie, um 150 ml dieses Dressings zu machen?

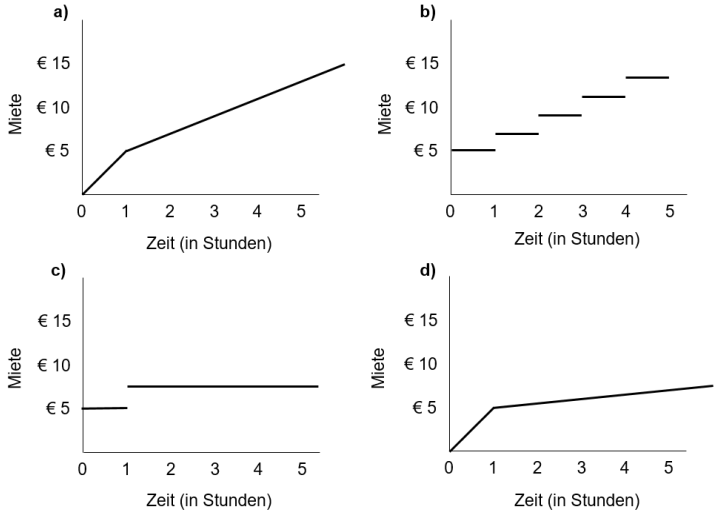
[offene Antwort: **90**]

18 Bestimmen Sie x, wenn $10x - 15 = 5x + 20$

[offene Antwort: **7**]

Tabelle 27 *(fortgesetzt)*

19 [a] In einem Park werden Fahrräder vermietet. Die erste Stunde (oder ein Teil davon) kostet 5 Euro und jede weitere angefangene Stunde kostet 2 Euro. Welches Diagramm zeigt dies?
Setzen Sie das Kreuz am entsprechenden Diagramm.



[**b**]

20 [a] In Zedland wurden Meinungsumfragen durchgeführt, um die Unterstützung für den Präsidenten bei der kommenden Wahl herauszufinden. Vier Zeitungsherausgeber machten separate landesweite Umfragen. Die Ergebnisse der Umfragen durch die vier Zeitungen werden unten angegeben.
Das Ergebnis welcher Zeitung ist am ehesten geeignet, um die Unterstützung für den Präsidenten vorauszusagen, wenn die Wahl am 25. Januar stattfindet?

- ○ a Zeitung 1: 36,5% (Umfrage durchgeführt am 6. Januar, bei einer Stichprobe von 500 zufällig ausgewählten Stimmberechtigten)
- ○ b Zeitung 2: 41,0% (Umfrage durchgeführt am 20. Januar, bei einer Stichprobe von 500 zufällig ausgewählten Stimmberechtigten)

k

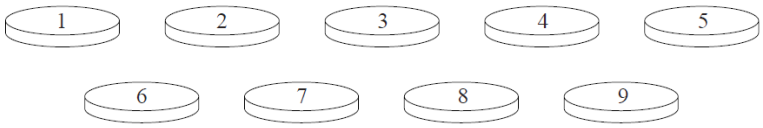
- ○ c Zeitung 3: 39,0% (Umfrage durchgeführt am 20. Januar, bei einer Stichprobe von 1000 zufällig ausgewählten Stimmberechtigten)
- ○ d Zeitung 4: 44,5% (Umfrage durchgeführt am 20. Januar, bei einer Stichprobe von 1000 Lesern, die angerufen haben, um zu sagen, wen sie wählen werden)

21 In zwei Kisten befinden sich 54 kg Äpfel. Die zweite Kiste Äpfel wiegt 12 kg mehr als die erste Kiste. Wie viele Kilogramm Äpfel sind in jeder Kiste?

Erste Kiste: [offene Antwort: **21**]
Zweite Kiste: [offene Antwort: **33**]

Tabelle 27 *(fortgesetzt)*

22 Die neun abgebildeten Spielsteine werden in einem Sack gemischt.



Madeleine zieht einen Spielstein aus dem Sack. Wie groß ist die Wahrscheinlichkeit, dass sie einen Spielstein mit einer geraden Zahl zieht?

○ a $\frac{1}{9}$

○ b $\frac{2}{9}$

k ○ c $\frac{4}{9}$

○ d $\frac{1}{2}$

23 Im letzten Jahr besuchten 1172 Schüler und Schülerinnen das Gymnasium Neufeld. Dieses Jahr sind es 15% mehr als im letzten. Wie viele Schüler und Schülerinnen gehen dieses Jahr ungefähr ins Gymnasium Neufeld?

○ a 1800

○ b 1600

○ c 1500

k ○ d 1400

○ e 1200

24 Welcher dieser Ausdrücke ist gleichbedeutend mit y^3 ?

○ a $y + y + y$

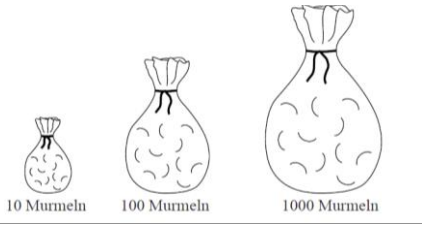
k ○ b $y \cdot y \cdot y$

○ c $3y$

○ d $y^2 + y$

Tabelle 27 *(fortgesetzt)*

25 In jedem dieser Beutel gibt es nur eine rote Murmel.



Sie sollen ohne hinzusehen aus einem der Beutel eine Murmel herausnehmen. Bei welchem Beutel ist die Chance am größten, dass Sie die rote Murmel ziehen?

k ○ a Bei dem Beutel mit den 10 Murmeln.

○ b Bei dem Beutel mit den 100 Murmeln.

○ c Bei dem Beutel mit den 1000 Murmeln.

○ d Die Chance ist bei allen Beuteln gleich.

26 [a] Ein Stapel von 200 gleichen Bögen Papier ist 2,5 cm dick. Wie dick ist ein einzelner Bogen?

○ a 0,008 cm

k ○ b 0,0125 cm

○ c 0,05 cm

○ d 0,08 cm

27 [a] Die Länge eines Rechtecks beträgt 6 cm und sein Umfang 16 cm. Wie groß ist der Flächeninhalt des Rechtecks in Quadratzentimetern?

[offene Antwort: **12**]

28 [a] An Manuelas Schule führt der Physiklehrer Tests durch, bei denen 100 Punkte zu erreichen sind. Manuela hat bei ihren ersten vier Physiktests durchschnittlich 60 Punkte erreicht. Beim fünften Test erreichte sie 80 Punkte.
Was ist Manuelas Punktedurchschnitt in Physik nach allen fünf Tests?

[offene Antwort: **64**]

29 Wenn $3(x + 5) = 30$, dann ist $x =$

○ a 2

k ○ b 5

○ c 10

○ d 95

30 Wenn $x = 2$, welchen Wert hat $\frac{7x+4}{5x-4}$?

[offene Antwort: **3**]

Anmerkungen. k: korrekte Option, [a] Items, die in allen Stichproben erhoben wurden, alle Items stammen aus TIMSS bzw. PISA-Erhebungen, die Darstellung bei der Erhebung erfolge in größerer Schrift und mit größeren Abbildungen als hier dargestellt.

Tabelle 28

Instruktion und Aufgaben zum deduktiven Schließen (auf den nächsten Seiten fortgesetzt).

Aufgaben zum Denken

Nun bitte ich Sie noch ein paar Denkaufgaben zu bearbeiten.

Damit Sie durch das Verwenden bestimmter Inhalte der Aufgaben bei der Bearbeitung nicht benachteiligt werden, sind die Aufgaben mit sogenannten Pseudowörtern gestellt. Diese Wörter haben keine Bedeutung, sondern sind Platzhalter für echte Wörter. So können Sie sich allein auf das Denken konzentrieren.

Bei den Aufgaben geht es darum, zu entscheiden, welche Aussagen sich aus Regeln ableiten lassen. Zunächst wird Ihnen eine Regel gezeigt und bei manchen Aufgaben noch ein Geschehnis. Ihre Aufgabe besteht nun darin, die Aussagen anzukreuzen, die sich aus dieser Regel und gegebenenfalls dem Geschehnis ableiten lassen.

Beispiel:

Regel: **Wenn es riepst, dann klögt es.**

Geschehnis: Es hat geriepst.

Welche der Aussagen lässt/lassen sich logisch aus der Regel und dem Geschehnis ableiten?

- [x] a Es hat geklögt.
- [ ] b Es hat nicht geklögt.
- [ ] c Weder a) noch b).

Die Regel sagt, dass es klögt, wenn es riepst. Da es geriepst hat, muss es also auch klögen.

Geschehnis: Es hat <u>nicht</u> geriepst.

Welche der Aussagen lässt/lassen sich logisch aus der Regel und dem Geschehnis ableiten?

- [ ] a Es hat geklögt.
- [ ] b Es hat nicht geklögt.
- [x] c Weder a) noch b).

Die Regel sagt nichts darüber aus, was passiert, wenn es <u>nicht</u> riepst. Es könnte auch klögen, ohne dass es geriepst hat. Man kann also nicht sagen, ob es geklögt hat oder nicht, wenn es nicht geriepst hat.

Nach jeder Frage bzw. Aufgabe bitte ich Sie, auch hier wieder anzugeben, wie sicher Sie sich bei Ihrer Wahl der Antwort sind.

Tabelle 28 *(fortgesetzt)*

Welche der Aussagen lässt/lassen sich logisch aus der Regel und dem Geschehnis ableiten? Bitte wählen Sie für jedes Geschehnis aus.

1[a] **Regel: Wenn es schnuckt, dann glackst es.**

[a] Geschehnis: Es hat geschnuckt.

k ○ a Es hat geglackst.

○ b Es hat nicht geglackst.

○ c Weder a) noch b).

[a] Geschehnis: Es hat nicht geschnuckt.

○ a Es hat geglackst.

○ b Es hat nicht geglackst.

k ○ c Weder a) noch b).

[a] Geschehnis: Es hat geglackst.

○ a Es hat geschnuckt.

○ b Es hat nicht geschnuckt.

k ○ c Weder a) noch b).

[a] Geschehnis: Es hat nicht geglackst.

○ a Es hat geschnuckt.

k ○ b Es hat nicht geschnuckt.

○ c Weder a) noch b).

Welche der Aussagen lässt/lassen sich logisch aus der Regel und dem Geschehnis ableiten? Bitte wählen Sie für jedes Geschehnis aus.

2[a] **Regel: Wenn es preift, dann zwogt es nicht.**

[a] Geschehnis: Es hat gepreift.

○ a Es hat gezwogt.

k ○ b Es hat nicht gezwogt.

○ c Weder a) noch b).

[a] Geschehnis: Es hat nicht gepreift.

○ a Es hat gezwogt.

○ b Es hat nicht gezwogt.

k ○ c Weder a) noch b).

Tabelle 28 *(fortgesetzt)*

[a] Geschehnis: Es hat gezwogt.

- ○ a Es hat gepreift.
- **k** ○ b Es hat nicht gepreift.
- ○ c Weder a) noch b).

[a] Geschehnis: Es hat <u>nicht</u> gezwogt.

- ○ a Es hat gepreift.
- ○ b Es hat nicht gepreift.
- **k** ○ c Weder a) noch b).

Welche der Aussagen lässt/lassen sich logisch aus der Regel und dem Geschehnis ableiten? Bitte wählen Sie für jedes Geschehnis aus.

3 Regel: Wenn es märgt, dann ist es sehr wahrscheinlich, dass es guht.

Geschehnis: Es hat gemärgt.

- **k** ○ a Es hat sehr wahrscheinlich geguht.
- ○ b Es hat sehr wahrscheinlich nicht geguht.
- ○ c Weder a) noch b).

Geschehnis: Es hat <u>nicht</u> gemärgt.

- ○ a Es hat sehr wahrscheinlich geguht.
- ○ b Es hat sehr wahrscheinlich nicht geguht.
- **k** ○ c Weder a) noch b).

Geschehnis: Es hat geguht.

- ○ a Es hat sehr wahrscheinlich gemärgt.
- ○ b Es hat sehr wahrscheinlich nicht gemärgt.
- **k** ○ c Weder a) noch b).

Geschehnis: Es hat <u>nicht</u> geguht.

- ○ a Es hat sehr wahrscheinlich gemärgt.
- ○ b Es hat sehr wahrscheinlich nicht gemärgt.
- **k** ○ c Weder a) noch b).

Tabelle 28 *(fortgesetzt)*

Welche der Aussagen lässt/lassen sich logisch aus der Regel und dem Geschehnis ableiten? Bitte wählen Sie für jedes Geschehnis aus.

4 Regel: Wenn es beift, dann ist es sehr wahrscheinlich, dass es nicht rühst.

Geschehnis: Es hat gebeift.

○ a Es hat sehr wahrscheinlich gerühst.

k ○ b Es hat sehr wahrscheinlich nicht gerühst.

○ c Weder a) noch b).

Geschehnis: Es hat nicht gebeift.

○ a Es hat sehr wahrscheinlich gerühst.

○ b Es hat sehr wahrscheinlich nicht gerühst.

k ○ c Weder a) noch b).

Geschehnis: Es hat gerühst.

○ a Es hat sehr wahrscheinlich gebeift.

○ b Es hat sehr wahrscheinlich nicht gebeift.

k ○ c Weder a) noch b).

Geschehnis: Es hat nicht gerühst.

○ a Es hat sehr wahrscheinlich gebeift.

○ b Es hat sehr wahrscheinlich nicht gebeift.

k ○ c Weder a) noch b).

Welche der Aussagen lässt/lassen sich logisch aus der Regel und dem Geschehnis ableiten? Bitte wählen Sie für jedes Geschehnis aus.

5 Regel: Wenn es wievt, dann ist es sehr unwahrscheinlich, dass es loscht.

Geschehnis: Es hat gewievt.

○ a Es ist sehr wahrscheinlich, dass es geloscht hat.

k ○ b Es ist sehr unwahrscheinlich, dass es geloscht hat.

○ c Weder a) noch b).

Geschehnis: Es hat nicht gewievt.

○ a Es ist sehr wahrscheinlich, dass es geloscht hat.

○ b Es ist sehr unwahrscheinlich, dass es geloscht hat.

k ○ c Weder a) noch b).

Tabelle 28 *(fortgesetzt)*

Geschehnis: Es hat geloscht.

- ○ a Es ist sehr wahrscheinlich, dass es gewievt hat.
- ○ b Es ist sehr unwahrscheinlich, dass es gewievt hat.

k ○ c Weder a) noch b).

Geschehnis: Es hat <u>nicht</u> geloscht.

- ○ a Es ist sehr wahrscheinlich, dass es gewievt hat.
- ○ b Es ist sehr unwahrscheinlich, dass es gewievt hat.

k ○ c Weder a) noch b).

Welche der Aussagen lässt/lassen sich logisch aus der Regel ableiten?

6 Regel: Alle Schimten sind Tichter.

- ○ a Alle Tichter sind Schimten.
- ○ b Manche, aber nicht alle Tichter sind Schimten.

k ○ c Manche oder alle Tichter sind Schimten.

- ○ d Keine Tichter sind Schimten.
- ○ e Keine der genannten Aussagen lässt sich aus der Regel ableiten.

7 Regel: <u>Manche</u> Utfelle sind Sähzen.

- ○ a Alle Sähzen sind Utfelle.
- ○ b Manche, aber nicht alle Sähzen sind Utfelle.

k ○ c Manche oder alle Sähzen sind Utfelle.

- ○ d Keine Sähzen sind Utfelle.
- ○ e Keine der genannten Aussagen lässt sich aus der Regel ableiten.

8 Regel: <u>Manche</u> Kribse sind <u>keine</u> Tagel.

- ○ a Manche Kribse sind Tagel.
- ○ b Die meisten Kribse sind Tagel.
- ○ c Alle Tagel sind keine Kribse.
- ○ d Manche Tagel sind Kribse.

k ○ e Keine der genannten Aussagen lässt sich aus der Regel ableiten.

Tabelle 28 *(fortgesetzt)*

Welche der Aussagen lässt/lassen sich logisch aus der Regel ableiten?

9 **Regel: Alle Brahken sind Wugen.**
Alle Wugen sind Präutel.

k	○	a	Alle Brahken sind Präutel.
	○	b	Alle Präutel sind Brahken.
	○	c	Alle Wugen sind Brahken.
	○	d	Alle Präutel sind Wugen.
	○	e	Keine der genannten Aussagen lässt sich aus der Regel ableiten.

10 **Regel: Alle Äfzell sind Lafter.**
<u>Keine</u> Äfzell sind Münen.

	○	a	Alle Lafter sind Münen.
	○	b	Alle Münen sind Lafter.
	○	c	Kein Lafter ist eine Müne.
	○	d	Keine Müne ist ein Lafter.
k	○	e	Keine der genannten Aussagen lässt sich aus der Regel ableiten.

Anmerkungen. k = korrekte Antwortoption, [a] Items, die in allen Stichproben eingesetzt wurden.

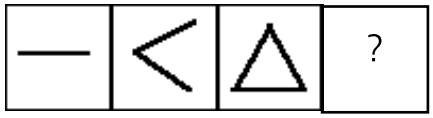
Tabelle 29

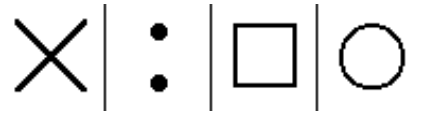
Instruktion und figurale Reihenvervollständigungsaufgaben aus dem Mensa Online IQ-Test (auf der nächsten Seite fortgesetzt). Mit freundlicher Genehmigung von Mensa Netherlands.

Aufgaben zum Denken 2

Schließlich bitte ich Sie, die folgenden 10 Rätsel zu lösen. Auch hier gibt es für Sie keine Benachteiligung durch Ihre persönliche Erfahrung oder Sprache:
Wählen Sie für die folgenden drei Figuren auf der linken Seite die passende vierte Figur aus, von den Figuren, die rechts angeboten werden (a, b, c oder d). Bitte geben Sie auch hier wieder an, wie sicher Sie sich bei der Lösung sind.

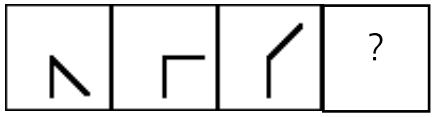
1





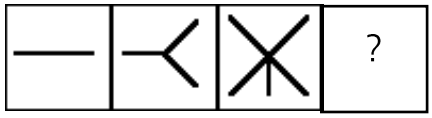
○ **a** ○ **b** ○ **c**[k] ○ **d**

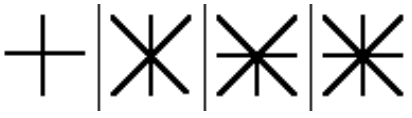
2



○ **a** ○ **b**[k] ○ **c** ○ **d**

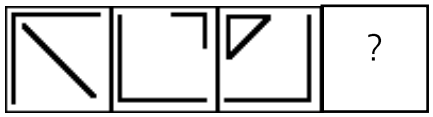
3





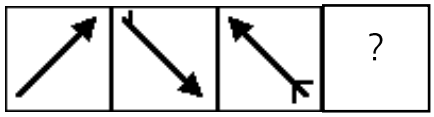
○ **a** ○ **b** ○ **c** ○ **d**[k]

4



○ **a** ○ **b** ○ **c** ○ **d**[k]

5



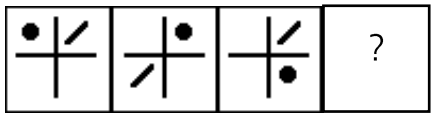
○ **a** ○ **b** ○ **c** ○ **d**[k]

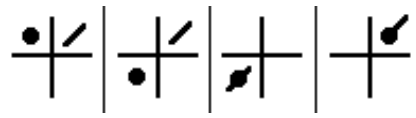
6

○ **a** ○ **b** ○ **c**[k] ○ **d**

Tabelle 29 *(fortgesetzt)*

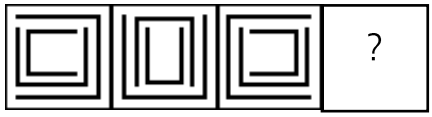
7





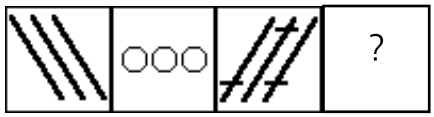
○ **a** ○ **b** ○ **c**[k] ○ **d**

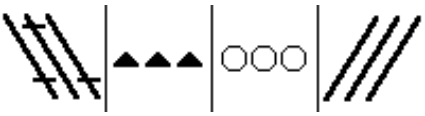
8



○ **a** ○ **b** ○ **c** ○ **d**[k]

9





○ **a** ○ **b**[k] ○ **c** ○ **d**

10

○ **a**[k] ○ **b** ○ **c** ○ **d**

Anmerkungen. [k] korrekte Antwortoption, die Items wurden bei den Stichproben 1 (2 SWS, B.A.) und 2 (2 SWS, M.A.) eingesetzt.

Schwierigkeiten

- für Items des Statistical Reasoning Assessments bei wohlwollender Bewertung (Vergabe eines Punktes beim Ankreuzen der richtigen Antwortoption, unabhängig davon, ob zusätzlich eine falsche Antwortoption angekreuzt wurde)
- für Mathematik-Items aus TIMSS und PISA
    - Originalangaben, wo verfügbar
    - aus den Stichproben 3 und 4 für alle 31 Items
    - aus den Stichproben 1 und 2 für 10 Items
- für Items des deduktiven Schließens
    - aus den Stichproben 3 und 4 für alle 25 Items
    - aus den Stichproben 1 und 2 für 8 Items
- für figurale Reihenvervollständigungsitems aus den Stichproben 1 und 2

Tabelle 30

Schwierigkeiten der SRA-Items für beide Zeitpunkte sowie die Differenz der Schwierigkeiten zwischen dem Beginn und dem Ende des Semesters für alle Stichproben bei wohlwollender Bewertung (Antworten, die die korrekte Antwortoption enthalten, gelten als korrekt).

S	Z	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	64.3	92.4	34.5	81.9	33.1	83.5	91.3	89.8	86	63.5	71.6	63.4	29.7	25.4	16.1	73	49	11.7	20.1	26.4
		(263)	(263)	(267)	(265)	(266)	(266)	(265)	(266)	(264)	(266)	(264)	(265)	(266)	(264)	(261)	(263)	(261)	(266)	(264)	(265)
	2	50.9	90.7	32.9	78.3	36.2	64.2	96.3	95.1	89.5	77.2	78.1	69.8	22.8	34.6	8.7	70	57.8	10.6	13.8	19.4
		(161)	(161)	(161)	(161)	(160)	(162)	(162)	(162)	(162)	(162)	(160)	(162)	(162)	(162)	(161)	(160)	(161)	(160)	(160)	(160)
	D	**-13.3**	**-1.7**	**-1.5**	**-3.6**	3.2	**-19.3**	5	5.2	3.5	13.6	6.5	6.4	**-6.9**	9.2	**-7.4**	**-3**	8.7	**-1**	**-6.3**	**-7**
2	1	52.9	94.1	29.4	82.4	33.3	70.6	90.9	88.2	88.2	67.6	82.4	70.6	18.2	8.8	12.9	88.2	55.9	17.6	14.7	23.5
		(34)	(34)	(34)	(34)	(33)	(34)	(33)	(34)	(34)	(34)	(34)	(34)	(33)	(34)	(31)	(34)	(34)	(34)	(34)	(34)
	2	46.4	92.9	21.4	78.6	28.6	53.6	89.3	85.7	89.3	67.9	82.1	53.6	10.7	3.6	14.3	85.7	75	10.7	7.1	14.3
		(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)	(28)
	D	**-6.5**	**-1.3**	**-8**	**-3.8**	**-4.8**	**-17**	**-1.6**	**-2.5**	1.1	0.2	**-0.2**	**-17**	**-7.5**	**-5.3**	1.4	**-2.5**	19.1	**-6.9**	**-7.6**	**-9.2**
3	1	41.9	74.2	22.6	83.9	41.9	54.8	93.5	83.9	71	54.8	58.1	54.8	32.3	32.3	3.2	61.3	45.2	9.7	19.4	38.7
		(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)	(31)
	2	62.1	93.1	48.3	58.6	20.7	79.3	86.2	96.6	96.6	72.4	82.8	44.8	17.2	31	0	82.8	58.6	17.2	27.6	41.4
		(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)	(29)
	D	20.1	18.9	25.7	**-25.3**	**-21.2**	24.5	**-7.3**	12.7	25.6	17.6	24.7	**-10**	**-15**	**-1.2**	**-3.2**	21.5	13.5	7.6	8.2	2.7
4	1	52.9	88.2	39.2	80.4	49	86.3	86.3	90.2	92.2	66.7	86.3	70.6	21.6	31.4	10.2	70.6	54.9	11.8	11.8	21.6
		(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(51)	(49)	(51)	(51)	(51)	(51)	(51)
	2	60.5	93	39.5	74.4	30.2	58.1	90.7	97.6	97.7	72.1	92.9	67.4	18.6	27.9	9.3	67.4	55.8	9.5	7.1	11.9
		(43)	(43)	(43)	(43)	(43)	(43)	(43)	(42)	(43)	(43)	(42)	(43)	(43)	(43)	(43)	(43)	(43)	(42)	(42)	(42)
	D	7.5	4.8	0.3	**-6**	**-18.8**	-28.1	4.4	7.4	5.5	5.4	6.6	**-3.1**	**-3**	**-3.5**	**-0.9**	**-3.1**	0.9	**-2.2**	**-4.6**	**-9.7**

Anmerkungen. *n* jeweils in Klammern angegeben. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., Z: Zeitpunkt mit 1 = zu Beginn der Intervention, 2 = am Ende der Intervention und D = Differenz der Schwierigkeiten wobei positive Werte eine Erhöhung des Schwierigkeitsindex anzeigen (d. h. mehr Studierende lösen das Item anteilig am Ende des Semesters; negative Werte sind **fett** gesetzt).

Tabelle 31

Übersicht der Schwierigkeiten der Aufgaben zu Mathematik.

Nr.	Quelle	Code	Bereich	P_o	P_3	P_4	P_1	P_2
1	TIMSS	Q5	Pr	67	94 (35)	87 (54)		
2	TIMSS	Q1	A	41	74 (35)	81 (53)	74 (258)	62 (32)
3	TIMSS	Q7	A	57	86 (35)	87 (54)		
4	TIMSS	J16	G	59	91 (35)	87 (54)	96 (259)	90 (31)
5	TIMSS	Q9	Z	50	69 (35)	74 (54)	80 (249)	84 (31)
6a	PISA	PM903	V	5^K	42 (26)	35 (48)		
6b	PISA	PM903	V	5^K	29 (24)	55 (38)		
7	TIMSS	V2	D	14	58 (33)	60 (52)		
8	TIMSS	J18	A	35	62 (34)	55 (51)		
9	TIMSS	I7	Z	73	91 (35)	87 (54)		
10	TIMSS	P10	A	57	83 (35)	93 (54)		
11	TIMSS	I1	A	27	39 (33)	56 (52)	63 (248)	36 (25)
12	TIMSS	P12	Z	73	89 (35)	94 (54)		
13	TIMSS	L14	Pr	18	45 (33)	49 (49)		
14	TIMSS	R14	Pr	38	88 (34)	94 (48)	83 (231)	87 (31)
15	TIMSS	N14	Z	64	91 (35)	87 (53)		
16	TIMSS	L10	D	87	91 (34)	96 (54)		
17	PISA	PM924	Gr	3^K	77 (35)	79 (52)		
18	TIMSS	L16	A	43	79 (29)	80 (41)		
19	TIMSS	M9	D	/	32 (34)	59 (54)	49 (256)	44 (32)
20	PISA	M702	U	5^K	69 (32)	68 (53)	75 (249)	84 (32)
21	TIMSS	T1	A	25	65 (31)	76 (50)		
22	TIMSS	N18	D	55	94 (34)	91 (54)		
23	TIMSS	K6	Z	53	73 (33)	85 (54)		
24	TIMSS	P15	A	73	89 (35)	94 (53)		
25	TIMSS	M3	D	83	91 (35)	87 (53)		
26	TIMSS	R7	Z	48	69 (35)	77 (53)	79 (258)	78 (32)
27	TIMSS	K5	M	29	48 (27)	71 (45)	81 (240)	89 (27)
28	PISA	M468	U	4^K	65 (31)	84 (49)	73 (246)	80 (30)
29	TIMSS	O7	A	79	97 (35)	91 (53)		
30	TIMSS	N13	A	50	91 (32)	90 (49)		

Anmerkungen. P_o = originale Schwierigkeit aus TIMSS bzw. PISA, P_3 = Schwierigkeit aus Stichprobe 3 (4 SWS Statistik, B.Sc.), P_4 = Schwierigkeit aus Stichprobe 4 (4(+4) SWS Statistik, B.Sc.), P_1 = Schwierigkeit aus Stichprobe 1 (2 SWS Statistik, B.A.), P_2 = Schwierigkeit aus Stichprobe 2 (2 SWS Statistik, M.A.), Pr = Proportionalität, A = Algebra, G = Geometrie, Z = Zahlen und Zahlenverständnis, V = Veränderung und Zusammenhänge, D = Darstellung und Analyse von Daten, Gr = Größen, U = Unsicherheit, M = Messen und Maßeinheiten, ^{K} = Kompetenzstufe, (N in Klammern).

Tabelle 32

Übersicht der Schwierigkeiten der Aufgaben zu deduktivem Schließen.

Nr. und Inhalt	P_3	P_4	P_1	P_2
Junktoren: Implikation				
1	97 (35)	95 (55)	98 (161)	97 (31)
2	46 (35)	36 (55)	67 (161)	45 (31)
3	20 (35)	25 (55)	50 (161)	32 (31)
4	77 (35)	69 (55)	50 (161)	68 (31)
5	91 (35)	95 (55)	97 (161)	97 (31)
6	51 (35)	56 (55)	73 (161)	45 (31)
7	74 (35)	67 (55)	50 (161)	52 (31)
8	15 (34)	24 (55)	37 (161)	23 (31)
Junktoren: Implikation (induktiv)				
9	97 (34)	98 (55)		
10	41 (34)	38 (55)		
11	32 (34)	25 (55)		
12	41 (34)	44 (55)		
13	100 (35)	96 (55)		
14	49 (35)	51 (55)		
15	20 (35)	35 (55)		
16	26 (35)	22 (55)		
17	97 (35)	98 (55)		
18	47 (34)	51 (55)		
19	17 (35)	35 (55)		
20	34 (35)	31 (55)		
Quantoren: Syllogismen				
21	26 (35)	36 (55)		
22	21 (34)	22 (55)		
23	21 (34)	11 (55)		
24	56 (34)	71 (55)		
25	41 (34)	42 (53)		

Anmerkungen. P_3 = Schwierigkeit aus Stichprobe 3 (4 SWS Statistik, B.Sc.), P_4 = Schwierigkeit aus Stichprobe 4 (4(+4) SWS Statistik, B.Sc.), P_1 = Schwierigkeit aus Stichprobe 1 (2 SWS Statistik, B.A.), P_2 = Schwierigkeit aus Stichprobe 2 (2 SWS Statistik, M.A.), (N in Klammern).

Tabelle 33

Übersicht der Schwierigkeiten der figuralen Reihenvervollständigungsitems.

Nr.	P_1	P_2
1	91 (160)	79 (28)
2	89 (160)	86 (28)
3	12 (159)	11 (28)
4	71 (159)	74 (27)
5	76 (160)	86 (28)
6	60 (160)	68 (28)
7	43 (159)	36 (28)
8	62 (158)	75 (28)
9	21 (160)	46 (28)
10	44 (158)	36 (28)

Anmerkungen. P_1 = Schwierigkeit aus Stichprobe 1 (2 SWS Statistik, B.A.), P_2 = Schwierigkeit aus Stichprobe 2 (2 SWS Statistik, M.A.), (N in Klammern).

Deskriptive Statistiken

- für die Kurz- und Langversion der Aufgabensammlung zur Mathematik
- für die Kurz- und Langversion der Aufgabensammlung zum deduktiven Schließen
- für die Summe der figuralen Reihenfortsetzungsaufgaben
- und Cronbach α für die Skalen des Survey of Attitudes Towards Statistics (SATS)
- in Form geglätteter Histogramme für die Summe der SRA Punkte

Alle Skalenwerte basieren auf Items, von denen höchstens 30% fehlende Angaben enthalten.

Tabelle 34

Deskriptive univariate Statistiken zu der Kurz- (10 Items) und Langversion (31 Items) der Aufgabensammlung zur Mathematik nach Stichproben.

Version	S	*M*	*SD*	*Md*	*IQB*	Min	Max	*n*
lang	3	21.66	4.52	21	[18 26.5]	15	29	35
	4	23.00	5.59	24.5	[20 27]	8	30	54
kurz	1	7.22	1.89	7	[6 9]	2	10	260
	2	7.00	2.17	7	[6 9]	2	10	32
	3	6.12	2.23	6	[5 8]	2	10	34
	4	7.21	2.16	8	[6 9]	2	10	53

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc.

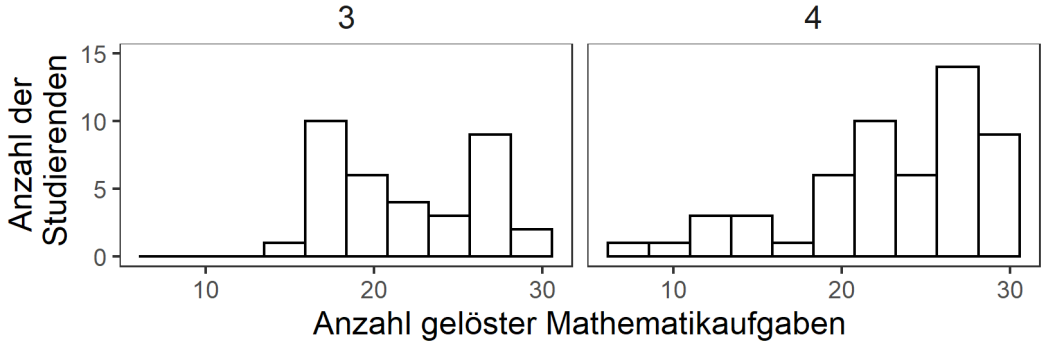


Abbildung 16. Häufigkeitsverteilung gelöster Aufgaben in den Stichproben 3 (4 SWS, B.Sc.) und 4 (4(+4) SWS, B.Sc.) der Mathematikaufgaben-Langversion.

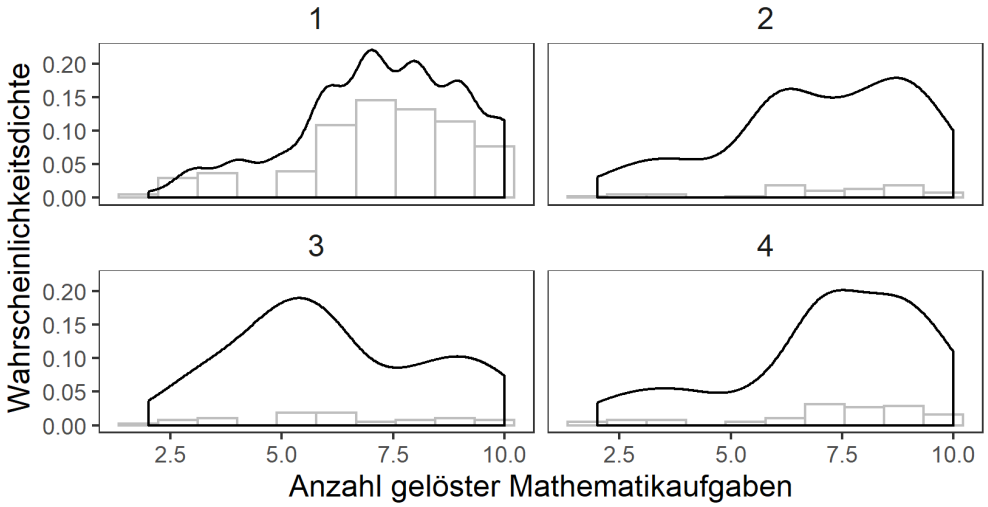


Abbildung 17. Wahrscheinlichkeitsverteilung (geglättetes Histogramm) gelöster Aufgaben in den Stichproben 1 (2 SWS, B.A.), 2 (2 SWS, M.A.), 3 (4 SWS, B.Sc.) und 4 (4(+4) SWS, B.Sc.) der Mathematikaufgaben-Kurzversion.

Tabelle 35

Deskriptive univariate Statistiken zu der Kurz- (8 Items) und Langversion (25 Items) der Aufgabensammlung zum deduktiven Schließen (Logik) nach Stichproben.

Version	S	*M*	*SD*	*Md*	*IQB*	Min	Max	*n*
lang	3	12.26	3.78	12	[9 14]	6	22	35
	4	12.71	4.09	13	[8 16]	6	21	55
kurz	1	5.21	1.34	5	[4 6]	1	8	161
	2	4.58	1.03	4	[4 5]	3	7	31
	3	4.71	1.23	4	[4 5.5]	2	8	35
	4	4.67	1.16	4	[4 5.5]	2	8	55

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc.

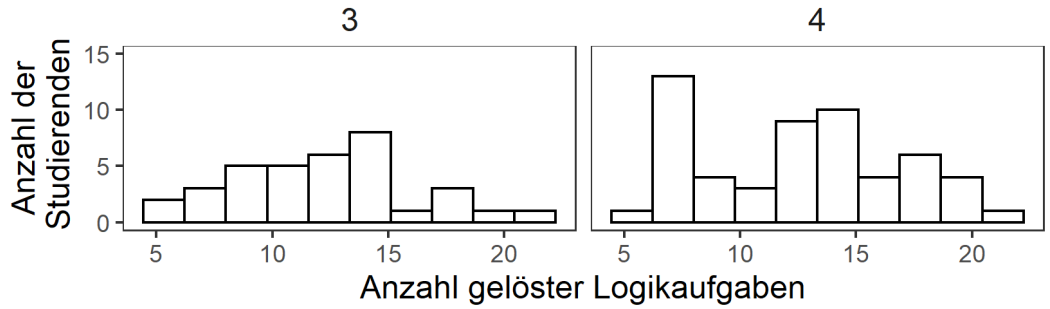


Abbildung 18. Häufigkeitsverteilung gelöster Aufgaben in den Stichproben 3 (4 SWS, B.Sc.) und 4 (4(+4) SWS, B.Sc.) der Logikaufgaben-Langversion.

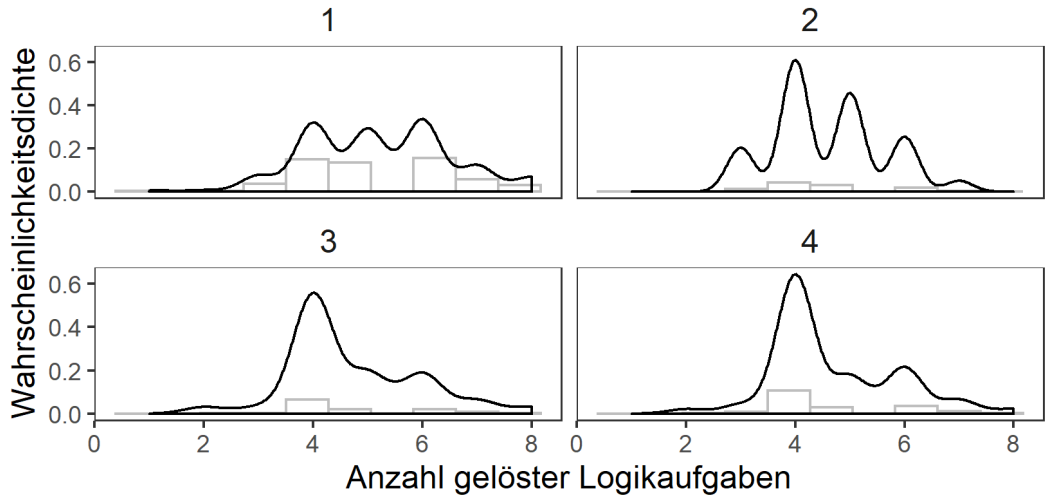


Abbildung 19. Wahrscheinlichkeitsverteilung (geglättetes Histogramm) gelöster Aufgaben in den Stichproben 1 (2 SWS, B.A.), 2 (2 SWS, M.A.), 3 (4 SWS, B.Sc.) und 4 (4(+4) SWS, B.Sc.) der Logikaufgaben-Kurzversion.

Tabelle 36

Deskriptive univariate Statistiken zu den figuralen Reihenfortsetzungsitems.

S	M	SD	Md	IQB	Min	Max	n
1	5.67	1.82	6	[4 7]	0	9	160
2	5.93	1.84	6.5	[4 7]	1	8	28

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc.

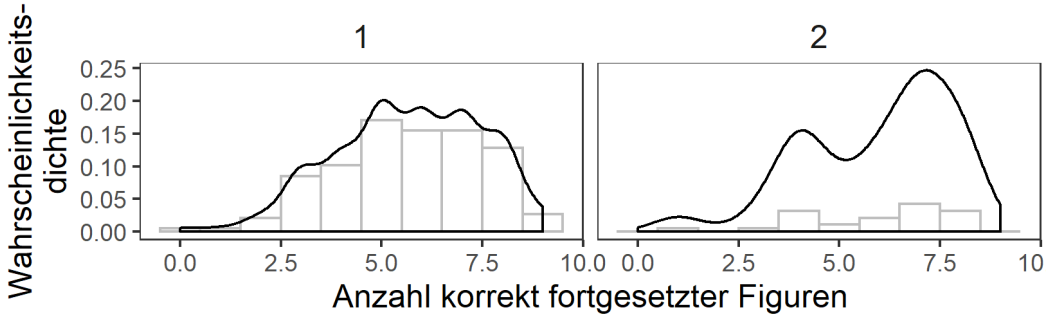


Abbildung 20. Wahrscheinlichkeitsverteilung (geglättetes Histogramm) gelöster Aufgaben in den Stichproben 1 (2 SWS, B.A.) und 2 (2 SWS, M.A.) der figuralen Reihenfortsetzungsitems.

Tabelle 37

Deskriptive univariate Statistiken und das Cronbach α für die Skalen des Survey of Attitudes Towards Statistics (SATS) für alle Gruppen zu Beginn und am Ende des Semesters.

Skala	S	Z	M	SD	Md	IQB	Min	Max	n	α
Affekt	1	1	3.72	1.17	3.83	[3;4.5]	1.00	7.00	266	.84
(6 Items)		2	3.84	1.18	3.83	[3;4.83]	1.00	6.20	161	.74
	2	1	3.43	1.01	3.42	[2.88;3.83]	1.33	6.33	34	.78
		2	3.34	0.96	3.50	[2.83;3.83]	1.00	6.17	28	.70
	3	1	3.80	0.93	3.50	[3.17;4.5]	2.00	5.67	31	.76
		2	3.67	1.23	3.50	[2.83;4.17]	1.50	6.17	29	.86
	4	1	4.57	1.28	4.67	[3.82;5.42]	1.50	6.83	51	.87
		2	3.97	1.40	4.17	[2.83;4.92]	1.33	6.50	43	.89
Kom-	1	1	4.31	1.16	4.33	[3.5;5.17]	1.00	7.00	265	.86
petenz		2	4.63	1.09	4.67	[4;5.5]	1.67	6.83	160	.77
(6 Items)	2	1	4.16	1.09	3.92	[3.38;4.83]	2.00	6.67	34	.89
		2	4.53	1.02	4.50	[4.08;5.17]	1.67	6.83	27	.81
	3	1	4.52	1.09	4.33	[3.92;5.25]	2.33	6.67	31	.80
		2	4.74	0.94	4.83	[4.17;5.33]	2.67	6.33	29	.75
	4	1	5.04	1.06	5.00	[4.25;6]	2.83	6.83	51	.82
		2	4.78	0.97	4.83	[4.17;5.42]	3.00	6.67	43	.77
Wert	1	1	4.54	0.97	4.56	[4;5.11]	1.44	8.56	266	.75
(9 Items)		2	4.30	1.09	4.33	[3.67;5.11]	1.33	6.78	161	.85
	2	1	4.88	0.94	4.83	[4.22;5.67]	3.33	6.75	34	.82
		2	4.36	0.92	4.28	[3.67;4.81]	2.89	6.56	28	.79
	3	1	5.00	1.06	5.22	[4.72;5.61]	1.78	6.78	31	.84
		2	4.84	0.80	4.78	[4.22;5.44]	3.22	6.11	29	.74
	4	1	5.00	0.81	4.89	[4.33;5.78]	3.11	6.44	51	.78
		2	5.11	0.87	5.22	[4.44;5.78]	3.11	6.89	43	.82
Schwie-	1	1	4.59	0.82	4.57	[4;5.14]	2.43	7.00	264	.73
rigkeit		2	4.60	0.88	4.71	[4.07;5.29]	2.43	7.00	159	.72
(7 Items)	2	1	5.00	0.61	5.00	[4.71;5.43]	3.71	6.43	34	.53
		2	4.85	0.73	4.86	[4.29;5.29]	3.57	6.57	27	.67
	3	1	4.72	0.57	4.71	[4.5;5]	3.29	6.29	31	.38
		2	4.74	0.80	4.71	[4.29;5.29]	3.14	6.29	29	.67
	4	1	4.38	0.72	4.43	[4;4.86]	2.71	6.14	51	.58
		2	4.37	0.69	4.29	[3.86;4.79]	3.00	6.00	43	.55

Tabelle 37 *(fortgesetzt)*

Skala	S	Z	M	SD	Md	IQB	Min	Max	n	α
Inte-	1	1	4.34	1.25	4.50	[3.5;5.25]	1.00	7.00	269	.86
resse		2	3.90	1.39	4.00	[3;5]	1.00	7.00	161	.86
(4 Items)	2	1	4.96	0.85	5.00	[4.31;5.5]	3.25	6.50	34	.70
		2	4.36	1.42	4.75	[3.75;5]	1.75	7.00	27	.87
	3	1	4.81	1.27	5.00	[4;5.71]	1.25	7.00	31	.90
		2	4.55	1.25	4.75	[3.75;5.5]	2.25	6.50	29	.88
	4	1	5.01	1.07	5.25	[4.25;5.75]	2.25	7.00	51	.80
		2	4.80	1.04	4.75	[4.25;5.62]	2.50	7.00	43	.77
Anstren-	1	1	5.25	1.10	5.50	[4.75;6]	1.00	7.00	268	.81
gung		2	4.75	1.30	4.75	[4;5.75]	1.00	7.00	160	.72
(4 Items)	2	1	5.53	0.97	5.50	[5.25;6]	3.00	7.00	34	.81
		2	5.45	1.22	5.71	[4.75;6.31]	1.25	7.00	28	.82
	3	1	6.19	1.06	6.50	[5.75;6.88]	1.50	7.00	31	.93
		2	5.70	1.00	5.75	[5.25;6.5]	2.25	7.00	29	.86
	4	1	5.26	1.21	5.50	[4.5;6.12]	2.25	7.00	51	.80
		2	5.13	1.20	5.50	[5.25;5.75]	1.25	6.50	43	.79

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., Z: Zeitpunkt mit 1 = zu Beginn des Semesters, 2 = am Ende des Semesters

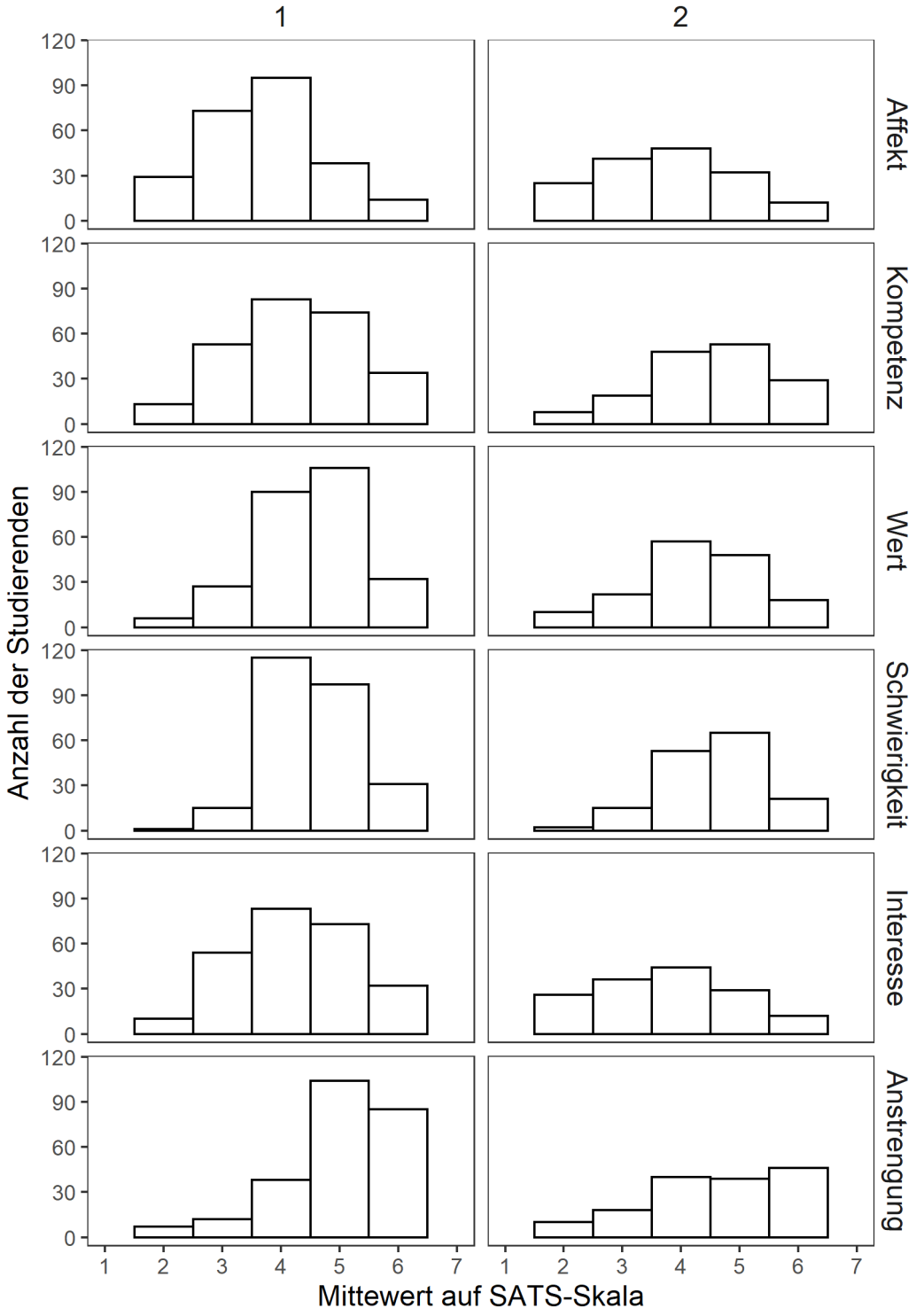


Abbildung 21. Häufigkeitsverteilung der Mittelwerte für die Skalen des Survey of Attutudes Towards Statistics in der Stichprobe 1 (2 SWS, B.A.) zu Beginn (1) und am Ende (2) des Semesters.

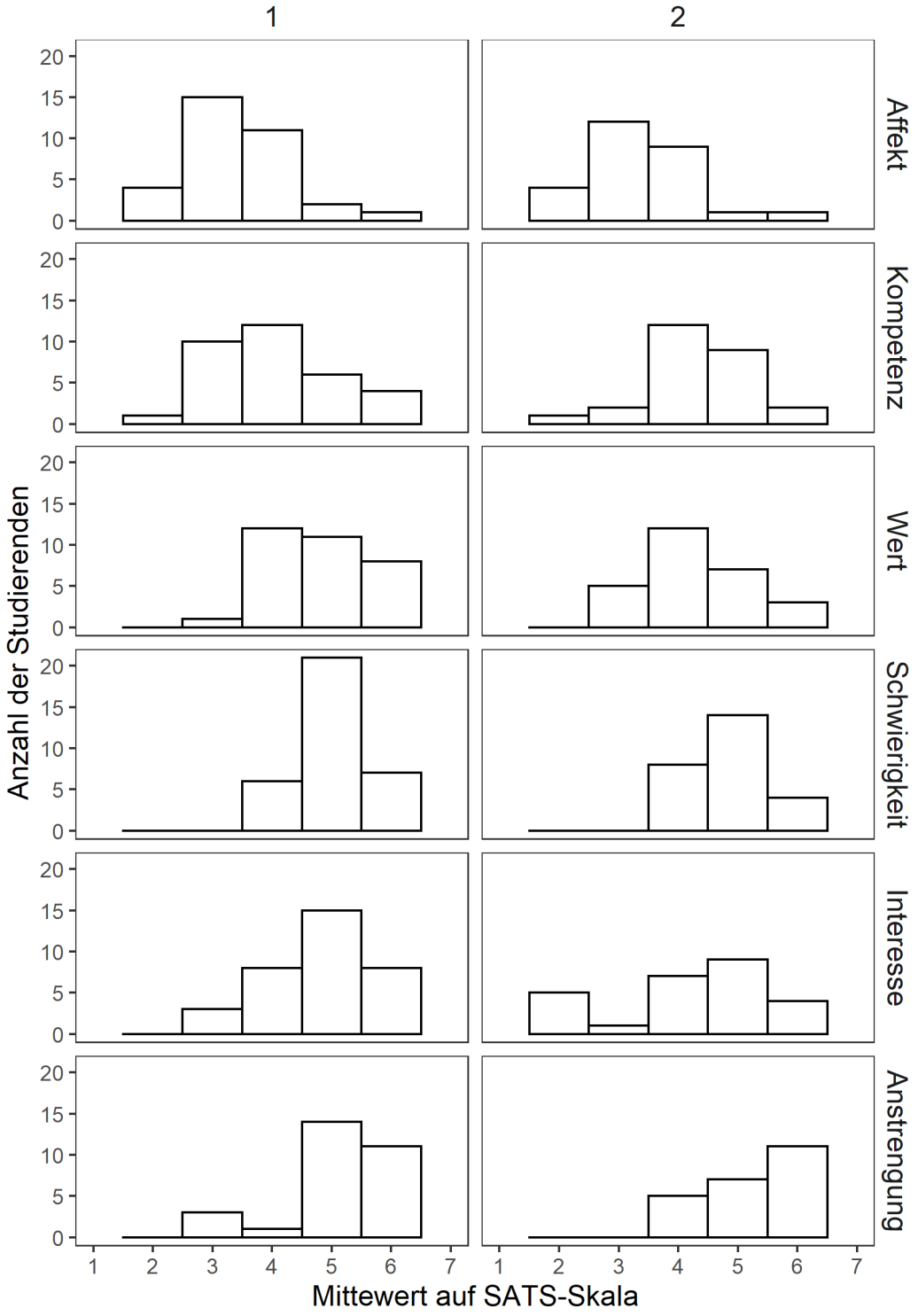


Abbildung 22. Häufigkeitsverteilung der Mittelwerte für die Skalen des Survey of Attutudes Towards Statistics in der Stichprobe 2 (2 SWS, M.A.) zu Beginn (1) und am Ende (2) des Semesters.

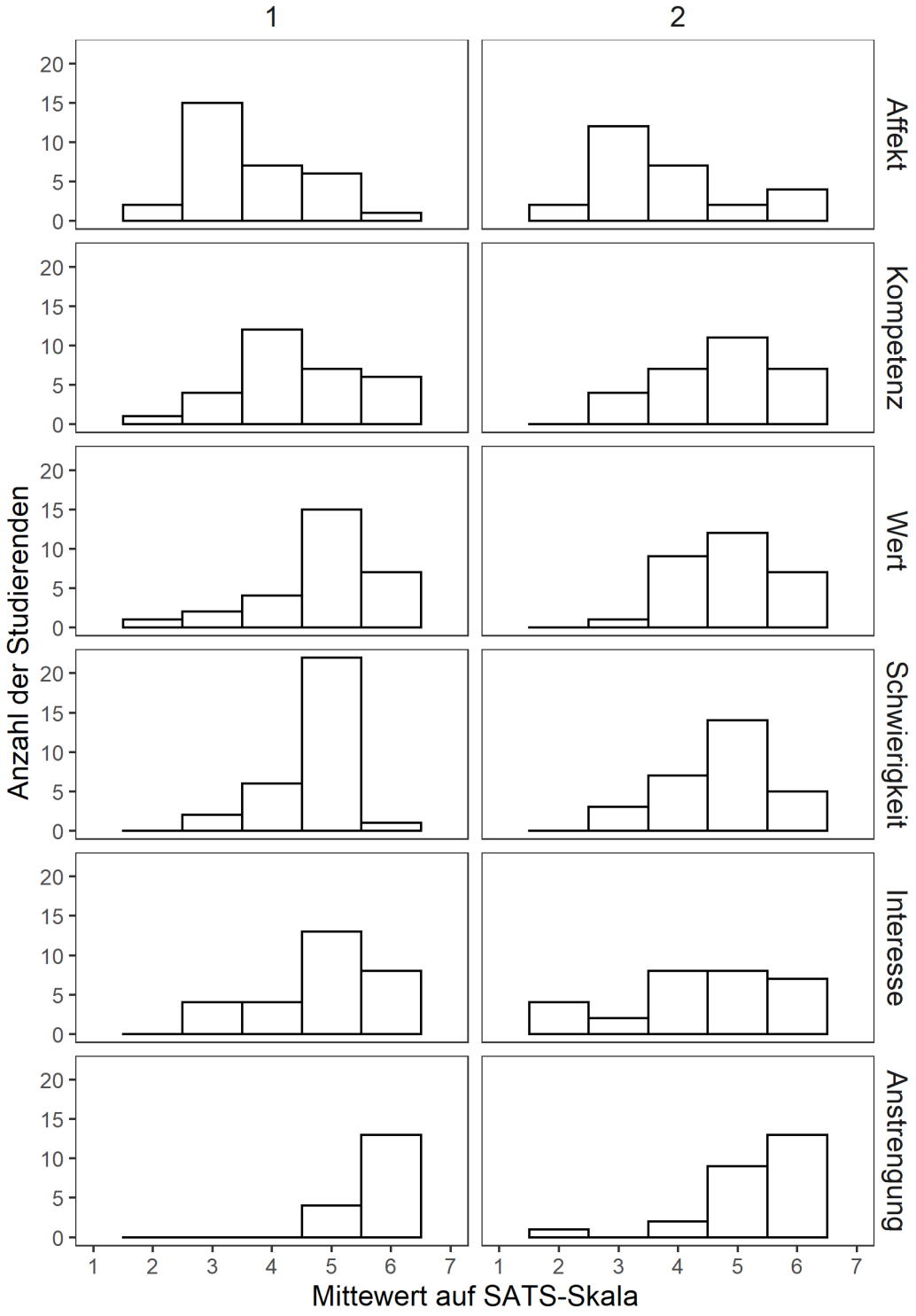


Abbildung 23. Häufigkeitsverteilung der Mittelwerte für die Skalen des Survey of Attutudes Towards Statistics in der Stichprobe 3 (4 SWS, B.Sc.) zu Beginn (1) und am Ende (2) des Semesters.

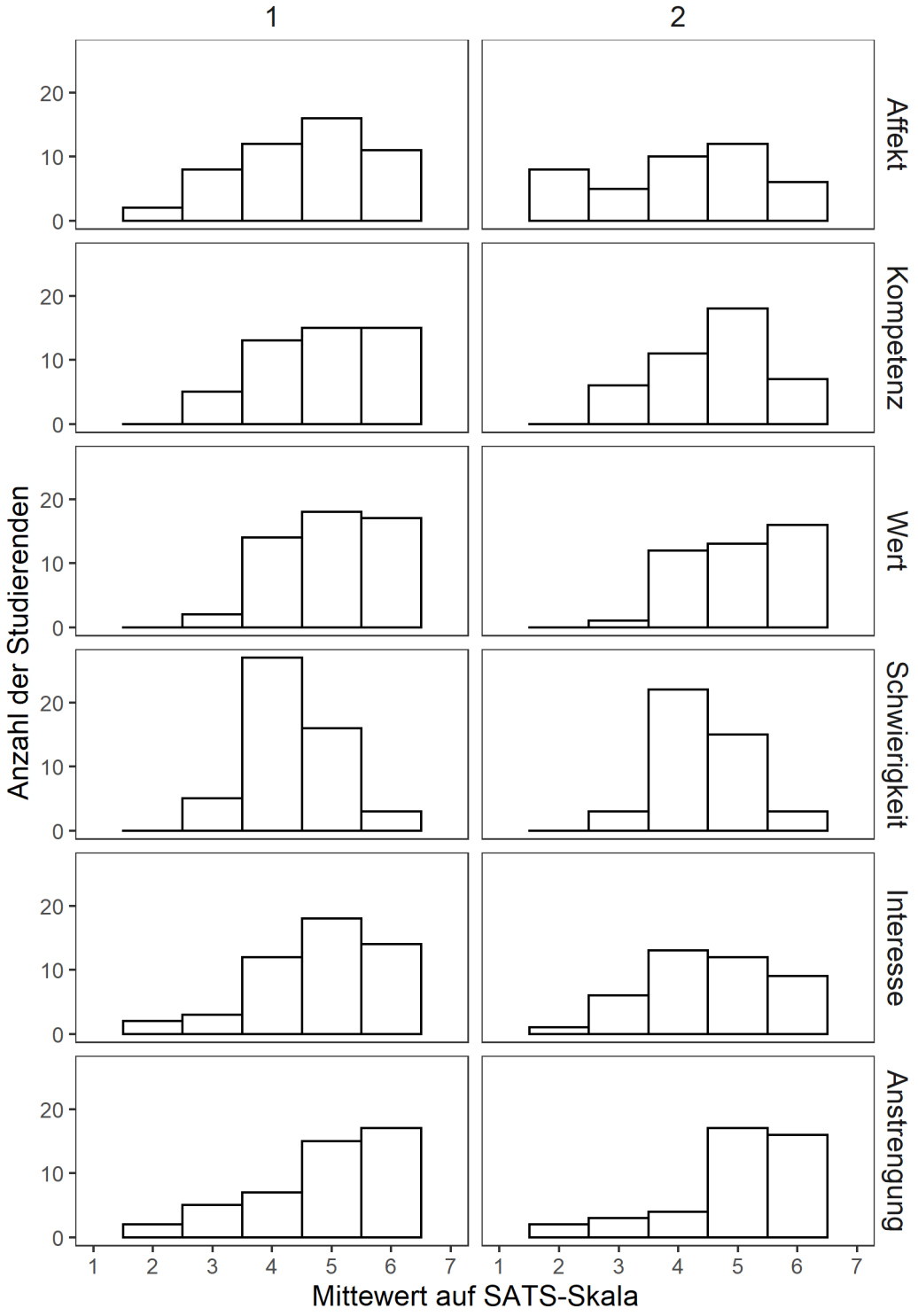


Abbildung 24. Häufigkeitsverteilung der Mittelwerte für die Skalen des Survey of Attutudes Towards Statistics in der Stichprobe 4 (4(+4) SWS, B.Sc.) zu Beginn (1) und am Ende (2) des Semesters.

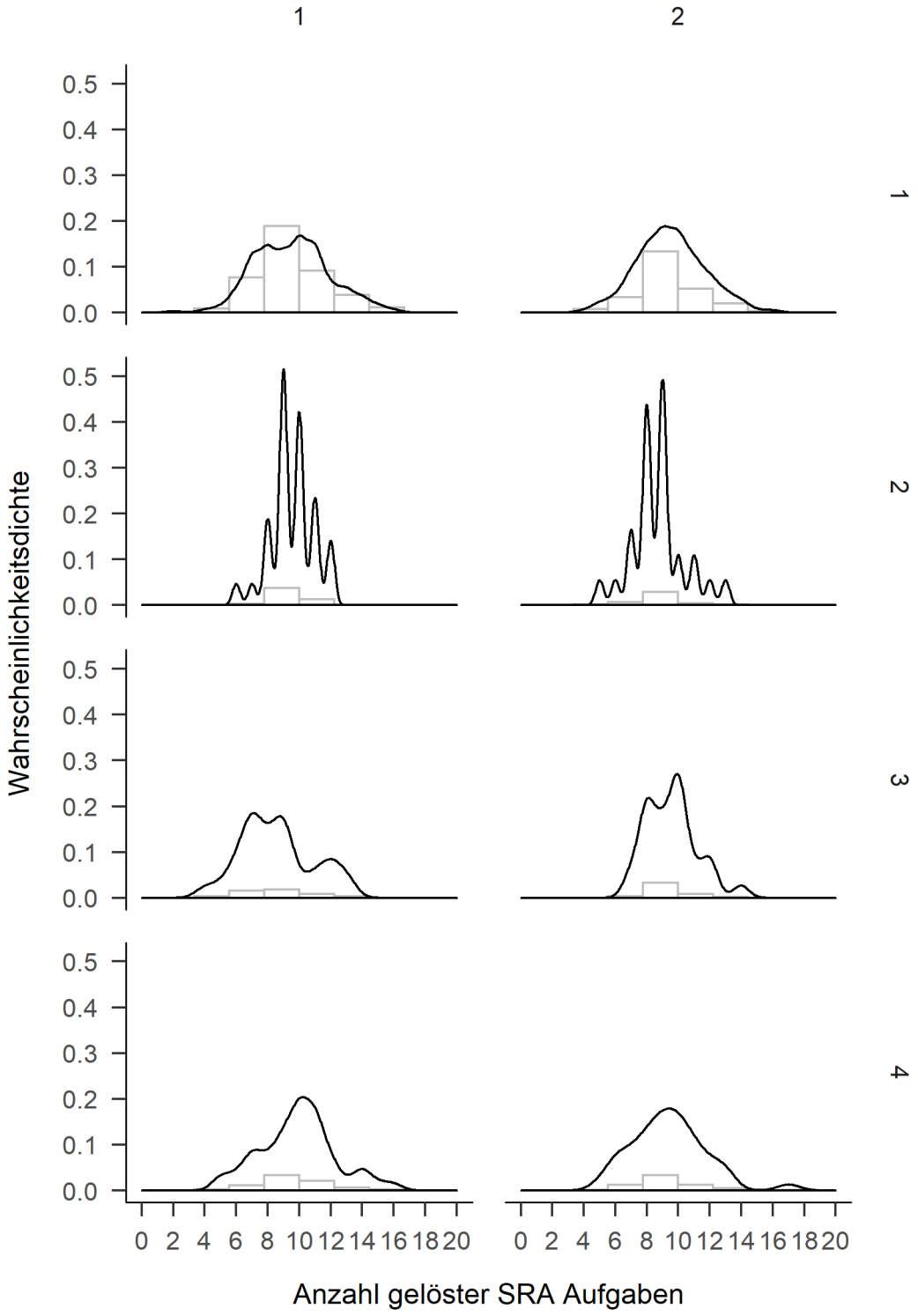


Abbildung 25. Wahrscheinlichkeitsverteilung (geglättetes Histogramm) gelöster SRA Aufgaben in den Stichproben 1 (2 SWS, B.A.), 2 (2 SWS, M.A.), 3 (4 SWS, B.Sc.) und 4 (4(+4) SWS, B.Sc.) zu Beginn (1) und am Ende (2) des Semesters.

SRA: Iteminterkorrelationen

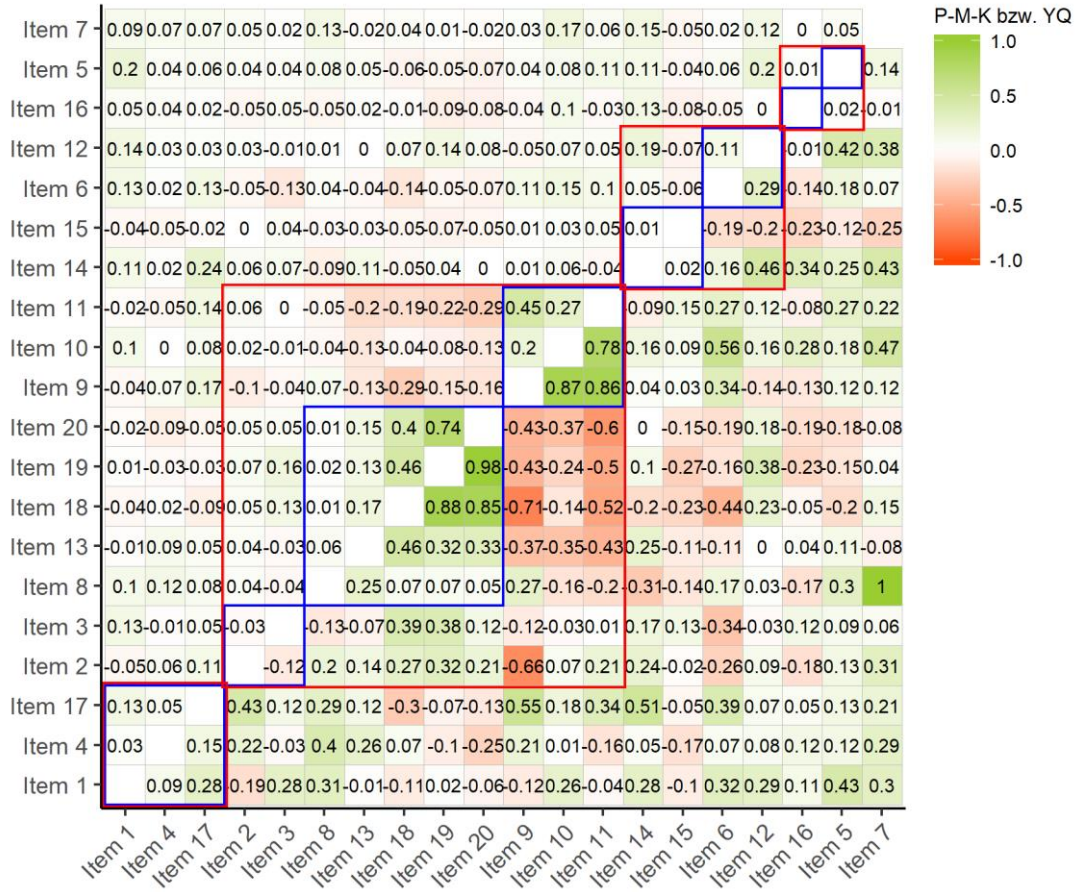


Abbildung 26. Iteminterkorrelationen für Stichprobe 1 (2 SWS, B.A.) mit Produkt-Moment-Korrelationskoeffizienten oberhalb der Diagonalen und Yules Q Koeffizienten unterhalb der Diagonalen zu Beginn des Semesters. Farblich umrandete Felder markieren zusammengehörige Items nach Subskalen (blau) und Typen des statistical reasoning (rot).

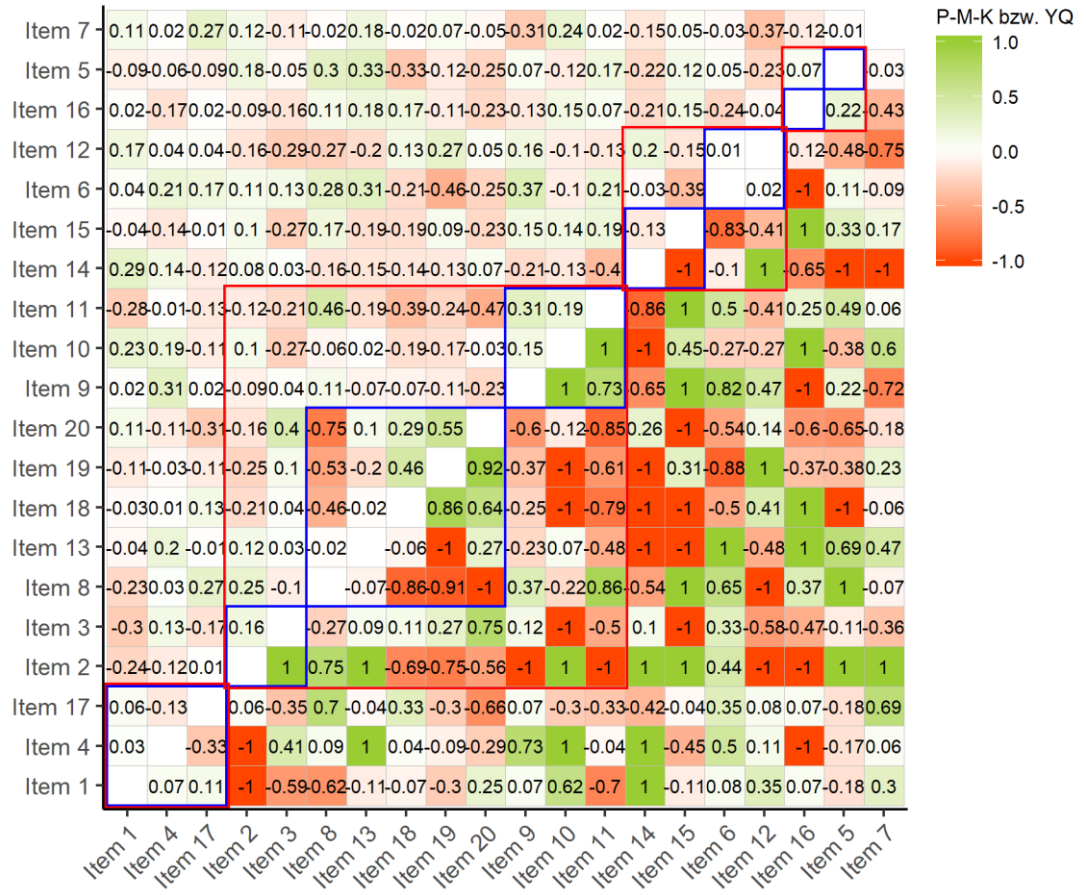


Abbildung 27. Iteminterkorrelationen für Stichprobe 2 (2 SWS, M.A.) mit Produkt-Moment-Korrelationskoeffizienten oberhalb der Diagonalen und Yules Q Koeffizienten unterhalb der Diagonalen zu Beginn des Semesters. Farblich umrandete Felder markieren zusammengehörige Items nach Subskalen (blau) und Typen des statistical reasoning (rot).

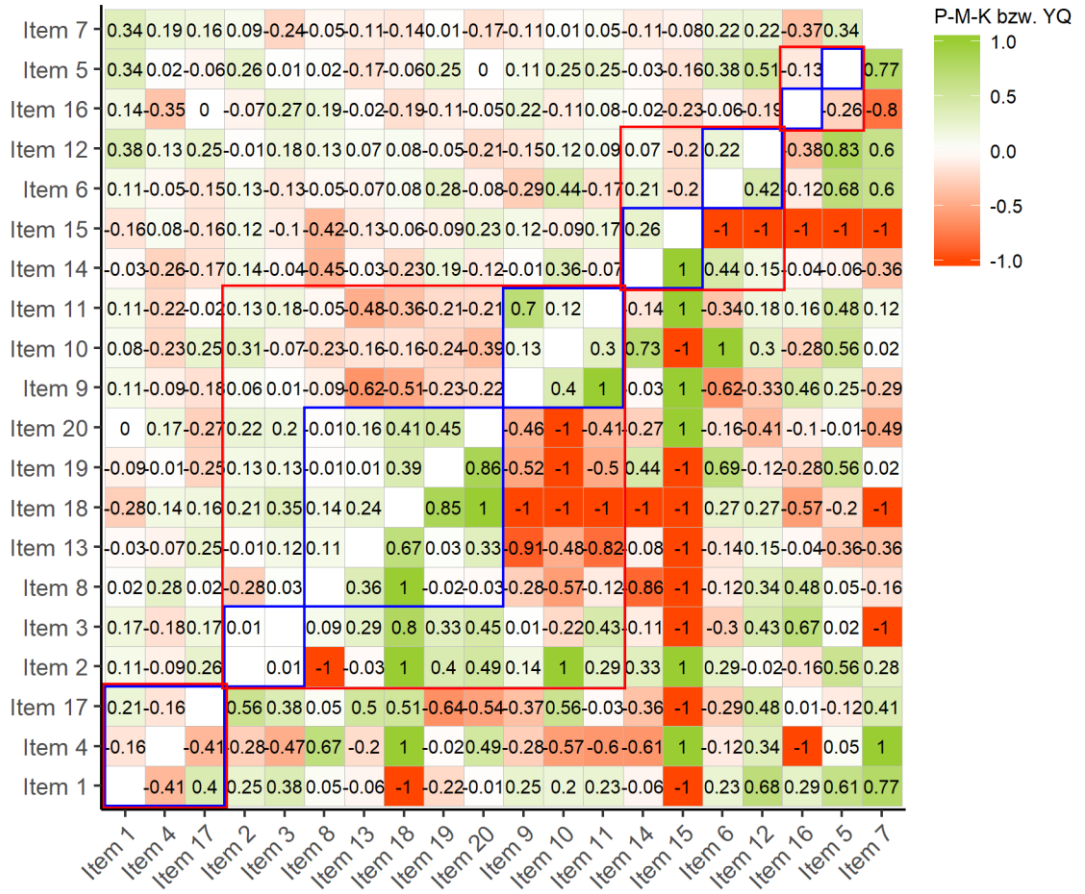


Abbildung 28. Iteminterkorrelationen für Stichprobe 3 (4 SWS, B.Sc.) mit Produkt-Moment-Korrelationskoeffizienten oberhalb der Diagonalen und Yules Q Koeffizienten unterhalb der Diagonalen zu Beginn des Semesters. Farblich umrandete Felder markieren zusammengehörige Items nach Subskalen (blau) und Typen des statistical reasoning (rot).

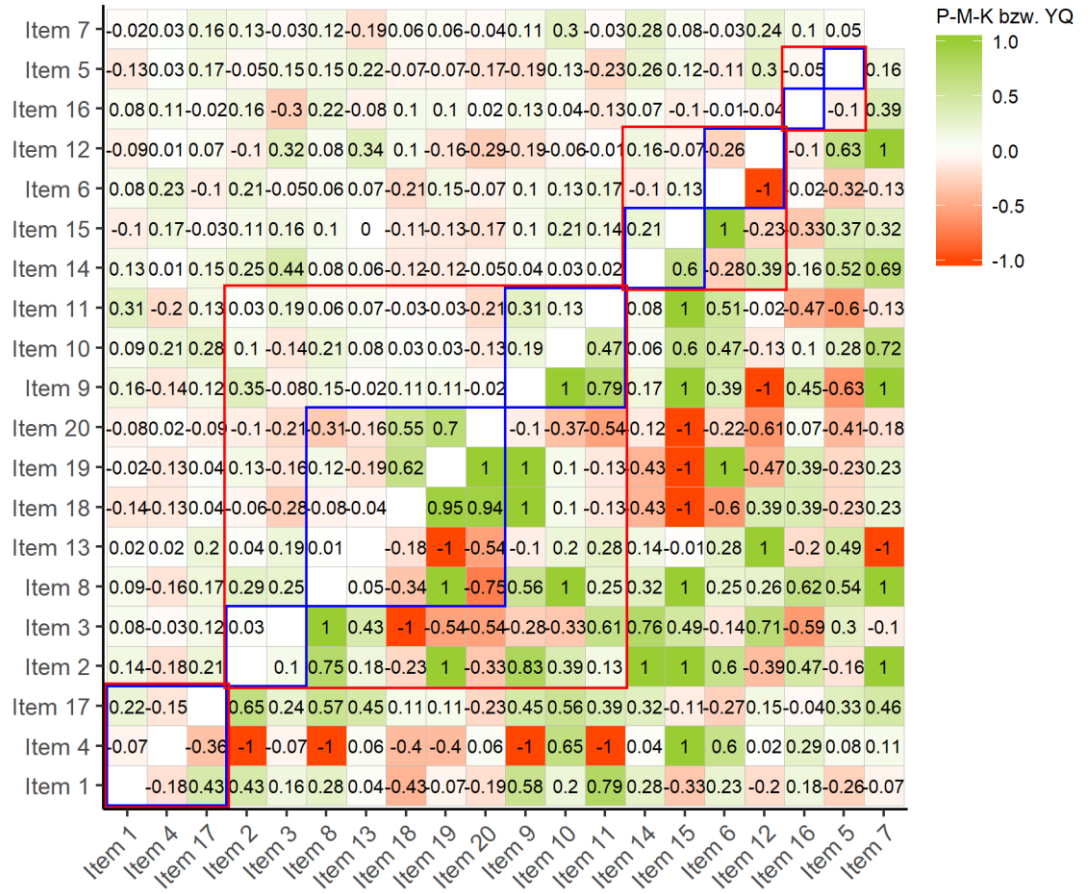


Abbildung 29. Iteminterkorrelationen für Stichprobe 4 (4(+4) SWS, B.Sc.) mit Produkt-Moment-Korrelationskoeffizienten oberhalb der Diagonalen und Yules Q Koeffizienten unterhalb der Diagonalen zu Beginn des (zweiten) Semesters. Farblich umrandete Felder markieren zusammengehörige Items nach Subskalen (blau) und Typen des statistical reasoning (rot).

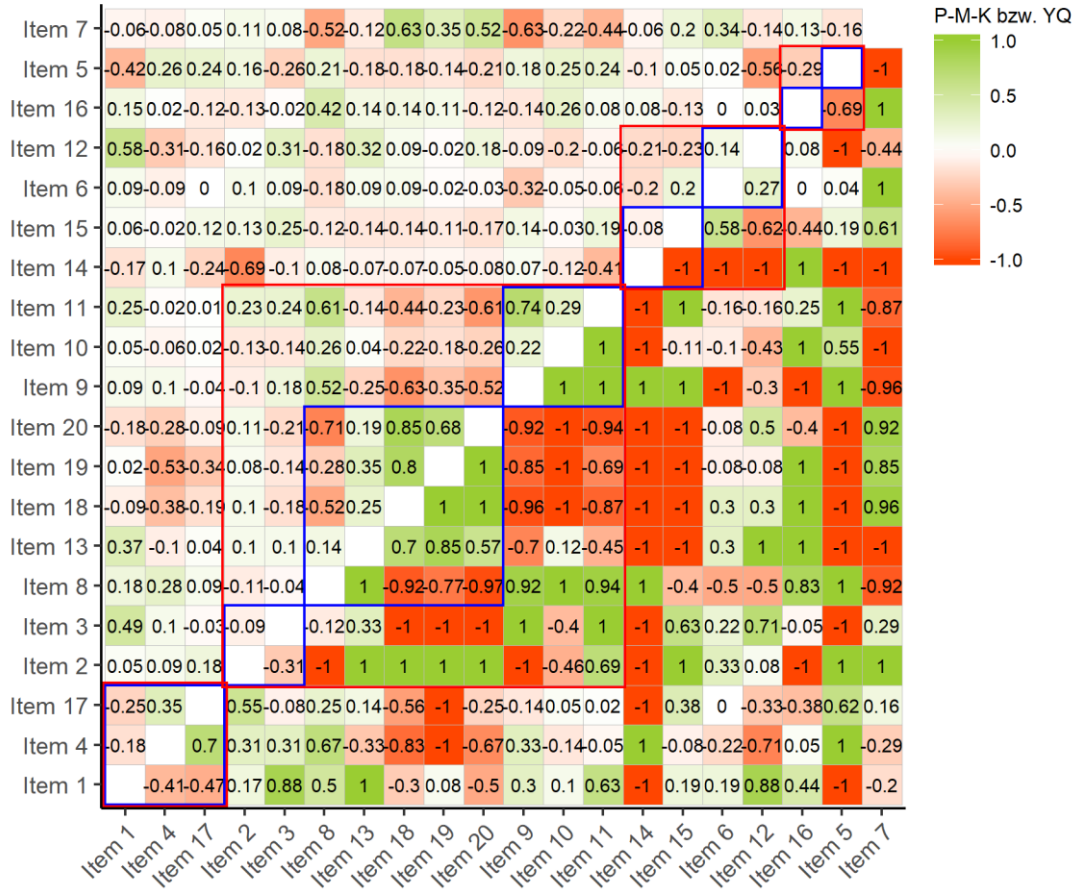


Abbildung 30. Iteminterkorrelationen für Stichprobe 2 (2 SWS, M.A.) mit Produkt-Moment-Korrelationskoeffizienten oberhalb der Diagonalen und Yules Q Koeffizienten unterhalb der Diagonalen am Ende des Semesters. Farblich umrandete Felder markieren zusammengehörige Items nach Subskalen (blau) und Typen des statistical reasoning (rot).

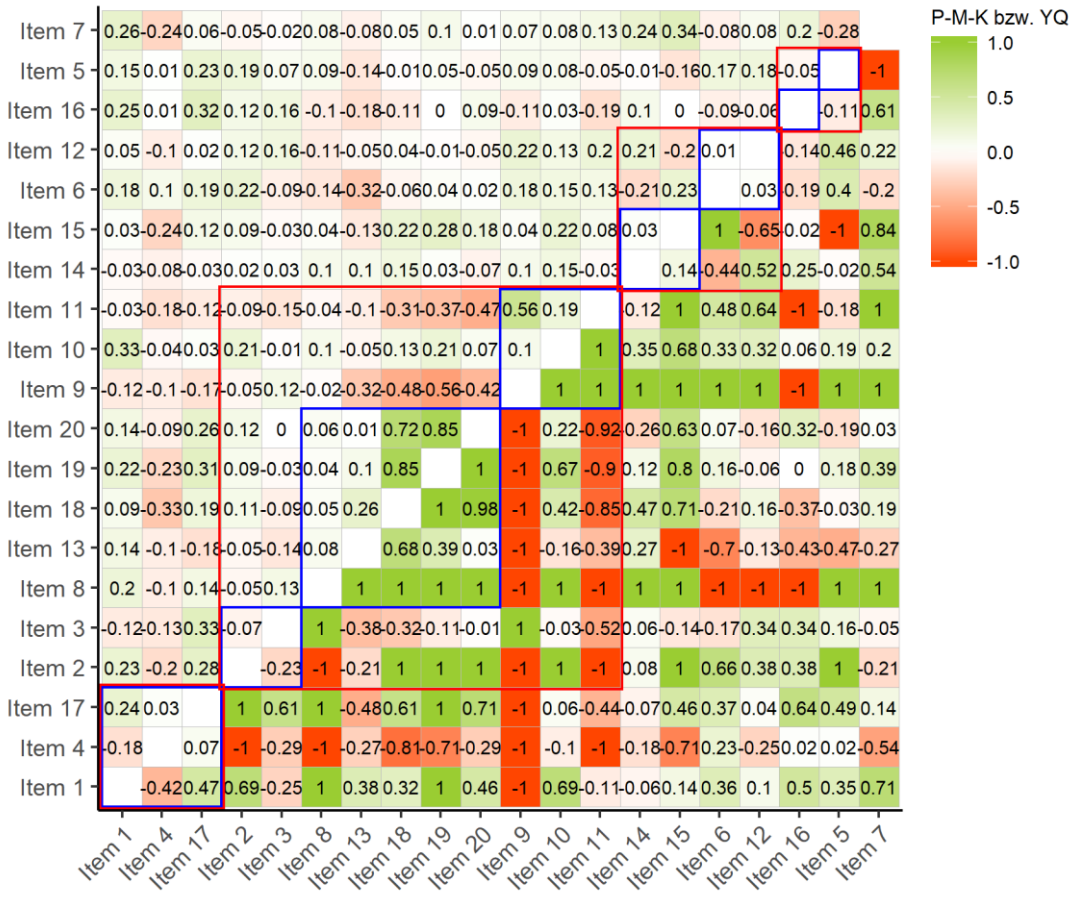


Abbildung 31. Iteminterkorrelationen für Stichprobe 4 (4 SWS, B.Sc.) mit Produkt-Moment-Korrelationskoeffizienten oberhalb der Diagonalen und Yules Q Koeffizienten unterhalb der Diagonalen am Ende des Semesters. Farblich umrandete Felder markieren zusammengehörige Items nach Subskalen (blau) und Typen des statistical reasoning (rot).

SRA: Interne Konsistenz

Tabelle 38

Cronbach α, die höchsten Cronbach α bei Weglassen der einzelnen Items der Skala sowie die durchschnittlichen und der Median der Korrelationen zwischen den Items der Gesamtskala, der Typen sowie der Subskalen des SRA zu Beginn und nach Ende der Intervention (auf nächster Seite fortgesetzt).

		n					*r*	
Skala	S	*Min*	*Max*	*k*	α	α_{max}	*M*	*Md*
zu Beginn								
Gesamt	1	261	267	19	.38	.41	.03	.02
	2	31	34	19	-.80	-.45	-.02	-.03
	3	31	31	19	.27	.34	.01	.00
	4	49	51	19	.44	.49	.04	.04
Kennzahlen	1	261	265	3	.21	.22	.08	.06
	2	34	34	3	-.02	.11	-.01	.03
	3	31	31	3	-.04	.34	-.04	-.16
	4	51	51	3	.05	.36	.00	-.07
Unsicherheit	1	263	267	10	.25	.31	.03	.01
	2	33	34	10	-.53	-.10	-.04	-.06
	3	31	31	10	.05	.23	.01	.01
	4	51	51	10	.32	.41	.06	.03
Stichproben	1	261	266	4	.16	.29	.04	.02
	2	31	34	4	-.40	.13	-.08	-.08
	3	31	31	4	.30	.38	.06	.14
	4	49	51	4	.04	.26	.01	.03
Zusammenhänge	1	263	266	2	.02	.01	.01	.01
	2	33	34	2	.12	.07	.07	.07
	3	31	31	2	-.30	.02	-.13	-.13
	4	51	51	2	-.10	.00	-.05	-.05
zentr. Tendenz	1	261	265	3	.21	.22	.08	.06
	2	34	34	3	-.02	.11	-.01	.03
	3	31	31	3	-.04	.34	-.04	-.16
	4	51	51	3	.05	.36	.00	-.07

Tabelle 38 *(fortgesetzt)*

		n					*r*	
Skala	S	*Min*	*Max*	*k*	α	α_{max}	*M*	*Md*
Interpretieren v.	1	263	267	2	-.06	.00	-.03	-.03
Wahrsch.	2	34	34	2	.23	.16	.16	.16
	3	31	31	2	.01	.01	.01	.01
	4	51	51	2	.05	.03	.03	.03
Berechnen v.	1	265	267	5	.59	.66	.21	.13
Wahrsch.	2	33	34	5	-.28	.50	-.06	-.02
	3	31	31	5	.52	.57	.19	.15
	4	51	51	5	.38	.61	.12	-.02
Unabhängigkeit	1	264	266	3	.57	.61	.31	.28
	2	34	34	3	.45	.47	.22	.19
	3	31	31	3	.60	.82	.32	.13
	4	51	51	3	.40	.46	.21	.19
Variabilität	1	261	265	2	.01	.00	.00	.00
	2	31	34	2	-.30	.02	-.13	-.13
	3	31	31	2	.30	.26	.26	.26
	4	49	51	2	.33	.21	.21	.21
Stichprobengröße	1	265	266	2	.19	.11	.11	.11
	2	34	34	2	.02	.01	.01	.01
	3	31	31	2	.36	.22	.22	.22
	4	51	51	2	-.66	.07	-.26	-.26
am Ende								
Gesamt	1	160	162	19	.24	.30	.01	.00
	2	28	28	19	-.02	.15	-.01	-.02
	3	29	29	18	-.24	.00	-.01	-.03
	4	42	43	19	.42	.50	.03	.03
Kennzahlen	1	161	161	3	-.08	.30	-.04	-.09
	2	28	28	3	-.28	.45	-.07	-.21
	3	29	29	3	-.52	-.06	-.13	-.13
	4	43	43	3	.10	.39	.03	.03

Tabelle 38 *(fortgesetzt)*

		n					*r*	
Skala	S	*Min*	*Max*	*k*	α	α_{max}	*M*	*Md*
Unsicherheit	1	160	162	10	.11	.23	.02	.00
	2	28	28	10	.11	.28	.02	.04
	3	29	29	10	-.05	.13	-.01	-.04
	4	42	43	10	.25	.39	.03	.01
Stichproben	1	161	162	4	.09	.16	.01	.03
	2	28	28	4	-.08	.11	-.07	-.14
	3	29	29	3	.08	.44	.05	.16
	4	43	43	4	.04	.14	.01	.02
Zusammenhänge	1	160	160	2	.04	.02	.02	.02
	2	28	28	2	-.78	.08	-.28	-.28
	3	29	29	2	.02	.01	.01	.01
	4	43	43	2	-.10	.00	-.05	-.05
zentr. Tendenz	1	161	161	3	-.08	.30	-.04	-.09
	2	28	28	3	-.28	.45	-.07	-.21
	3	29	29	3	-.52	-.06	-.13	-.13
	4	43	43	3	.10	.39	.03	.03
Interpretieren v.	1	161	161	2	.02	.01	.01	.01
Wahrsch.	2	28	28	2	.23	.14	.14	.14
	3	29	29	2	.09	.06	.06	.06
	4	43	43	2	-.13	.00	-.07	-.07
Berechnen v.	1	160	162	5	.64	.73	.24	.16
Wahrsch.	2	28	28	5	.44	.80	.18	.22
	3	29	29	5	.52	.71	.09	-.05
	4	42	43	5	.67	.79	.29	.09
Unabhängigkeit	1	160	162	3	.56	.69	.33	.22
	2	28	28	3	.65	.84	.42	.29
	3	29	29	3	.46	.49	.26	.26
	4	42	43	3	.38	.66	.29	.19
Variabilität	1	161	162	2	-.04	.00	-.03	-.03
	2	28	28	2	-.14	.01	-.08	-.08
	3 [a]							
	4	43	43	2	.06	.03	.03	.03

Tabelle 38 *(fortgesetzt)*

Skala	S	*n*		*k*	α	α_{max}	*r*	
		Min	*Max*				*M*	*Md*
Stichprobengröße	1	162	162	2	.03	.01	.01	.01
	2	28	28	2	.24	.14	.14	.14
	3	29	29	2	.44	.29	.29	.29
	4	43	43	2	.03	.01	.01	.01

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., Typen des statistical thinking mit Kennzahlen: „Beurteilen statistischer Kennzahlen“, Unsicherheit: „Schließen unter Unsicherheit“, Stichproben: „Beurteilen (der Güte) der Stichproben“, Zusammenhänge: „Beurteilen von Zusammenhängen“, Subskalen des statistical reasoning mit zentrale Tendez: „Verstehen, wie ein angemessenes Maß der zentralen Tendenz ausgewählt wird“, Berechnen v. Wahrsch.: „Wahrscheinlichkeiten korrekt berechnen“, Interpretieren v. Wahrsch.: „Wahrscheinlichkeiten korrekt interpretieren“, Stichprobengröße: „Bedeutung großer Stichproben verstehen“, Unabhängigkeit: „(stochastische) Unabhängigkeit verstehen“, Variabilität: „Variablilität innerhalb von Stichproben verstehen“, die Subskalen „zwischen Korrelation und Kausalität unterscheiden“ und „Vierfeldertafeln korrekt interpretieren“ enthalten jeweils nur ein Item, *n*: kleinstes (*Min*) und größtes (*Max*) für die paarweisen Korrelationen, *k* = Anzahl der Items, α: Cronbachs α, α_{max}: höchstes Cronbach α beim Weglassen einzelner Items, *r*: Mittelwert (*M*) und Median (*Md*) der Iteminterkorrelationen der gesamten Skala.

[a] Item 15 wurde in dieser Stichprobe von keiner Person gelöst und wurde daher aus den Analysen ausgeschlossen.

SRA: Faktorenanzahl

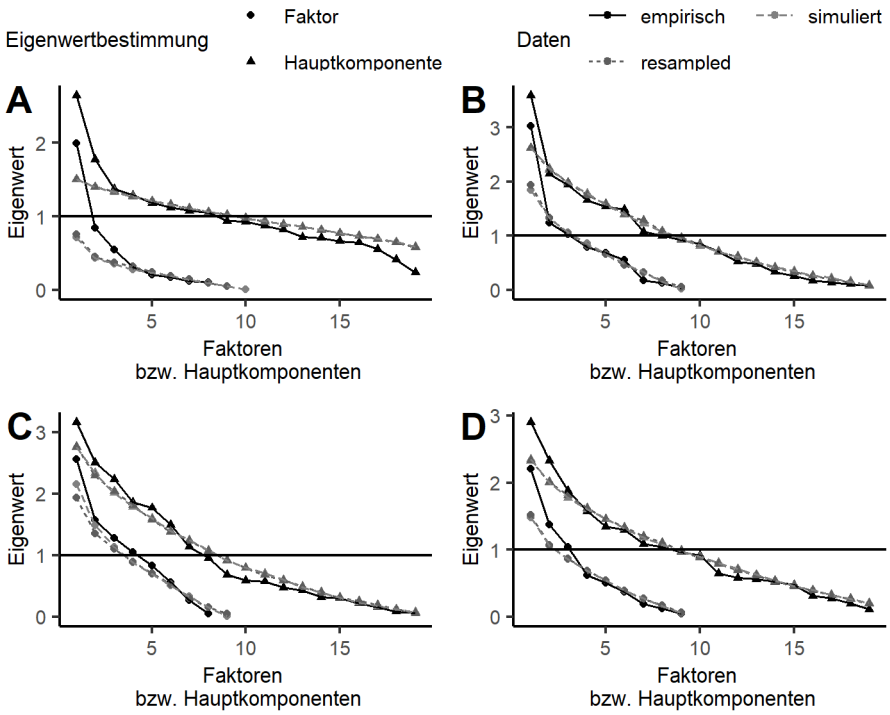


Abbildung 32. Screeplots und Ergebnisse der Parallelanalysen anhand der Produkt-Moment-Korrelationsmatrizen der Stichproben zu Beginn des Semesters, A. Stichprobe 1 (2 SWS, B.A.), B. Stichprobe 2 (2 SWS, M.A.), C. Stichprobe 3 (4 SWS, B.Sc.), D. Stichprobe 4 (4(+4) SWS, B.Sc.)

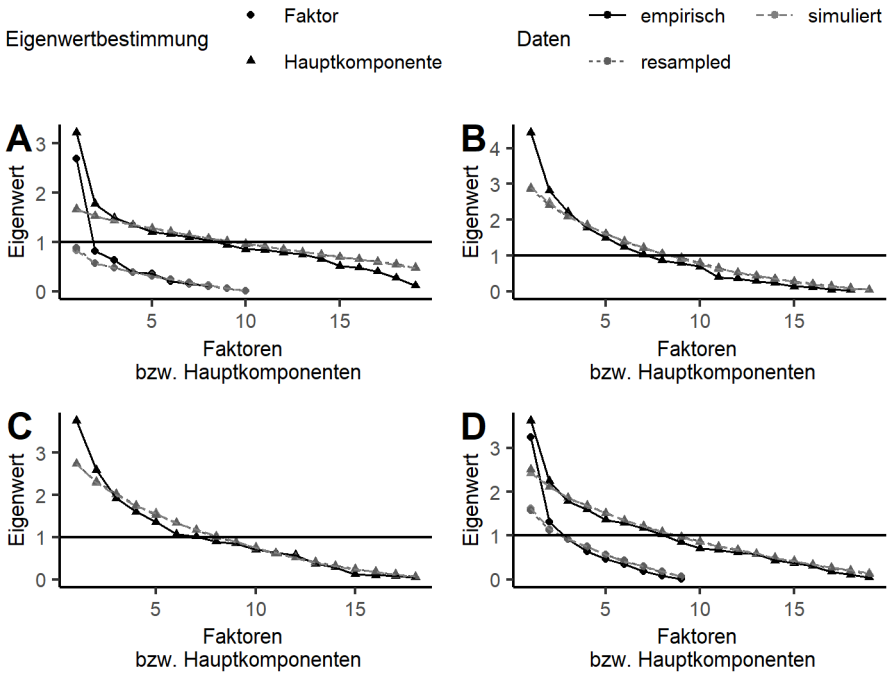


Abbildung 33. Screeplots und Ergebnisse der Parallelanalysen anhand der Produkt-Moment-Korrelationsmatrizen der Stichproben am Ende des Semesters, A. Stichprobe 1 (2 SWS, B.A.), B. Stichprobe 2 (2 SWS, M.A.), C. Stichprobe 3 (4 SWS, B.Sc.), D. Stichprobe 4 (4(+4) SWS, B.Sc.)

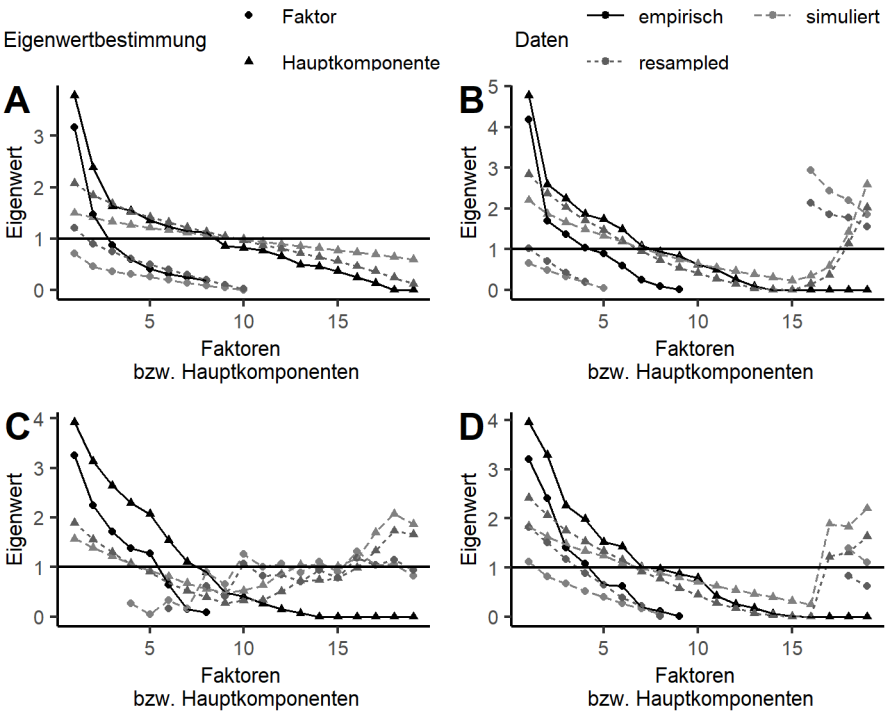


Abbildung 34. Screeplots und Ergebnisse der Parallelanalysen anhand der tetrachorischen Korrelationsmatrizen der Stichproben zu Beginn des Semesters, A. Stichprobe 1 (2 SWS, B.A.), B. Stichprobe 2 (2 SWS, M.A.), C. Stichprobe 3 (4 SWS, B.Sc.), D. Stichprobe 4 (4(+4) SWS, B.Sc.)

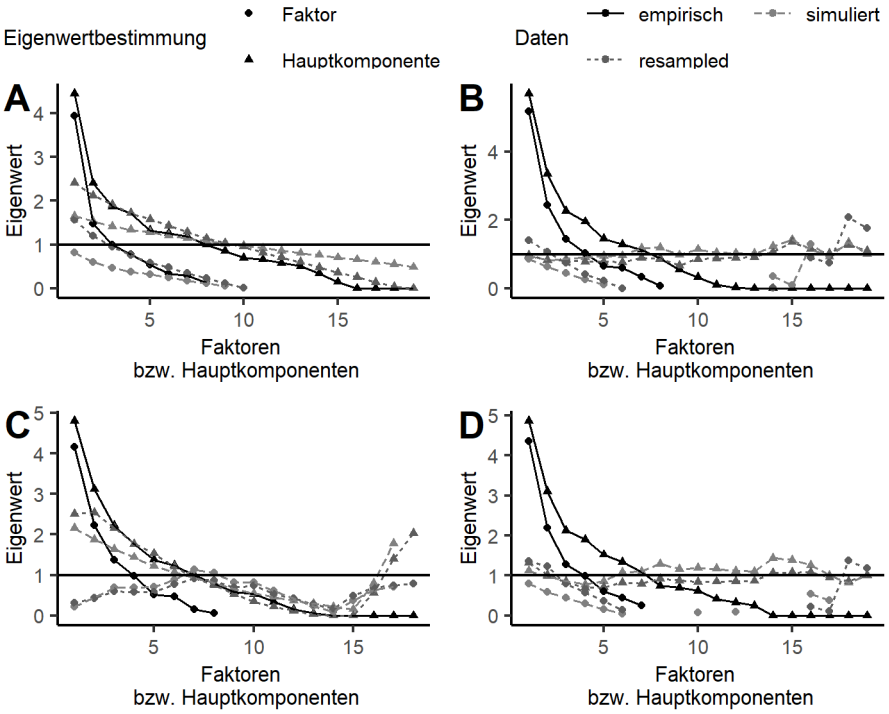


Abbildung 35. Screeplots und Ergebnisse der Parallelanalysen anhand der tetrachorischen Korrelationsmatrizen der Stichproben am Ende des Semesters, A. Stichprobe 1 (2 SWS, B.A.), B. Stichprobe 2 (2 SWS, M.A.), C. Stichprobe 3 (4 SWS, B.Sc.), D. Stichprobe 4 (4(+4) SWS, B.Sc.)

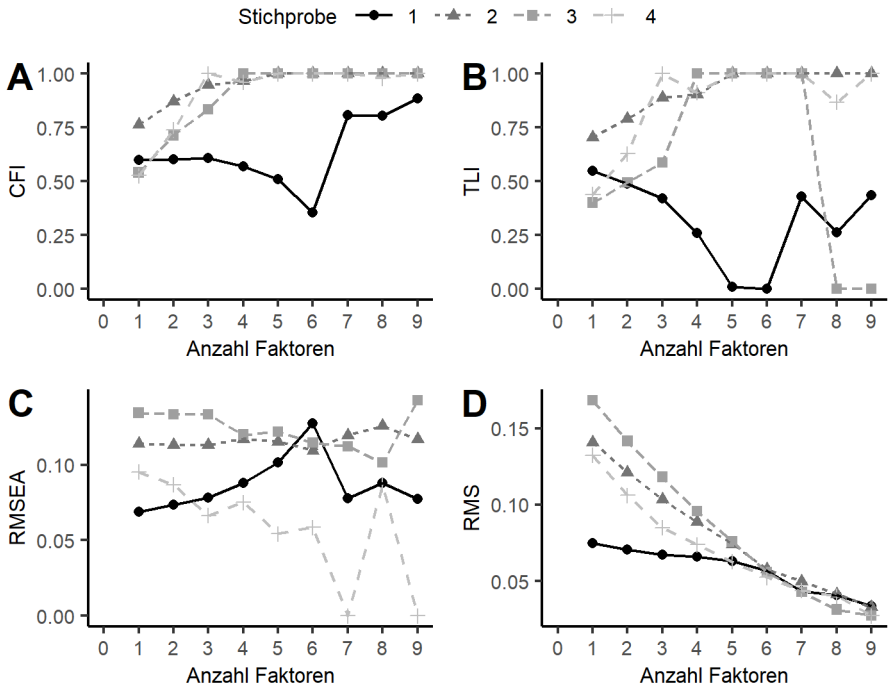


Abbildung 36. Güteindizes der Faktorenanalysen anhand der Produkt-Moment-Korrelationsmatrizen der Stichproben zu Beginn des Semesters, A. CFI, B. TLI, C. RMSEA, D. RMS.

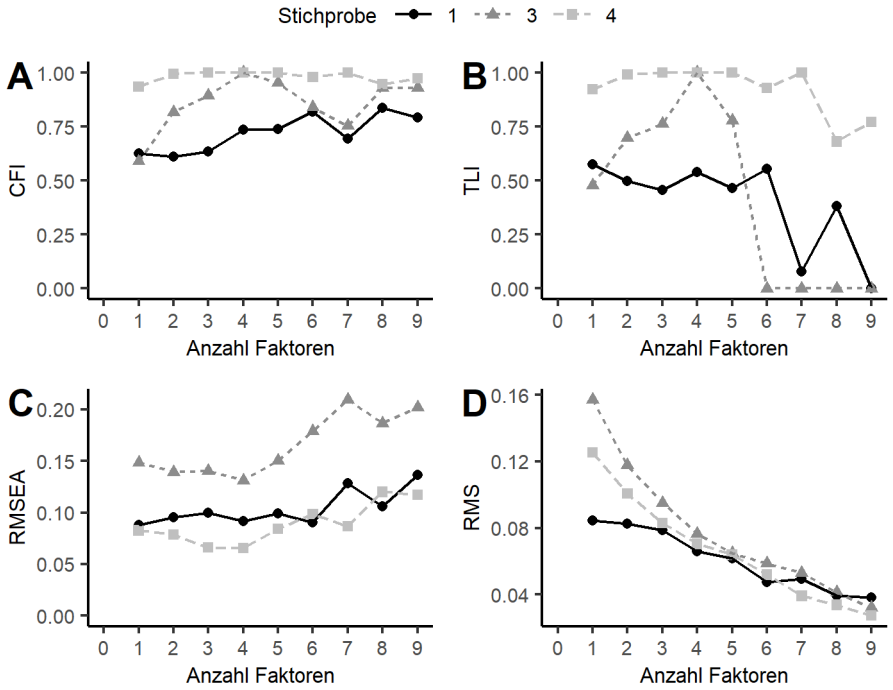


Abbildung 37. Güteindizes der Faktorenanalysen anhand der Produkt-Moment-Korrelationsmatrizen der Stichproben am Ende des Semesters, A. CFI, B. TLI, C. RMSEA, D. RMS.

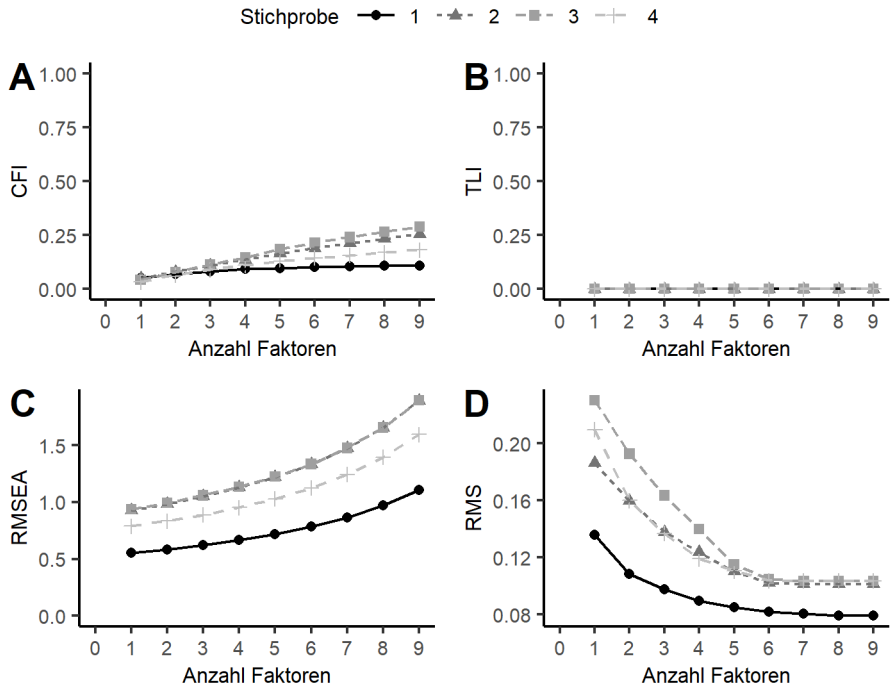


Abbildung 38. Güteindizes der Faktorenanalysen anhand der tetrachorischen Korrelationsmatrizen der Stichproben zu Beginn des Semesters, A. CFI, B. TLI, C. RMSEA, D. RMS.

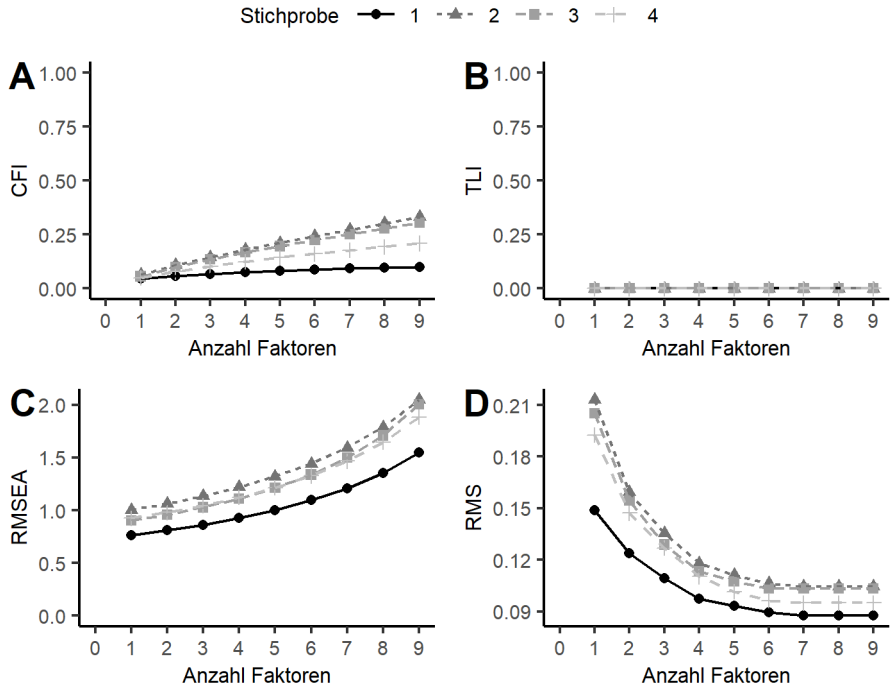


Abbildung 39. Güteindizes der Faktorenanalysen anhand der tetrachorischen Korrelationsmatrizen der Stichproben am Ende des Semesters, A. CFI, B. TLI, C. RMSEA, D. RMS.

Tabelle 39

Anzahl ermittelter Komponenten bzw. Faktoren anhand der Analysen der Produkt-Moment-Korrelationsmatrix.

		PC			F							
	n	Sc	K-G	P	Sc	K-G	P	p_{χ^2}	CFI	TLI	RMSEA	RMS
Z 1												
S 1	277	8	8	2	3	1	3	/	/	/	/	7
			8						**(3)**			
S 2	34	3;6	7	1	1;3	3	1	1	3;5	3;5	/	7
			(1)						**(3)**			
S 3	31	3;5	7	3	4	4	2	1;2;4	4	4	/	8
			3						**(4)**			
S 4	51	6	8	2	3	3	3	3	3	3	7	7
			(8)						**3**			
Z 2												
S 1	163	4	8	2	3	1	3	/	/	/	/	6
			4;8						**(3)**			
S 2	32	4	6	3	/	/	/	/	/	/	/	/
			(4)						**/**			
S 3[a]	29	5	7	2	/	/	/	2;4	4	4	/	7
			(2)						**(4)**			
S 4	43	4	8	1	3	2	1	1;3	2	2	/	7
			(1)						**(2)**			

Anmerkungen. Z 1: zu Beginn des Semesters, Z 2: am Ende des Semesters, S 1 = 2 SWS, B.A., S 2 = 2 SWS, M.A., S 3 = 4 SWS, B.Sc., S 4 = 4(+4) SWS, B.Sc., PC: Hauptkomponenten, F: Faktoren, Sc: Screeplot-Kriterium, K-G: Kaiser-Guttman-Kriterium, P: Ergebnis der Parallelanalyse, p_{χ^2}: p-Wert der χ^2-Statistik, CFI: Comparative-Fit-Index, TLI: Tucker-Lewis-Non-Normed-Index, RMSEA: Root Mean Square Error of Approximation, RMS: Summe der Residuenquadrate im Verhältnis zu den Freiheitsgraden, bei deutlichen Mängeln und fehlender Differenzierung in den Kriterien ist keine Anzahl der Faktoren eingetragen (/), die Anzahl der ermittelten Komponenten bzw. Faktoren anhand der Kriterien für die Stichprobe insgesamt ist **fett** gesetzt, Werte in Klammern zeigen uneindeutige Anzahlen und/oder deutliche Mängel in einigen der Kriterien an.

[a] Item 15 wurde in dieser Stichprobe von keiner Person gelöst und wurde daher aus den Analysen ausgeschlossen.

Tabelle 40

Anzahl ermittelter Komponenten bzw. Faktoren anhand der Analysen der tetrachorischen Korrelationsmatrix.

		PC			F							
	n	Sc	K-G	P	Sc	K-G	P	p_{χ^2}	CFI	TLI	RMSEA	RMS
Z 1												
S 1	277	4	8	8	8	2	8	/	/	/	/	/
			8						**(8)**			
S 2	34	3;5	7	6	1;5	4	7	/	/	/	/	/
			(3)						**(4)**			
S 3	31	5	7	7	5	5	3	/	/	/	/	/
			7						**(5)**			
S 4	51	2;4	7	4	2;6	4	6	/	/	/	/	/
			4						**(6)**			
Z 2												
S 1	163	4	7	5	4	2	7	/	/	/	/	(7)
			(4)						**(7)**			
S 2	32	4	7	4	4	4	1	/	/	/	/	/
			4						**(4)**			
S 3[a]	29	6	6	2	1;4	3	1	/	/	/	/	/
			6						**(1)**			
S 4	43	4	7	7	4	3	3	/	/	/	/	/
			7						**(3)**			

Anmerkungen. Z 1: zu Beginn des Semesters, Z 2: am Ende des Semesters, S 1 = 2 SWS, B.A., S 2 = 2 SWS, M.A., S 3 = 4 SWS, B.Sc., S 4 = 4(+4) SWS, B.Sc., PC: Hauptkomponenten, F: Faktoren, Sc: Screeplot-Kriterium, K-G: Kaiser-Guttman-Kriterium, P: Ergebnis der Parallelanalyse, p_{χ^2}: p-Wert der χ^2-Statistik, CFI: Comparative-Fit-Index, TLI: Tucker-Lewis-Non-Normed-Index, RMSEA: Root Mean Square Error of Approximation, RMS: Summe der Residuenquadrate im Verhältnis zu den Freiheitsgraden, bei deutlichen Mängeln und fehlender Differenzierung in den Kriterien ist keine Anzahl der Faktoren eingetragen (/), die Anzahl der ermittelten Komponenten bzw. Faktoren anhand der Kriterien für die Stichprobe insgesamt ist **fett** gesetzt, Werte in Klammern zeigen uneindeutige Anzahlen und/oder deutliche Mängel in einigen der Kriterien an.

[a] Item 15 wurde in dieser Stichprobe von keiner Person gelöst und wurde daher aus den Analysen ausgeschlossen.

Tabelle 41

Güte der explorativen Faktorenanalysen mit WLS-Schätzung auf der Basis der Produkt-Moment-Korrelationsmatrizen, der tetrachorischen Korrelationsmatrizen zu beiden Messzeitpunkten für alle Stichproben mit jeweils 1 bis 9 Faktoren (auf nächsten Seiten fortgesetzt).

S	n	AF	χ^2	df	p	TLI	CFI	RMSEA	RMS
Beginn des Semesters									
Produkt-Moment-Korrelationsmatrizen									
1	267	1	334.56	152	<.001	0.55	0.6	0.07	0.07
	267	2	315.22	134	<.001	0.49	0.6	0.07	0.07
	267	3	296.28	117	<.001	0.42	0.61	0.08	0.07
	267	4	297.2	101	<.001	0.26	0.57	0.09	0.07
	267	5	309.03	86	<.001	0.01	0.51	0.1	0.06
	267	6	366.11	72	<.001	0	0.35	0.13	0.06
	267	7	146.64	59	<.001	0.43	0.81	0.08	0.04
	267	8	136.74	47	<.001	0.26	0.8	0.09	0.04
	267	9	88.44	36	<.001	0.44	0.88	0.08	0.03
2	34	1	165.62	152	.213	0.7	0.76	0.11	0.14
	34	2	141.53	134	.311	0.79	0.87	0.11	0.12
	34	3	120.06	117	.405	0.89	0.95	0.11	0.1
	34	4	102.98	101	.427	0.9	0.97	0.12	0.09
	34	5	84.38	86	.529	1	1	0.12	0.07
	34	6	66.44	72	.663	1	1	0.11	0.06
	34	7	55.71	59	.598	1	1	0.12	0.05
	34	8	44.45	47	.579	1	1	0.13	0.04
	34	9	31.4	36	.687	1	1	0.12	0.03
3	31	1	173.13	152	.115	0.4	0.54	0.13	0.17
	31	2	147.17	134	.206	0.5	0.71	0.13	0.14
	31	3	124.63	117	.297	0.59	0.83	0.13	0.12
	31	4	97.2	101	.588	1	1	0.12	0.1
	31	5	80.8	86	.638	1	1	0.12	0.08
	31	6	63.06	72	.765	1	1	0.11	0.06
	31	7	49.19	59	.815	1	1	0.11	0.04
	31	8	35.9	47	.881	0	1	0.1	0.03
	31	9	32.53	36	.635	0	1	0.14	0.03

Tabelle 41 *(fortgesetzt)*

S	n	AF	$\chi 2$	df	p	TLI	CFI	RMSEA	RMS
Beginn des Semesters									
Produkt-Moment-Korrelationsmatrizen									
4	51	1	186.4	152	.03	0.44	0.53	0.1	0.13
	51	2	153.01	134	.125	0.63	0.74	0.09	0.11
	51	3	116.58	117	.494	1	1	0.07	0.09
	51	4	103.99	101	.399	0.91	0.96	0.08	0.07
	51	5	77.94	86	.72	1	1	0.05	0.06
	51	6	65.52	72	.692	1	1	0.06	0.05
	51	7	40.88	59	.965	1	1	0	0.04
	51	8	48.54	47	.411	0.87	0.98	0.09	0.04
	51	9	22.68	36	.959	1	1	0	0.03
Ende des Semesters									
Produkt-Moment-Korrelationsmatrizen									
1	162	1	325.58	152	<.001	0.58	0.62	0.09	0.08
	162	2	313.77	134	<.001	0.5	0.61	0.1	0.08
	162	3	286.21	117	<.001	0.46	0.63	0.1	0.08
	162	4	223.57	101	<.001	0.54	0.74	0.09	0.07
	162	5	206.97	86	<.001	0.46	0.74	0.1	0.06
	162	6	155.74	72	<.001	0.55	0.82	0.09	0.05
	162	7	200.13	59	<.001	0.08	0.69	0.13	0.05
	162	8	121.87	47	<.001	0.38	0.84	0.11	0.04
	162	9	132.25	36	<.001	0	0.79	0.14	0.04
3	29	1	159.79	135	.071	0.48	0.59	0.15	0.16
	29	2	129.02	118	.23	0.7	0.82	0.14	0.12
	29	3	108.4	102	.314	0.76	0.89	0.14	0.1
	29	4	85.22	87	.534	1	1	0.13	0.08
	29	5	75.87	73	.386	0.78	0.95	0.15	0.06
	29	6	69.73	60	.183	0	0.84	0.18	0.06
	29	7	62.94	48	.073	0	0.75	0.21	0.05
	29	8	41.33	37	.287	0	0.93	0.19	0.04
	29	9	31.28	27	.26	0	0.93	0.2	0.03

Tabelle 41 *(fortgesetzt)*

S	n	AF	χ^2	df	p	TLI	CFI	RMSEA	RMS
Ende des Semesters									
Produkt-Moment-Korrelationsmatrizen									
4	43	1	159.14	152	.329	0.92	0.94	0.08	0.13
	43	2	134.61	134	.469	0.99	0.99	0.08	0.1
	43	3	108.23	117	.707	1	1	0.07	0.08
	43	4	91.37	101	.743	1	1	0.07	0.07
	43	5	83.67	86	.551	1	1	0.08	0.06
	43	6	74.38	72	.401	0.93	0.98	0.1	0.05
	43	7	55.79	59	.595	1	1	0.09	0.04
	43	8	52.96	47	.255	0.68	0.95	0.12	0.03
	43	9	39	36	.336	0.77	0.97	0.12	0.03
Beginn des Semesters									
Tetrachorische Korrelationsmatrizen									
1	267	1	12078.1	152	<.001	0	0.05	0.55	0.14
	267	2	11826.86	134	<.001	0	0.07	0.58	0.11
	267	3	11639.61	117	<.001	0	0.08	0.62	0.1
	267	4	11506.07	101	<.001	0	0.09	0.66	0.09
	267	5	11429.37	86	<.001	0	0.1	0.72	0.08
	267	6	11361.12	72	<.001	0	0.1	0.78	0.08
	267	7	11311.42	59	<.001	0	0.1	0.87	0.08
	267	8	11255.62	47	<.001	0	0.11	0.97	0.08
	267	9	11226.02	36	<.001	0	0.11	1.11	0.08
2	34	1	3404.27	152	<.001	0	0.05	0.93	0.19
	34	2	3281.67	134	<.001	0	0.08	0.98	0.16
	34	3	3163.4	117	<.001	0	0.11	1.05	0.14
	34	4	3053.8	101	<.001	0	0.14	1.13	0.12
	34	5	2943.38	86	<.001	0	0.16	1.22	0.11
	34	6	2843.01	72	<.001	0	0.19	1.33	0.1
	34	7	2753.98	59	<.001	0	0.21	1.47	0.1
	34	8	2667.24	47	<.001	0	0.23	1.65	0.1
	34	9	2580.5	36	<.001	0	0.26	1.89	0.1

Tabelle 41 *(fortgesetzt)*

S	n	AF	χ^2	df	p	TLI	CFI	RMSEA	RMS
Beginn des Semesters									
Tetrachorische Korrelationsmatrizen									
3	31	1	3066.43	152	<.001	0	0.04	0.94	0.23
	31	2	2935.49	134	<.001	0	0.08	0.99	0.19
	31	3	2811.73	117	<.001	0	0.11	1.06	0.16
	31	4	2694.35	101	<.001	0	0.15	1.14	0.14
	31	5	2566.12	86	<.001	0	0.18	1.22	0.12
	31	6	2459.57	72	<.001	0	0.22	1.33	0.1
	31	7	2369.26	59	<.001	0	0.24	1.48	0.1
	31	8	2282.31	47	<.001	0	0.27	1.66	0.1
	31	9	2195.37	36	<.001	0	0.29	1.89	0.1
4	51	1	4122.64	152	<.001	0	0.03	0.79	0.21
	51	2	3974.52	134	<.001	0	0.06	0.83	0.16
	51	3	3858.12	117	<.001	0	0.09	0.89	0.14
	51	4	3756.7	101	<.001	0	0.11	0.95	0.12
	51	5	3667.12	86	<.001	0	0.13	1.03	0.11
	51	6	3584.07	72	<.001	0	0.14	1.12	0.1
	51	7	3521.62	59	<.001	0	0.16	1.24	0.1
	51	8	3460.11	47	<.001	0	0.17	1.39	0.1
	51	9	3398.59	36	<.001	0	0.18	1.59	0.1
Ende des Semesters									
Tetrachorische Korrelationsmatrizen									
1	162	1	13693.79	152	<.001	0	0.04	0.76	0.15
	162	2	13503.74	134	<.001	0	0.06	0.81	0.12
	162	3	13337.16	117	<.001	0	0.07	0.86	0.11
	162	4	13206.02	101	<.001	0	0.07	0.93	0.1
	162	5	13100.43	86	<.001	0	0.08	1	0.09
	162	6	13003.76	72	<.001	0	0.09	1.1	0.09
	162	7	12923.51	59	<.001	0	0.09	1.21	0.09
	162	8	12865.75	47	<.001	0	0.1	1.36	0.09
	162	9	12807.99	36	<.001	0	0.1	1.55	0.09

Tabelle 41 *(fortgesetzt)*

S	n	AF	χ^2	df	p	TLI	CFI	RMSEA	RMS
Ende des Semesters									
Tetrachorische Korrelationsmatrizen									
2	28	1	3049.94	152	<.001	0	0.06	1	0.21
	28	2	2895.44	134	<.001	0	0.11	1.06	0.16
	28	3	2764.49	117	<.001	0	0.14	1.13	0.14
	28	4	2641.25	101	<.001	0	0.18	1.22	0.12
	28	5	2527.35	86	<.001	0	0.21	1.32	0.11
	28	6	2414.89	72	<.001	0	0.24	1.44	0.11
	28	7	2308.65	59	<.001	0	0.27	1.59	0.1
	28	8	2207.17	47	<.001	0	0.3	1.79	0.1
	28	9	2105.69	36	<.001	0	0.33	2.05	0.1
3	29	1	2353.1	135	<.001	0	0.06	0.9	0.21
	29	2	2235.75	118	<.001	0	0.1	0.96	0.15
	29	3	2135.88	102	<.001	0	0.13	1.03	0.13
	29	4	2043.49	87	<.001	0	0.17	1.11	0.11
	29	5	1961.81	73	<.001	0	0.2	1.21	0.11
	29	6	1881.65	60	<.001	0	0.22	1.34	0.1
	29	7	1808.57	48	<.001	0	0.25	1.5	0.1
	29	8	1735.5	37	<.001	0	0.28	1.71	0.1
	29	9	1662.43	27	<.001	0	0.3	2.01	0.1
4	43	1	4567.57	152	<.001	0	0.05	0.93	0.19
	43	2	4416.97	134	<.001	0	0.08	0.98	0.15
	43	3	4288.46	117	<.001	0	0.1	1.05	0.13
	43	4	4172.71	101	<.001	0	0.12	1.12	0.11
	43	5	4067.97	86	<.001	0	0.14	1.22	0.1
	43	6	3969	72	<.001	0	0.16	1.33	0.1
	43	7	3879.49	59	<.001	0	0.18	1.47	0.1
	43	8	3793.76	47	<.001	0	0.19	1.65	0.1
	43	9	3708.02	36	<.001	0	0.21	1.88	0.1

Anmerkungen. S: Stichprobe mit 1 = 2 SWS, B.A., 2 = 2 SWS, M.A., 3 = 4 SWS, B.Sc., 4 = 4(+4) SWS, B.Sc., AF: Anzahl der Faktoren, CFI: Comparative-Fit-Index, TLI: Tucker-Lewis-Non-Normed-Index, RMSEA: Root Mean Square Error of Approximation, bei der Bestimmung der Fit-Indizes TLI und CFI wurden als untere Grenze 0 und obere Grenze 1 festgesetzt (siehe Brown, 2015 für Vorgehen bei CFI).

Statistisch-mathematische Probleme bei der Ermittlung der Komponenten bzw. Faktoren

Parallelanalysen

Analyse der Produkt-Moment-Korrelationsmatrix nach Anzahl der Hauptkomponenten (Beginn des Semesters):

- S1: /
- S2: standard deviation is zero
- S3: standard deviation is zero
- S4: /

Analyse der Produkt-Moment-Korrelationsmatrix nach Anzahl der Faktoren (Beginn des Semesters):

- S1: /
- S2: standard deviation is zero, matrix not positive definite; estimated weights probably incorrect
- S3: standard deviation is zero, estimated weights are probably incorrect
- S4: standard deviation is zero

Analyse der Produkt-Moment-Korrelationsmatrix nach Anzahl der Hauptkomponenten (Ende des Semesters):

- S1: /
- S2: in factor scores correlation matrix is singular, matrix not positive definite; estimated weights are probably incorrect, standard deviation is zero
- S3: standard deviation is zero
- S4: standard deviation is zero, estimated weights for factor scores probably incorrect

Analyse der Produkt-Moment-Korrelationsmatrix nach Anzahl der Faktoren (Ende des Semesters):

- S1: /
- S2: error in solve.defaults(S): system is computationally singular
- S3: standard deviation is zero, estimated weights for factor scores probably incorrect
- S4: standard deviation is zero, estimated weights for factor scores probably incorrect

Analyse der tetrachorischen Korrelationsmatrix nach Anzahl der Hauptkomponenten (Beginn des Semesters):

- S1: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect
- S2: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect; Item 2 deleted (no variance)
- S3: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect; Item 15 deleted (no variance)
- S4: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect

Analyse der tetrachorischen Korrelationsmatrix nach Anzahl der Faktoren (Beginn des Semesters), WLS-Schätzung:

- S1: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect
- S2: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect; Item 2 deleted (no variance)
- S3: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect; Item 15, Item 18 deleted (no variance)
- S4: cell entry of 0; matrix was not positive definite

Analyse der tetrachorischen Korrelationsmatrix nach Anzahl der Hauptkomponenten (Ende des Semesters):

- S1: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect
- S2: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect; Items 14, 2, 8, 19, 15, 13 deleted (no variance)
- S3: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect; Items 9, 8, 6, 2 deleted (no variance)
- S4: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect; Items 8, 11, 15, 9, 19 deleted (no variance)

Analyse der tetrachorischen Korrelationsmatrix nach Anzahl der Faktoren (Ende des Semesters), WLS-Schätzung:

- S1: cell entry of 0; matrix was not positive definite

- S2: cell entry of 0; matrix was not positive definite; estimated weights probably incorrect; Items 18, 14, 9, 19, 13, 20 deleted (no variance)
- S3: cell entry of 0; matrix was not positive definite; Items 8, 9, 2 deleted (no variance)
- S4: cell entry of 0; matrix was not positive definite; Items 9, 8, 19, 11 deleted (no variance)

Ermittlung der Indizes für explorative Faktorenanalysen (ermittelt durch die WLS-Schätzung)

Analyse der Produkt-Moment-Korrelationsmatrix nach Anzahl der Faktoren (Beginn des Semesters):

- S1: /
- S2: estimated weights for factor scores probably incorrect; Ultra-Heywood Case
- S3: estimated weights for factor scores probably incorrect
- S4: loading greater than 1 detected

Analyse der Produkt-Moment-Korrelationsmatrix nach Anzahl der Faktoren (Ende des Semesters):

- S1: /
- S2: system computationally singular
- S3: estimated weights for factor scores probably incorrect; Ultra-Heywood Case
- S4: estimated weights for factor scores probably incorrect; Ultra-Heywood Case

Analyse der tetrachorischen Korrelationsmatrix nach Anzahl der Faktoren (Beginn des Semesters):

- S1: Matrix not positive definite
- S2: cell entry of 0, matrix not positive definite, convergence not obtained
- S3: cell entry of 0, matrix not positive definite, convergence not obtained
- S4: cell entry of 0, matrix not positive definite

Analyse der tetrachorischen Korrelationsmatrix nach Anzahl der Faktoren (Ende des Semesters):

- S1: cell entry of 0, matrix not positive definite, convergence not obtained
- S2: cell entry of 0, matrix not positive definite, convergence not obtained
- S3: cell entry of 0, matrix not positive definite, convergence not obtained
- S4: cell entry of 0, matrix not positive definite, convergence not obtained

SRA: Ladungen

Tabelle 42

Übersicht der Ladungen auf Komponenten bzw. Faktoren für Stichprobe 2 zu Beginn und am Ende des Semesters.

	Beginn					Ende							
	PCA	FA (PMK)			FA (TCK)	PCA				FA (TCK)			
	K1	F1	F2	F3	F1	K1	K2	K3	K4	F1	F2	F3	F4
1	-.18	-.22	.18	.48	-.24	-.07	.85	.21	.06	.28	.78	.11	.03
4	.05	-.07	.33	.08	.08	-.47	-.33	-.12	.03	.33	-.36	-.29	.07
17	.18	.14	.08	.14	.19	-.20	-.37	-.16	.45	.06	-.38	-.40	.15
2	.29	.23	.05	-.30	.46	.09	-.08	-.12	.87	.07	.00	-.65	-.02
3	-.21	-.27	.20	-.64	-.18	-.39	.68	-.24	.16	.17	.55	.01	.45
8	.85	.86	.06	-.16	.82	-.35	-.06	.75	-.05	.77	.01	.05	-.23
13	.12	.02	.22	-.24	.11	.45	.32	.31	.16	-.28	.34	.20	.00
18	-.60	-.48	-.13	.07	-.62	.91	-.01	-.06	.07	-.79	.08	.15	-.09
19	-.70	-.53	-.41	-.02	-.70	.87	.04	.18	.06	-.58	.10	.37	-.04
20	-.81	-.83	-.06	-.25	-.75	.80	-.05	-.31	.05	-.85	-.01	.01	-.09
9	.31	.17	.23	.16	.38	-.64	.03	.30	-.03	.79	.00	.01	-.01
10	.17	.14	-.05	.21	.07	.00	-.19	.61	.06	.46	-.13	.37	-.21
11	.65	.60	-.09	.02	.64	-.44	.10	.51	.42	.75	.15	-.04	.04
14	-.26	-.28	.23	.18	-.38	-.14	-.09	-.07	-.90	-.23	-.15	.60	.19
15	.24	.37	-.52	.07	.11	-.36	.07	-.33	.23	.05	-.04	.15	.58
6	.42	.15	.85	-.02	.45	.04	.19	-.23	.21	-.23	.19	-.16	.25
12	-.31	-.31	.11	.48	-.33	.11	.78	-.11	.04	-.24	.69	-.18	-.06
16	.13	.20	-.28	.06	.21	.28	.12	.63	-.16	.25	.29	.01	-.42
5	.44	.38	-.07	-.27	.49	-.17	-.72	.14	.27	.21	-.67	.11	.10

Anmerkungen. PC: Hauptkomponentenanalyse, FA: Faktorenanalyse, TCK: Tetrachorische Korrelationsmatrix als Grundlage, PMK: Produkt-Moment-Korrelationsmatrix als Grundlage, Anteil aufgeklärter Varianz beträgt 19% (PCA Beginn) bis 59% (PCA Ende).

Tabelle 43

Übersicht der Ladungen auf Komponenten bzw. Faktoren für Stichprobe 3 zu Beginn und am Ende des Semesters.

	Beginn												Ende						
	PCA			FA (PMK)				FA (TCK)					PCA		FA (PMK)				FA (TCK)
	K1	K2	K3	F1	F2	F3	F4	F1	F2	F3	F4	F5	K1	K2	F1	F2	F3	F4	F1
1	-.19	.52	-.24	.10	.38	-.20	-.12	-.17	-.08	.43	-.13	.32	.40	.40	-.11	.08	.91	.04	.45
4	.25	-.24	-.09	-.04	.00	.25	-.21	.11	-.47	.08	.15	-.43	-.20	-.14	-.25	-.17	-.25	-.47	-.26
17	.02	.39	-.38	-.25	.08	-.44	-.10	.36	.05	.11	-.46	.38	.47	.06	-.21	.26	-.27	.44	.45
2	.11	.35	.39	-.01	.26	.11	.30	.02	.36	.31	.14	.18	-.08	.63	.12	.63	-.07	.03	.14
3	.09	.13	-.27	-.02	.14	.06	-.19	.02	-.04	.06	.22	.65	.47	.10	-.34	-.06	.31	.00	.48
8	.10	-.05	-.70	-.06	.10	-.04	-.70	.07	-.73	.10	-.06	.02	.15	.83	-.08	.93	.07	.02	.43
13	.56	.02	-.27	-.62	-.12	-.11	-.12	.67	-.07	-.13	-.08	.19	.23	-.56	-.09	-.38	-.10	.24	.09
18	.75	-.01	-.06	-.46	.11	.36	-.13	.57	.05	.04	.35	.20	-.78	.07	.76	.23	-.24	-.07	-.75
19	.56	.05	.32	-.17	.28	.58	.10	.16	.12	.21	.66	-.03	-.77	-.25	.93	-.15	-.06	.08	-.73
20	.53	-.32	.21	-.10	.01	.73	-.04	.14	-.06	-.07	.68	.13	-.75	-.03	.77	-.06	.15	-.13	-.68
9	-.84	-.11	.07	.99	-.01	-.03	-.02	-.78	.03	-.06	-.07	.08	.57	-.19	-.26	-.14	.14	.34	.54
10	-.24	.62	.30	-.03	.29	-.45	.46	-.03	.50	.34	-.35	-.06	.23	-.63	.00	-.40	-.05	.34	.11
11	-.73	.09	.02	.71	.16	-.08	-.03	-.67	.06	.11	-.06	.24	.61	-.32	-.38	-.10	-.27	.47	.59
14	-.03	.20	.64	-.06	.01	-.06	.65	.00	.62	.04	-.01	-.12	-.18	.07	.02	.14	-.24	-.07	-.18
15	-.09	-.38	.58	.19	-.29	.28	.41	.03	.50	-.23	.25	.09							
6	.27	.59	.29	-.25	.48	-.02	.28	.18	.19	.54	.05	-.24	.44	-.31	.01	-.16	.11	.54	.47
12	.08	.67	-.25	-.14	.57	-.22	-.16	.16	-.17	.62	-.17	.10	.48	.03	.03	.10	.22	.56	.53
16	-.30	-.06	-.35	.16	-.09	-.16	-.21	-.27	-.16	-.18	-.09	.41	.24	.53	-.12	.37	.19	.03	.40
5	-.02	.66	.09	.17	.86	.12	-.01	-.21	-.02	.73	.16	.00	.07	-.05	-.27	.09	-.45	-.08	-.03

Anmerkungen. PC: Hauptkomponentenanalyse, FA: Faktorenanalyse, TCK: Tetrachorische Korrelationsmatrix als Grundlage, PMK: Produkt-Moment-Korrelationsmatrix als Grundlage, Anteil aufgeklärter Varianz beträgt 21% (FA TCK Ende) bis 48% (FA TCK Beginn).

Tabelle 44

Übersicht der Ladungen auf Komponenten bzw. Faktoren für Stichprobe 4 zu Beginn und am Ende des Semesters.

	Beginn													Ende					
	PCA				FA (PMK)			FA (TCK)						PCA	FA (PMK)		FA (TCK)		
	K1	K2	K3	K4	F1	F2	F3	F1	F2	F3	F4	F5	F6	K1	F1	F2	F1	F2	F3
1	-.14	.23	-.15	.47	-.09	.36	-.09	-.16	.26	.42	-.06	-.05	-.07	.32	.26	.31	.20	.48	.20
4	-.22	.10	-.14	-.69	-.18	-.19	-.12	-.21	-.15	-.34	-.09	.44	-.04	-.23	-.19	-.18	-.28	-.40	.17
17	.11	.40	.35	.28	.10	.40	.28	.17	.30	.29	.37	.07	.10	.40	.35	.38	.31	.11	.61
2	.08	.62	-.02	.18	.04	.56	-.03	-.02	.59	.16	-.11	.05	.14	.24	.21	.38	-.13	.51	.32
3	-.35	.01	.48	.31	-.21	.16	.44	-.31	-.05	.17	.17	-.17	.41	.00	.00	.12	-.02	-.16	.33
8	.00	.57	.24	.08	-.02	.45	.18	-.12	.56	.03	.17	.04	.11	.13	.10	.04	-.48	.18	.14
13	-.16	.09	.41	.04	-.09	.08	.34	-.17	-.06	.13	.54	.14	-.06	.20	.14	-.39	.34	.01	-.53
18	.84	.00	.12	.03	.82	-.02	.15	.73	-.05	-.01	.15	-.01	-.06	.85	.85	-.04	.78	.21	-.16
19	.83	.20	-.08	.04	.80	.19	-.07	.66	.11	.03	-.12	.05	-.01	.93	.98	.10	.83	.17	.06
20	.77	-.17	-.14	-.07	.73	-.2	-.15	.60	-.13	-.09	-.23	-.08	-.01	.85	.84	.02	.74	.00	.20
9	.10	.49	-.27	.34	.07	.53	-.22	.04	.41	.40	-.18	.11	-.06	-.68	-.61	.43	-.79	.21	.10
10	.00	.62	-.01	-.26	-.03	.32	-.01	.10	.17	.10	.15	.60	.05	.17	.16	.39	.17	.54	.03
11	-.25	.16	-.21	.61	-.13	.38	-.09	-.09	-.04	.69	.00	.07	.03	-.55	-.45	.29	-.64	.36	-.17
14	-.09	.31	.46	.06	-.06	.24	.35	-.11	.26	-.06	.03	-.08	.52	.04	.03	.02	.10	.22	-.23
15	-.29	.37	-.01	-.25	-.2	.17	.01	-.01	-.14	-.02	-.12	.35	.49	.27	.23	.26	.52	.14	.03
6	-.23	.35	-.47	-.16	-.18	.21	-.35	-.21	.06	.15	-.29	.41	-.11	-.02	.02	.48	-.10	.27	.38
12	-.07	-.13	.71	.03	.00	-.10	.67	-.09	-.08	-.09	.62	-.14	.05	-.10	-.06	.22	-.21	.41	-.04
16	.25	.36	-.08	-.25	.12	.14	-.10	.07	.56	-.31	-.06	.03	-.19	.13	.09	.12	.05	-.06	.44
5	-.01	.18	.66	-.35	-.02	-.02	.5	-.05	.07	-.34	.44	.10	.26	.02	.03	.29	-.03	.19	.32

Anmerkungen. PC: Hauptkomponentenanalyse, FA: Faktorenanalyse, TCK: Tetrachorische Korrelationsmatrix als Grundlage, PMK: Produkt-Moment-Korrelationsmatrix als Grundlage, Anteil aufgeklärter Varianz beträgt 19% (PCA Ende) bis 46% (PCA Beginn).

SRA: Veränderungen

Abbildung 40. Boxplots, Mittelwerte (graue große Punkte) und verbundene Wertepaare (hellgraue dünne Linien) für die Anzahl der erreichten Punkte im SRA nach Stichproben (S 1: Sonderpädagogikstudierende, S 2: Masterstudierende, S 3: Psychologiestudierende im ersten Semester und S 4: Psychologiestudierende im zweiten Semester)

SRA: Korrelationen zu relevanten Merkmalen

Tabelle 45

Mittelwerte, Standardabweichungen und Produkt-Moment-Korrelationen für Skalen bzw. Punktesummen aus Stichprobe 2 (2 SWS, M.A.).

		M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Statistik (1)	9.56	1.37																
2	Statistik (2)	8.71	1.70	**.42**															
3	ded. Schließen	4.58	1.03	.29	.19														
4	Figurenreihen	5.93	1.84	**-.54**	-.11	.13													
5	Mathematik	7.00	2.17	.31	**.45**	-.34	-.20												
6	Affekt (1)	3.43	1.01	**.59**	**.50**	**.50**	**-.48**	**.35**											
7	Kompetenz (1)	4.16	1.09	**.50**	**.59**	**.51**	-.15	**.51**	**.81**										
8	Wert (1)	4.88	0.94	.23	**.52**	-.17	-.30	**.59**	**.42**	**.51**									
9	Schwierigkeit (1)	5.00	0.61	-.27	-.10	-.16	**.48**	.01	**-.68**	-.29	-.32								
10	Interesse (1)	4.96	0.85	**.37**	**.62**	**-.42**	-.33	**.52**	.30	.28	**.77**	**-.35**							
11	Anstrengung (1)	5.53	0.97	.14	.09	-.23	-.27	-.19	-.24	-.29	.12	-.22	.31						
12	Affekt (2)	3.34	0.96	**.81**	**.36**	.23	**-.83**	**.35**	**.77**	**.59**	**.46**	**-.59**	**.40**	.20					
13	Kompetenz (2)	4.53	1.02	**.60**	**.47**	.26	**-.47**	**.51**	**.77**	**.89**	**.59**	**-.39**	**.44**	-.12	**.76**				
14	Wert (2)	4.36	0.92	.32	**.48**	**-.40**	**-.57**	**.42**	.21	.24	**.78**	-.22	**.82**	**.43**	**.50**	**.53**			
15	Schwierigkeit (2)	4.85	0.73	-.29	-.19	.05	**.44**	-.13	**-.50**	-.33	**-.42**	**.84**	**-.54**	**-.47**	**-.59**	**-.48**	**-.42**		
16	Interesse (2)	4.36	1.42	**.53**	.24	**-.46**	**-.68**	**.41**	.28	.18	**.63**	**-.43**	**.80**	**.43**	**.65**	**.51**	**.82**	**-.61**	
17	Anstrengung (2)	5.45	1.22	-.20	-.25	**-.40**	**-.54**	.06	-.01	-.09	.32	-.30	.21	.32	.24	.19	**.53**	**-.42**	**.43**

Anmerkungen. *N* beträgt 13, moderate bis hohe Effekte (ab ≈ ±.40) sind **fett** gesetzt, ded. Schließen: deduktives Schließen, der Messzeitpunkt ist in Klammern angegeben (1: zu Beginn des Semesters, 2: am Ende des Semesters).

Tabelle 46

Mittelwerte, Standardabweichungen und Produkt-Moment-Korrelationen für Skalen bzw. Punktesummen aus Stichprobe 3 (4 SWS, B.Sc.).

		M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Statistik (1)	8.55	2.32																
2	Statistik (2)	9.55	1.64	.36															
3	ded. Schl. (kurz)	4.71	1.23	.66	.22														
4	Figurenreihen			.	.														
5	Mathematik (kurz)	6.12	2.23	.56	.36	.29													
6	Affekt (1)	3.80	0.93	.15	.17	.09	.	.33											
7	Kompetenz (1)	4.52	1.09	.08	.07	.08	.	.35	.73										
8	Wert (1)	5.00	1.06	-.13	.01	-.05	.	.14	.15	.57									
9	Schwierigkeit (1)	4.72	0.57	-.01	.07	.05	.	-.04	.08	.23	.47								
10	Interesse (1)	4.81	1.27	-.22	.09	-.16	.	-.09	.50	.66	.71	.55							
11	Anstrengung (1)	6.19	1.06	-.20	.15	-.04	.	.03	.27	.64	.84	.60	.76						
12	Affekt (2)	3.67	1.23	.23	.09	.05	.	.19	.51	.44	.13	-.35	.25	.12					
13	Kompetenz (2)	4.74	0.94	.42	.01	.15	.	.23	.39	.31	-.03	-.50	.09	-.18	.84				
14	Wert (2)	4.84	0.80	-.13	-.20	-.32	.	-.17	-.55	-.32	.31	-.26	-.16	.04	.16	.11			
15	Schwierigkeit (2)	4.74	0.80	-.15	-.17	-.07	.	-.15	-.24	-.26	-.23	.25	-.38	-.03	-.53	-.61	-.09		
16	Interesse (2)	4.55	1.25	-.03	-.19	-.34	.	-.16	-.13	.09	.15	-.31	.16	.06	.65	.59	.59	-.30	
17	Anstrengung (2)	5.70	1.00	.16	.34	.13	.	.20	.34	.48	.47	.25	.33	.58	.29	.05	.12	-.01	.07

Anmerkungen. *N* beträgt 19 und 16 (grau unterlegt), moderate bis hohe Effekte (ab ≈ ±.40) sind **fett** gesetzt, ded. Schließen: deduktives Schließen, der Messzeitpunkt ist in Klammern angegeben (1: zu Beginn des Semesters, 2: am Ende des Semesters).