2026, 3. Jg., Heft 1
ISSN 2944-4969
https://doi.org/10.26204/KLUEDO/13198

Franz Unterholzner & Hans-Georg Müller

Metakognition und Sprachbetrachtung. Zur Operationalisierung metakognitiven Monitorings in der sprachbezogenen Kompetenzmessung

Der Beitrag dokumentiert die Operationalisierung metakognitiven Monitorings im Rahmen der Pilotierung von Testaufgaben zur Sprachbetrachtung der österreichweiten Kompetenzmessung iKM^{PLUS} (N = 1184; dritte und siebte Klassenstufe). In der Erhebung wurde Monitoring mittels globaler und lokaler Selbsteinschätzungen (Confidence Judgements) erfasst. Die Ergebnisse zeigen eine stabile Selbstüberschätzung sowie Zusammenhänge zwischen Leistung und Einschätzungspräzision, jedoch keine altersabhängige Verbesserung der Monitoring-Fähigkeit. Die Studie belegt die prinzipielle Integrierbarkeit prozedural-metakognitiver Messansätze zur Kontextualisierung von Kompetenzwerten der Sprachbetrachtung.

Schlagworte: Metakognition, Monitoring, Kompetenzmessung, Sprachbewusstheit, Sprachbetrachtung

Metacognition and Language Awareness: On the Operationalization of Metacognitive Monitoring in Competence Assessment

This paper documents the operationalization of metacognitive monitoring in a pilot study for the nationwide competence assessment of language awareness in Austria. Drawing on iKM^{PLUS} data (N = 1184; grades 3, 7), monitoring was assessed using global and local performance postdictions (confidence judgements). The results indicate a typical tendency towards overconfidence and significant associations between performance and judgement accuracy, but no age-related improvement in monitoring ability. The study demonstrates the integrability of procedural metacognitive measurement approaches for contextualizing language awareness competence scores in principle.

Keywords: metacognition, monitoring, competence assessment, language awareness

1 Einleitung

Im Zuge der Entwicklung weg von einem formalistisch orientierten Grammatikunterricht hin zum kompetenzorientierten Verständnis von Sprachbetrachtung hielt der Begriff der „Sprachbewusstheit“ (Gornik 2022, 43–45) Einzug in die Deutschdidaktik. Hier wurde auch diskutiert, inwiefern Sprachbewusstheit als eine spezifische, auf Sprache bezogene Ausprägung von Metakognition verstanden werden kann (Funke 2005, 9), denn das sprachbewusste Betrachten eigener sprachlicher Handlungen und Produkte ist seinem Wesen nach metakognitiv (Unterholzner & Müller 2023, 21). Dies legt es nahe, sprachbezogene Aspekte von Metakognition in die Konzeptualisierung metasprachlicher Bewusstheit zu integrieren (Ossner 2007, 138–140).

Diese Überlegung liegt dem Versuch zugrunde, Metakognition in Pilotierungen zur Kompetenzmessung von Sprachbewusstheit in den entsprechenden Bonusmodulen der iKM^{PLUS} zu integrieren, was den ersten solchen Versuch in österreichischen Large-Scale-Messungen darstellt (für Deutschland Händel et al. 2013 für NEPS; Artelt et al. 2009 für PISA), und auch den unseres Wissens ersten deutschsprachigen Versuch überhaupt, Aspekte von Metakognition im Rahmen von Sprachbetrachtung zu messen.

Im Fall der hier vorgestellten iKM^{PLUS}-Pilotierungen wurden sowohl deklarative als auch prozedurale Aspekte von Metakognition erhoben, wobei in diesem Beitrag ausschließlich auf das dem prozeduralen Bereich zuzuordnende metakognitive Monitoring eingegangen wird.

2 Metakognitives Monitoring

Der Ausdruck „Metakognition" bezeichnet alle selbstbezogenen Formen der Kognition, die auf der Wahrnehmung des eigenen Denkens, Wahrnehmens und Handelns basieren. Als metakognitiv können demnach alle Wissensbestände und kognitiven Prozesse verstanden werden, die der Selbstevaluation und Selbstregulation dienen. Seit der frühen Zweiteilung in metakognitives Wissen und metakognitive Erfahrungen (Flavell 1979) bzw. *knowledge about cognition* und *regulation of cognition* (Brown et al. 1983) wurde das Konstrukt Metakognition weiter kollaborativ verfeinert. Für einen aktuellen systematischen Überblick siehe bspw. Loaiza et al. (2023).

Einer innerhalb des Metakognitions-Paradigmas weitgehend anerkannten, empirisch abgestützten (Schraw & Dennison 1994) und im Detail in verschiedenen Ausformungen ausgestalteten Dichotomie folgend, kann (verbalisierbares) metakognitives Wissen weitgehend dem deklarativen Gedächtnis zugeordnet werden, während die metakognitiven Fähigkeiten überwiegend auf das prozedurale Gedächtnis zurückgreifen (Schneider et al. 2022, 275; Chen & McDunn 2022, 2–3; Artelt &Schneider 2015, 3; Schneider & Löffler 2015, 293).

Die in der Folge nicht weiter thematisierte deklarative Komponente der Metakognition kann grob in Wissen über sich selbst als Person, die Aufgabe sowie über Strategien eingeteilt werden (Flavell 1979, 907). So könnte etwa das Wissen darüber, dass man am Ende von Texten zu besonders vielen Rechtschreibfehlern neigt (Wissen über Person und Aufgabe), dazu führen, dass man sich in Prüfungssituationen das letzte Drittel des eigenen Textes für die formale Textrevision zuerst vornimmt (Strategiewissen).

Deklarativ-metakognitives Strategiewissen kann naturgemäß nur aus Wissen im entsprechenden Lernbereich entwickelt werden. Nur wer bspw. metasprachliches Wissen über syntaktische Zusammenhänge hat, kann entscheiden, ob es bei Fragen zur satzinternen Großschreibung mehr Sinn ergibt, eine Artikel- oder Attribuierungsprobe auszuführen oder gar im Wörterbuch nachzusehen. Deklarative Metakognition wird daher als weitgehend domänenspezifisch eingeschätzt, wobei nicht gänzlich unumstritten ist, ob es mit zunehmender Lernerfahrung in vielen Bereichen zu einer Domänengeneralisierung selbstreflektierten strategischen Handelns kommt (Geurten et al. 2018). Üblicherweise werden solche Strategieentscheidungen getestet, indem Proband:innen mehrere

Strategien nach ihrem funktionalen Potenzial für eine adäquate Aufgabenbearbeitung einschätzen und reihen müssen (bspw. Händel et al. 2013 oder Seeger et al. 2021). Auch im Zuge der hier beschriebenen Pilotierungsstudien wurden metakognitiv orientierte Strategieabfragen getestet. Die Ergebnisse dieser Versuche sind aber nicht Thema des vorliegenden Beitrags.

Die prozedurale Komponente ist nicht eigentlich auf Bewusstheit im Sinne verbalisierbarer Selbstwahrnehmungen angewiesen, sehr wohl aber auf erhöhte Aufmerksamkeit (Müller & Unterholzner 2025, 7). Ein plötzliches Innehalten beim Schreiben, weil man das Gefühl hat, gerade einen Buchstabendreher eingebaut zu haben, wäre etwa ein Beispiel für prozedurale Metakognition.

Die prozedurale Komponente der Metakognition unterliegt ihrerseits einer Zweiteilung in Monitoring einerseits und Control/Self-Regulation andererseits (Nelson & Narens 1990). Die Subkomponente der Selbstregulation bezieht sich auf Umsteuerungsprozesse als Konsequenz eines im Monitoring detektierten Veränderungsbedarfs. In der Operationalisierung kommen hier Online-Verfahren zum Einsatz, die solche Umsteuerungsprozesse auf Verhaltensebene beobachtbar machen (Schneider et al., 2022, 280). Hier ist in der Gesamtbeurteilung der Studienlage eine Entwicklungstendenz zu beobachten, wonach Monitoring bei Lernenden des mittleren Volksschulalters bereits gut ausgebildet ist, während die potenziell daran anknüpfende Selbstregulation allerdings bis in höhere Altersstufen meist trotzdem noch ausbleibt. Die Komplexität der Entwicklungsaufgabe für Lernende besteht demnach vor allem im Erlernen eines effizienten Zusammenspiels von Monitoring und Selbstregulation (Schneider 2015, 284–286). Da Selbstregulation im Zuge dieser Studie nicht operationalisiert wurde, wird hierauf ebenfalls nicht weiter eingegangen.

Metakognitives Monitoring als evaluierende Aufmerksamkeit für das eigene Agieren greift weniger auf einen zu erlernenden Wissensschatz zurück, vielmehr auf ein grundsätzliches Empfinden von Flüssigkeit und Mühelosigkeit beim Bearbeiten einer Aufgabe (Koriat 1997). Da dieses Empfinden wenig von fachwissensspezifischen Bedingungen abhängig ist, kann für Monitoring in höherem Maß Domänengeneralität angenommen werden als für deklarativ-strategische Anteile von Metakognition (Geurten et al. 2018, 76).

Die Operationalisierung der Monitoring-Komponente geschieht, indem in Bezug auf eine Leistungsaufgabe auch eine Selbsteinschätzung erhoben wird. Das Ausmaß an Passung zwischen Leistung und Selbsteinschätzung gilt dabei als Indikator für die Qualität und Aktivität des Monitorings (zu etablierten Maßen dafür siehe Schneider 2015, 265). Globale Einschätzungen zu einem ganzen Test können prospektiv als *performance prediction* oder retrospektiv als *performance postdictions* ausgestaltet sein. Auch lokale Item-by-Item-Einschätzungen zum individuellen Sicherheitsgefühl in Bezug auf eine Aufgabe können als Vorhersagen oder als Selbsteinschätzung direkt nach Absolvierung einer einzelnen Testaufgabe ausgestaltet sein. Letztere werden als *confidence judgements* bezeichnet (Destan & Roebers 2015, 348). Dabei zeigt sich im Zuge der Berechnung eines Maßes der Übereinstimmung zwischen tatsächlicher Leistung und darauf bezogener Selbsteinschätzung (*calibration measures*; Schraw et al. 2013), dass globale Selbsteinschätzungen weniger deutlich mit der Schwierigkeit des Tests variieren und

stärker vom allgemeinen Selbstkonzept abhängen als lokale (Händel et al. 2020), sowie dass Postdictions sensitiver auf die Aufgabe ansprechen als die wiederum deutlicher mit dem allgemeinen Selbstkonzept zusammenhängenden Predictions (van Loon et al. 2022; Lingel et al. 2019, 589).

Die Entwicklung der Monitoring-Komponente erreicht nach den ersten drei, vier Schuljahren von Kindern bereits ein Plateau und unterliegt danach einem wenig deutlichen Entwicklungstrend (Schneider et al. 2022, 285). Eine Konstante in Monitoring-Studien ist eine deutliche, stabile Selbstüberschätzung. Bei jüngeren Kindern bis zu den ersten ein, zwei Schuljahren ist ein höheres Maß an besonders stabiler Selbstüberschätzung zu beobachten, was möglicherweise weniger auf eine noch wenig ausgeprägte Selbsteinschätzungsfähigkeit als auf *wishful thinking* zurückzuführen ist (Schneider et al. 2022, 285).

In Bezug auf Verbesserungen der zu optimistischen Selbsteinschätzung mit zunehmendem Alter ist die Studienlage insgesamt etwas uneindeutig. Einige rezente Studien mit Kindern des Primarstufenalters finden keinen deutlichen Zusammenhang zwischen zunehmender schulischer Erfahrung bzw. höherem Alter und erhöhter Präzision bzw. reduzierter Selbstüberschätzung (bspw. van Loon et al. 2022, 2). Ein über eine größere Altersspanne hinweg schon stabil zu beobachtender Trend besteht jedenfalls darin, dass die Selbsteinschätzungen älterer Kinder und Jugendlicher in der Gesamttendenz der Populationen pessimistischer werden. Vor allem steigt das Unsicherheitsgefühl im Zuge von inkorrekt beantworteten Leistungsaufgaben (Schneider et al. 2022, 286). Es handelt sich also eher um einen undifferenzierten, wenig vom eigenen situationalen Leistungserleben beeinflussten, allgemeinen Trend hin zu weniger Optimismus im Laufe der Schulkarriere.

Ferner ist die Forschung zum metakognitiven Monitoring auch von der Frage geprägt, inwiefern die grundsätzliche Selbstüberschätzung sowie der hier stabile statistische Zusammenhang zwischen präziserem Monitoring und höherer Testleistung (weniger deutliche Selbstüberschätzung) tatsächlich zeigt, dass Monitoring- und Leistungsfähigkeit miteinander in Zusammenhang stehen (Fleming & Lau 2014). Dafür spricht, dass Wissen eine Vorbedingung für gutes Monitoring zu sein scheint – und zwar in dem Sinne, dass Individuen mit weniger Wissen schlicht auch über weniger Entscheidungskriterien für eine präzise Selbsteinschätzung verfügen (*unskilled-and-unaware-of-it-effect*; Kruger & Dunning 1999).

Gegen den Zusammenhang von Monitoring- und Leistungsfähigkeit sprechen mehrere Aspekte, die eine systematische Verzerrung der Datenlage (Bias) betreffen: Erstens haben Proband:innen mit hohen Testwerten grundsätzlich eine höhere Wahrscheinlichkeit, präzise zu sein, weil sie mit ihren höheren Leistungswerten durchschnittlich näher bei den stabil hohen Selbsteinschätzungswerten landen (Juslin et al. 2000). Zweitens besteht die Problematik, dass eine schwächere Testleistung und ein erhöhtes Unsicherheitsgefühl und damit eine schwächere Selbsteinschätzung statistisch konfundiert sind (Vuorre & Metcalfe 2022), was – drittens – auch mit dem unterschiedlichen Einfluss der Ratewahrscheinlichkeit von Aufgaben in Hinblick auf mehr oder weniger unsichere Bearbeitung zu tun hat (Lingel et al., 2019, 589–590). Viertens liegt nahe, dass vor allem

Proband:innen des unteren Leistungsspektrums in Testungen besonders gering motiviert sind und deshalb auch in Bezug auf die eigene Selbsteinschätzung wenig motiviert sind, genauer darüber zu reflektieren (Bol & Hacker 2012, 3).

3 Methode

3.1 Testung und Stichprobe

In der iKM^{PLUS}, der *Individuellen Kompetenzmessung PLUS* des IQS (Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen), werden in Österreich jährlich ca. 80000 Schüler:innen der dritten und siebten Klassenstufe in den verpflichtenden Basismodulen Deutsch/Lesen, Englisch und Mathematik getestet. Zusätzlich zu den Basismodulen können Lehrkräfte freiwillige Bonusmodule mit ihren Klassen durchführen, darunter die Bonusmodule *Einsicht in Sprache durch Sprachbetrachtung* (dritte Klasse, in der Folge kurz: *Sprachbetrachtung*) und *Sprachbewusstsein* (siebte Klasse). Für weiterführende Informationen zum Kompetenzmodell und zur didaktischen Ausrichtung der Testungen siehe Illetschko et al. (2025, 26–31).

Die Testung der Metakognitions-Elemente fand im Rahmen von Pilotierungsstudien zu neuen Aufgabenstellungen statt. Neben deklarativ-metakognitiven Strategieaufgaben wurden folgende Elemente zu metakognitivem Monitoring in die Testdesigns integriert:

1) Vier direkt an einzelne Testaufgaben gekoppelte Selbsteinschätzungsfragen;
2) eine am Ende des Tests gestellte Gesamt-Selbsteinschätzungsfrage.

Die Pilotierung der Primarstufe wurde im Frühjahr 2023 in 32 dritten Klassen in allen neun österreichischen Bundesländern sowohl im ländlichen als auch im städtischen Raum durchgeführt. Hier wurden abzüglich der Selbsteinschätzungs-Items 139 neue Testaufgaben pilotiert. Die Pilotierung der Sekundarstufe wurde im Januar 2024 in 30 siebten und 15 achten Klassen durchgeführt (ebenfalls repräsentativer Querschnitt der österreichischen Bildungslandschaft). Es wurden 108 Testaufgaben pilotiert. In die hier vorgelegte Studie flossen 623 Datensätze der dritten Klasse sowie die 561 Datensätze der siebten Klasse ein (n = 1184).

In den Pilotierungsstudien kamen aus 32 (und selten 33) Leistungsaufgaben bestehende Tests zum Einsatz. Sie beinhalteten jeweils vier (selten drei oder fünf) Selbsteinschätzungs-Items, die direkt im Anschluss an ausgewählte Aufgaben – auf lokaler Testebene also – fragten: „Denkst du, du hast die letzte Aufgabe richtig lösen können?“. Die Proband:innen konnten als Antwortoption ausschließlich „Ja“ oder „Nein“ auswählen. Eine solche binäre Selbsteinschätzung entspricht dem Vorgehen im überwiegenden Teil von Monitoring-Studien (Lingel et al. 2019, 590). Der Hauptgrund dafür ist, dass binäre Selbsteinschätzungen im Gegensatz zu skalierten Antwortmöglichkeiten einerseits undifferenzierte Mittelwerttendenzen verhindern und sich andererseits deutlich leichter auf die ebenfalls binäre Richtig-falsch-Struktur der Leistungsaufgaben beziehen lassen, bspw. in Confidence-accuracy-Tabellen wie in Tabelle 1 (Lingel et al. 2019, 590). Dem Pilotcharakter der Studie entsprechend wurde daher der methodisch einfache Zugang gewählt. Um die angesprochene Passung zu gewährleisten, wurden die

Selbsteinschätzungs-Items darüber hinaus ausschließlich an Multiple-Choice-Aufgaben angebunden, die nicht Teil größerer Testlets waren, was die Berechnung eines mathematisch unverzerrten Verhältnismaßes zwischen Test- und Selbsteinschätzungs-Item weiter erleichterte.

Die Testaufgaben, denen eine direkt folgende metakognitive Selbsteinschätzung angefügt wurde, betrafen Einschätzungs-, Anwendungs- und Zuordnungsaufgaben auf den Ebenen Morphosyntax, Semantik und Pragmatik. Beispielsweise mussten die Proband:innen pragmatische Funktionen einzelner Textabschnitte einschätzen, semantische Verwandtschaften kennzeichnen oder analoge Beziehungen zu vorgegebenen Sprachbeispielen fremder Sprachen herstellen, aber etwa auch einzelne Satzglieder oder Wortarten bestimmen. In beiden Klassenstufen bezog sich gut die Hälfte der Aufgaben auf Items, die den curricularen Vorgaben der Klassenstufe entsprachen, während die andere Hälfte klassenstufentypische Anwendungsaufgaben darstellte, die sprachbewusstes Handeln, aber kein besonderes sprachsystematisches Wissen erforderten. In nur einem Viertel der Aufgaben waren fachterminologische Kenntnisse erforderlich, von denen der überwiegende Anteil in Aufgaben für die Klassenstufe 7 lag. Gut die Hälfte der Aufgaben konnte durch Anwenden grammatischer Prozeduren (z. B. Umstellen) gelöst werden.

Versuchsweise durchgeführte Regressionsanalysen zeigten keinerlei auch nur annähernd signifikante Zusammenhänge zwischen den Kompetenz- und Selbsteinschätzungsmaßen einerseits und den fachlichen Charakteristika der Aufgaben (Fachterminologie, gramm. Prozeduren, logisches Schließen) andererseits. Die Irrtumswahrscheinlichkeiten lagen in allen Analysen zwischen p_{min} = .245 und p_{max} = .978. Diese Befunde machen es unwahrscheinlich, dass die gemessenen Selbsteinschätzungen von den Charakteristika der gewählten Aufgaben abhängen. Da die beiden Klassenstufen jedoch unterschiedliche Aufgabensets bearbeiteten, ist eine Restunsicherheit nicht auszuschließen und kann nur anhand von Daten in Folgestudien näher spezifiziert werden.

Am Ende der Testung wurde außerdem mit Blick auf die insgesamt 32 Testaufgaben gefragt: „Was meinst du: Wie viele Aufgaben hast du richtig bearbeitet?" Die Antwort konnte auf einer Skala markiert (dritte Klasse) bzw. als Zahlenwert eingegeben werden (7. Klasse) und entspricht einer Selbsteinschätzung auf globaler Testebene.

3.2 Statistische Auswertung

Aus den erhobenen Leistungsaufgaben und Selbsteinschätzungs-Items wurden im Rahmen der explorativen Datenanalyse Messwerte abgeleitet, die im Folgenden erläutert werden (für einen Überblick vgl. Tabelle 2 im Ergebnisteil).

Der Messwert *Testleistung global (TL-G)* gibt den prozentualen Anteil richtig gelöster Leistungsaufgaben wieder. Der Messwert *Selbsteinschätzung global (SE-G)* gibt an, wie viel Prozent der Leistungsaufgaben die Proband:innen nach Beendigung des Tests als richtig gelöst einschätzten. Den Ausprägungsgrad der *Selbstüberschätzung global (SÜ-G)* gibt der gleichnamige Messwert an, der sich aus der Differenz von *Selbsteinschätzung global* und *Testleistung* ergibt. Da es dabei natürlich auch zu Unterschätzungen der eigenen Testleistung kam, gibt der Messwert DIFF-G die Abweichung der Selbsteinschätzung von der Testleistung global als absoluten Betrag dieser Differenz an. *DIFF-G* gibt also das

Ausmaß des *Selbsteinschätzungsfehlers global* an und kann als Messwert für die individuelle Qualität der globalen Selbsteinschätzung gelesen werden, während die *SÜ-G* eher eine Tendenz der Gesamtpopulation verdeutlicht.

Analog zu den globalen Test- und Selbsteinschätzungsleistungen bezeichnen die als „*lokal*" gekennzeichneten Werte in Tab. 2 die Einschätzungen, die die Proband:innen jeweils unmittelbar nach Bearbeitung einer einzelnen Leistungsaufgabe abgaben. Der Messwert *Testleistung lokal (TL-L)* stellt dabei den prozentualen Testscore für diejenigen Leistungsaufgaben dar, für die direkt im Anschluss eine Selbsteinschätzung erbeten wurde. Die Messwerte *Selbsteinschätzung lokal (SE-L)*, *Selbstüberschätzung lokal (SÜ-L)* und *Selbsteinschätzungsfehler lokal (DIFF-L)* wurden aus der Summe der lokalen Selbsteinschätzungen und ansonsten analog zu den oben dargestellten globalen Messwerten ermittelt.

Die Anzahl der *Hits, False Alarms, Missings* und *Correct Rejections* in direkter Nachfolge einzelner Leistungsaufgaben wurde entsprechend der Kreuzklassifikation in Tabelle 1 für alle Proband:innen individuell erfasst. In Tabelle 2 sind die durchschnittlichen Anteile der vier Felder an der Gesamtzahl der lokalen Selbsteinschätzungen auf Ebene der Gesamtstichprobe dargestellt. Ebenfalls findet man hier die Anteile der korrekten (*Hits* + *Correct Rejections*) sowie der inkorrekten Selbsteinschätzungen (*Missings* + *False Alarms*) und darüber hinaus die Anteile für grundsätzlich optimistische (*Hits + Missings*) sowie grundsätzlich pessimistische (*False Alarms + Correct Rejections*) Selbsteinschätzungen.

Tab. 1: Confidence Accuracy Table

		Selbsteinschätzung	
		als richtig gelöst vermutet	als falsch gelöst vermutet
Leistungs-aufgabe	richtig gelöst	Hit	False Alarm
	falsch gelöst	Missing	Correct Rejection

Die aus dem Verhältnis korrekter und inkorrekter Selbsteinschätzungen berechnete Variable *Selbsteinschätzungsscore (SE-L-Score)* reflektiert schließlich die Präzision der Selbsteinschätzung auf Ebene der einzelnen Proband:innen und kann damit als Messwert erfolgreichen Monitorings unmittelbar im Anschluss an die Bearbeitung einzelner Testaufgaben gelesen werden. Da sie auf der Idee beruht, zu erfassen, wie gut Proband:innen in der Selbsteinschätzung zwischen erfolgreicher und nicht erfolgreicher Aufgabenbearbeitung diskriminieren können, ergibt sie sich aus der Summe der Anteile korrekter Selbsteinschätzungen (*Hits + Correct Rejections*) abzüglich der Summe der Anteile inkorrekter Selbsteinschätzungen (*Missings* + *False Alarms*). Auf diese Weise entsteht dieser *Discrimination Score*, bei dem +1 für eine 100 % korrekte und -1 für eine 0 % korrekte Selbsteinschätzung stehen.[1]

[1] Dieser Discrimination Score (bspw. van Loon et al. 2024; Schraw et al. 2013; Destan & Roebers 2015, 357) reduziert Bias gegenüber einfachen Φ-Korrelationen zwischen Leistungs- und Selbsteinschätzungswerten (Fleming & Lau

Auf Basis der verschiedenen in Tabelle 2 angegebenen Werte wurden Korrelationen berechnet, die den Zusammenhang zwischen Testleistung und Präzision des Monitorings und den Zusammenhang zwischen Alter bzw. Klassenstufe und Präzision des Monitorings beleuchten sollten. Die untersuchungsleitenden Erwartungen waren:

1) Höhere Testleistungen korrelieren auf lokaler und globaler Ebene positiv mit höherer Monitoring-Präzision. Das heißt, TL-G und TL-L korrelieren
   a) jeweils negativ mit der Selbstüberschätzung SÜ-G und SÜ-L,
   b) jeweils negativ mit dem Selbsteinschätzungsfehler DIFF-G und DIFF-L.
   c) TL-L korreliert positiv mit dem SE-L-Score.
2) Die in Punkt 1 angegebenen Korrelationen sind für die lokale Ebene höher als für die globale, weil die lokalen Selbsteinschätzungen eher das unmittelbare Gefühl der Aufgabenbearbeitung spiegeln, während die globalen Selbsteinschätzungen stärker durch ein allgemeines und stabiles Selbstkonzept bestimmt sind.
3) In Klassenstufe 7 ist eine signifikant höhere Monitoring-Präzision gegeben als in Klassenstufe 3. Das heißt:
   a) Die DIFF-G ist in Klassenstufe 7 signifikant niedriger als in Klassenstufe 3.
   b) In Klassenstufe 7 ist ein signifikant höherer SE-Score gegeben als in Klassenstufe 3.

4 Ergebnisse

4.1 Deskriptive Statistik

Tabelle 2 gibt zentrale Kenngrößen der Messwerte wieder. Da sowohl die Aufgabenanzahl als auch die Anzahl der Selbsteinschätzungen pro Testheft schwankten, wurden alle Messwerte als Prozentwerte ausgegeben. Lediglich der SE-L-Score gibt keinen Prozentwert wieder, sondern ein Verhältnis von korrekten zu inkorrekten Einschätzungen zwischen -1 und +1 (siehe oben).

TL-G: Der Mittelwert unter .5 und die leichte positive Schiefe zeigen, dass die Tests im Mittel etwas zu schwer ausgefallen sind. Die negative Wölbung (Kurtosis) deutet auf eine flachgipflige Verteilung hin. Ein probeweise durchgeführter K-S-Test zeigte eine signifikante Abweichung von der Normalverteilung an, sodass in späteren Gruppenanalysen, insbesondere beim Vergleich der Leistungen der dritten und siebten Klasse, ausschließlich mit nichtparametrischen Tests gearbeitet wurde. Auch für alle anderen im Folgenden diskutierten Messwerte kann keine Normalverteilung vorausgesetzt werden.

2014, 443). Bei gegebener Itemstruktur (zu wenige Items und daraus resultierend zu viele Proband:innen mit nur richtigen oder falschen Items, was eine Division durch Null ergibt) konnten weitere Bias reduzierende Maße wie Sensitivity, Specificity, Odds-Ratio; Gamma, Kappa, Meta-d' (Schraw et al. 2013) nicht berechnet werden (für Gamma Lingel et al. 2019, 594).

Tab. 2: Deskriptive Werte

	N	Min.	Max.	Mean	Std. abw.	Schiefe	Kurtosis
TL-G	1184	0	94	41.69	18.04	.274	-.542
SE-G	890	0	100	65.86	24.01	-.866	.227
SÜ-G	890	-84	91	22.34	25.01	-.312	.572
DIFF-G	890	0	91	27.84	18.69	.695	-.122
TL-L	1165	0	100	54.90	29.45	-.160	-.792
SE-L	1165	0	150	70.84	34.20	-.539	-.570
SÜ-L	1165	-100	150	15.94	40.42	.108	.195
DIFF-L	1165	0	150	32.44	28.94	.873	.457
Anteil korrekter SE	1166	.00	1.00	.6051	.2816	-.303	-.611
Anteil inkorrekter SE	1166	.00	1.00	.3949	.2816	.303	-.611
Anteil Hits	1166	.00	1.00	.4252	.312	.221	-.940
Anteil False Alarms	1166	.00	1.00	.1234	.199	1.751	3.03
Anteil Correct Rejections	1166	.00	1.00	.1798	.244	1.293	1.02
Anteil Missings	1166	.00	1.00	.2715	.277	.821	-.080
Anteil zuversichtlicher SE	1166	.00	1.00	.6967	.328	-.732	-.651
Anteil nicht zuversichtlicher SE	1166	.00	1.00	.3033	.328	.732	-.651
SE-Score	1166	-1.00	1.00	.2102	.563	-.303	-.611

Test- und Selbsteinschätzungsleistung (global): Dass der Mittelwert der SE-G deutlich höher ist als der TL-G und auch die Schiefe negativ ausfällt, verdeutlicht, dass sich die Proband:innen am Ende des Tests durchschnittlich zu positiv einschätzten. Da vereinzelte Selbstunterschätzungen das Maß an Selbstüberschätzung der Gesamtstichprobe etwas reduzieren, fällt DIFF-G noch etwas deutlicher aus als die SÜ-G.

Bei der Interpretation der SE-G muss berücksichtigt werden, dass für einen nicht unerheblichen Teil der Versuchspopulation keine Antworten vorliegen, was insbesondere die 3. Klasse betrifft, in der 246 Proband:innen die abschließende Frage „Was meinst du: Wie viele Aufgaben hast du richtig bearbeitet?" nicht beantworteten – mehrheitlich, weil sie das Ende des Tests nicht erreichten. Da die Testleistung dieser Teilpopulation signifikant unterhalb des Gesamtdurchschnitts lag, muss an dieser Stelle mit Bias gerechnet werden, der möglicherweise auf Zeitdruck oder die im Rahmen von Kompetenzmessung ungewohnte Aufforderung zurückzuführen ist. Die Ergebnisse decken sich mit den bekannten Befunden (Kruger & Dunning 1999), dass schwächere Lernende der Tendenz nach schlechtere Selbsteinschätzungsleistungen zeigen. Der hier vorliegende Ausfall von Daten, der eben vor allem die schwächeren Lernenden betrifft, führt also mutmaßlich zu einer Überschätzung der Monitoringpräzision der jüngeren Proband:innen im Vergleich zu der nicht von diesem Ausfall betroffenen älteren Population.

Test- und Selbsteinschätzungsleistung (lokal): Die TL-L liegt insgesamt etwas oberhalb der TL-G und hat eine höhere Standardabweichung, was bedeutet, dass die zugehörigen Testaufgaben tendenziell zu den leichteren gehörten, aber auch eine höhere Streuung aufwiesen. Die Werte von SE-L und SÜ-L zeigen, dass die Versuchspopulation ihre Leistung unmittelbar im Anschluss an die Bearbeitung der jeweiligen Testaufgabe ebenfalls

tendenziell zu positiv einschätzte, der Selbstüberschätzungseffekt aber insgesamt etwas geringer ausfiel als bei den globalen Selbsteinschätzungen. Dennoch war die lokale Selbsteinschätzung keineswegs adäquater als die globale, was der Wert zum Selbsteinschätzungsfehler DIFF-L zeigt. Im Gegenteil war der lokale DIFF-L sogar problematischer als der globale DIFF-G, wie die jeweiligen Abweichungen zeigen. Dieser zunächst paradox anmutende Umstand wird dadurch verursacht, dass die durchschnittliche Fehleinschätzung auf lokaler Ebene stärker streute, sodass hier auch mehr Selbstunterschätzungen auftraten. Dieser Umstand verursacht auch die höheren Standardabweichungen für *SÜ-L* und *DIFF-L*.

4.2 Korrelationsanalysen

Tabelle 3 stellt die Korrelationsmatrix aus globaler Testleistung und den globalen Selbsteinschätzungen dar. Sie zeigt zunächst eine knapp mittelhohe positive Korrelation von TL-G und SE-G sowie eine mittelhohe negative Korrelation mit SÜ-G und DIFF-G. Sämtliche Korrelationen sind mit höchstens 0.1-prozentiger Irrtumswahrscheinlichkeit sehr gut gegen den Zufall abgesichert. Die Befundlage deutet somit darauf hin, dass höhere Testleistungen systematisch mit adäquateren Selbsteinschätzungen einhergehen. Dass die Testleistung dabei am höchsten mit der Variable des absoluten Selbsteinschätzungsfehlers DIFF-G korreliert (r = -.433**), zeigt, dass der Zusammenhang nicht ausschließlich als Artefakt der systematisch zu positiven Selbsteinschätzung der Versuchspopulation zu werten ist, die ja Personen mit tatsächlich hohen Testleistungen systematisch begünstigen würde (Juslin et al. 2000). Vielmehr muss davon ausgegangen werden, dass mit zunehmender Leistung auch die Selbsteinschätzungen tendenziell adäquater werden.

Tab. 3: Korrelationen globale Testebene

		SE-G	SÜ-G	DIFF-G
TL-G	Pearson Korrelation	.322**	-.418**	-.433**
	Sig. (2-tailed)	.000	.000	.000
	N	890	890	890
SE-G	Pearson Korrelation		.725**	.422**
	Sig. (2-tailed)		.000	.000
	N		890	890
SÜ-G	Pearson Korrelation			.720**
	Sig. (2-tailed)			.000
	N			890

Anmerkungen: **: Die Korrelation ist auf 0,01-Niveau signifikant (zweiseitig).

Die Betrachtung der lokalen Selbsteinschätzungen zeigt eine ganz ähnliche Befundlage. Die wichtigsten Korrelationen sind in Tabelle 4 dargestellt. Anders als auf globaler Ebene korreliert SÜ-L stärker mit TL-L als DIFF-L mit TL-L. Dass damit eine geringere Selbstüberschätzung höher mit einer steigenden Testleistung korreliert (-.559**) als der eigentliche Fehlerwert der Selbsteinschätzung DIFF-L (-.453**), spricht für ein etwas höheres Maß an Bias in SÜ-L. Dafür spricht auch, dass der gegen Bias besser abgesicherte

SE-L-Score ebenfalls mit .453** auf sehr ähnlichem Niveau mit der TL-L korreliert, und mit der DIFF-L (-.739**) deutlicher negativ korreliert als mit der SÜ-L (-.328**).

Tab. 4: Korrelationen

		SE-L	SÜ-L	DIFF-L	SE-L-Score
TL-L	Pearson-Korrelation	.200**	-.559**	-.453**	.453**
	Sig. (2-seitig)	.000	.000	.000	.000
	N	1165	1165	1165	1165
SE-L	Pearson-Korrelation		.700**	.002	.002
	Sig. (2-seitig)		.000	.935	.935
	N		1165	1165	1165
SÜ-L	Pearson-Korrelation			-.328**	-.328**
	Sig. (2-seitig)			.000	.000
	N			1165	1165
DIFF-L	Pearson-Korrelation				-.739**
	Sig. (2-seitig)				.000
	N				1165

Anmerkungen: **: Die Korrelation ist auf 0,01-Niveau signifikant (zweiseitig).

Insgesamt kann zunächst festgehalten werden, dass TL-L und SE-L-Score praktisch denselben korrelativen Zusammenhang wie auf globaler Ebene zeigen.[2]

4.3 Klassenstufeneffekte

Da eines der Hauptinteressen dieser Untersuchung in der Frage liegt, ob und inwieweit sich (meta)kognitive Leistungen mit zunehmendem Alter der Versuchspersonen zunehmend entfalten, werden im Folgenden die Leistungs- und Verhaltensvariablen der beiden Teilkohorten der dritten bzw. siebten Klasse im Vergleich dargestellt. Tabelle 5 gibt zunächst Mittelwerte und Standardabweichungen beider Teilpopulationen im Vergleich zur Gesamtstichprobe wieder.[3] Die Gruppenunterschiede wurden mithilfe von Mann-Whitney-Tests auf Signifikanz geprüft. Die zugehörigen Ergebnisse werden in der folgenden Darstellung zusammen mit den jeweiligen Signifikanzniveaus wiedergegeben. Auf eine Darstellung der zugehörigen Prüfstatistiken wird verzichtet.

Die TL-G und TL-L in Tabelle 5 dürfen nicht als tatsächliche Leistungsunterschiede der Teilpopulationen interpretiert werden, da die Erhebungen in beiden Teilpopulationen mit unterschiedlichen Testaufgaben erfolgten und damit nicht im Detail verglichen werden

[2] Dass der Korrelationskoeffizient für den Selbsteinschätzungsscore positiv ausfällt, liegt daran, dass der Discrimination Score kein Abweichungsmaß darstellt. Es gilt also: Je höher der Wert, desto höher die Präzision der Einschätzung.

[3] Angesichts der fehlenden Normalverteilung erscheint es zunächst unsinnig, die Mittelwerte und nicht die Mediane der Teilpopulationen wiederzugeben. Die Entscheidung begründet sich aus der Anschaulichkeit der Mittelwerte, die besser zur Befundlage der durchgeführten Mann-Whitney-Tests passt. So zeigte etwa die Abweichung von Leistung und Selbsteinschätzung für beide Klassen denselben Median, während der Mann-Whitney-Test einen signifikanten Gruppenunterschied belegte (s. u.). Bei der Qualität der lokalen Selbsteinschätzung war es genau umgekehrt: Hier schien der Median Gruppenunterschiede zu indizieren, die im Mann-Whitney-Test nicht bestätigt werden konnten.

können. Die Leistungswerte zeigen aber, dass der prozentuale Anteil richtig gelöster Leistungsaufgaben in beiden Klassen ähnlich hoch lag, sodass nicht mit größeren Verzerrungen durch stark unterschiedliche Testschwierigkeiten und entsprechend unterschiedliche motivationale Dispositionen gerechnet werden muss.

Tab. 5: Testleistungen und Selbsteinschätzungen nach Klassenstufen

Klassenstufe	3			7			Total		
	Mean	N	Std. Abw.	Mean	N	Std. Abw.	Mean	N	Std. Abw.
TL-G	43.11	623	.187	40.11	561	.172	41.69	1184	.180
DIFF-G	26.08	377	.181	29.14	513	.190	27.84	890	.187
TL-L	58.14	605	.308	51.40	560	.275	54.90	1165	.294
SE-L-Score	.2021	606	.593	.2189	560	.529	.2102	1166	.563
Hits	.4759	606	.325	.3704	560	.288	.4252	1166	.312
Correct Rejections	.1251	606	.210	.2390	560	.263	.1798	1166	.244
Zutreffende SE gesamt	.6011	606	.296	.6094	560	.265	.6051	1166	.282
Missings	.2933	606	.297	.2478	560	.253	.2715	1166	.277
False Alarms	.1056	606	.199	.1427	560	.198	.1234	1166	.199
Unzutreffende SE gesamt	.3989	606	.297	.3906	560	.265	.3949	1166	.282
Zuversichtliche SE gesamt	.7693	606	.308	.6182	560	.331	.6967	1166	.328
Unzuversichtliche SE gesamt	.2307	606	.308	.3818	560	.331	.3033	1166	.328

Dagegen sind die Messwerte der dritten und siebten Klasse für die Qualität der globalen bzw. lokalen Selbsteinschätzung (DIFF-G bzw. SE-L-Score) durchaus vergleichbar. Der Unterschied zwischen dritter und siebter Klasse für DIFF-G erweist sich anhand eines Mann-Whitney-Tests zwar als signifikant ($p = .015$), muss aber dennoch mit äußerster Vorsicht interpretiert werden, da es in Klasse 3 zu größeren und mutmaßlich unzufällig auftretenden Datenausfällen gekommen war, die insbesondere im unteren Leistungsspektrum auftraten. Verbunden mit den in Abschnitt 3.2 dargestellten Korrelationen von Testleistung und Selbsteinschätzung kann damit plausibel vermutet werden, dass der Messwert DIFF-G in Tabelle 5 die Abweichung der Testleistung (global) und der tatsächlichen Selbsteinschätzung (global) in der dritten Klasse systematisch zu niedrig einschätzt und den Gruppenunterschied zwischen dritter und siebter Klasse folglich systematisch überschätzt (siehe Ergebnisdarstellung SE-G).

Für diese Interpretation spricht auch der lokale SE-L-Score, der nicht von dem genannten Datenausfall betroffen war und keinen signifikanten Unterschied zwischen den Klassen belegt. Der zugehörige Mann-Whitney-Test weist mit einer Irrtumswahrscheinlichkeit von $p = .989$ die numerischen Unterschiede zwischen beiden Kohorten als höchstwahrscheinlich zufällig aus. Verbunden mit den Befunden aus Kap. 3.1, in denen globale und lokale Selbsteinschätzungen durchaus vergleichbare Effekte gezeigt hatten, sprechen die Daten damit insgesamt eher gegen als für einen Alterseffekt bei der Fähigkeit zur adäquaten Selbsteinschätzung.

Während sich die Selbsteinschätzungen beider Kohorten damit quantitativ nicht gesichert unterscheiden, deuten die Werte ab Zeile 5 aus Tabelle 5 (Hits) durchaus auf qualitative Unterschiede im Antwortverhalten beider Klassen hin. So legen die jüngeren Versuchspersonen ein deutlich zuversichtlicheres Antwortverhalten an den Tag, das sich sowohl in einem höheren Anteil von Hits als auch in einem höheren Anteil von Missings niederschlägt, während die siebte Klasse die eigene Leistung insgesamt deutlich pessimistischer einschätzt, was im Detail sowohl zu einem höheren Anteil von False Alarms als auch zu einem höheren Anteil von Correct Rejections führt. Die zugehörigen Mann-Whitney-Tests weisen diese Gruppenunterschiede als höchst signifikant ($p < .001$)[4] aus.

Tab. 6: Korrelationen

		DIFF-G	
		3. Klasse	7. Klasse
TL-G	Pearson Korrelation	-.468**	-.398**
	Sig. (2-tailed)	.000	.000
	N	377	513

Tab. 7:Korrelationen

		SE-L-Score	
		3. Klasse	7. Klasse
TL-L	Pearson Korrelation	.602**	.262**
	Sig. (2-tailed)	.000	.000
	N	605	560

Insgesamt korrelieren Testleistung und Präzision der Selbsteinschätzung in der dritten Klasse etwas stärker. Im Fall der globalen Selbsteinschätzung muss wiederum von der Existenz eines Verzerrungseffekts ausgegangen werden, da insbesondere schwächere Testleistungen von einem Ausfall der Selbsteinschätzung betroffen sind (siehe oben). Folglich zeigt sich, dass die Stärke der Korrelation für die lokalen Selbsteinschätzungen, die mit dem Selbsteinschätzungsscore die Diskriminierungsfähigkeit zwischen erfolgreicher und nicht erfolgreicher Aufgabenbearbeitung stärker betont, in noch deutlicherem Maß abnimmt. Von der dritten zur siebten Klasse hin zeigt sich also insgesamt eine Tendenz zur Entkoppelung der Präzision der Selbsteinschätzung von der Leistungsfähigkeit.

5 Diskussion

In der Gesamtbetrachtung zeigt sich für das metakognitive Monitoring im Rahmen von Sprachbetrachtung eine grundsätzliche Selbstüberschätzung auf allen Ebenen. Dies entspricht dem Stand der Forschung (Schneider et al. 2022, 285). Die Erwartungen a, b und c in Punkt 1 (siehe Abschnitt 2.2) wurden in jeder Hinsicht klar bestätigt. Tatsächlich

[4] Eine Ausnahme bildet lediglich die Variable *Missings*, bei der das Signifikanzniveau auf $p = .021$ verbleibt.

korrelieren die Selbstüberschätzungen SÜ-G und SÜ-L (nahezu) stark negativ mit der Testleistung (global: -.418**, lokal: -.559**). Auch die Selbsteinschätzungsfehler DIFF-G und DIFF-L korrelieren eindeutig negativ – mittel bis nahezu stark – mit den Testwerten (global: -.433**, lokal: -.453**). Ebenso der lokale Selbsteinschätzungsscore SE-L-Score: Hier zeigt sich mit .453** eine sehr ähnliche mittlere bis starke Korrelation mit TL-L. Insgesamt deuten diese Werte darauf hin, dass sich die Tendenz zur Selbstüberschätzung mit zunehmender Leistungsfähigkeit deutlich reduziert. Der Unskilled-and-unaware-of-it-Effekt (Kruger & Dunning 1999) zeigt sich also auch in den vorliegenden Daten. Offen bleibt, ob sich in diesen Daten nicht möglicherweise auch spiegelt, dass leistungsschwächere Proband:innen in Kompetenztestungen weniger motiviert sind (Bol & Hacker 2012, 3).

In Hinblick auf Punkt 2 der Erwartungen zeigt ein erster Vergleich der Maße an Selbstüberschätzung, der Erwartung entsprechend, dass das Maß der globalen Selbstüberschätzung SÜ-G mit 22.34 höher ausfiel als für die lokale Ebene SÜ-L (15.94). Allerdings dreht sich dieser Eindruck um, wenn der absolute Einschätzungsfehler DIFF herangezogen wird – wenn zu pessimistische Selbsteinschätzungen also nicht wertmindernd auf den Durchschnittswert wirken. Folglich fällt DIFF-G mit 27.84 sogar niedriger aus als DIFF-L (32.44), bei dem sich auch eine deutlich größere Standardabweichung zeigt. Dies lässt die Interpretation zu, dass die lokalen Selbsteinschätzungen direkt nach Aufgabenbearbeitung zwar nicht präziser sind als die globalen, sehr wohl aber weniger eindeutig positiv gefärbt, also mit mehr Unsicherheitsgefühl behaftet sind, insofern also wohl tatsächlich stärker vom Eindruck der gerade ausgeführten Testaufgabenbearbeitung, und weniger von einem allgemeinen Selbstkonzept beeinflusst sind (Händel et al. 2020, 71). Dafür spricht auch die auf lokaler Ebene deutlicher ausgeprägte Korrelation zwischen Testleistung und Selbstüberschätzung (lokal: -.559**, global: -.418**). Allerdings zeigen die absoluten Fehlerwerte DIFF, die auf lokaler und globaler Ebene sehr ähnlich mit der Testleistung korrelieren (lokal: -.453**, global: -.433**), dass sich trotz eines Anstiegs negativer Einschätzungen die Präzision der lokalen Selbsteinschätzung gegenüber der globalen Einschätzung nicht erhöht.

Erwartung 3a war davon ausgegangen, dass sich die Präzision der globalen Selbsteinschätzung von der dritten auf die siebte Klasse erhöhen würde. Tatsächlich fällt der absolute Einschätzungsfehler DIFF-G in der dritten Klasse allerdings mit 26,08 sogar etwas niedriger aus als in der siebten Klasse mit 29,14. Problematisch an diesen Werten ist die hohe Ausfallquote im Fall der Selbsteinschätzung in der dritten Klasse, die daraus resultiert, dass vor allem schwächere Proband:innen das Ende des Tests nicht erreichten oder nicht wussten, was im Fall der Selbsteinschätzungsfrage zu tun war. Da alle anderen Daten darauf hinweisen, dass Selbsteinschätzungen im Fall schwächerer Proband:innen besonders unpräzise ausfallen, ist davon auszugehen, dass der Präzisionsvorteil der dritten Klasse gegenüber der siebten Klasse in unbestimmbarem Maß aus diesem Verzerrungseffekt resultiert. Trotzdem muss festgestellt werden, dass die Erwartung zur höheren Selbsteinschätzungspräzision der siebten Klasse nicht eingetreten ist. Von weiteren Interpretationen dieses Faktums sollte in Anbetracht der unsicheren Datenlage abgesehen werden.

Für die Überprüfung von Annahme 3b wurde der SE-L-Score berechnet, der für die dritte Klasse einen Messwert von .2021 und für die siebte Klasse einen etwas höheren von .2189 aufwies. Dieser Unterschied in den Scores ist mit p = .989 sehr deutlich nicht signifikant. Es konnte also kein präzisionssteigernder Alterseffekt beobachtet werden.

In Zusammenhang mit dem Fehlen eines eindeutigen Alterseffekts sei darauf hingewiesen, dass die zusammenfassende Forschungsliteratur in Bezug auf die Entwicklung der Monitoring-Fähigkeiten von einer uneindeutigen Forschungslage spricht und eher davon ausgeht, dass in den späteren Primarstufenjahren keine deutlichen Verbesserungen zu beobachten sind. Sehr wohl aber ist zu beobachten, dass ältere Proband:innen allgemein zu pessimistischeren Selbsteinschätzungen tendieren (Schneider et al. 2022, 286). Dies zeigt sich auch in den hier vorliegenden Ergebnissen deutlich: Von der dritten auf die siebte Klasse reduzieren sich die zuversichtlichen Selbsteinschätzungen von ca. 77 % auf 62 %. Insgesamt ist also eine Verschiebung hin zu einem weniger zuversichtlichen Antwortverhalten zu finden, das interessanterweise trotzdem nicht zu höherer Präzision führt. Denn im gleichen Maß, wie sich die korrekt-pessimistischen Einschätzungen (Correct Rejections) von der dritten zur siebten Klasse erhöhen, sinkt der Anteil der korrekt-optimistischen Selbsteinschätzungen (Hits) von der dritten zur siebten Klasse. Es gibt im Detail also einen Trend zu mehr Pessimismus, aber keinen Trend zu mehr Präzision.

Für die weiteren Korrelationsanalysen kann festgestellt werden, dass sich die Testleistung und die Präzision der Selbsteinschätzung von der dritten zur siebten Klasse etwas entkoppeln. So korreliert der Selbsteinschätzungsscore mit der Testleistung in der dritten Klasse mit .602** stark, in der siebten Klasse aber mit .262** nur noch schwach. Eine mögliche Interpretation dieser Beobachtung ist, dass mit zunehmendem Alter alle Leistungsgruppen lernen, sich etwas besser einzuschätzen, da die schwächeren Proband:innen ebenso zunehmend Lernerfahrungen und Expertise erworben haben und im Fall der besseren Proband:innen bereits ein Plateau erreicht wurde. Die Reduktion der Korrelation zwischen Testleistung und Präzision der Selbsteinschätzung könnte also darauf zurückzuführen sein, dass die schwächeren Lernenden in Bezug auf die Präzision der Selbsteinschätzung mit zunehmendem Alter zu den besseren Lernenden aufholen. Diese Einschätzung deckt sich mit den Daten von van Loon et al. (2022, 5).

6 Fazit und Ausblick

Im Zuge der Pilotierung neuer Testaufgaben im Kompetenzbereich Deutsch Sprachbetrachtung/Sprachbewusstsein der iKM^{PLUS} wurden Formate zur Erhebung deklarativ-metakognitiven Wissens und prozedural-metakognitiver Fähigkeiten erprobt. In diesem Beitrag wurden die daraus resultierenden Ergebnisse der metakognitiven Monitoring-Aufgaben als Indikator für prozedural-metakognitive Fähigkeiten berichtet.

Insgesamt zeigte sich, dass die Operationalisierung von Monitoring-Aufgaben im Rahmen der Kompetenzmessung Deutsch Sprachbetrachtung möglich ist und erwartbare Ergebnisse produziert. Dabei erscheint es einerseits vor dem Hintergrund der wenig spezifischen Ergebnisse zum Zusammenhang von Testleistung und Selbsteinschätzung derzeit ein weiter Weg, bis eine metakognitive Monitoring-Fähigkeit in ein

kompetenzmessendes Konstrukt von Sprachbewusstheit integriert werden kann. Überlegungen zur „metakognitiven Kompetenz“ und zum „inneren Monitor“ als Teil von Sprachbewusstheit (Ossner 2007) bleiben derzeit also theoretischer Natur. Andererseits erscheint es aus didaktischer Perspektive durchaus ratsam, Lehrkräften Metakognitionswerte für diagnostische Interpretationen zur Verfügung zu stellen – allein schon, um damit eine Fokusverschiebung auf metakognitive Aspekte zu erzielen, was hohes didaktisches Potenzial hätte (Schneider 2015, 307). Es könnte sich daher trotzdem lohnen, Werte zu metakognitiven Monitoringfähigkeiten im Kompetenzmessungsparadigma mitzuerheben und an Lehrende und Lernende für die kontextualisierenden Ergebnisrückmeldungen weiterzuleiten.

7 Literatur

Artelt, C., Beinicke, A., Schlagmüller, M. & Schneider, W. (2009). Diagnose von Strategiewissen beim Textverstehen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *41*(2), 96–103. https://doi.org/10.1026/0049-8637.41.2.96

Artelt, C. & Schneider, W. (2015). Cross-Country Generalizability of the Role of Metacognitive Knowledge in Students' Strategy Use and Reading Competence. *Teachers College Record: The Voice of Scholarship in Education*, *117*(1), 1–32. https://doi.org/10.1177/016146811511700109

Bol, L. & Hacker, D. J. (2012). Calibration research: where do we go from here? *Frontiers in psychology*, *3*, 229. https://doi.org/10.3389/fpsyg.2012.00229

Brown, A. L., Bransford, J. D., Ferrara, R. A. & Campione, J. C. (1983). Learning, remembering and understanding. In J. H. Flavell & E. M. Markham (Hrsg.), *Handbook of Child Psychology: Cognitive Development* (S. 77–166). Wiley.

Chen, S. & McDunn, B. A. (2022). Metacognition: History, measurements, and the role in early childhood development and education. *Learning and Motivation*, *78*, 101786. https://doi.org/10.1016/j.lmot.2022.101786

Destan, N. & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, *10*(3), 347–374. https://doi.org/10.1007/s11409-014-9133-z

Flavell, J. H. (1979). Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry. *American Psychologist*, *34*(10), 906–911.

Fleming, S. M. & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Funke, R. (2005). *Sprachliches im Blickfeld des Wissens: Grammatische Kenntnisse von Schülerinnen und Schülern*. Vollst. zugl.: Flensburg, Univ., Habil.-Schr., 2002. Niemeyer.

Geurten, M., Meulemans, T. & Lemaire, P. (2018). From domain-specific to domain-general? The developmental path of metacognition for strategy selection. *Cognitive Development*, *48*, 62–81. https://doi.org/10.1016/j.cogdev.2018.08.002

Gornik, H. (2022). Sprachreflexion, Sprachbewusstheit, Sprachwissen, Sprachgefühl und die Kompetenz der Sprachthematisierung: ein Einblick in ein Begriffsfeld. In H. Gornik & I. Rautenberg (Hrsg.), *Deutschunterricht in Theorie und Praxis (DTP): Bd. 6. Sprachreflexion und Grammatikunterricht* (2. überarbeitete und erweiterte Aufl., S. 39–55). Schneider Hohengehren.

Händel, M., Artelt, C. & Weinert, S. (2013). Assessing metacognitive knowledge: Development and evaluation of a test instrument. *Journal for educational research online*, *5*(2), 162–188. https://doi.org/10.25656/01:8429

Händel, M., Bruin, A. B. H. de & Dresel, M. (2020). Individual differences in local and global metacognitive judgments. *Metacognition and Learning*, *15*(1), 51–75. https://doi.org/10.1007/s11409-020-09220-0

Illetschko, M., Unterholzner, F., Österbauer, V., Oberauner, V., Winter, S., Greil, T., Krelle, M., Jost, J., Fladung, I., Hoffmann, L., Kremmel, B. & Eberharter, K. (2025). *Kompetenzmodellierung in der iKM^{PLUS}.* IQS – Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen. https://doi.org/10.17888/IQSREPORT-2025-1

Juslin, P., Winman, A. & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: a critical examination of the hard-easy effect. *Psychological review*, *107*(2), 384–396. https://doi.org/10.1037/0033-295x.107.2.384

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121–1134. https://doi.org/10.1037//0022-3514.77.6.1121

Lingel, K., Lenhart, J. & Schneider, W. (2019). Metacognition in mathematics: do different metacognitive monitoring measures make a difference? *ZDM*, *51*(4), 587–600. https://doi.org/10.1007/s11858-019-01062-8

Loaiza, Y., Patiño, M., Umaña, O. & Duque, P. (2023). What is New in Metacognition Research? Answers from Current Literature. *Educación y Educadores*, *25*(3), Artikel e2535, 1–24. https://doi.org/10.5294/edu.2022.25.3.5

Müller, H.-G. & Unterholzner, F. (2025). How to stumble purposefully: Metacognitive and metalinguistic self-regulation on the path from knowledge to skill. *Pedagogical Linguistics.* Vorab-Onlinepublikation. https://doi.org/10.1075/pl.24013.mul

Nelson, T. O. & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. *The Psychology of Learning and Motivation: Advances in Research and Theory*(26), 125–173.

Ossner, J. (2007). Sprachbewusstheit: Anregung des inneren Monitors. In H. Willenberg (Hrsg.), *Kompetenzhandbuch für den Deutschunterricht* (S. 134–147). Schneider Hohengehren.

Schneider, W. (2015). *Memory development from early childhood through emerging adulthood*. Springer. https://doi.org/10.1007/978-3-319-09611-7

Schneider, W. & Löffler, E. (2015). The Development of Metacognitive Knowledge in Children and Adolescents. In J. Dunlosky & S. K. Tauber (Hrsg.), *Oxford library of psychology: Bd. 1. The Oxford handbook of metamemory* (S. 491–519). Oxford University Press.

Schneider, W., Tibken, C. & Richter, T. (2022). The development of metacognitive knowledge from childhood to young adulthood: Major trends and educational implications. In J. J. Lockman (Hrsg.), *Advances in Child Development and Behavior* (Bd. 63, S. 273–307). Elsevier. https://doi.org/10.1016/bs.acdb.2022.04.006

Schraw, G. & Dennison, R. S. (1994). Assessing Metacognitive Awareness. *Contemporary Educational Psychology*, *19*(4), 460–475. https://doi.org/10.1006/ceps.1994.1033

Schraw, G., Kuch, F. & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction*, *24*, 48–57. https://doi.org/10.1016/j.learninstruc.2012.08.007

Seeger, J., Lenhard, W. & Wisniewski, K. (2021). Metakognitives Strategiewissen in sprachbezogenen Situationen: Interne Struktur und Validität des ScenEx. *Diagnostica*, Artikel 0012-1924/a000275, 1–11. https://doi.org/10.1026/0012-1924/a000275

Unterholzner, F. & Müller, H.-G. (2023). Metakognition als Brücke zwischen sprachlichem Wissen und Können. *Didaktik Deutsch*, *28*(55), 20–38. https://doi.org/10.21248/dideu.677

van Loon, M. H., Bayard, N. S., Steiner, M. & Roebers, C. M. (2022). The accuracy and annual rank-order stability of elementary school children's self-monitoring judgments. *Journal of Applied Developmental Psychology*, *80*, 1–9. https://doi.org/10.1016/j.appdev.2022.101419

van Loon, M. H., Orth, U. & Roebers, C. (2024). The structure of metacognition in middle childhood: Evidence for a unitary metacognition-for-memory factor. *Journal of experimental child psychology*, *241*, 105857. https://doi.org/10.1016/j.jecp.2023.105857

Vuorre, M. & Metcalfe, J. (2022). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning*, *17*(2), 269–291. https://doi.org/10.1007/s11409-020-09257-1

8 Korrespondenzangaben

HS-Prof. Dr. Franz Unterholzner, Pädagogische Hochschule Salzburg, Institut für Fachdidaktiken und Fachwissenschaften (IQS bis 2025-09)

Prof. Dr. Hans-Georg Müller, Universität Potsdam, Institut für Germanistik

Korrespondenz an: franz.unterholzner@phsalzburg.at