Verstehen von Argumenten in wissenschaftlichen Texten: Reliabilität und Validität des Argumentstrukturtests (AST)

zur Veröffentlichung angenommen bei der Zeitschrift *Diagnostica* (2018)

Hannes Münchow, Tobias Richter, Sarah von der Mühlen,
Universität Würzburg

Sebastian Schmid
Universität Regensburg

Katherine E. Bruns & Kirsten Berthold
Universität Bielefeld

Autorenhinweis

Der Argumentstrukturtest ist bis auf Weiteres für den nicht-kommerziellen Einsatz in Forschung und Lehre über den Erst- oder den Zweitautor erhältlich. Die hier dargestellte Forschung wurde durch das Bundesministerium für Bildung und Forschung (FKZ 01PK11017B und 01PK15009B) gefördert. Wir danken Frau Elisabeth Schmidt für hilfreiche Diskussionen zu dem Instrument in einer früheren Projektphase.

Zusammenfassung

Informelle Argumente sind in wissenschaftlichen Texten allgegenwärtig. Um solche Argumente verstehen und bewerten zu können, müssen Studierende ihre Struktur entschlüsseln. Zur Erfassung dieser Kompetenz wurde der computergestützte Argumentstrukturtest (AST) für Studierende sozial- und erziehungswissenschaftlicher Fächer sowie Lehramtsstudierende entwickelt. Die Testpersonen lesen kurze Texte mit informellen Argumenten und identifizieren ihre funktionalen Komponenten (z. B. Behauptung, Begründung, Schlussregel). Anhand einer Stichprobe von 225 Studierenden wurde der AST einer ersten Überprüfung seiner Reliabilität und Validität unterzogen. Dabei erwies sich der AST als intern valide, mit einer breiten Streuung der Itemschwierigkeiten. In einem explanatorischen Item-Response-Modell konnten die Itemschwierigkeiten sehr präzise durch theoretisch relevante Itemmerkmale, die das Argumentverstehen beeinflussen, vorhergesagt werden. Korrelationen mit verbaler Intelligenz und Schul- und Studienleistungen sprechen darüber hinaus für die Kriteriumsvalidität des Instruments.

Schlüsselwörter: Argumentkomponenten, Argumentstruktur, Argumentverstehen, epistemische Kompetenzen, Studierende

Abstract

Informal arguments are omnipresent in scientific texts. In order to understand and evaluate such arguments, students have to decode their structure. To measure this competency, the computer-assisted argument structure test (AST) was developed for students of social and educational sciences as well as student teachers. The test takers read short texts containing informal arguments and identify their functional components (e. g., claim, reason, warrant). On the basis of a sample of 225 students, the AST was subjected to a first examination of its reliability and validity. The AST proved to be reliable, with a wide range of item difficulties. In an explanatory item response model, the item difficulties were predicted very precisely through theoretically relevant item features that are known to influence argument comprehension. Correlations with verbal intelligence as well as school and study performance provided evidence for the criterion validity of the instrument.

Key words: argument components, argument comprehension, argument structure, epistemic competencies, university students

Der Umgang mit wissenschaftlicher Fachliteratur ist in nahezu allen Studienfächern von großer Bedeutung. Neuere Konzeptionen von Scientific Literacy berücksichtigen dabei neben rezeptiven Prozessen der Informationsaufnahme auch epistemische Prozesse, bei denen die Einschätzung der Plausibilität oder Glaubwürdigkeit von Informationen relevant ist (Britt, Richter & Rouet, 2014). Aus diesen Überlegungen heraus lassen sich je nach Verarbeitungsziel und -modus vier unterschiedliche Strategien im Umgang mit wissenschaftlichen Texten ableiten: rezeptiv-systematisch (z. B. Organisieren/Strukturieren von Informationen; Wild, 2000), rezeptiv-heuristisch (z. B. Scannen eines Texts auf der Suche nach bestimmten Informationen; Bazerman, 1985), epistemisch-systematisch (z. B. Bewertung der argumentativen Konsistenz; Richter & Schmid, 2010) und epistemisch-heuristisch (z. B. Nutzung von Quelleninformationen; Korpan, Bisanz, Bisanz & Henderson, 1997). Während rezeptive Strategien bereits intensiv beforscht wurden (z. B. Bazerman, 1985; Wild & Schiefele, 1994), gibt es bislang kaum empirische Untersuchungen zu epistemischen Strategien beim Lesen von wissenschaftlicher Fachliteratur.

Wissenschaftliche Argumente verstehen und bewerten zu können, stellt dabei domänenübergreifend eine wichtige epistemisch-systematische Kompetenz für die Rezeption wissenschaftlicher Originalliteratur dar. Nach dem Argumentationsmodell von Toulmin (1958) bestehen Argumente aus mehreren funktionalen Argumentkomponenten: Behauptung, Gründe, Schlussregel, Stützung der Schlussregel und Einschränkung. Gründe können empirischer, theoretischer oder praktischer Natur sein; sie liefern Belege, die die Behauptung untermauern. Die Schlussregel gibt an, warum die Gründe die Behauptung stützen sollen. Die Stützung der Schlussregel begründet diese empirisch, praktisch oder theoretisch. Die Einschränkung schließlich grenzt den Geltungsanspruch der Behauptung ein (z. B. durch den Verweis auf Ausnahmen).

Ein Schlüssel für das Verstehen und die Bewertung von Argumenten ist die Fähigkeit, die Struktur eines Arguments, das heißt die verschiedenen Argumentkomponenten korrekt zu erkennen (Britt & Larson, 2003; Larson, Britt & Larson, 2004). Zur Erfassung dieser Fähigkeit wurde im Rahmen vorausgehender Studien der Argumentstrukturtest (AST; von der Mühlen, Richter, Schmid, Schmidt & Berthold, 2016) entwickelt und erfolgreich eingesetzt. Dabei zeigte sich, dass Psychologiestudierende zu Beginn ihres Studiums im Vergleich zu wissenschaftlich arbeitenden Psychologinnen und Psychologen Schwierigkeiten in der Unterscheidung dieser Komponenten haben und damit einhergehend auch schlechter darin sind, starke von schwachen Argumenten zu unterscheiden, Argumentationsfehler zu erkennen und Gegenargumente zu entwickeln (von der Mühlen et al., 2016; s. auch Schroeder, Richter & Hoever, 2008; Wu & Tsai, 2007). Stattdessen stützen sie ihre Argumentbewertung auf spontane Einschätzungen der Plausibilität (epistemisches Monitoring; Richter, Schroeder & Wöhrmann, 2009; Schroeder et al., 2008) und vernachlässigen die interne Konsistenz von Argumenten (vor allem die Relevanz und Vollständigkeit der angegebenen Gründe; Shaw, 1996; von der Mühlen et al., 2016). Ziel der vorliegenden Untersuchung war die Überprüfung des AST im Hinblick auf seine Reliabilität, Konstrukt- und Kriteriumsvalidität.

Konzeption und Aufbau des AST

Der AST ist ein computergestütztes Diagnostikum zur Erfassung der Fähigkeit, funktionale Argumentkomponenten zu erkennen und korrekt zuzuordnen. Der Test besteht aus acht kurzen Texten (M = 104 Wörter, SD = 24 Wörter). In jedem dieser Texte wird ein wissenschaftliches Argument präsentiert, das sich in Argumentkomponenten nach Toulmin (1958) gliedert. Jede Argumentkomponente entspricht einem Satz beziehungsweise Teilsatz. Die acht Argumente basieren auf wissenschaftlicher Literatur zu verschiedenen psychologischen Themen und wurden für den AST adaptiert. Britt und Larson (2003) zufolge hat die Abfolge der einzelnen Ar-

gumentkomponenten innerhalb eines Arguments einen Einfluss darauf, wie leicht die Argumentkomponenten identifiziert werden können. Bei den leichteren *Claim-first-Argumenten* steht die Behauptung zu Beginn des Arguments, während bei den schwereren *Reason-first-Argumenten* die Begründung an erster Stelle steht. Der AST beinhaltet jeweils vier Reason-first-Argumente und Claim-first-Argumente. Um sowohl im hohen als auch im niedrigen Kompetenzbereich diskriminieren zu können, besteht der AST zudem aus einfachen und komplexen Argumenten. Zwei der acht Argumente beinhalten lediglich drei der fünf Argumentkomponenten (*einfach*), während die restlichen Argumente alle Argumentkomponenten enthalten (*komplex*). Die beiden einfachen Argumente sind jeweils Claim-first-Argumente.

Beim AST wird den Probandinnen und Probanden ein Argument zuerst als Fließtext dargeboten (s. Abbildung 1A). Anschließend wird das Argument nach Argumentkomponenten absatzweise segmentiert und nummeriert auf dem Bildschirm dargestellt (s. Abbildung 1B). Die Aufgabe der Probandinnen und Probanden besteht darin, bei jedem Argument die nummerierten Sätze jeweils bestimmten Argumentkomponenten zuzuordnen. Da bei allen acht Argumenten jeweils nach allen fünf Argumentkomponenten gefragt wird, enthält der AST insgesamt 40 Items. Um Abhängigkeiten zwischen den Items zu vermeiden, kann jede Argumentkomponente innerhalb eines Argumentes mehrfach zugeordnet werden. Dadurch wirkt sich ein Fehler bei einem Item nicht automatisch auf die Bearbeitung eines anderen Items in demselben Argument aus. Es gibt bei der Bearbeitung keine Zeitbeschränkung. Die Anzahl der korrekt zugeordneten Argumentkomponenten (richtig bearbeitete Items) dient als Testwert, der die Kompetenz widerspiegelt, die Struktur informeller Argumente zu verstehen.

Untersuchungsziele und Hypothesen

Ziel der vorliegenden Untersuchung war es, Reliabilitäts- und Validitätsaspekte des Argumentstrukturtests (AST) zu prüfen. Die deskriptiven Kennwerte des AST sind in Tabelle 1 wiedergegeben.

Zur Prüfung der *Konstruktvalidität* wurden Effekte von Itemmerkmalen untersucht, für die sich aus der Forschung zum Argumentverstehen Vorhersagen zu systematischen Effekten auf die Itemschwierigkeit ableiten lassen. So sollten Items zu einfachen Argumenten besser zu lösen sein als Items zu komplexen Argumenten. Ebenso sollten die Probandinnen und Probanden bei der Bearbeitung von Claim-first-Argumenten im Mittel eine höhere Leistung aufweisen als bei Reason-first-Argumenten. Auf der Ebene der Argumentkomponenten erwarteten wir, dass Gründe und Einschränkungen am besten identifiziert werden können, weil sie in der Regel sprachlich (z. B. durch entsprechende Konnektoren) markiert sind (Larson et al., 2004). Mithilfe eines explanatorischen Item-Response-Modells (Wilson & De Boeck, 2004) wurde geprüft, ob diese systematisch variierten Itemcharateristika die Itemschwierigkeiten beim AST vorhersagen können. Eine Kongruenz der vorgesagten und der empirischen Itemschwierigkeiten wäre ein starker Beleg für die Konstruktvalidität des Tests.

Mit Blick auf den Aspekt der *Kriteriumsvalidität* wurde erwartet, dass positive Zusammenhänge zwischen der Fähigkeit, Argumentkomponenten zu differenzieren und anderen epistemisch-systematischen Kompetenzen (Bewertung der Plausibilität von Argumenten, Erkennen von Argumentationsfehlern) existieren. Zudem sollte die Leistung im AST positiv mit kriterialen Leistungsmaßen wie Schul- und Studienleistungen und der verbalen Intelligenz korrelieren. Zuletzt wurde erwartet, dass es signifikante Zusammenhänge zwischen der Leistung im AST und den epistemologischen Überzeugungen zum Bereich *Psychologie als Wissenschaft* der Probandinnen und Probanden gibt (Stahl & Bromme, 2007). Probandinnen und Probanden mit besserer Leistung im AST sollten wissenschaftlich-psychologisches Wissen eher als strukturiert (vs. unstrukturiert) und als veränderlich (vs. unveränderlich) beschreiben.

--- hier Tabelle 1 einfügen ---

Methode

Stichprobe

An der Validierungsstudie für den AST nahmen 225 Studierende der Universitäten Kassel und Würzburg teil (77 % Frauen, 23 % Männer). Die Studierenden waren im Mittel 23.6 Jahre alt (*SD* = 5.4), die mittlere Studienzeit betrug 3.3 Semester (*SD* = 2.9). Insgesamt studierten 142 (63 %) der getesteten Personen Psychologie, 73 Lehramt (32 %) und 10 (4 %) sonstige Studienfächer (z. B. Zahnmedizin, Soziale Arbeit). Eine Studentin machte keine Angaben über das Studienfach. Die Probandinnen und Probanden wurden unter Verwendung von Online-Rekrutierungssystemen der beiden Universitäten angeworben. Voraussetzung für die Teilnahme an der Studie waren neben dem Studierendenstatus auch ausreichend gute Deutschkenntnisse. Von allen Teilnehmenden gaben 95 % (214) als Muttersprache Deutsch an. Von den restlichen 5 % (11) wurden hauptsächlich Russisch, Spanisch oder Türkisch als Muttersprache angegeben. Eine Person gab ihre Muttersprache nicht an. Bei Ausreißerprüfungen zeigte jedoch keine der betreffenden Personen auffällige Werte in einer oder mehrerer der erhobenen Variablen.

Durchführung der Validierungsstudie

Die Probandinnen und Probanden wurden zu Beginn der etwa 90 Minuten dauernden, computergestützten Untersuchung über Zweck, Dauer und Vorgehen der Studie informiert und gaben schriftlich ihr Einverständnis für die Teilnahme an der Studie (informed consent). Die Untersuchungsteilnahme wurde mit 12 Euro oder (bei Psychologiestudierenden) mit 4 Euro und der Bescheinigung einer Versuchspersonenstunde vergütet.

Die Probandinnen und Probandenwurden in Gruppen von bis zu acht Personen getestet. Neben demografischen Fragen zur Person sowie Fragen zum Leseverhalten bei wissenschaftlichen Texten bearbeiteten die Teilnehmenden eine Testbatterie, die den AST, einen Test zur Plausi-

bilitätsbewertung von Argumenten und einen Test zum Erkennen von typischen Argumentationsfehlern beinhaltete. Als zusätzliche Maße der Kriteriumsvalidität wurden die epistemologischen Überzeugungen sowie die verbale Intelligenz der Probandinnen und Probanden ebenfalls erfasst.

Instrumente

Argumentbewertungstest

Da neben dem Erkennen und Zuordnen von Argumentkomponenten auch die Einschätzung der Plausibilität von Argumenten zu wichtigen epistemisch-systematischen Lesekompetenzen gezählt werden kann, wurde ein Argumentbewertungstest (ABT) zur Einschätzung der Plausibilität von wissenschaftlichen Argumenten als Maß der Kriteriumsvalidität einbezogen. Der ABT besteht aus einem Text mit 30 kurzen Argumenten zu je ein bis zwei Sätzen. Zehn der Argumente wurden dabei so konzipiert, dass sie typische Argumentationsfehler aufweisen, wie beispielsweise einen Zirkelschluss oder falsche Analogien. Aufgabe der Teilnehmenden ist zum einen, die jeweils einzeln präsentierten Argumente entweder als plausibel oder unplausibel einzuschätzen. Zum anderen werden die Probandinnen und Probanden gebeten, zu den als unplausibel bewerteten Argumenten jeweils den entsprechenden Argumentationsfehler aus einer Liste anzugeben. Die Anteile der korrekt als unplausibel zugeordneten Argumente sowie der korrekt zugewiesenen Argumentationsfehler dienen dabei als Maße für die Fähigkeit, Argumente hinsichtlich ihrer Plausibilität einschätzen zu können. Die deskriptiven Kennwerte des ABT werden in Tabelle 2 berichtet.

Epistemologische Überzeugungen

Epistemologische Überzeugungen beschreiben implizite Annahmen einer Person über die Struktur, Stabilität und Generierung von Wissen (Hofer & Pintrich, 2002). Der Einfluss epistemologischer Überzeugungen auf formelle und informelle Lernprozesse gilt als empirisch gut belegt (vgl. Mayer & Rosman, 2016). In der vorliegenden Studie wurden epistemologische Überzeugungen über die Psychologie als Wissenschaft mit dem Fragebogen CAEB (Connotative Aspects of Epistemic Beliefs; Stahl & Bromme, 2007) erfasst und mit der Leistung im AST korreliert. Der CAEB besteht aus 24 Paaren von gegensätzlichen Adjektiven (z. B. „simpel" – „komplex") in den Dimensionen Strukturierheit beziehungsweise Veränderbarkeit von Wissen, die in Form eines semantischen Differentials angeordnet sind. Mittels einer siebenstufigen Likertskala können die Probandinnen und Probanden angeben, welches der beiden Adjektive eines Items die Psychologie als Wissenschaft besser beschreibt.

Verbale Intelligenz

Zur Erfassung der verbalen Intelligenz der Probandinnen und Probandenwurden die Subtests Satzergänzung, Analogien und Gemeinsamkeiten aus dem Grundmodul des I-S-T 2000R (Amthauer, Brocke, Liepmann & Beauducel, 2001) verwendet. Die Leistungsscores (Anteil korrekt bearbeiteter Items) in den drei Subtest werden zu einem Index für die verbale Intelligenz aggregiert. Der IST 2000R ist ein reliables und valides Verfahren zur Intelligenzmessung (s. Tabelle 2 für deskriptive Kennwerte).

--- hier Tabelle 2 einfügen ---

Ergebnisse

Fehlende Werte

Einzelne fehlende Werte bei den Subtests des I-S-T 2000R (< 0.1 % der Werte) wurden durch den Mittelwert der jeweiligen Person ersetzt. Bei jeweils einer Person fehlten Angaben zur Schulabschlussnote sowie zu den epistemologischen Überzeugungen. Daten zum momentanen Leistungsdurchschnitt im Studium gaben 77 (34 %) der Probandinnen und Probanden an.

Item-/Skalenkennwerte und interne Konsistenz

Tabelle 1 zeigt die Itemschwierigkeiten und -trennschärfen, die Akkuratheitswerte für den AST und die einzelnen Argumente. Um eine Überschätzung der internen Konsistenz des AST aufgrund möglicher technischer Abhängigkeiten (Items geschachtelt in Argumenten) zu vermeiden, wurden für diese Berechnungen die Items innerhalb der Argumente aggregiert und die interne Konsistenz über die acht Argumente berechnet. Der AST wies insgesamt eine zufriedenstellende interne Konsistenz (Cronbachs $\alpha = .76$) bei einer mittleren Itemschwierigkeit von .69 ($SD = .16$) auf. Bei dem Verfahren scheint es sich demnach um einen zuverlässigen Test zu handeln, der vor allem im mittleren bis hohen Fähigkeitsbereich gut trennt.

Validitätsschätzungen

Konstruktvalidität

Zur Konstruktvalidierung wurde geprüft, wie gut sich die beobachteten Itemschwierigkeiten mithilfe von Itemmerkmalen vorhersagen lassen und ob sich auf Basis theoretischer Annahmen aus der Forschung zum Argumentverstehen erleichternde beziehungsweise erschwerende Effekte auf die Itembearbeitung erwarten lassen. In einem ersten Schritt wurde zur Ermittlung der Itemschwierigkeiten ein 1-PL-Modell (Rasch-Modell) geschätzt, und es wurde geprüft, ob die dem Rasch-Modell inhärente Annahme der Unabhängigkeit von Items aufrechterhalten werden

kann, obgleich sich die Items gruppenweise demselben Argument zuordnen lassen. In einem zweiten Schritt wurde ein explanatorisches Modell in Form eines linear-logistischen Testmodells (LLTM; Fischer, 1974) geschätzt. Die Modelle in Schritt 1 und 2 wurden als Generalisierte Linear-gemischte Modelle (Generalized Linear Mixed Models; GLMM) mit dem R-Paket lme4 (Bates et al., 2017) und Maximum-Likelihood geschätzt. In einem dritten Schritt wurden dann die vorhergesagten Itemschwierigkeiten aus Schritt 2 mit den auf Basis des 1-PL-Modells geschätzten Itemschwierigkeiten aus Schritt 1 korreliert (Wilson & De Boeck, 2004).

Schritt 1: 1-PL-Modell. Das im ersten Schritt zur Ermittlung der empirischen Itemschwierigkeiten geschätzte 1-PL-Modell wies gemäß dem Andersen Likelihood-Ration-Test (2 Teilstichproben geteilt am arithmetischen Mittel der Testwerte) eine gute Modellpassung auf, $\chi^2(df = 38, N = 225) = 46.81$, $p = .130$ (ermittelt mit dem R-Paket ltm; Rizopoulos, 2018). Bei der Konzeption des AST wurde davon ausgegangen, dass die Items als unabhängig betrachtet werden können, auch wenn sich jeweils mehrere Items auf ein Argument beziehen. Um diese Annahme zu prüfen, wurde das 1-PL-Modell mit einem liberaleren Modell verglichen, in das zusätzlich Zufallseffekte der Argumente (random intercepts) aufgenommen wurden. Im Modellvergleich zeigte sich, dass das liberalere Modell nicht signifikant mehr Varianz aufklärte als Modell 2, $\chi^2(df = 1, N = 225) = .07$, $p = .985$. Die Daten geben damit keine Hinweise auf statistische Abhängigkeiten zwischen Items, die demselben Argument zuzuordnen sind.

Schritt 2: Explanatorisches Item-Response-Modell. Zur Überprüfung von Aspekten der Konstruktvalidität wurde ein explanatorisches Item-Response-Modell in Form eines LLTMs mit den dummykodierten Prädiktoren (feste Effekte) Reason-first- versus Claim-first-Argumente, komplexe versus einfache Argumente und den jeweils erfragten Argumentkomponenten (einbezogen in Form von vier dummykodierten Prädiktoren) geschätzt. Zusätzlich wurden die Argumentlänge (Zeichenzahl, zentriert) und die Position des Arguments im Test (zentriert) als Kontrollvariablen in Form von Prädiktoren mit festen Effekten einbezogen. Aus Gründen der

Anschaulichkeit berichten wir die Ergebnisse hier in Form von mittleren Lösungswahrscheinlichkeiten (modellbasiert geschätzt und rücktransformiert aus den Logit-Werten des LLTM mit dem Paket lsmeans; Lenth, 2016). Es zeigten sich die theoretisch erwarteten Effekte. Reason-first-Argumente waren schwieriger (P = .67, SE = .02) als Claim-first-Argumente (P = .82, SE = .01), $p < .001$, und komplexe Argumente waren schwieriger (P = .64, SE = .01) als einfache Argumente (P = .85, SE = .01), $p < .001$. Die Argumentkomponenten Gründe (P = .80, SE = .01) und mehr noch Einschränkungen (P = .96, SE = .01) waren erwartungsgemäß am leichtesten, gefolgt von der Stützung der Schlussregel (P = .70, SE = .02), Behauptungen (P = .67, SE = .02) und schließlich der Schlussregel selbst (P = .59, SE = .02). Mit Ausnahme von Behauptungen und der Stützung der Schlussregel unterschieden sich in paarweisen Vergleichen sämtliche Argumentkomponenten in ihren mittleren Lösungswahrscheinlichkeiten, $p < .001$. Die Itemschwierigkeiten des AST hingen also in sehr systematischer Weise von Itemmerkmalen ab, die bisherigen Untersuchungen zufolge das Erkennen der Argumentstruktur beeinflussen sollten und bei der Konstruktion der Items systematisch variiert wurden.

Schritt 3: Korrelation der empirischen Itemschwierigkeiten mit den vorhergesagten Itemschwierigkeiten aus dem explanatorischen Item-Response-Modell. Im nächsten Schritt wurden die im LLTM auf Basis der Itemmerkmale vorhergesagten Itemschwierigkeiten mit den empirischen Itemschwierigkeiten korreliert, die sich mit einem 1-PL-Modell schätzen lassen. Anhand der Itemmerkmale im LLTM ließen sich die empirischen Itemschwierigkeiten aus dem 1-PL-Modell sehr gut vorhersagen; die erklärte Varianz (R^2) betrug .82 (s. Abbildung 2). Tabelle 3 berichtet die inkrementelle Varianzaufklärung in den Itemschwierigkeiten, die durch jeden einzelnen Prädiktor aus dem explanatorischen Item-Response-Modell geleistet wurde.

--- hier Tabelle 3 einfügen ---

Zusätzlich wurde eine multiple Regressionsanalyse mit den Itemschwierigkeiten als abhängige Variable sowie den Itemmerkmalen Argumenttyp (komplex vs. einfach), Position

(Reason-first vs. Claim-first) und den erfragten Argumentkomponenten (Behauptung vs. Begründung vs. Schlussregel vs. Stützung der Schlussregel vs. Einschränkung; einbezogen in Form von 4 dummykodierten Prädiktoren) sowie den (zentrierten) Kontrollvariablen Argumentlänge (erfasst durch Zeichenzahl) und Position des Arguments im Test als Prädiktoren durchgeführt. Die Ergebnisse entsprechen dabei sehr weitgehend den Ergebnissen aus dem LLTM. Der Anteil der erklärten Varianz (R^2) betrug .76, $F(8,31) = 16.45$, $p < .001$. Wie erwartet waren Reason-first-Argumente schwieriger als Claim-first-Argumente ($b = .15$, $SE_b = .04$, $t = -3.50$, $p < .01$) und komplexe Argumente waren schwieriger als einfache Argumente ($b = .14$, $SE_b = .06$, $t = -2.53$, $p < .05$). Ebenfalls analog zu dem LLTM waren die Effekte der Argumentkomponenten. Im Vergleich zur Stützung der Schlussregel konnten Gründe ($b = .31$, $SE_b = .05$, $t = -5.71$, $p < .001$) und Einschränkungen ($b = .15$, $SEb = .05$, $t = -2.77$, $p < .01$) leichter identifiziert werden. Für die anderen Komponenten ergaben sich keine signifikanten Effekte. Wie erwartet ohne prädiktiven Wert blieben die Argumentlänge ($b = .0002$, $SE_b = .0002$, $t = -1.13$, $p = .267$) und die Position des Argumentes im Test ($b = .02$, $SE_b = .02$, $t = 0.94$, $p = .353$).

Kriteriumsvalidität

Tabelle 2 stellt die Korrelationen der Leistung im AST mit den anderen Leistungsmaßen sowie den epistemologischen Überzeugungen der Probandinnen und Probanden dar. Der AST korrelierte dabei substanziell mit anderen epistemisch-systematischen Lesekompetenzen (Argumentbewertung, Erkennen von Argumentationsfehlern) und mit Maßen der verbalen Intelligenz. Kleinere Korrelationen ergaben sich mit der Schulabschlussnote, dem momentanen Notendurchschnitt im Studium und den epistemologischen Überzeugungen der Probandinnen und Probanden. In allen Fällen entsprachen die Vorzeichen der Korrelationen den Erwartungen. Die beobachteten Korrelationen lassen sich als Hinweis auf die Kriteriumsvalidität des Tests interpretieren.

Diskussion

Die berichteten Befunde legen nahe, dass der AST einen reliablen sowie konstrukt- und kriteriumsvaliden Test zur Messung der Fähigkeit darstellt, Argumentkomponenten nach Toulmin (1958) erkennen und zuordnen zu können. So konnten in dieser Untersuchung ausreichend gute Itemkennwerte und eine zufriedenstellende interne Konsistenz des Tests ermittelt werden. In einem exploratorischen Item-Response-Modell (LLTM) konnte darüber hinaus gezeigt werden, dass sich die Itemschwierigkeiten durch theoriegeleitet variierte relevante schwierigkeitsgenerierende Merkmale sehr präzise vorhersagen lassen. Diese Ergebnisse sind ein starker Beleg für die Konstruktvalidität des AST. Des Weiteren konnten Belege für die Kriteriumsvalidität in Form von plausiblen Zusammenhängen der Testleistung mit Intelligenz, Schul- und Studienleistungen als Belege für die Kriteriumsvalidität gefunden werden. Diese vielversprechenden Validitätshinweise werden durch weitere Untersuchungen komplettiert, in denen die Trainierbarkeit der durch den AST erfassten Kompetenzen (von der Mühlen, Richter, Schmid & Berthold, 2018) und die der Testbearbeitung zugrunde liegenden kognitiven Prozesse mithilfe von Eye-Tracking-Daten näher untersucht wurden.

Literatur

Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (2001). *I-S-T 2000 R – Intelligenz-Struktur-Test 2000 R.* Göttingen: Hogrefe.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B. & Sigmann, H. (2017). *lme4: Linear mixed-effects models using Eigen and S4* (R-package version 1.1–14) [Computer software]. Verfügbar unter: http://cran.r-project.org/package=lme4

Bazerman, C. (1985). Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written Communication, 2*, 3–23. https://doi.org/10.1177/0741088385002001001

Britt, M. A. & Larson, A. (2003). Construction of argument representations during on-line reading. *Journal of Memory and Language, 48*, 749–810. http://dx.doi.org/10.1016/S0749-596X(03)00002-0

Britt, M. A., Richter, T. & Rouet, J. F. (2014). Scientific Literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist, 49*, 104–122. https://doi.org/10.1080/00461520.2014.916217

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Hofer, B. K. & Pintrich, P. R. (Eds.). (2002). *Personal epistemology: The psychology of beliefs about knowledge and knowing*. Mahwah, NJ: Erlbaum.

Korpan, C. A., Bisanz, G. L., Bisanz, J. & Henderson, J. M. (1997). Assessing literacy in science: Evaluation of scientific news briefs. *Science Education, 81*, 515–532. https://doi.org/10.1002/(SICI)1098-237X(199709)81:5<515::AID-SCE2>3.0.CO;2-D

Larson, M., Britt, M. A. & Larson, A. A. (2004). Disfluencies in comprehending argumentative texts. *Reading Psychology, 25*, 205–224. https://doi.org/10.1080/02702710490489908

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Soft-*

ware, 69, 1–33. Verfügbar unter https://www.jstatsoft.org/article/view/v069i01/v69i01.pdf

Mayer, A.-K. & Rosman, T. (2016). Epistemologische Überzeugungen und Wissenserwerb in akademischen Kontexten. In A.-K. Mayer & T. Rosman (Hrsg.), *Denken über Wissen und Wissenschaft: Epistemologische Überzeugungen* (S. 7–23). Lengerich: Pabst.

Mühlen, S. von der, Richter, T., Schmid, S. & Berthold, K. (2018). How to improve argumentation comprehension in university students: Experimental test of a training approach. Instructional Science. https://doi.org/10.1007/s11251-018-9471-3

Mühlen, S. von der, Richter, T., Schmid, S., Schmidt, L. M. & Berthold, K. (2016). Judging the plausibility of arguments in scientific texts: A student-scientist comparison. Thinking & Reasoning, *22*, 221–246. https://doi.org/10.1080/13546783.2015.1127289

Richter, T. & Schmid, S. (2010). Epistemological beliefs and epistemic strategies in self-regulated learning. *Metacognition and Learning*, *5*, 47–65. http://dx.doi.org/10.1007/s11409-009-9038-4

Richter, T., Schroeder, S. & Wöhrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology, 96*, 538–558. http://dx.doi.org/10.1037/a0014038

Rizopoulos, D. (2018). *ltm: Latent Trait Models under IRT (*R-package version 1.1-1) [Computer software]. Verfügbar unter: https://github.com/drizopoulos/ltm

Schroeder, S., Richter, T. & Hoever, I. (2008). Getting a picture that is both accurate and stable: Situation models and epistemic validation. *Journal of Memory and Language*, *59*, 237–259. https://doi.org/10.1016/j.jml.2008.05.001

Shaw, F. W. (1996). The cognitive processes in informal reasoning. *Thinking and Reasoning*, 2, 51–80. http://dx.doi.org/10.1080/135467896394564

Stahl, E. & Bromme, R. (2007). The CAEB: An instrument for measuring connotative aspects of epistemological beliefs. *Learning and Instruction, 17*, 773–785. https://doi.org/10.1016/j.learninstruc.2007.09.016

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Wild, K.P. (2000). *Lernstrategien im Studium: Strukturen und Bedingungen*. Münster: Waxmann.

Wild, K.-P. & Schiefele, U. (1994). Lernstrategien im Studium: Ergebnisse zur Faktorenstruktur und Reliabilität eines neuen Fragebogens. *Zeitschrift für Differentielle und Diagnostische Psychologie, 15*, 185–200.

Wilson, M. & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43–74). New York: Springer.

Wu, Y. T. & Tsai, C. C. (2007). High school students' informal reasoning on a socio-scientific issue: Qualitative and quantitative analyses. *International Journal of Science Education*, 29, 1163–1187. https://doi.org/10.1080/09500690601083375

Tabelle 1. Deskriptive Statistiken der Argumente und Items im Argumentstrukturtest

	Position	Komplexität	Items	M_p	Trennschärfe	*P*
Argumentstrukturtest			40	.67		
Argument 1	claim-first	komplex	5	.76		
Behauptung					.36	.75
Gründe					.32	.83
Schlussregel					.32	.61
Stützung					.36	.68
Einschränkungen					.37	.94
Argument 2	reason-first	komplex	5	.48		
Behauptung					.41	.25
Gründe					.37	.65
Schlussregel					.32	.26
Stützung					.43	.46
Einschränkungen					.26	.78
Argument 3	claim-first	einfach	5	.88		
Behauptung					.42	.93
Gründe					.19	.95
Schlussregel					.33	.73
Stützung					.43	.86
Einschränkungen					.32	.93
Argument 4	reason-first	komplex	5	.65		
Behauptung					.33	.62
Gründe					.35	.70
Schlussregel					.35	.47
Stützung					.49	.51
Einschränkungen					.41	.95
Argument 5	reason-first	komplex	5	.52		
Behauptung					.52	.33
Gründe					.40	.52
Schlussregel					.37	.34
Stützung					.32	.47
Einschränkungen					.27	.92
Argument 6	reason-first	komplex	5	.50		
Behauptung					.49	.44
Gründe					.41	.71
Schlussregel					.42	.44
Stützung					.53	.51
Einschränkungen					.30	.96
Argument 7	claim-first	komplex	5	.66		
Behauptung					.49	.59
Gründe					.42	.78
Schlussregel					.45	.48
Stützung					.48	.53
Einschränkungen					27	.95
Argument 8	claim-first	einfach	5	.89		
Behauptung					.28	.96
Gründe					.23	.96
Schlussregel					.37	.78
Stützung					.37	.87
Einschränkungen					.25	.93

Anmerkungen: $N = 225$. M_p = mittlere Lösungswahrscheinlichkeit. P = Schwierigkeit (Lösungswahrscheinlichkeit). Alle Itemkennwerte sind deskriptiv (nicht modellbasiert) ermittelt.

Tabelle 2. Korrelationen der Antwortakkuratheit im AST mit anderen Leistungsmaßen und epistemologischen Überzeugungen

		M	*SD*	Cronbachs α	1	2	3[b]	4	5	6	7
1	AST Gesamt	.67	.16	.76							
2	Schulabschlussnote[a]	2.11	1.29	–	.17*						
3	Aktueller Notendurchschnitt[b]	2.02	.50	–	.20	.25*					
4	Verbaler IQ	.52	.12	.82	.40**	-.23**	-.35**				
5	Argumentbewertungstest[c,d]	.70	.01	.64	.26**	-.05	-.24*	.22**			
6	Zuordnung Argumentationsfehler[c,d]	.36	.22	–	.44**	-.20**	-.46**	.52**	.50**		
7	Strukturierheit psycholog. Wissens[a,e]	5.34	.99	.72	.20**	-.02	-.19	.16*	.15*	.12	
8	Veränderbarkeit psycholog. Wissens[a,e,f]	2.71	.77	.71	-.33**	.09	.24*	-.20*	-.31**	-.35**	-.47**

*Anmerkungen: N = 225. * p < .05; ** p < .01.*

[a] *n* = 224. [b] *n* = 77. [c] Selbstentwickelter Test (nicht Teil des AST). [d] Anteil korrekt bearbeiteter Items. [e] Subskalen Structure bzw. Variability des CAEB (Stahl & Bromme, 2007); Antwortbereich von 1–7. [f] Polung der Skala: 1 = Veränderlich; 7 = Unveränderlich.

Tabelle 3. Regressionskoeffizienten und Anteil der aufgeklärten Varianz der Prädiktoren des Explanatorischen Item-Response-Modells

Prädiktoren	Itemschwierigkeiten (logit)			
	b	*t*	*p*	ΔR^2
Argumentlänge	-.002	-1.71	.097	.02
Argumentposition	-.113	-1.35	.186	.01
Argumenttyp[a]	-.950	-3.06	.005	.06
Argumentkomplexität[b]	1.036	2.49	.018	.04
Argumentkomponenten[c]				.42
Gründe vs. Behauptung	.769	1.92	.064	.02
Schlussregel vs. Behauptung	-.552	-1.38	.177	.01
Stützung der Schlussregel vs. Behauptung	.068	.17	.865	.00
Einschränkungen vs. Behauptung	2.404	6.01	.000	.30
Gesamtes R^2				.80

Anmerkungen: Argumentlänge und Argumentposition jeweils am Mittelwert zentriert.

[a] Claim-first- vs. Reason-first-Argumente; eingefügt als dummy-kodierte Variable mit Claim-first-Argumenten als Referenzkategorie. [b] Komplexe vs. einfache Argumente; eingefügt als dummykodierte Variable mit komplexen Argumenten als Referenzkategorie. [c] Argumentkomponenten eingefügt als vier dummykodierte Variablen mit Behauptung als Referenzkategorie.

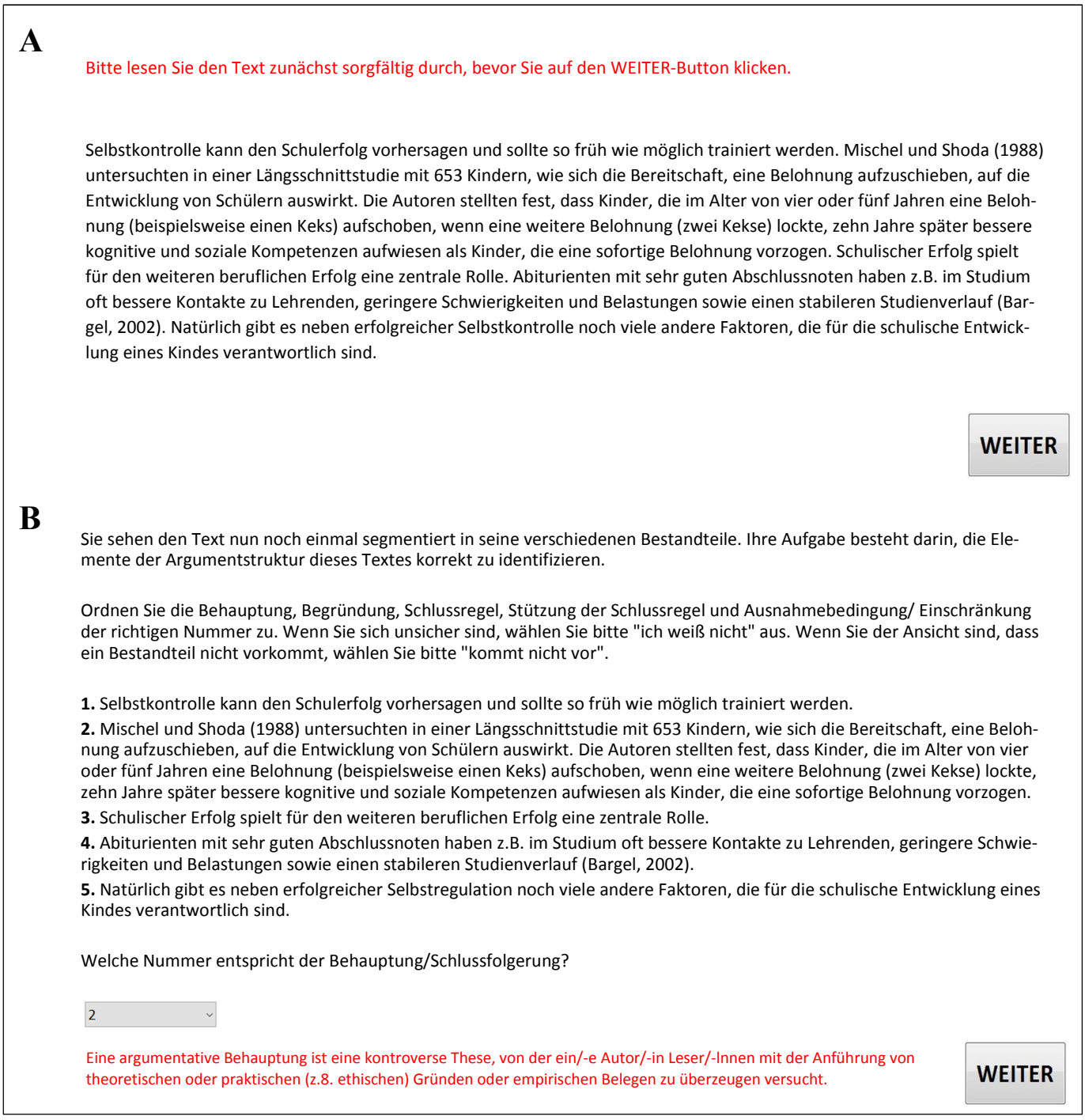

Abbildung 1. Beispielitem des AST mit (A) Fließtext des Argumentes und (B) segmentierter Darstellung des Argumentes während der Zuordnungssaufgabe.

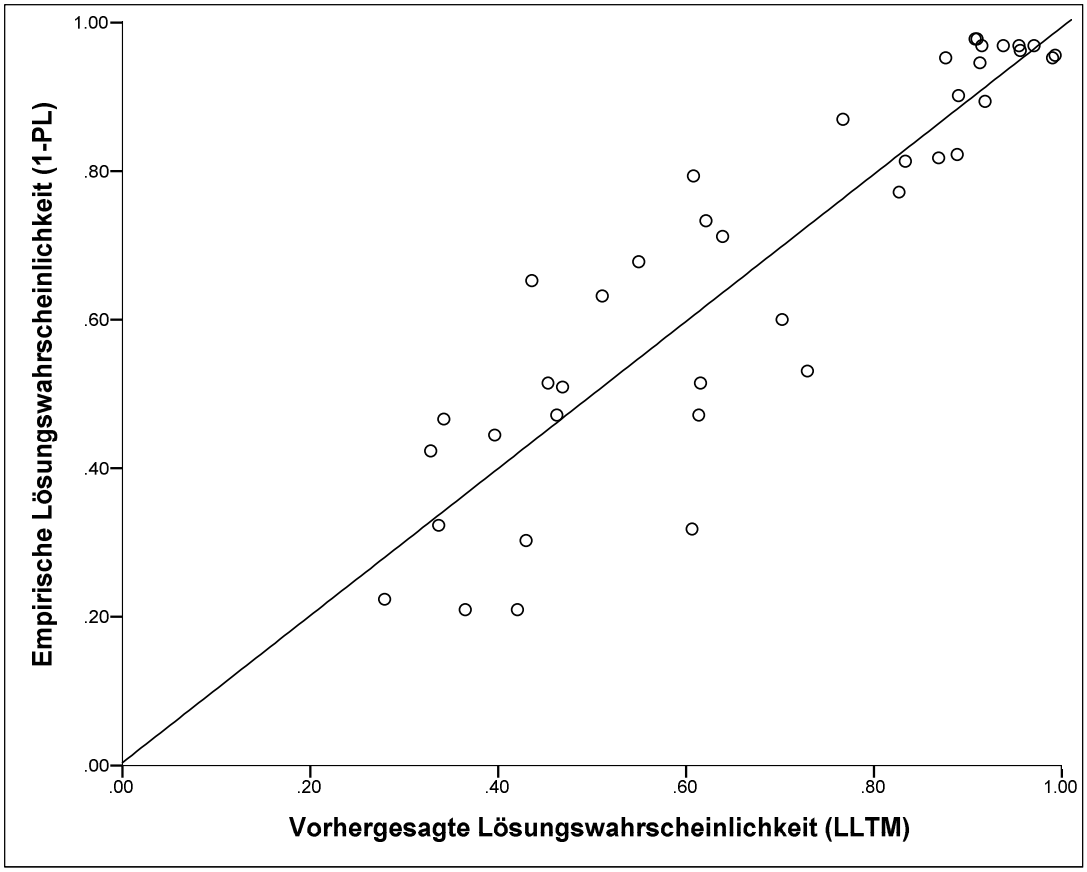

Abbildung 2. Scatterplot der vorhergesagten und empirischen Itemschwierigkeiten des AST.