Zusammenhänge zwischen dem Blickverhalten und der Antwortsicherheit beim Lösen von Aufgaben zum Graphenverständnis

Hanna Blumenthal*, Pascal Klein*

* Georg-August-Universität Göttingen, Didaktik der Physik, Friedrich-Hund-Platz 1, 37077 Göttingen, h.blumenthal@stud.uni-goettingen.de

Kurzfassung

Die Antwortsicherheit beim Bearbeiten physikalischer Aufgaben ist ein wichtiger Indikator zur Ermittlung des Lernstands und zur Überprüfung des Verständnisses von Konzepten. Bislang wird die Antwortsicherheit durch Abfragen nach der Bearbeitung der Aufgabe erfasst. Erste Ergebnisse von Eye-Tracking-Studien zeigen, dass auch das Blickverhalten während der Bearbeitung der Aufgabe zur Vorhersage der Antwortsicherheit verwendet werden kann. In dieser Arbeit wurden die Zusammenhänge zwischen der Antwortsicherheit und dem Blickverhalten anhand von Aufgaben zum Verständnis von Graphen überprüft und die Vorhersagekraft des Blickverhaltens für die Antwortsicherheit untersucht. Dafür wurden die Augenbewegungen von Versuchspersonen ($N = 114$) während der Bearbeitung der Items des Test of Understandig Graphs in Kinematics und die von den Versuchspersonen angegebene Antwortsicherheit in linearen gemischten Modellen analysiert. Es wurde zwischen verschiedenen Itemkategorien unterschieden, welche sich aus der unterschiedlichen Verwendung von Graphen, Texten und Werten in der Frage und den Antwortoptionen ergeben. Die Analyse bestätigt signifikante Zusammenhänge zwischen der Antwortsicherheit und dem Blickverhalten. Allerdings konnte nur ein geringer Zuwachs der Vorhersagekraft für die Antwortsicherheit gefunden werden, wenn neben der Leistung der Versuchspersonen und der Bearbeitungszeit zusätzlich das Blickverhalten als Prädiktor verwendet wurde.

1. Einleitung

Die Antwortsicherheit von Lernenden beim Lösen von Aufgaben ist innerhalb des Lernprozesses ein wichtiger Indikator für den Lernstand der Lernenden. Erstens deuten unsichere Antworten darauf hin, dass ein Konzept noch nicht vollständig verstanden wurde und zweitens kann aus falschen Antworten, die mit hoher Sicherheit getroffen wurden, auf das Vorhandensein von Fehlkonzepten geschlossen werden (Hasan, Bagayoko & Kelley, 1999).

Außerdem gehen Lindsey und Nagel (2015) davon aus, dass für ein vollständiges Verständnis eines Inhalts sich die Lernenden zusätzlich über ihr eigenes Wissen bewusst sein müssen. Bezogen auf die Antwortsicherheit bedeutet dies, dass Lernende metakognitiv fähig sein müssen zu entscheiden, ob sie eine Frage richtig beantworten konnten oder nicht.

Momentan wird die Antwortsicherheit nach Bearbeitung der Aufgabe durch eine zusätzliche Abfrage aufgenommen. Wäre eine Ableitung der Antwortsicherheit aus dem Blickverhalten möglich, müsste keine Unterbrechung zur Ermittlung der Antwortsicherheit während des Lernprozesses stattfinden (Smith, Legg, Matovic & Kinsey, 2018) und eine simultane Erhebung der Antwortsicherheit während der Bearbeitung der Aufgabe und noch bevor die Antwort gegeben wird, wäre denkbar.

Die Einsatzmöglichkeiten dieses Verfahrens wären vielfältig. In adaptiven Lernumgebungen könnte es genutzt werden, um die Lernumgebung noch präziser an die lernende Person anzupassen. Außerdem könnten während des Bearbeitens von Aufgaben Hinweise eingeblendet werden, sobald eine Unsicherheit der lernenden Person erkannt wurde. Schließlich könnte es in Eye-Tracking-Studien zur Erhebung der Antwortsicherheit genutzt werden.

In jüngster Zeit gab es bereits mehrere Untersuchungen, die Antwortsicherheit mithilfe von maschinellem Lernen aus dem Blickverhalten abzuleiten. Bei den Studien wurde die Antwortsicherheit kategorial als sicher und unsicher erhoben. Nach einer Trainingsphase konnte die Antwortsicherheit in den Studien mit einer Genauigkeit von 78 Prozent oder mehr vorhergesagt werden (vgl. Ishimaru, Maruichi, Dengel & Kise 2021; Smith et al., 2018; Yamada, Kise & Augereau, 2017). In der Studie von Ishimaru et al. (2021) zeigte ein Vergleich der Vorhersagekraft der Antwortsicherheit aus dem Blickverhalten und aus der Bearbeitungszeit, dass die Vorhersage aus dem Blickverhalten lediglich um 6 Prozent besser war als die Vorhersage aus der Bearbeitungszeit.

Innerhalb der Physikdidaktik wurde die Verbindung zwischen der Antwortsicherheit und dem Blickverhalten erst wenig untersucht. Klein et al. (2020) analysierten die visuelle Aufmerksamkeit von Lernenden beim Bearbeiten des modifizierten Test of

Understanding Graphs in Kinematics (TUG-K) von Zavala, Tejeda, Barniol und Beichner (2017) und setzten die Blickzeiten der Lernenden auf Frage und Antwortoptionen mit ihrer Antwortsicherheit in Verbindung. Die Blickzeit auf die Frage korrelierte signifikant mit der Antwortsicherheit. Für die Blickzeit auf die Optionen konnte dies nicht festgestellt werden. Die Antwortsicherheit wirkte sich auf die Blickzeiten auf die Frage bzw. auf die Optionen mit einer geringen Effektstärke aus.

In einer weiteren Veröffentlichung zu denselben Daten untersuchten Klein, Becker, Küchemann & Kuhn (2021), ob die Items des TUG-K von den Lernenden bearbeitet werden, wie es Zavala et al. (2017) beabsichtigten. Denn der TUG-K besitzt unterschiedliche Itemkategorien und daraus folgend übergeordnete Aufgaben, die sich aus den verschiedenen Kombinationen von Texten, Graphen und Werten in Frage und Antwort ergeben. Diese Aufgaben sind das Auswählen eines passenden Graphen zu einer textlichen Beschreibung, das Auswählen eines passenden Graphen zu einem gegebenen Graphen, das Auswählen einer textlichen Beschreibung zu einem gegebenen Graphen und das Auswählen eines Wertes zu einem gegebenen Graphen. Klein et al. führten eine Cluster-Analyse durch, in der sie die Sprünge des Blicks zwischen der Frage und den Antwortoptionen sowie die Sprünge zwischen den einzelnen Antwortoptionen als Variablen zum Clustering nutzten. Es ergaben sich drei Cluster, die den oben vorgestellten unterschiedlichen Itemkategorien des TUG-K entsprachen, wobei das Auswählen einer textlichen Beschreibung bzw. eines Wertes zu einem gegebenen Graphen ein Cluster bildeten. Als interessantes Nebenergebnis konnten die Forschenden feststellen, dass sich die Anzahl der Sprünge sicherer und unsicherer Lernender stärker unterschieden als die Anzahl der Sprünge von Personen, die richtig oder falsch antworteten.

Die Ergebnisse der bisherigen Studien zur Vorhersage der Antwortsicherheit aus dem Blickverhalten mithilfe von maschinellem Lernen deuten darauf hin, dass Zusammenhänge zwischen dem Blickverhalten und der Antwortsicherheit bestehen. Eine erste Beschreibung einiger Zusammenhänge fand in der Studie von Klein et al. (2020) statt. Ziel dieser Untersuchung war die genauere Herausarbeitung und Benennung der Zusammenhänge und das Bestimmen ihrer Vorhersagekraft für die Antwortsicherheit. Die Forschungsfragen lauten:

1: Welche Zusammenhänge bestehen zwischen der Antwortsicherheit und dem Blickverhalten beim Lösen physikalischer Multiple-Choice-Aufgaben?

2: Inwieweit kann die Varianz der Antwortsicherheit neben der Leistung durch das Blickverhalten aufgeklärt werden?

Für die erste Forschungsfrage wurden vier Hypothesen aufgestellt, die auf den Ergebnissen der bisherigen Studien basieren:

H1: Sichere Personen antworten schneller als unsichere Personen.

H2: Sichere Personen betrachten sowohl die Frage als auch die Antwortoptionen kürzer als unsichere Personen.

H3: Sichere Personen springen weniger häufig zwischen Frage und Antwortoptionen als unsichere Personen.

H4: Sichere Personen betrachten vorrangig die gewählte Antwortoption. Unsichere Personen springen mehr zwischen den Antwortoptionen.

2. **Methode**

Die Daten für diese Studie stammen aus der Erhebung von Klein et al. (2020).

2.1. **Stichprobe und Datenerhebung**

Die Stichprobe bestand aus 114 (58 weiblich, 56 männlich) Lernenden der Oberstufe. Diese bearbeiteten den TUG-K in seiner modifizierten Form von Zavala et al. (2017) an einem Computer. Dabei wurden ihre Augenbewegungen aufgezeichnet. Nach jedem Item sollten die Lernenden ihre Antwortsicherheit angeben (Klein et al., 2020).

2.2. **Material**

Der TUG-K besteht aus 26 Items zum Verständnis kinematischer Graphen. Wie bereits oben erwähnt, können die Items des TUG-K verschiedenen Itemkategorien zugeordnet werden, die sich aus der unterschiedlichen Verwendung von Texten, Graphen und Werten in der Frage und den Antwortoptionen ergeben. Diese sind in Tabelle 1 dargestellt.

Kat.	Aufbau
1	Text → Graph
2	Text + Graph → Text
3	Text + Graph → Werte
4	Text + Graph → Graphen
5	Text + Graphen → Text (zweistufige Aufgabe)

Tab. 1: Aufbauweisen der verschiedenen Itemkategorien (Kat.) des TUG-K

Der Test wurde in die Tobii Studio Eye-Tracking-Software eingebunden, welche gleichzeitig mit dem stationären Eye-Tracker verbunden war. Zu jedem Item wurden drei Folien erstellt. Als Erstes wurde das Item mit der Frage und den Antwortoptionen präsentiert. Auf der nächsten Folie musste die gewählte Antwortoption angegeben werden. Auf der letzten Folie wurde die Antwortsicherheit abgefragt. Es wurde eine sechsstufige Likert-Skala eingesetzt. Diese Skala reichte von absolut sicher (1) bis völlig unsicher (6) (Klein et al., 2020).

2.3. **Definition der Areas of Interest (AOIs)**

Um die Blickdaten bezüglich der Forschungsfragen auswerten zu können, wurden für jedes Item mehrere Areas of Interest (AOIs) definiert.

Die erste AOI (T) wurde über die gesamte Folie gelegt. Eine zweite AOI bedeckt die Frage (Q). Diese besteht je nach Itemkategorie aus Text oder Text und Graph. Über die Antwortoptionen wurde eine dritte AOI (O) gelegt. Innerhalb der O-AOI wurden fünf kleinere AOIs (A - E) definiert, die jeweils eine der Antwortoptionen beinhalten. Die Antwortoptionen bestehen je nach Itemkategorie aus Texten, Werten oder Graphen. In Abbildung 1 ist die Definition der AOIs für ein Item beispielhaft dargestellt.

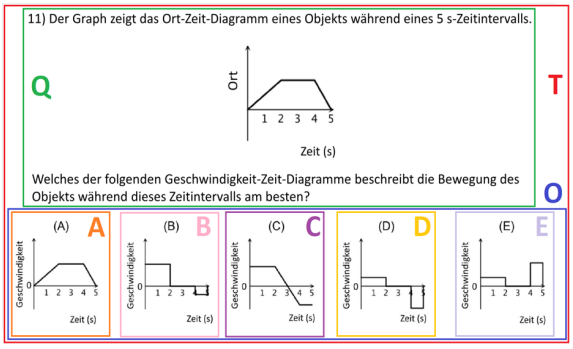

Abb. 1: Definition der Areas of Interest (T,Q,O,A-E) für das Item 11 des TUG-K

2.4. **Eye-Tracking Apparatur**

Für die Studie wurde ein 22-Zoll-Bildschirm mit einer Auflösung von 1920 × 1080 Pixeln und einer Bildwiederholfrequenz von 75 Hz verwendet. Zur Aufnahme der Augenbewegungen wurde der stationäre Eye-Tracker Tobii Pro X3-120 genutzt. Dieser besitzt laut Herstellerangaben eine Genauigkeit von weniger als 0.4° des Blickwinkels und die Position der Augen wird mit 120 Hz erfasst. Als eine Sakkade wurde eine Augenbewegung erkannt, bei der die Beschleunigung der Augen 8500°s^{-2} überstieg und die Geschwindigkeit der Augen größer als 30°s^{-2} war. Der durchschnittliche Abstand zwischen dem Monitor und den Augen der Teilnehmenden betrug 62 cm (Klein et al., 2020; 2021).

2.5. **Eye-Tracking-Maße**

Zum Überprüfen der Hypothesen wurden das in den Hypothesen beschriebene Blickverhalten durch verschiedene Eye-Trackingmaße gemessen und daraus Blickvariablen erstellt.

Erstens wurden die Blickzeiten auf das gesamte Item (TVD_T), auf die Frage (TVD_Q) und auf die Antwortoptionen (TVD_O) bestimmt. Die Blickzeit auf das gesamte Item entspricht der Bearbeitungszeit.

Zweitens wurden als Blickvariable für die Anzahl der Blickwechsel zum einen die Sprünge zwischen der Frage und den Antwortoptionen (Jumps_QO) sowie zwischen den einzelnen Antwortoptionen (Jumps_AE) verwendet. Zum anderen wurden die Besuche auf der Frage und den Antwortoptionen (VC_QO) beziehungsweise auf der Frage und den einzelnen Antwortoptionen (VQ_QAE) sowie die Besuche auf den einzelnen Antwortoptionen (VC_AE) genutzt.

Drittens wurde das von Rodemer, Graulich und Bernholt (2020) entwickelte Fixation/Transition Ratio (FTR) verwendet. Das FTR beschreibt das Verhältnis zwischen den Fixationen und den Sprüngen. Der Vorteil dieses Maßes ist die Möglichkeit, das Sprungverhalten unabhängig von der Bearbeitungszeit mit der Antwortsicherheit in Verbindung zu setzen, denn die Anzahl der Fixationen auf eine AOI korreliert hoch mit der Blickzeit auf die AOI (Tullis & Albert 2013). Das FTR kann demnach als Maß für die durchschnittliche Zeit zwischen zwei Sprüngen angesehen werden. Es wurden das FTR zwischen der Frage und den Antwortoptionen (FTR_QO) sowie das FTR zwischen der Frage und den einzelnen Antwortoptionen (FTR_QAE) bestimmt.

Viertens wurde der Anteil der Blickzeit auf die gewählte Option von der Blickzeit auf alle Antwortoptionen verwendet.

2.6. **Daten und Datenauswertung**

In die Analyse wurden die Antwortsicherheit der Lernenden, die Blickvariablen sowie die Leistung der Lernenden einbezogen. Die Angaben zur Antwortsicherheit wurden auf eine Skala von 0 (völlig unsicher) bis 1 (völlig sicher) linear transformiert. Als Leistung wurde der Anteil der richtig gelösten Aufgaben bestimmt. Die Blickvariablen und die Leistung der Lernenden wurden z-transformiert.

Um die Zusammenhänge zwischen der Antwortsicherheit und dem Blickverhalten zu untersuchen, wurden lineare gemischte Modelle verwendet. Die Antwortsicherheit S wurde als abhängige Variable gewählt. Die Personen und die Items wurden als zufällige Effekte in Form von zufälligen Schnittpunkten für die einzelnen Personen und Items miteinbezogen. Das Modell mit den zufälligen Parametern lautet

$$S_{ij} = b_0 + P_{0_i} + I_{0_j} + \epsilon_{ij}. \quad (1)$$

Hierbei beschreibt S_{ij} die Antwortsicherheit der i-ten Person beim j-ten Item. Der Schnittpunkt für das Modell ist b_0. P_{0i} bzw. I_{0j} geben die Verschiebung des Schnittpunkts für die Person i und das Item j an. Der verbleibende Fehler wird als ϵ_{ij}bezeichnet.

Zum Testen der Hypothesen wurde das Modell jeweils um eine Blickvariable BV als Prädiktor für die Antwortsicherheit erweitert

$$S_{ij} = b_0 + P_{0_i} + I_{0_j} + b_1 * BV_{ij} + \epsilon_{ij}. \quad (2)$$

Die Hypothesen wurden außerdem für jede Itemkategorie einzeln überprüft. Dazu wurden die Itemkategorien zunächst als Effekt-kodierte Variable dargestellt und anschließend die Gewichte der Blickvariablen für die einzelnen Itemkategorien bestimmt.

Um die zweite Forschungsfrage zu beantworten, wurde zunächst ein allgemeines Modell entwickelt, das die Leistung sowie relevante Blickvariablen als Prädiktoren beinhaltet. Dazu wurde ein Modell mit der Leistung als Prädiktor sukzessive um Blickvari-

ablen ergänzt. Es wurde jeweils eine Blickvariable dem Modell hinzugefügt und mittels der Devianz verglichen, welche Blickvariable das Modell am stärksten und signifikant verbessert. Es wurden Blickvariablen hinzugefügt, bis die Erweiterung um eine Blickvariable zu keiner signifikanten Verbesserung des Modells führte. Um Aussagen über die Varianzaufklärung zu treffen, wurde für die Modelle ein korrigiertes pseudo R^2 nach Zhang 2017 berechnet.

Analog wurde für jede Itemkategorie ein eigenes spezielles Modell aufgestellt.

Schließlich wurden das allgemeine Modell und die speziellen Modelle auf die einzelnen Items angewendet, um die Passung der Modelle auf die Items zu überprüfen. Dazu wurden für jedes Item vier multiple Regressionen durchgeführt. Bei der ersten Regression wurde nur die Leistung als Prädiktor verwendet. Für die zweite Regression wurden die Leistung und die Blickzeit auf das gesamte Item als Prädiktoren gewählt. Für die dritte Regression wurden die Leistung und die Blickvariablen des allgemeinen Modells als Prädiktoren eingesetzt. Bei der vierten Regression wurden die Leistung und die Blickvariablen, die das spezielle Modell der Itemkategorie enthält, als Prädiktoren verwendet. Für diese vier Regressionen wurde jeweils das korrigierte R^2 bestimmt.

Für die Analyse der vorliegenden Daten wurde das lmerTest Package Version 3.1-3 (Kuznetsova, Brockhoff & Christensen, (2017) in R (R Core Team, 2022) verwendet. Dieses nutzt Satterthwaites Approximationsmethode um die Anzahl der Freiheitsgrade und somit die Signifikanz eines Parameters zu bestimmten. Als Signifikanzniveau wurde $p = .05$ gewählt. Das pseudo R^2 wurde mit dem rsq Package Version 2.2 (Zhang, 2021) bestimmt.

3. Ergebnisse

Zwischen der Antwortsicherheit und den einzelnen Blickvariablen bestehen jeweils signifikante Zusammenhänge. Die Gewichte der Blickvariablen für das Modell (2) sind in Tabelle 2 angegeben. Für Blickvariablen, die Blickzeiten oder Anzahlen von Sprüngen bzw. Besuchen beschreiben, ist der Zusammenhang zur Antwortsicherheit negativ. Für den Anteil der Blickzeit auf die gewählte Option und für die FTRs besteht ein positiver Zusammenhang zur Antwortsicherheit.

Die Ergebnisse der Berechnungen für die einzelnen Itemkategorien sind in Tabelle 5 im Anhang dargestellt. Auch hier können größtenteils signifikante Zusammenhänge zwischen den Blickvariablen und der Antwortsicherheit festgestellt werden. Kein signifikanter Zusammenhang zwischen der Blickzeit auf die Frage und der Antwortsicherheit kann für die Itemkategorie 1 ausgemacht werden. Für die zwei FTRs bestehen nur signifikante Zusammenhänge zur Antwortsicherheit für die Itemkategorien 3 und 4.

	BV	b_1	$SE(b_1)$	p
H1	TVD_T	-0.076	0.004	$< .001$
H2	TVD_Q	-0.058	0.004	$< .001$
	TVD_O	-0.072	0.003	$< .001$
H3	Jumps_QO	-0.063	0.004	$< .001$
	VC_QO	-0.068	0.004	$< .001$
	FTR_QO	0.013	0.004	$< .001$
H4	An_TVD	0.033	0.003	$< .001$
	Jumps_AE	-0.054	0.004	$< .001$
	VC_AE	-0.072	0.003	$< .001$
	VC_QAE	-0.075	0.004	< 0.01
	FTR_QAE	0.017	0.004	$< .001$

Tab. 2: Gewichte b_1 der einzelnen Blickvariablen als Prädiktoren für die Antwortsicherheit

Die ausgewählten Blickvariablen des allgemeinen Modells sind in Tabelle 3 angegeben.

Die Leistung erklärt 14 Prozent der Varianz der Antwortsicherheit. Die Ergänzung des Modells um die Blickzeit auf das gesamte Item führt zu einer Steigerung der Varianzaufklärung von weiteren 8 Prozent. Durch das Hinzufügen der weiteren Blickvariablen steigt die Varianzaufklärung jeweils im Promillebereich. Insgesamt können 23 Prozent der Varianz der Antwortsicherheit aufgeklärt werden.

Prädiktoren	R^2
Leistung	.137
+TVD_T	.214
+VC_AE	.221
+An_TVD	.223
+TVD_O	.225
+FTR_QO	.228

Tab. 3: Schrittweise Entwicklung eines allgemeinen Modells zur Beschreibung der Antwortsicherheit durch die Leistung und das Blickverhalten und deren Varianzaufklärung pro Stufe durch ein korrigiertes pseudo R^2

Die ausgewählten Blickvariablen für die speziellen Modelle der Itemkategorien sind in Tabelle 4 angegeben. Durch die Leistung können zwischen 10 Prozent der Varianz bei den Itemkategorien 2 und 3 und 26 Prozent bei Itemkategorie 4 erklärt werden. Durch die Blickvariablen kann die Varianzaufklärung um weitere 6 bis 13 Prozent gesteigert werden. Insgesamt werden Varianzaufklärungen zwischen 20 Prozent bei Itemkategorie 2 und 31 Prozent bei Itemkategorie 4 erreicht.

Die Varianzaufklärungen der Modelle angewendet auf die einzelnen Items sind in Tabelle 6 im Anhang dargestellt.

Wenn die Leistung als einziger Prädiktor verwendet wurde, schwankt die Aufklärung der Varianz der Antwortsicherheit zwischen 1 Prozent bei Item 4 und 33 Prozent bei Item 21.

Wurden die Regressionen um die Blickzeit auf das gesamte Item als Prädiktor erweitert, können bis zu weitere 22 Prozent der Varianz erklärt werden. Die Varianzaufklärung der Antwortsicherheit liegt für dieses Modell zwischen 11 Prozent und 37 Prozent.

Die Verwendung der Variablen des allgemeinen Modells für die Regression führt zu einer Steigerung der Varianzaufklärung im Vergleich zu der Regression, in der die Leistung und die Blickzeit auf das gesamte Item als Prädiktoren verwendet wurde, um bis zu weitere 8 Prozent. Jedoch gibt es auch sieben Fälle, in denen die Varianzaufklärung abnimmt. Besonders bei den Items der Kategorie 5 scheinen die weiteren Variablen zu keinem Gewinn der Varianzaufklärung zu führen. Für die Itemkategorien 1 und 2 werden maximal weitere 3 Prozent der Varianz erklärt. Für die meisten Items der Kategorie 3 liegt die Steigerung im Bereich um weitere 5 Prozent. Für die Itemkategorie 4 liegen die Differenzen der Varianzaufklärung in einer Spanne von 1 Prozent bis 6 Prozent.

Ein ähnliches Bild zeigt sich, wenn die Varianzaufklärung der speziellen Modelle mit der Varianzaufklärung des Modells mit der Leistung und der Blickzeit auf das gesamte Item verglichen wird. Für die Itemkategorie 5 sind die Werte sehr ähnlich. Für die Itemkategorien 1 und 2 können Steigerungen um bis zu weitere 4 Prozent ausgemacht werden. Bei den Itemkategorien 3 und 4 gibt es wenige Items mit einer Steigerung um 6 Prozent, bei den anderen sind die Werte beider Modelle ähnlich.

4. Diskussion

Zur Beantwortung der ersten Forschungsfrage, die nach Zusammenhängen zwischen der Antwortsicherheit und dem Blickverhalten fragt, wurden die aufgestellten Hypothesen überprüft.

Die erste Hypothese, dass sichere Personen schneller antworten als unsichere Personen, kann bestätigt werden. Es können signifikante Zusammenhänge zwischen der Blickzeit auf das gesamte Item und der Antwortsicherheit festgestellt werden. Dies gilt für alle Itemkategorien. Weil die Blickzeit auf das gesamte Item der Bearbeitungszeit entspricht, kann geschlossen werden, je kürzer das Item bearbeitet wurde, desto sicherer waren sich die Lernenden.

Auch die zweite Hypothese, dass sichere Lernende sowohl die Frage als auch die Antwortoptionen kürzer betrachten als unsichere Personen, kann im Allgemeinen zunächst bestätigt werden. Sowohl die Blickzeit auf die Frage als auch die Blickzeit auf die Antwortoptionen besitzen einen signifikanten Zusammenhang zur Antwortsicherheit. Mit steigenden Blickzeiten sinkt die Antwortsicherheit. Es besteht jedoch eine Ausnahme. Für Items der Kategorie 1, die nur Text in der Frage besitzen, ist die Blickzeit auf die Frage kein signifikanter Prädiktor für die Antwortsicherheit.

Die Überprüfung der dritten Hypothese, dass sichere Personen weniger häufig zwischen Frage und Antwortoptionen springen als unsichere Personen, erfolgte durch mehrere Blickvariablen und kann eingeschränkt bestätigt werden. Die Analyse unter Einbeziehung der Daten zu allen Items zeigt jeweils einen Zusammenhang der Antwortsicherheit mit den Sprüngen zwischen der Frage und den Antwortoptionen, mit der Anzahl der Besuche auf der Frage und den Antwortoptionen sowie mit dem FTR zwischen Frage und Antwortoptionen. Mit steigender Anzahl von Sprüngen oder Besuchen nimmt die Antwortsicherheit ab. Für das FTR besteht ein positiver Zusammenhang. Je mehr Fixationen durchschnittlich zwischen den Sprüngen lagen, desto höher war die Antwortsicherheit. Wurden die Itemkategorien einzeln betrachtet, so bleiben die Sprünge zwischen der Frage und den Antwortoptionen sowie die Besuche auf diesen signifikante Prädiktoren. Für das FTR zwischen Frage und Antwortoptionen besteht nur für die Itemkategorien 3 und 4 ein signifikanter Zusammenhang zur Antwortsicherheit. Da das FTR als Maß für die durchschnittliche Zeit zwischen zwei Sprüngen genutzt werden kann, kann aus den Ergebnissen geschlossen werden, dass sichere Personen zwar weniger mit ihrem Blick zwischen Frage und Antwortoptionen springen, dies für die drei Itemkategorien 1, 2 und 5 jedoch auf der längeren Bearbeitungszeit unsicherer Personen basiert.

Zur Überprüfung der vierten Hypothese, dass sichere Personen vorrangig die gewählte Antwortoption betrachten und unsichere Personen mehr zwischen den einzelnen Antwortoptionen springen, wurden für mehrere Blickvariablen die Zusammenhänge zur Antwortsicherheit untersucht. Auch diese Hypothese kann eingeschränkt bestätigt werden. Der Anteil der Blickzeit auf die gewählte Option von der Blickzeit auf alle Optionen steht in einem signifikanten Zusammenhang zur Antwortsicherheit. Für alle Itemkategorien nimmt die Antwortsicherheit mit steigender Fokussierung auf die gewählte Option zu. Sowohl die Sprünge zwischen den einzelnen Antwortoptionen als auch die Besuche auf den einzelnen Antwortoptionen zeigen zudem signifikante Zusammenhänge zur Antwortsicherheit. Mit steigender

Kategorie 1		Kategorie 2		Kategorie 3		Kategorie 4		Kategorie 5	
Prädiktoren	R^2	Prädiktoren	R^2	Prädiktoren	R^2	Prädiktoren	R^2	Prädiktoren	R^2
Leistung	.135	Leistung	.104	Leistung	.101	Leistung	.256	Leistung	.196
+TVD_O	.210	+TVD_O	.183	+VC_QAE	.212	+TVD_T	.298	+TVD_T	.264
		+Jumps_QO	.196	+TVD_T	.229	+An_TVD	.313	+VC_QO	.273

Tab. 4: Schrittweise Entwicklung von speziellen Modellen zur Beschreibung der Antwortsicherheit durch die Leistung und das Blickverhalten für die einzelnen Itemkategorien und deren Varianzaufklärung pro Stufe durch ein korrigiertes pseudo R^2

Anzahl der Sprünge bzw. Besuche nimmt die Antwortsicherheit bei allen Itemkategorien ab. Für FTR zwischen der Frage und den einzelnen Antwortoptionen besteht allerdings nur für die Itemkategorien 3 und 4 ein signifikanter positiver Zusammenhang zur Antwortsicherheit. Dies deutet darauf hin, dass für die anderen Itemkategorien der häufigere Blickwechsel zwischen den Antwortoptionen auf der längeren Bearbeitungszeit basiert.

Die in dieser Untersuchung festgestellten Zusammenhänge zwischen der Antwortsicherheit und der Anzahl der Sprünge zwischen der Frage und den Antwortoptionen bzw. der Sprünge zwischen den einzelnen Antwortoptionen bestätigen die Unterschiede im Sprungverhalten sicherer und unsicherer Personen, die Klein et al. (2021) in ihrer Clusteranalyse beobachteten.

Zusammenfassend kann zu den Zusammenhängen zwischen dem Blickverhalten und der Antwortsicherheit festgehalten werden, dass Zusammenhänge zwischen Blickzeiten bzw. Sprüngen des Blicks und der Antwortsicherheit bestehen. Diese jedoch wahrscheinlich auf der längeren Bearbeitungszeit unsicherer Personen beruhen.

Zur Beantwortung der zweiten Forschungsfrage, die nach der Varianzaufklärung der Antwortsicherheit durch das Blickverhalten fragt, wurde ein allgemeines Modell zur Beschreibung der Antwortsicherheit auf Basis der Leistung und des Blickverhaltens entwickelt. Die Varianzaufklärung des allgemeinen Modelles verbessert sich durch das Hinzufügen der Blickzeit auf das gesamte Item als Prädiktor deutlich. Die Erweiterungen um die weiteren Blickvariablen als Prädiktoren führten jeweils nur zu einer geringen Verbesserung im Promillebereich. Weil die Blickzeit auf das gesamte Item mit der Bearbeitungszeit gleichgesetzt werden kann und diese auf anderen Wegen ermittelt werden kann, lässt sich aus dieser Analyse schließen, dass das Blickverhalten wenig zur Varianzaufklärung der Antwortsicherheit beitragen kann.

Auf Ebene der Itemkategorien konnten Unterschiede der Varianzaufklärung zwischen den Itemkategorien festgestellt werden. Die Leistung scheint besonders für die Itemkategorie 4 ein gewichtiger Prädiktor zu sei. Die größte Verbesserung durch das Hinzufügen weiterer Variablen neben der Leistung konnte für die Itemkategorie 3 ausgemacht werden.

Die Analyse auf Itemebene bestätigt, dass für die Itemkategorie 4 viel Varianz allein durch die Leistung erklärt werden kann. Außerdem kann durch das Hinzufügen der Blickzeit auf das gesamte Item als Prädiktor in den meisten Fällen eine deutliche Steigerung der Varianzaufklärung erreicht werden. Im Vergleich zu diesem Modell mit der Leistung und der Blickzeit auf das gesamte Item, die der Bearbeitungszeit entspricht, als Prädiktoren konnten für das allgemeine Modell und für die speziellen Modelle keine eindeutig besseren Varianzaufklärungen festgestellt werden. Eine Ausnahme bilden die Items der Kategorie 3. Für diese Kategorie scheint das Blickverhalten gemäß dem allgemeinen Modell die Varianzaufklärung verbessern zu können. Generell schwankte der Wert der Varianzaufklärung jedoch stark zwischen den Items – auch innerhalb einer Itemkategorie.

Zusammenfassend deuten die Ergebnisse zur Varianzaufklärung der Antwortsicherheit darauf hin, dass das Blickverhalten, wie es hier durch die Blickvariablen dargestellt wurde, neben der Leistung und der Bearbeitungszeit für die meisten Itemkategorien nur wenig zur Varianzaufklärung beitragen kann.

Bereits die Überprüfung der Zusammenhänge zwischen dem Blickverhalten und der Antwortsicherheit ließ vermuten, dass die festgestellten Zusammenhänge größtenteils auf dem Zusammenhang zwischen der Bearbeitungszeit und der Antwortsicherheit beruhen. Diese Abhängigkeit sollte in weiteren Studien genauer untersucht werden.

Die Ergebnisse dieser Untersuchung decken sich mit den Erkenntnissen von Ishimaru et al. (2021). Diese fanden nur eine geringe Verbesserung der Vorhersagekraft der Antwortsicherheit, wenn neben der Bearbeitungszeit auch das Blickverhalten in die Vorhersage der Antwortsicherheit einbezogen wurde.

In dieser Untersuchung wurde das Blickverhalten nicht sequenziell erfasst. Durch eine Erhebung der Blickpfade könnten die verschiedenen Strategien zum Lösen von Multiple-Choice-Aufgaben in die Vorhersage der Antwortsicherheit einbezogen werden.

Außerdem wurde das Blickverhalten relativ grob für verschiedene AOIs untersucht, um die Ergebnisse einfach auf andere Multiple-Choice-Aufgaben anwenden zu können. Möglicherweise lohnt sich eine genauere Untersuchung des Blickverhaltens auf die Graphen, um die Antwortsicherheit präziser vorhersagen zu können.

Zudem wurde eine Tendenz festgestellt, dass die Itemkategorie 3, die aus Text und einem Graphen in der Frage und Werten als Antwortoptionen besteht, sich besser als die anderen Itemkategorien eignet, um die Antwortsicherheit aus dem Blickverhalten vorherzusagen. Eine mögliche Erklärung der besseren Eignung dieser Kategorie basiert auf den Erkenntnissen von Mitchum und Kelly (2010). Diese stellen die Constructive Matching Strategy, bei welcher zunächst die Aufgabe gelöst und dann die eigene Lösung in den Optionen gesucht wird, der Response Elimination Strategy, bei welcher die einzelnen Antwortoptionen ausgeschlossen werden bis eine übrig bleibt, als Strategien zum Lösen von Multiple-Choice-Aufgaben gegenüber und fanden eine genauere Einschätzung der Antwortsicherheit, wenn die Constructive Matching Strategy genutzt wurde.

Bei den Aufgaben der Kategorie 3 müssen entweder Werte aus einem Graphen berechnet werden oder

bestimmte Bereiche in einem Graphen identifiziert werden, sodass sich bei dieser Itemkategorie die Constructive Matching Strategy anbietet. Zunächst wird der gesuchte Wert berechnet oder der Bereich im Graphen ausfindig gemacht und anschließend wird die eigene Lösung mit den Antwortoptionen verglichen. Personen, deren Lösung mit einer der Optionen übereinstimmt, werden sich nach Mitchum und Kelly ihrer Antwort sicher sein. Diese Personen werden fokussierter sein und weniger mit ihrem Blick springen. Personen, die ihre Lösung nicht in den Optionen finden, werden häufiger mit ihrem Blick springen, um eine der Optionen auszuwählen und sich nach Mitchum und Kelly unsicher sein. Für diese Vermutung sprechen die gefundenen Zusammenhänge zwischen den FTRs und der Antwortsicherheit für diese Itemkategorie sowie kurze Blickzeiten auf die Antwortoptionen, die in der Veröffentlichung von Klein et al. (2020) angegeben sind. Die Aufgaben der anderen Itemkategorien können nicht oder nur schwer ohne Betrachtung der Optionen gelöst werden. Hier wird vermutlich vorrangig die Response Elimination Strategy Anwendung finden, sodass auch die sicheren Personen alle Optionen betrachten. Deswegen wird sich ihr Blickverhalten für diese Aufgaben weniger von dem Blickverhalten unsicherer Personen unterscheiden. Zur Überprüfung dieser Vermutung könnten die Blickpfade der Teilnehmenden analysiert werden. Bei einer Bestätigung dieser Vermutung könnte für die Erstellung von Items abgeleitet werden, dass sich Items, die sich ohne Betrachtung der Optionen lösen lassen, besser zur Erhebung der Antwortsicherheit aus dem Blickverhalten eignen.

Nach den Ergebnissen dieser Studie lohnt sich der Einsatz von Eye-Tracking momentan nicht, um die Antwortsicherheit vorherzusagen. Es sollte auf die Leistung und die Bearbeitungszeit zurückgegriffen werden. Die genannten weiteren Forschungsansätze könnten jedoch dazu führen, dass Eye-Tracking gewinnbringend zur Erfassung der Antwortsicherheit eingesetzt werden kann.

5. Literatur

Hasan, S., Bagayoko, D. & Kelley, E. L. (1999). Misconceptions and the Certainty of Response Index (CRI). *Phys. Educ., 34* (5), 294–299.

Ishimaru, S., Maruichi, T., Dengel, A. & Kise, K. (2021). *Confidence-Aware Learning Assistant.* Zugriff auf https://arxiv.org/pdf/2102.07312

Lindsey, B. A. & Nagel, M. L. (2015). Do students know what they know? Exploring the accuracy of students' self-assessments. *Phys. Rev ST Phys. Educ. Res., 11* (2), 59.

Klein, P., Becker, S., Küchemann, S. & Kuhn, J. (2021). Test of understanding graphs in kinematics: Item objectives confirmed by clustering eye movement transitions. *Phys. Rev. Phys. Educ. Res*., *17* (1).

Klein, P., Lichtenberger, A., Küchemann, S., Becker, S., Kekule, M., Viiri, J., . . . Kuhn, J. (2020). Visual attention while solving the test of understanding graphs in kinematics: an eye-tracking analysis. *Eur. J. Phys., 41* (2), 025701.

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *J. Stat. Softw., 82* (13).

Mitchum, A. L. & Kelley, C. M. (2010). Solve the problem first: constructive solution strategies can influence the accuracy of retrospective confidence judgments. *J. Exp. Psychol. Learn. Mem. Cogn., 36* (3), 699–710.

R Core Team. (2022). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. Zugriff auf https://www.R-project.org/

Rodemer, M., Graulich, N. & Bernholt, S. (2020). Decoding Case Comparisons in Organic Chemistry: Eye-Tracking Students' Visual Behavior. *J. Chem. Educ., 97*, 3530–3539.

Smith, J., Legg, P., Matovic, M. & Kinsey, K. (2018). Predicting User Confidence During Visual Decision Making. *ACM Trans. Inf. Syst., 8* (2), 1–30.

Tullis, T. & Albert, B. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (2. Aufl.). Waltham, MA: Morgan Kauffmann/Elsevier.

Yamada, K., Kise, K. & Augereau, O. (2017). Estimation of confidence based on eye gaze. In S. Lee, L. Takayama & K. Truong (Hrsg.), *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (S. 217–220). New York: ACM.

Zavala, G., Tejeda, S., Barniol, P. & Beichner, R. J. (2017). Modifying the test of understanding graphs in kinematics. *Phys. Rev. Phys. Educ. Res., 13* (2), 285.

Zhang, D. (2017). A Coefficient of Determination for Generalized Linear Models. *Am. Stat., 71* (4), 310–316.

Zhang, D. (2021). R-squared and related measures [Software-Handbuch]. Zugriff auf https://cran.r-project.org/web/packages/rsq/index.html

Danksagung

Wir bedanken uns für die Unterstützung und Beratung bei Prof. Dr. Sascha Schroeder (Institut für Psychologie, Universität Göttingen).

Anhang

BV	Kat	B_k	$SD(B_k)$	p
TVD_T	1	−0.062	0.009	<.001
	2	−0.075	0.006	<.001
	3	−0.096	0.006	<.001
	4	−0.059	0.008	<.001
	5	−0.069	0.008	<.001
TVD_Q	1	−0.014	0.010	.150
	2	−0.049	0.007	<.001
	3	−0.085	0.006	<.001
	4	−0.048	0.009	<.001
	5	−0.062	0.008	<.001
TVD_O	1	−0.067	0.009	<.001
	2	−0.076	0.006	<.001
	3	−0.092	0.006	<.001
	4	−0.050	0.008	<.001
	5	−0.054	0.008	<.001
Jumps_QO	1	−0.032	0.01ß	.001
	2	−0.066	0.006	<.001
	3	−0.082	0.006	<.001
	4	−0.058	0.009	<.001
	5	−0.046	0.008	<.001
VC_QO	1	−0.041	0.010	<.001
	2	−0.066	0.006	<.001
	3	−0.085	0.006	<.001
	4	−0.057	0.008	<.001
	5	−0.068	0.008	<.001
FTR_QO	1	−0.005	0.010	.622
	2	0.010	0.007	.134
	3	0.023	0.006	<.001
	4	0.028	0.009	.002
	5	−0.001	0.009	.868
An_TVD	1	0.032	0.010	.001
	2	0.018	0.007	.006
	3	0.042	0.006	<.001
	4	0.043	0.009	<.001
	5	0.035	0.009	<.001
Jumps_AE	1	−0.050	0.010	<.001
	2	−0.051	0.006	<.001
	3	−0.072	0.006	<.001
	4	−0.031	0.008	<.001
	5	−0.051	0.008	<.001
VC_AE	1	−0.059	0.009	<.001
	2	−0.076	0.006	<.001
	3	−0.092	0.006	<.001
	4	−0.048	0.008	<.001
	5	−0.062	0.008	<.001
VC_QAE	1	−0.058	0.009	<.001
	2	−0.077	0.006	<.001
	3	−0.095	0.006	<.001
	4	−0.053	0.008	<.001
	5	−0.066	0.008	<.001
FTR_QAE	1	0.001	0.010	.953
	2	0.011	0.007	.108
	3	0.031	0.006	<.001
	4	0.022	0.009	.015
	5	0.003	0.009	.710

Tab. 5: Gewichte B_k der einzelnen Blickvariablen pro Itemkategorie als Prädiktoren für die Antwortsicherheit

Kat.	Item	Korrigiertes R^2			
		L	L+TVD_T	all.	spez.
1	1	.163	.182	.172	.195
	9	.057	.113	.130	.137
	23	.190	.249	.256	.287
2	3	.204	.268	.246	.256
	8	.143	.153	.165	.154
	10	.070	.129	.159	.178
	17	.115	.281	.288	.313
	19	.054	.183	.210	.208
	24	.117	.267	.258	.248
	25	.134	.166	.187	.201
3	2	.153	.241	.291	.239
	4	.012	.166	.206	.190
	5	.061	.276	.355	.342
	6	.241	.336	.324	.332
	7	.120	.132	.194	.128
	13	.173	.356	.412	.351
	16	.038	.258	.301	.259
	18	.184	.321	.384	.380
4	11	.283	.358	.374	.364
	14	.244	.315	.373	.369
	15	.172	.212	.228	.205
	21	.327	.334	.327	.335
5	12	.188	.373	.361	.387
	20	.141	.212	.208	.225
	22	.207	.318	.316	.312
	26	.252	.247	.266	.252

Tab. 6: Varianzaufklärung der Modelle Leistung (L) und Leistung und Bearbeitungszeit (L+TVD_T), des allgemeinen Modells (all.) und der speziellen Modelle (spez.) für die einzelnen Items. Sortiert nach den Itemkategorien (Kat.)