Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MASTERARBEIT

Die Wirkung sprachlicher und kognitiv-fachlicher Aufgabenmerkmale in Physik auf ihre Bearbeitung durch Schüler*innen

vorgelegt von

Berenike Kesten

Fakultät für Erziehungswissenschaften

Fachbereich Didaktik der gesellschaftswissenschaftlichen und mathematisch-naturwissenschaftlichen Fächer

Studiengang: Lehramt an Gymnasien Physik - Französisch

Matrikelnummer: 6571496

Erstgutachter: Prof. Dr. Dietmar Höttecke

Zweitgutachterin: Carina von der Geest

Inhaltsverzeichnis

Abbildungsverzeichnis

Zusammenfassung

Diese Masterarbeit stellt eine Präpilotierung im Kontext des VAMPS-Forschungsprojekts dar. Dieses untersucht, ob und in welchem Maße sprachliche und kognitiv-fachliche Merkmale von Leistungsaufgaben in Mathematik und Physik deren Schwierigkeit beeinflussen. In der vorliegenden Untersuchung wurden physikalische Leistungsaufgaben, die mit Hilfe eines theoretischen Modells systematisch in Bezug auf kognitiv-fachliche Aufgabenmerkmale variiert wurden, Schüler*innen zur Bearbeitung und zur Einschätzung der Schwierigkeit vorgelegt. Dabei wurden durch kriteriengeleitete Beobachtung Hinweise zur Überarbeitung der Aufgaben gesammelt. Es zeigte sich allerdings, dass die Schwierigkeitseinschätzungen von Schüler*innen nicht die Niveaustufen des Modells reproduzierten. Deswegen hat sich im Zuge der Auswertung der Daten die Frage gestellt, ob die Annahme, dass Schüler*innen als Expert*innen für Aufgabenschwierigkeit gesehen werden können, verworfen werden muss. Aufgrund dessen wurden anschließend die Begründungen der Schüler*innen zur Einschätzung der Aufgabenschwierigkeit qualitativ ausgewertet. Dabei zeigte sich, dass Schüler*innen offenbar nur grob eine zutreffende Einschätzung vornehmen können, die nicht zur Einordung der Niveaustufen des Modells ausreichend ist.

Abstract

This master thesis is a pre-pilot-testing in the context of the VAMPS research project which addresses the question if and to what extent linguistic and subject specific cognitive features of test items in mathematics and physics assessments affect item difficulty. In the present study, test items in physics, which had been varied systematically into three levels based on a model of cognitive an content demands, were presented to student in order to complete the tasks and estimate its level of difficulty. Criteria-guided observation was used to collect information on the revision of the tasks. However, it turned out that the students' assessment of difficulty did not reproduce the levels of the model. Therefore, the assumption that students can be considered as experts for task difficulty was questioned during the data analysis. Consequently, the students' reasons for the assessment of the level of difficulty were then qualitatively evaluated. This revealed that students apparently can only make a rough estimate that is insufficient for the classification into the model's levels.

1. Einleitung

Schüler*innen werden in ihrer Schullaufbahn permanent mit Aufgaben konfrontiert: Im Unterricht beispielsweise zum Wiederholen und Festigen von Lernstoff oder in Klassenarbeiten und Abschlussprüfungen zur Überprüfung von Leistungsstand und Lernzuwachs. Test- oder Leistungsaufgaben sind ein wichtiges Instrument zur Diagnose, Bewertung und Rückmeldung von Schüler*innenleistungen. Zentrale Abiturprüfungen und Untersuchungen wie PISA und TIMMS sollen den Leistungsstand von Schüler*innen und damit das Ergebnis des Bildungssystems anhand geeigneter Aufgaben messen.

Dabei soll mit Hilfe von Leistungsaufgaben eine Eigenschaft der Person gemessen werden, zum Beispiel ein Wissenstand, eine bestimmte Kompetenz oder eine Fähigkeit. Doch die gezeigte Leistung einer Person in einem Test zu einem bestimmten Zeitpunkt ist nicht nur von diesen zu messenden Eigenschaften abhängig, sondern von weiteren unterschiedlichen und komplex interagierenden Faktoren. Dies kann ein Messergebnis mitunter stark verzerren. Um aus der gezeigten Leistung in einer Testaufgabe tatsächlich belastbare Rückschlüsse auf die Fähigkeiten einer Person zu ziehen, ist es überaus bedeutsam, Informationen darüber zu sammeln, welche Faktoren die gezeigte Leistung in welchem Maße beeinflussen. Wie beeinflussen verschiedene Merkmale der Aufgabe, also Merkmale, die nicht in der Testperson liegen, das Ergebnis? Gibt es bestimmte Aufgabenmerkmale, die ihre Schwierigkeit besonders beeinflussen? Welchen Einfluss haben beispielsweise sprachliche Merkmale der Aufgabenstellung in Physikaufgaben? Solche Effekte zu verstehen hilft dabei, sie zu kontrollieren, nicht intendierte Schwierigkeiten abzubauen und so aus der Messung einer gezeigten Leistung möglichst valide, objektive und reliable Aussagen über den Leistungs- oder Kompetenzstand einer Person ableiten zu können. Das ist notwendig, um Aufgaben lerngruppengerecht zu gestalten, Lernhürden zu vermeiden und zu verhindern, dass möglicherweise systematisch bestimmte Schüler*innengruppen in Leistungssituationen bevorzugt bzw. benachteiligt werden.

In bisherigen Untersuchungen hat sich bereits angedeutet, dass Sprache im naturwissenschaftlichen Unterricht eine wichtige Rolle spielt, die für Schüler*innen eine Hürde darstellen kann, weswegen in der Fachdidaktik eine Entwicklung hin zu einem

sprachsensiblen Fachunterricht mehr und mehr angestrebt wird. Vor dem Hintergrund einer bisher eher uneindeutigen Forschungslage soll der Einfluss der Sprache auf die Schwierigkeit von fachlichen Leistungsaufgaben nun systematisch untersucht werden. Dies ist Ziel des Forschungsprojekts „VAMPS: Variation von Aufgaben – Mathematik, Physik, Sprache“. Dabei wird die Schwierigkeit von Mathematik- und Physikaufgaben in Abhängigkeit verschiedener kognitiv-fachlicher und sprachlicher Merkmale untersucht. Dazu wurden gezielt Leistungsaufgaben entwickelt, die mit Hilfe von Modellen zur Aufgabenschwierigkeit systematisch bezüglich ihrer sprachlichen und kognitiv-fachlichen Anforderungen variiert wurden. Die vorliegende Arbeit beschreibt eine Untersuchung, die als Präpilotierung im Kontext dieses VAMPS Projektes durchgeführt wurde, um die neu entwickelten physikalischen Leistungsaufgaben mit Schüler*innen zu testen.

Diese Arbeit befasst sich zunächst mit Leistungsaufgaben aus theoretischer Perspektive und mit dem Konstrukt der Aufgabenschwierigkeit. Anschließend werden Aufgabenmerkmale, die die Schwierigkeit von Physikaufgaben aus kognitiv-fachlicher Perspektive beeinflussen, betrachtet und die daraus abgeleitete Konstruktion der Leistungsaufgaben für das VAMPS-Projekt beschrieben. Diese Aufgaben wurden in der hier vorgestellten Erhebung 58 Schüler*innen aus drei verschiedenen Schulen vorgelegt. Dabei wurde der Bearbeitungsprozess kriteriengeleitet beobachtet und die Schüler*innen zu ihrer Einschätzung der Aufgabenschwierigkeit befragt. Ein wichtiges Ziel war es, nicht intendierte Schwierigkeiten und Probleme bei den konstruierten Aufgaben aufzudecken, sowie erste Erkenntnisse über die Qualität des verwendeten Modells der schwierigkeitserzeugenden Aufgabenmerkmale bezüglich kognitiv-fachlicher Anforderungen zu gewinnen, um die Aufgaben weiter zu überarbeiten und somit das Testinstrument für die Hauptuntersuchung zu verbessern.

Dies alles dient dem Ziel der Hauptuntersuchung, den Einfluss von Aufgabeneigenschaften, insbesondere den Einfluss der Sprache auf die Aufgabenschwierigkeit für Schüler*innen, besser zu verstehen. Gerade für ein Unterrichtsfach wie Physik, das allgemein bei Schüler*innen oft als unbeliebt und schwer gilt, sind detaillierte Erkenntnisse über Aufgabenschwierigkeiten und

Einflussfaktoren von großem Wert für Didaktik und Lehrpersonen. Je besser schwierigkeitsgenerierende Merkmale erforscht sind, desto besser können Leistungstests den tatsächlichen Leistungsstand von Schüler*innen messen und desto besser kann der Schwierigkeitsgrad lerngruppengerecht angepasst werden, angemessene Hilfestellungen angeboten werden und nicht intendierte Hürden vermieden werden. All dies verfolgt das übergeordnete Ziel einer Verbesserung der Unterrichtsqualität und der Bildungsgerechtigkeit.

2. Theoretische Grundlage

2.1. Lern- und Leistungsaufgaben

Aufgaben sind zentrale Werkzeuge für Lehrkräfte zur Gestaltung des Unterrichtes. Aufgaben im naturwissenschaftlichen Unterricht können sehr Unterschiedliches von den Schüler*innen fordern: die Spanne reicht von einfachen Reproduktionsaufgaben, bei denen gelerntes Wissen erinnert und wiedergegeben werden muss bis hin zu Transferaufgaben, die eine Reorganisation des Wissens erfordern (Hopf, Schecker, & Wiesner, 2011). Sie können in vielfältigen didaktischen Zusammenhängen genutzt werden, beispielsweise, um Inhalte zu entwickeln, typische Probleme zu lösen, Problemlösen selbst zu unterrichten und Unterricht entlang der Prinzipien naturwissenschaftlichen Arbeitens zu strukturieren (Fischer & Draxler, 2002). Unter „(Physik-) Aufgaben“ werden zunächst nach Hopf et. al (2011) alle abgegrenzten Arbeitsaufträge verstanden, „die Lerner zu einer aktiven Auseinandersetzung mit einem physikalischen Sachverhalt veranlassen.“ (ebd., S. 123) Sie sollen „eine Überlegungskette [...] in Gang setzen, deren Ergebnis in schriftlicher oder mündlicher Form präsentiert wird.“ (ebd.) Leutner et al. (2008, S. 169f.) ergänzen, dass dieses „im Hinblick auf den Inhalt als (mehr oder weniger) richtig beurteilt werden kann.“ Gemäß diesen Definitionen sollen Aufgaben im Folgenden als Arbeitsaufträge verstanden werden, die einen kognitiven Prozess bei dem Bearbeitenden hervorrufen, dessen Ergebnis im Hinblick auf seine Richtigkeit beurteilt werden kann. Die Betrachtungen dieser Arbeit konzentrieren sich dabei auf schriftlich gestellte Aufgaben.

Das Thema Aufgaben erhält in den letzten Jahren in Fach- und allgemeinen Didaktiken wachsende Aufmerksamkeit. Die Lehr- Lernforschung beschäftigt sich zunehmend mit der Qualität von Aufgaben und der unterrichtlichen Aufgabenkultur. Auch für Leistungsstudien wie TIMMS oder PISA und für die Überprüfung der eingeführten Bildungsstandards spielen Test- und Diagnoseaufgaben eine bedeutende Rolle (Luthiger, 2012, S. 3). Mit geeigneten Aufgaben sollen Schüler*innen im kompetenzorientierten Unterricht einen Lernprozesszyklus durchlaufen, in dem 1.) Kompetenzen erworben und/oder verändert werden sollen; 2.) Kompetenzen

festgestellt und beschrieben werden sollen und 3.) anhand dieser diagnostischen Informationen zielgerichtet der weitere Kompetenzaufbau (stützend oder vertiefend) ermöglicht werden soll (Luthiger, 2012, S. 4). Daraus ergibt sich die Notwendigkeit zweier verschiedener Aufgabentypen mit grundsätzlichem funktionalem Unterschied: Für den ersten und dritten Schritt sind Aufgaben nötig, die dem Aufbau und der Entwicklung von Kompetenzen dienen, für den zweiten Schritt braucht es Aufgaben zur Überprüfung und Diagnose von Kompetenzen. Somit ergibt sich eine Unterteilung in *Lernaufgaben* (zur Entwicklung von Kompetenzen) und *Leistungsaufgaben* (zur Überprüfung von Kompetenzen). Eine detaillierte Übersicht über diese Einteilung verschiedener Aufgaben nach ihrer Funktion im Lehr-Lernprozess ist in Abbildung 1 dargestellt. Die Unterscheidung in Lern- und Leistungsaufgaben scheint auch aus psychologischer Sicht sinnvoll; so hat bereits Weinert (1999) in Längsschnittstudien auf die „völlig unterschiedlichen psychologischen Gesetzmäßigkeiten“ von „Lernen und Leisten verwiesen“. Da Lern- und Leistungsaufgaben unterschiedliche Ziele verfolgen, lassen sich ebenso unterschiedliche Qualitätsmerkmale ableiten, die im Folgenden knapp vorgestellt werden.

Lernaufgaben erfüllen, wie oben erwähnt, den Zweck der (Weiter-)Entwicklung von Kompetenzen. Sie sollen Möglichkeiten zur intensiven Auseinandersetzung mit den zu

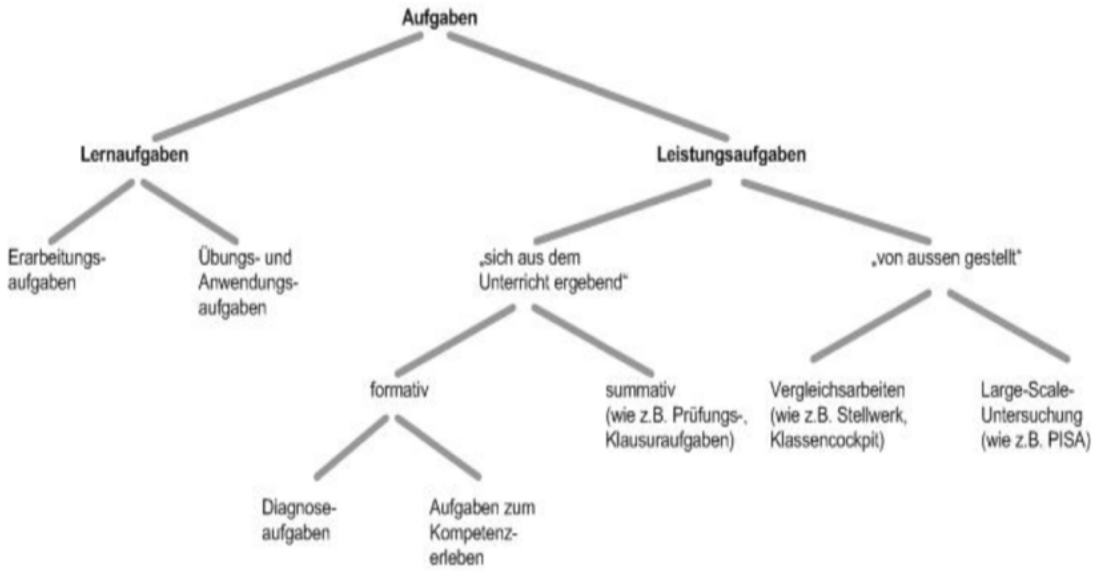


Abbildung 1: Einteilung der Aufgaben nach ihrer Funktion im Lehr-Lernprozess (Abraham & Müller, 2009, S. 6)

lernenden Inhalten bieten und dadurch einen Lernprozess bei den Bearbeitenden hervorrufen. Lernaufgaben zielen dabei auf den Bearbeitungsprozess ab, der einen Lernzuwachs bewirken soll und nicht auf die korrekte Lösung: Eine Aufgabe, die durch eine*n Schüler*in falsch gelöst wurde, kann durchaus trotzdem einen Lernzuwachs bewirkt haben. Lernaufgaben sind vor allem Aufgaben zum Erarbeiten, Üben und Anwenden. Anhand der Ziele von Lernaufgaben lassen sich verschiedene Qualitätsmerkmale formulieren, wie sie z. B. Luthiger (2012, S. 8) anhand von 4 Prinzipien vorschlägt: Prinzip der Orientierung an Lerntätigkeiten (Aufgaben sollen aus den Strukturen des Lernens heraus entwickelt werden), Prinzip der Passung (Differenzierung und Wahlmöglichkeiten), Prinzip des sozialen Austauschs (Weiterverarbeitung und sozialer Kontext fördern die Auseinandersetzung mit dem Inhalt) und Prinzip der Struktur (klare Strukturierung zur Unterstützung des Autonomie- und Kompetenzerlebens).

Leistungsaufgaben (oder auch Testaufgaben) dagegen sind ein diagnostisches Instrument, welches nicht auf den Lernprozess, sondern auf dessen Ergebnis fokussiert ist. Sie können formativ (z. B. Diagnoseaufgaben) oder summativ (z.B. Prüfungs- und Klausuraufgaben) sein. Auch Aufgaben in Vergleichsarbeiten oder Large-Scale-Untersuchungen wie z.B. PISA sind Beispiele für Leistungsaufgaben (Übersicht siehe Abbildung 1). Da Leistungsaufgaben ein grundsätzlich anderes Ziel verfolgen als Lernaufgaben, sind auch andere Qualitätsmerkmale festzustellen. In Leistungsaufgaben zeigen Schüler*innen eine Leistung: im Allgemeinen handelt es sich hierbei um die Lösung der Aufgabe. Aus der gezeigten Leistung soll eine Information über die oder den Lernende*n gewonnen werden: beispielsweise Informationen über das Erreichen einer bestimmten Kompetenz, Fähigkeit oder eines Wissensstandes. Allgemeiner formuliert: Leistungsaufgaben sollen Aufschluss über eine Eigenschaft der bearbeitenden Person geben. Die Qualität einer Leistungsaufgabe ist umso höher, je besser sie die gewünschte Eigenschaft messen kann. Die Qualitätsmerkmale sind dabei im Schulkontext nicht

anders als bei wissenschaftlichen Messinstrumenten: Objektivität, Validität und Reliabilität (Leutner, Fischer, Kauertz, Schabram, & Fleischer, 2008).

Vor allem das Merkmal der Validität stellt bei der Konstruktion von Leistungsaufgaben eine Herausforderung dar. Ein valider Test misst das, was auch gemessen werden soll. In den meisten schulischen Kontexten, in denen Leistungsaufgaben zum Einsatz kommen, soll der „Output" von Schule gemessen werden, also die Wirkung des Unterrichts auf eine*n Schüler*in. Der gewünschte Output von Schule ist mittlerweile laut Bildungsplan der Kultusminister*innenkonferenz der Erwerb von Kompetenzen, für das Fach Physik beispielsweise in den vier Bereichen Fachwissen, Erkenntnisgewinnung, Kommunikation und Bewertung (KMK, 2004). Leistungsaufgaben sollen im schulischen Kontext also in der Regel fachspezifische Kompetenzen messen. Dabei ergibt sich allerdings das grundsätzliche Problem, dass etwas gemessen werden soll, was nicht direkt messbar ist. Kompetenzen als latente Fähigkeiten und Fertigkeiten können, genauso wie andere Persönlichkeitseigenschaften, nicht direkt in Leistungsaufgaben gemessen werden, sondern lediglich eine Performanz, also die gezeigte Leistung eines Schülers oder einer Schülerin zu einem bestimmten Zeitpunkt. Aus dieser Performanz soll anschließend auf die zu messende Kompetenz geschlossen werden, in dem man davon ausgeht, dass die gezeigte Leistung diese Kompetenz als „verdeckte Ursache" habe (Kauertz, 2008, S. 11). Dazu braucht es ein Kompetenzmodell, also eine „messbare Beschreibung von Kompetenz" (ebd., S. 13). Dieses Kompetenzmodell kann in Aufgabeneigenschaften übersetzt werden, in dem Aufgaben konstruiert werden, zu deren Lösung die zu messende Kompetenz (laut Modell) notwendig ist. Wird so eine Leistungsaufgabe anschließend Schüler*innen zur Kompetenzmessung vorgelegt, erhält man zunächst eine Information über die gezeigte Leistung der Schüler*innen. Im einfachsten Fall, zum Beispiel bei Multiple-Choice-Aufgaben, gibt es nur zwei verschiedene Möglichkeiten: Entweder die Aufgabe wurde korrekt gelöst oder die Aufgabe wurde nicht korrekt gelöst. Diese Leistung muss dann mit Hinblick auf das Kompetenzmodell interpretiert werden. Ist das zu Grunde liegende Modell zutreffend, bildet die theoretische Kompetenz die „wahre Kompetenz" eines Schülers oder einer Schülerin gut ab. Vereinfacht gesagt, wenn er oder sie die Aufgabe richtig gelöst hat,

kann man erwarten, dass die zu messende Kompetenz erworben wurde. Wenn er oder sie die Aufgabe nicht korrekt gelöst hat, erwartet man, dass diese Kompetenz nicht ausreichend erworben wurde. Natürlich existieren Kompetenzen nicht nach einem binären Schema „vorhanden“ oder „nicht vorhanden“, sondern können in verschiedenen Ausprägungen vorliegen (Kater-Wettstädt, 2015, S. 30). Die Aussagekraft von kompetenzdiagnostischen Tests und damit auch die Qualität von Leistungsaufgaben hängt aber entscheidend davon ab, wie zutreffend das zugrunde liegende Kompetenzmodell ist (Leutner, Fischer, Kauertz, Schabram, & Fleischer, 2008).

Die Kompetenzmodellierung ist ein weitreichendes Themengebiet, welches im Rahmen dieser Arbeit nicht weiter betrachtet wird. Die vorgestellte Untersuchung konzentriert sich auf ein weiteres, elementar wichtiges Konstrukt zur Qualitätssicherung von Leistungsaufgaben: Die Aufgabenschwierigkeit. Denn auch wenn das zugrunde liegende Kompetenzmodell zutreffend ist, bedeutet dies nicht zwangsläufig, dass der Test die zu messenden Kompetenzen valide erfassen kann. Leistungsaufgaben stehen durch den Versuch der indirekten Messung einer Kompetenz über eine Performanz vor einem weiteren Problem: Die gezeigte Performanz ist das Ergebnis eines mitunter komplexen Bearbeitungsprozesses, dessen Erfolg nicht ausschließlich von der Ausprägung der zu messenden Kompetenz abhängig ist. Eine Vielzahl weiterer Faktoren wie äußere Umstände, Eigenschaften des Bearbeitenden und unterschiedlichste Aufgabenmerkmale beeinflussen die Bearbeitung. Einige dieser Faktoren erzeugen dabei (mitunter nur für bestimmte Schüler*innen) eine zusätzliche Schwierigkeit, die verursacht, dass bestimmte andere Kompetenzen und Fähigkeiten mitgemessen werden, die gar nicht gemessen werden sollen und somit das eigentliche Messergebnis verzerren. Dies sollte zur Qualitätssicherung von Leistungsaufgaben möglichst vermieden werden. Allerdings ist es in der Praxis oft nicht möglich, Kompetenzen, die zur Lösung einer Aufgabe notwendig sind, klar voneinander zu unterscheiden. Beispielsweise könnte ein*e Schüler*in in einem Leistungstest zwar durchaus über die zu messende Kompetenz verfügen, die Aufgabenstellung aber einfach sprachlich falsch verstanden haben und deswegen die Aufgabe nicht korrekt lösen. In diesem Fall würde die Leistungsaufgabe ein falsches Messergebnis liefern. Dies ist nur ein Beispiel für

zahlreiche Faktoren, die das Messergebnis verzerren können. Solche Faktoren und ihren Einfluss auf den Erfolg des Bearbeitungsprozesses möglichst genau zu kennen, ist entscheidend, um die Qualität einer Leistungsaufgabe zu sichern. Einflussfaktoren auf die Lösungswahrscheinlichkeit einer Aufgabe werden durch das Konzept der Aufgabenschwierigkeit beschrieben, welches im folgenden Abschnitt genauer betrachtet wird.

2.2. Aufgabenschwierigkeit

Schwierigkeit bezeichnet laut Duden „etwas, was der Verwirklichung eines Vorhabens o. Ä. im Wege steht und nicht ohne Weiteres zu bewältigen ist." (Duden, 2020) Schwierigkeit bzw. „schwierig sein" kann aber ebenso eine Eigenschaft darstellen (ebd.). So unterschiedlich Aufgaben und ihre Merkmale und Einsatzmöglichkeiten im Unterricht sein können, so unterschiedlich sind auch die Anforderungen, die sie an Schüler*innen stellen. Der Begriff der Schwierigkeit kann als Eigenschaft einer Aufgabe verstanden werden. Die Aufgabenschwierigkeit bezieht sich dann analog zur Itemschwierigkeit in der Testtheorie auf die Wahrscheinlichkeit, dass eine Aufgabe korrekt gelöst wird. Aufgaben mit hoher Aufgabenschwierigkeit werden mit geringerer Wahrscheinlichkeit korrekt gelöst als Aufgaben mit einer geringen Aufgabenschwierigkeit (Astleitner, 2008, S. 65). Gemäß dieser Definition kann Aufgabenschwierigkeit mit Hilfe von Tests mit ausreichend hoher Teilnehmerzahl empirisch durch den Prozentsatz der korrekten Lösungen bestimmt werden. Aufgaben, die von deutlich über 50 % der Schüler*innen, die diese Aufgabe bearbeitet haben, korrekt gelöst werden, gelten allgemein als leicht. Bei Lösungsprozentsätzen um die 50 % gelten Aufgaben als mittelschwer und bei Lösungsprozentsätzen von deutlich unter 50 % als schwer[1] (Astleitner, 2008, S. 69). Anhand dieser Werte lässt sich die so genannte objektive (oder empirische) Aufgabenschwierigkeit bestimmen.

Wenn es nun darum geht, die Aufgabenschwierigkeit einer Aufgabe durch bestimmte Ursachen zu erklären, zeigt sich ein Konstrukt von hoher theoretischer Komplexität mit

[1] Anmerkung: die Begriffe „schwer" und „schwierig" als Charakteristikum für die Aufgabenschwierigkeit werden hier und im Folgenden synonym verwendet.

vielen verschiedenen möglichen Einflussfaktoren. Laut Astleitner (2008, S. 68) ist die Aufgabenschwierigkeit das Ergebnis eines komplexen Interaktionsprozesses, bei dem Kontextmerkmale, Merkmale des Bearbeitenden und Aufgabenmerkmale in einem zeitlichen Ablauf zusammenwirken (siehe Abbildung 2). Als Kontextmerkmale zählen hier alle äußeren Faktoren, wie zum Beispiel Bearbeitungszeit, Raumtemperatur etc., die einen Einfluss auf die Aufgabenschwierigkeit haben können. Merkmale des Bearbeitenden sind einerseits mehr oder weniger zufällige Faktoren, wie z.B. die Müdigkeit zum Zeitpunkt der Bearbeitung, motivationale und volitionale Faktoren, aber auch Kompetenzen, Fähigkeiten, Fertigkeiten usw., die zur Lösung der Aufgabe notwendig sind. In diesem Einflussfaktor stecken folglich die Eigenschaften, die mit einer Leistungsaufgabe eigentlich gemessen werden sollen. Der dritte Einflussbereich bezeichnet Merkmale der Aufgabe selbst, wie zum Beispiel das fachliche Anforderungsniveau, aber auch formale, sprachliche oder strukturelle Merkmale der Aufgabe. Diese drei Einflussfaktoren sind keinesfalls isoliert zu betrachten, sondern stehen in einem Interaktionsprozess.

Eine besondere Rolle spielt hier der Einfluss der Aufgabenmerkmale, da diese bei der Testkonstruktion gezielt gestaltet werden können. Aufgrund der Interaktion von

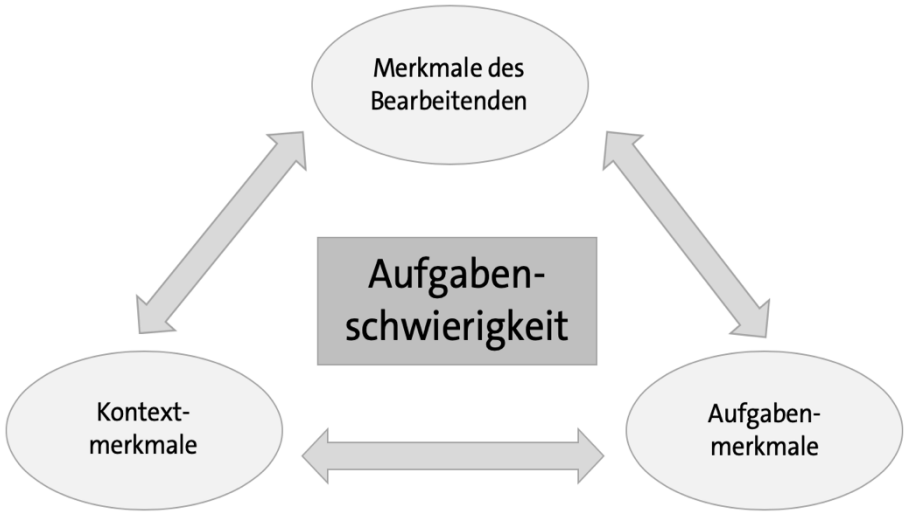


Abbildung 2: Einflüsse auf die Aufgabenschwierigkeit

Aufgabenmerkmalen mit den Merkmalen des Bearbeitenden können beispielsweise durch eine gut erforschte Schwierigkeit in einem konstruktvaliden Test fachlich-kognitive Aufgabenmerkmale gezielt variiert werden, um verschiedene Niveaustufen innerhalb einer Kompetenz bei den Lernenden messen zu können (Kauertz, 2008).

Aufgaben enthalten allerdings zwangsläufig eine Vielzahl verschiedener Merkmale, von denen einige auch Schwierigkeit erzeugen können, die unabhängig von der zu messenden Kompetenz ist, also eine konstruktferne Aufgabenschwierigkeit. Diese hat, im Gegensatz zu einigen Kontextmerkmalen, kein zufälliges Grundrauschen in der Schwierigkeitsvarianz zur Folge, sondern kann mitunter bestimmte Schüler*innengruppen systematisch benachteiligen. Ein Beispiel dafür ist ein Phänomen, welches in der Fachdidaktik als Sprach-Bias bezeichnet wird: Eine schriftlich gestellte Leistungsaufgabe, die Aufschluss über einen Kompetenzstand geben soll, misst zusätzlich zu dieser Kompetenz auch immer Lesefähigkeiten der Schüler*innen, aus dem simplen Grund, dass Lesefähigkeiten nötig sind, um überhaupt die Aufgabenstellung zu verstehen. So misst auch jede Fachaufgabe in Physik neben fachlichen Kompetenzen gleichzeitig sprachliche Kompetenzen wie Lesefähigkeit, Hörverstehen oder Sprachrezeption (je nachdem wie die Aufgabe gestellt ist). Das Medium, mit dem die Aufgabe transportiert wird, ist zwangsläufig die Sprache. So erzeugen sprachliche Fähigkeiten bei Leistungstests auch immer eine Varianz, die nicht durch die eigentlich zu messenden Eigenschaften entsteht. Diese Varianz ist mitunter abhängig von den sprachlichen Merkmalen der Aufgabe und kann durch Veränderung von Aufgabenmerkmalen beeinflusst werden.

Ziel eines validen Leistungstests muss es sein, möglichst genau den Einfluss verschiedener Faktoren auf die Aufgabenschwierigkeit zu kennen, einerseits um eine gewünschte Varianz durch die Niveaustufen der Kompetenz messbar zu machen und andererseits, um den Einfluss unerwünschter Faktoren zu minimieren. Da letzterer nicht vollkommen vermieden werden kann, ist es umso wichtiger, ihn genau zu verstehen: Unter welchen Umständen haben bestimmte Aufgabenmerkmale einen Einfluss auf die Aufgabenschwierigkeit? Wie groß ist dieser Einfluss? Für welche Schüler*innen ist er möglicherweise besonders hoch? Im Sinne von Bildungs- und Chancengerechtigkeit

spielt vor allem die letzte dieser Fragen eine wichtige Rolle. Je besser diese Einflussfaktoren bekannt sind, desto besser können sie kontrolliert oder bei der Interpretation der Ergebnisse berücksichtigt werden und desto validere Informationen erhält man über die Kompetenz, die eigentlich gemessen werden soll. Die Aussagekraft und damit die Qualität von Leistungsaufgaben hängt auch maßgeblich davon ab, wie gut schwierigkeitserzeugende Merkmale erforscht und bei der Konstruktion der Aufgabe bzw. bei der Auswertung und Interpretation der Ergebnisse berücksichtigt wurden. So liegt ein wichtiger Fokus bei der Qualitätssicherung von Leistungstests auf dem Einfluss der Aufgabenmerkmale auf die Aufgabenschwierigkeit. Deswegen konzentriert sich diese Arbeit im Folgenden auf diesen Bereich.

Aus der Komplexität der Aufgabenschwierigkeit resultiert allerdings eine generelle Begrenztheit der Aussagekraft von Leistungstests: Selbst wenn man exakt bestimmen könnte, welche Aufgabenmerkmale Schwierigkeit erzeugen und diese genau kontrollieren könnte (was man nicht erreichen wird, da sich Schwierigkeit je nach Lerner*in individuell unterscheidet, ausführlicher dazu siehe Abschnitt 2.3.1), wird man trotzdem aus einer Leistungsaufgabe nie vollkommen valide Informationen über eine bestimmte Kompetenz ableiten können. Es spielen immer auch äußere Kontextmerkmale eine Rolle, sowie weitere Merkmale des Lernenden, die nicht gemessen werden sollten und das Ergebnis verzerren. Dennoch verbessern Erkenntnisse über schwierigkeitserzeugende Merkmale einer Aufgabe die Messqualität von Leistungsaufgaben und Testinstrumenten und sind somit sowohl im Kontext von Schule (v. a. bei Klausuren oder Abschlussprüfungen) als auch in der wissenschaftlichen Forschung von hoher Bedeutung. Im nächsten Abschnitt werden deshalb schwierigkeitserzeugende Merkmale von Leistungsaufgaben näher in den Blick genommen.

2.3. Schwierigkeitserzeugende Merkmale von Leistungsaufgaben

Dem Merkmal der Aufgabenschwierigkeit kommt, wie im vorigen Abschnitt dargelegt, in Leistungsaufgaben eine hohe Bedeutung zu. Bevor sich diese Arbeit im Folgenden ausschließlich auf diesen Aufgabentyp konzentriert, sei an dieser Stelle betont, dass

Aufgabenschwierigkeit auch für Lernaufgaben ein sehr relevantes Merkmal darstellt: Erkenntnisse über die Aufgabenschwierigkeit und über die Faktoren, die einen Einfluss auf diese haben, helfen, Lernprozesse und somit Lernen und dessen Förderung gezielt und forschungsgeleitet in der Unterrichtspraxis zu gestalten. Aufgaben bezüglich ihres Schwierigkeitsgrades korrekt einzuschätzen bzw. mehrere Aufgaben korrekt nach Schwierigkeitsgrad zu ordnen hat eine bedeutende Funktion für Lehrkräfte bei der Steuerung von Lernprozessen, beispielsweise bei Binnendifferenzierung oder der Anpassung des Unterrichts an individuelle Lerner*innenbedürfnisse (Astleitner, 2008, S. 65). Die Erforschung schwierigkeitserzeugender Merkmale ist also für eine Verbesserung der Unterrichtsqualität in mehrerlei Hinsicht von Bedeutung.

2.3.1. Einschränkungen und Grenzen

Bei der Betrachtung von schwierigkeitserzeugenden Merkmalen gilt es zunächst drei zentrale, einschränkende Dinge zu beachten:

1. Aufgabenschwierigkeit kann zwar, wie bereits beschrieben, bei ausreichend hoher Teilnehmerzahl als Lösungswahrscheinlichkeit objektiv bestimmt werden, schwierigkeitsauslösende Merkmale einer Aufgabe sind allerdings subjektiv und keineswegs für alle Lernenden gleich, da Schwierigkeit immer in der Interaktion zwischen Lernendem und Aufgabe entsteht. Bestimmte Merkmale machen Aufgaben für Schüler*innen *individuell* schwierig. Diese Merkmale können für jede*n Schüler*in andere sein, weshalb bei Verallgemeinerungen Vorsicht geboten ist. Deshalb steht in der Theorie der Aufgabenschwierigkeit der objektiven Aufgabenschwierigkeit die subjektive Aufgabenschwierigkeit gegenüber, die nicht unbedingt mit der erst genannten übereinstimmen muss. Die subjektive Aufgabenschwierigkeit ist eine Einschätzung einer Person über die Schwierigkeit einer Aufgabe (Astleitner, 2008). Hier spielen also vor allem persönliche Eigenschaften der Person, die die Aufgabe beurteilt, eine Rolle.

Dennoch können in der Forschung Merkmale einer Aufgabe identifiziert werden, die einen Einfluss auf die empirische, also objektive Aufgabenschwierigkeit einer Aufgabe haben, beispielsweise Merkmale, die für besonders viele Schüler*innen oder für besondere Schüler*innengruppen Schwierigkeit erzeugen.

2. Bei der Analyse der Aufgabenschwierigkeit und dem Rückschließen auf schwierigkeitsrelevante Merkmale der Aufgabe ist immer zu bedenken, dass Aufgabenschwierigkeit, wie oben beschrieben, das Ergebnis eines komplexen Interaktionsprozesses ist (Astleitner, 2008, S. 68). Aufgabenmerkmale stellen nur einen (mitunter kleinen) Teil der schwierigkeitserzeugenden Faktoren dar. Bei der Analyse können die Einflüsse dieser verschiedenen Faktoren nicht immer klar voneinander abgegrenzt werden. Unter Kontrolle aller weiteren Einfluss- und Interaktionsfaktoren festzustellen, welchen Einfluss auf Aufgabenschwierigkeit genau ein bestimmtes Aufgabenmerkmal hat, ist ein Ideal, welches in der Praxis nicht zu erreichen ist.

3. Nicht alle schwierigkeitsrelevanten Merkmale einer Aufgabe lassen sich objektiv bestimmen. Während formale Merkmale einer Aufgabe wie Antwortformat, Textlänge, Art der Aufgabeninhalte etc. relativ leicht objektiv und niedrig interferent zu bestimmen sind, gestaltet sich dies bei prozessbezogenen Merkmalen, also kognitiven Anforderungen wie benötigtes Fachwissen, Informationsverarbeitung und allgemeinen kognitiven Prozessen deutlich schwieriger. Prozessbezogene Merkmale sind in ihrer Objektivität begrenzt, weil sie mit einem gewissen Interpretationsspielraum eingeschätzt werden müssen (Florian, Sandmann, & Schmiemann, 2014). Allerdings sind gerade solche Aufgabenmerkmale stärker an bestimmte Bearbeitungsprozesse geknüpft und daher deutlich aussagekräftiger (siehe Abschnitt 2.3.3). Merkmale, die sich auf den Inhalt der Aufgabe beziehen, können dabei oft noch mit ausreichender Objektivität bestimmt werden. Merkmale des kognitiven Verarbeitungsprozesses sind dagegen hochinterferent, also nur mit einem hohen Maß an Interpretation zu bestimmen, da es zu einer Aufgabe oft nicht nur einen einzigen, eindeutigen Lösungsprozess gibt und kognitive Prozesse außerdem oft nicht trennscharf voneinander unterschieden werden können (Kauertz, 2008, S. 24). Allein dadurch sind Modelle über schwierigkeitserzeugende Merkmale von Aufgaben grundsätzlich in ihrer Aussagekraft begrenzt.

2.3.2. Methoden der Bestimmung

Um genauer zu untersuchen, welche Merkmale einer Aufgabe Schwierigkeit erzeugen, sind verschiedene Methoden möglich. Einerseits können durch Beobachtung der Schüler*innen bei der Lösung der Aufgaben prozessorientierte Informationen gewonnen werden. Informationen über häufige Fehler, die Dauer der Bearbeitungszeit oder von den Schüler*innen angefragte Lernhilfen können bei Beobachtungsstudien Hinweise darauf geben, welche Merkmale der Aufgabe konkret Schwierigkeit auslösen (Astleitner, 2008). Hier können Informationen über die subjektive Schwierigkeit gesammelt werden, welche wiederum Hinweise auf Merkmale geben können, die die objektive Aufgabenschwierigkeit beeinflussen.

Eine weitere Methode, die einen entscheidenden Beitrag zur Erforschung von schwierigkeitsauslösenden Merkmalen leistet, stellt die Reanalyse von Aufgaben dar. Dabei wird die objektive Aufgabenschwierigkeit nach der Testbearbeitung durch eine ausreichend hohe Teilnehmerzahl mit verschiedenen Aufgabenmerkmalen in Verbindung gesetzt und mit den Erwartungen bezüglich der Schwierigkeit verglichen (Florian, Sandmann, & Schmiemann, 2014). Dabei können sowohl Merkmale berücksichtigt werden, die schon bei der Aufgabenentwicklung beachtet und variiert wurden, als auch Merkmale, die im Nachhinein induktiv abgeleitet und theoretisch begründet werden (edb). Dadurch erhält man empirisch abgesichertes Wissen mit größerer Allgemeingültigkeit über schwierigkeitsbestimmende Merkmale.

2.3.3. Bisherige Forschung: Theorie und Empirie

Diese Arbeit legt bei der Analyse schwierigkeitserzeugender Merkmale einen Schwerpunkt auf Physikaufgaben. Zwar können einige Aufgabenmerkmale sicherlich fächerübergreifend als schwierigkeitserzeugend angenommen werden (vor allem innerhalb einiger Fächergruppen wie z.B. naturwissenschaftliche Fächer), allerdings ergeben sich doch relevante Unterschiede unter den Fächern (Astleitner, 2008, S. 68). In diesem Kapitel werden zunächst allgemein einige Aufgabenmerkmale vorgestellt, die sich in der bisherigen Forschung als schwierigkeitserzeugend gezeigt haben. Im dann folgenden Abschnitt „ Aufgabenformat und -entwicklung für das VAMPS-Projekt“ wird

ein theoretisches Modell vorgestellt, mit dem bestimmte schwierigkeitserzeugende Merkmale zur gezielten Konstruktion von Physikaufgaben in verschiedenen Anforderungsniveaus genutzt wurden.

Eine Aufgabe soll, gemäß obiger Definition, einen kognitiven Prozess beim Bearbeitenden hervorrufen. Es ist also ein kognitiver Verarbeitungsprozess nötig, um eine Aufgabe korrekt zu lösen. Je aufwändiger dieser Verarbeitungsprozess ist, desto öfter passieren dabei Fehler und desto höher ist demnach die Schwierigkeit dieser Aufgabe (Kauertz, 2008, S. 46). Schwierigkeitserzeugende Merkmale einer Aufgabe sind also Merkmale, die den Aufwand der kognitiven Verarbeitung beeinflussen. Dies ist auf mehreren Ebenen möglich.

Zunächst seien nicht-fachliche, sondern formale Aufgabenmerkmale genannt, die fächerübergreifend die kognitive Verarbeitung beeinflussen. So hat sich bisher insbesondere ein Effekt des Antwortformats (z.B. Martinez, 1999) bzw. der Offenheit einer Aufgabe gezeigt. „Bei gleichem Inhalt ist eine Aufgabe umso schwieriger, je offener das Antwortformat ist“, stellten auch Fischer und Draxler (2006) fest. Eine offene Antwort zu formulieren ist also kognitiv aufwändiger und damit fehleranfälliger, als eine Antwort aus mehreren Möglichkeiten auszuwählen. Ebenso beeinflussen Stresseffekte bei der Bearbeitung den kognitiven Verarbeitungsprozess, was das Verhältnis von Testlänge zu Bearbeitungszeit zu einem relevanten Merkmal für Aufgabenschwierigkeit macht. Dies kann vor allem die Aufgabenschwierigkeit der letzten Aufgaben eines Tests erhöhen (Wuttke, 2006). Der Einfluss solcher formalen Merkmale auf die Aufgabenschwierigkeit ist allerdings als geringer anzusehen, als der Einfluss von Merkmalen, die die Inhaltsstruktur einer Aufgabe betreffen (Kauertz, 2008).

In der zweiten Ebene der Aufgabenmerkmale mit einem Einfluss auf die Aufgabenschwierigkeit stehen deshalb Merkmale, die die kognitive Anforderung aus fachlicher bzw. inhaltlicher Sicht beeinflussen. Als erstes sei hier die kognitive Tätigkeit genannt, die zur Lösung der Aufgabe vom Bearbeitenden ausgeführt werden muss. Anderson et al. (2001) benennen Tätigkeiten in sechs Stufen mit aufsteigendem kognitiven Aufwand: *Erinnern, Verstehen, Anwenden, Analysieren, Bewerten* und *Erschaffe*n. Eine Aufgabe, bei der Wissen erinnert werden muss, ist laut diesem Modell

also mit geringerem kognitiven Aufwand verbunden und damit als weniger schwierig anzusehen, als eine Aufgabe, bei der etwas bewertet oder erschaffen werden muss. Dieses Modell von Anderson et al. stellt eine Revision der Bloomschen Lernzieltaxonomie (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956) dar, die jahrzehntelang die Beschreibung von Lernzielen und Aufgabenstellungen beeinflusste. Dazu fügten Anderson et al. der Bloomschen Taxonomie eine zweite Dimension hinzu, die Wissensdimension, die in aufsteigender Reihenfolge die vier Kategorien *Faktenwissen, Konzeptionelles Wissen, Prozedurales Wissen* und *Metakognitives Wissen* enthält (Anderson & Krathwohl, 2001) und laut Modell ebenso einen Einfluss auf die kognitive Anforderung einer Aufgabe hat.

Maier, Kleinknecht, Metz und Bohl (2010) lehnen sich teilweise an dieses Modell an und formulieren ein neues, deutlich detaillierteres Modell zur Analyse des kognitiven Potentials von Aufgaben. Dieses Modell enthält ebenfalls die beiden Dimensionen *Wissensart* und *Kognitiver Prozess* (welcher der kognitiven Tätigkeit entspricht). Allerdings wird hier die hierarchische Stufung kognitiver Prozesse bei Anderson und Krathwohl (2001) in Anlehnung an die bloomsche Lernzieltaxonomie kritisch gesehen und kognitive Prozesse nur noch in vier verschiedene, klarer voneinander abgegrenzte Ausprägungen gruppiert: *Reproduktion, Naher Transfer, Ferner Transfer* und *Problemlösen*. Außerdem beschreiben Maier et al. (2010) sechs weitere Dimensionen als relevant für die kognitive Anforderung einer Aufgabe: *Wissenseinheiten, Offenheit, Lebensweltbezug, Sprachlogische Komplexität* und *Repräsentationsformen*. Die Dimensionen *Offenheit* und *Repräsentationsformen* beinhalten die bereits erwähnten formalen Merkmale einer Aufgabe. Des Weiteren unterstellen Maier et al. einen höheren kognitiven Aufwand bei einer Aufgabe mit realem Lebensweltbezug, also realer Problemlösung, als bei einer Aufgabe ohne oder mit konstruiertem Lebensweltbezug. Hervorzuheben ist an dieser Stelle die Dimension *Wissenseinheiten*. Hier zählen Maier et al. bei der Aufgabenanalyse die Wissenseinheiten ab, die nötig sind, um eine Aufgabe zu lösen und klassifizieren je nach Anzahl verschiedene kognitive Anforderungsniveaus. Je mehr Wissenseinheiten nötig sind, desto höher der kognitive Aufwand und damit auch die Aufgabenschwierigkeit. Ein sehr ähnliches Merkmal hat sich auch in einer

Untersuchung von Kauertz (2008) als ein Merkmal mit besonderer Relevanz für die Schwierigkeit einer Aufgabe gezeigt. Kauertz entwickelte ein „Inhaltsstrukturmodell“, um Zusammenhänge zwischen Merkmalen einer Physikaufgabe und ihrer Schwierigkeit zu beschreiben. Auch er verwendete in diesem Modell eine Dimension, in der die Anzahl der Wissenseinheiten abgezählt wurde. Er nannte diese Dimension *Komplexität* und verband sie zusätzlich mit der Art des Wissens, was sich auch in den beschriebenen Modellen von Maier et al. und Anderson et al. wiederfinden lässt. So beinhaltet Kauertz‘ Dimension *Komplexität* einerseits die Anzahl der inhaltlichen Elemente einer Aufgabe (Fakten bzw. Zusammenhänge), welche zur Aufgabenlösung identifiziert, erinnert oder gebildet werden müssen, und andererseits die Art der Verknüpfung dieser Elemente, wie Abhängigkeit, Bedingung und Kausalität (Kauertz, 2008, S. 47). „Je mehr Elemente eine Aufgabe enthält und je verknüpfter die Elemente in der Aufgabe sind, desto komplexer ist die Inhaltsstruktur“ (ebd.) formuliert er in seiner Studie. Es ergeben sich in seiner Untersuchung die sechs Komplexitätsniveaus: *ein Fakt, Fakten, Zusammenhang, Unverbundene Zusammenhänge, Verbundene Zusammenhänge* und *Übergeordnete Konzepte* mit aufsteigender Komplexität, also auch mit aufsteigendem kognitiven Aufwand. *Übergeordnete Konzepte* sind dabei zentrale physikalische Konzepte wie Energie oder System. Kauertz konnte in seinen Untersuchungen 23% der Varianz der Aufgabenschwierigkeit und damit dem mit Abstand größten Einfluss durch das Merkmal *Komplexität* erklären. Allerdings spiegelte sich die theoretisch vorhergesagte Komplexitätsstruktur nur teilweise in der empirischen Struktur der Aufgabeschwierigkeit wider (Luthiger, 2012). Das Modell kann also erste wichtige Hinweise liefern, seine Genauigkeit lässt sich allerdings weiter verbessern. Auch verschiedene andere Untersuchungen liefern Hinweise für den Zusammenhang von Aufgabenschwierigkeit und Komplexität (z.B. Aufschnaiter & Welzel, 1997 oder Fischer, 1994). Dies lässt sich mit der „Cognitive Load Theory“ erklären: Je höher die Komplexität einer Aufgabe, desto höher ist der mit der Bearbeitung verbundene kognitive Aufwand. Da nur eine begrenzte „Rechenkapazität“ des kognitiven Systems zur Verfügung steht, verfügt es „bei zu großer Elemente-Interaktivität nicht mehr über ausreichende Kapazität für die notwendigen Verarbeitungsprozesse“ (Kauertz, 2008, S. 46). Die Folge

sind Fehler und damit eine geringere Lösungswahrscheinlichkeit bzw. höhere Aufgabenschwierigkeit.

Zusätzlich zum Merkmal *Komplexität* formuliert Kauertz zwei weitere Dimensionen in seinem Inhaltsstrukturmodell zur Beschreibung schwierigkeitserzeugender Merkmale von Physikaufgaben und nennt diese *Leitidee* und *Kognitive Aktivität*. Mit der Dimension *Leitidee* konnte Kauertz 12 % der Schwierigkeitsvarianz erklären. *Leitidee* bezeichnet die übergeordneten Konzepte, die sich auch in den Bildungsstandards finden lassen: *Energie, Materie, Wechselwirkung* und *System*. Kauertz ergänzt in seinem Modell zusätzlich die Konzepte *Naturwissenschaftliche Arbeitsweisen* und *Formalismus*. Daraus lässt sich schließen, dass durch die Leitideen inhaltsspezifische Teilkompetenzen beschrieben werden können (Kauertz, 2008). Der thematische Inhalt einer Physikaufgabe hat demzufolge ebenso einen Einfluss auf die Aufgabenschwierigkeit.

Die dritte Dimension *Kognitive Aktivität* zeigte in Kauertz‘ Untersuchungen keinen signifikanten Einfluss auf die Aufgabenschwierigkeit. Die *Kognitive Aktivität* charakterisiert dabei in seiner Studie die Inhaltsstruktur der Aufgabe in Abhängigkeit davon, ob die Konzepte und Operatoren den Schüler*innen bekannt sein können. Dies wurde anhand der Curricula in die drei Bereiche *erinnern* (curriculare Elemente müssen nur reproduziert werden), *strukturieren* (Inhalt muss kognitiv verarbeitet werden, logische Beziehungen kommen nicht im Curriculum vor) und *explorieren* (Strukturen auf unbekannten Inhalt bzw. nicht-curriculare Konzepte anwenden) eingeteilt. Allerdings merkt Kauertz in Bezug auf die Bewertung dieses Ergebnisses an, dass eine mögliche Fehlerquelle und Erklärung für den fehlenden Nachweis eines signifikanten Einflusses darin liegen könnte, dass curriculare Inhalte und tatsächlich Gelerntes möglicherweise stark differenzieren (Kauertz, 2008, S. 119). Verschiedene kognitive Tätigkeiten wie *Verstehen, Anwenden, Bewerten* etc. wie bei Maier et al. wurden dabei nicht genauer differenziert, was eventuell ebenfalls eine Erklärung für den fehlenden Nachweis eines Einflusses dieser Kategorie sein könnte.

Auch die Kultusministerkonferenz hat Aussagen über schwierigkeitserzeugende Aufgabenmerkmale für Physikaufgaben getroffen. Sowohl in den Einheitlichen Prüfungsanforderungen für die Abiturprüfungen als auch in den Bildungsstandards

finden sich stets drei Anforderungsbereiche. Die KMK schreibt dazu: „Es handelt sich [...] um Merkmale von Aufgaben, die verschiedene Schwierigkeitsgrade [...] abbilden können." (KMK, 2004). Die Anforderungsbereiche werden allgemein (in aufsteigender Reihenfolge der Schwierigkeit) mit *Reproduzieren*, *Anwenden* und *Problemlösendes Denken* bezeichnet. Diese Anforderungsbereiche sind für die einzelnen fachspezifischen Kompetenzen in den Bildungsstandards konkreter formuliert. Auch hier werden, ähnlich wie in den Modellen von Maier, Anderson oder Bloom, Merkmale beschrieben, die Aufgabenschwierigkeit anhand der kognitiv unterschiedlich anspruchsvollen Tätigkeiten wie z.B. *Reproduzieren* (Anforderungsbereich I), *Anwenden* (Anforderungsbereich II) und *Bewerten* (Anforderungsbereich III) differenziert (KMK, 2004). Florian, Sandmann und Schmiemann (2014) haben in einer Reanalyse von Abituraufgaben im Fach Biologie, die von gut 2300 Prüflingen bearbeitet wurden, festgestellt, dass die kognitiven Anforderungen, angelehnt an diese Anforderungsbereiche sowie Anforderungen an die Nutzung von Fachwissen, Lösungswege und Informationsverarbeitung „zufrieden stellende Anteile der Schwierigkeitsvarianz der Abiturprüfung erklären" (Florian, Sandmann, & Schmiemann, 2014, S. 175).

Zusammenfassend lässt sich festhalten, dass sich vor allem bei verschiedenen Merkmalen, die sich auf die Inhaltstruktur einer Aufgabe beziehen, wie *Komplexität* (also Anzahl und Art der Verknüpfung der Elemente der Aufgabe) oder *Kognitive Tätigkeit* ein Einfluss auf die Aufgabenschwierigkeit zeigt. Aber auch formale Aufgabenmerkmale wie Antwortformat, Offenheit, Position im Testheft und Bearbeitungszeit haben einen belegten Einfluss auf die Aufgabenschwierigkeit und können somit Störeffekte bei der Leistungsmessung erzeugen.

Neben den beiden beschriebenen Ebenen der formalen Aufgabenmerkmale und der Aufgabenmerkmale, die die kognitive Verarbeitung aus fachlicher Sicht beeinflussen, werden nun in einer dritten Ebene Aufgabenmerkmale betrachtet, die ebenfalls einen Einfluss auf den kognitiven Verarbeitungsaufwand und damit auf die Aufgabenschwierigkeit haben, aber vor allem bei Leistungsaufgaben in den naturwissenschaftlichen Fächern unabhängig vom fachlichen Inhalt sind und somit

konstruktferne Aufgabenschwierigkeit erzeugen[2]: Es handelt sich um sprachliche Merkmale des Aufgabentextes. *Sprachologische Komplexität* wird in dem Modell des kognitiven Potentials von Aufgaben von Maier et al. (siehe oben) als eine der sechs relevanten Dimensionen genannt. Laut den Autoren kann „die sprachliche Darstellung der Aufgabenstellung oder der Aufgabeninformationen [...] wesentlich zum kognitiven Anforderungsniveau einer Aufgabe beitragen." (Maier, Kleinknecht, Metz, & Bohl, 2010). Auch Kauertz nennt sprachliche Merkmale wie Satzlänge, verwendete Wörter, Verwendung indirekter Rede etc., aus denen sich Störeffekte ergeben können: „Da über die Texte eine Kommunikation mit dem Bearbeiter [oder der Bearbeiterin] stattfindet und Kommunikation sehr anfällig für Fehler ist (z.B. Watzlawick et al., 1974), führen Fehlinterpretationen des Texts bereits zu Resultaten, die nicht durch das Modell erklärt werden." (Kauertz, 2008, S. 53) Um den Unterricht möglichst diskriminierungsfrei zu gestalten, fachliches Lernen gezielt zu unterstützen und die Validität fachlicher Leistungstests zu steigern, wird in der Fachdidaktik schon seit einiger Zeit im Kontext des sprachsensiblen Fachunterrichts eine Reduktion der sprachlichen Komplexität von Aufgaben angestrebt. Die Effekte auf die Aufgabenschwierigkeit sind bisher allerdings nicht ausreichend belegt: Zum Einfluss von sprachlichen Merkmalen auf Aufgabenschwierigkeit, insbesondere in Physik sowie anderen mathematischen und naturwissenschaftlichen Fächern, gibt es bereits einige Studien, deren unterschiedliche Ergebnisse allerdings den Schluss nahelegen, dass verschiedene linguistische, inhaltliche und Textmerkmale, die mit Aufgabenschwierigkeit zusammenhängen, in einer komplexen Weise miteinander interagieren, die zur Zeit noch nicht vollständig erforscht und verstanden ist (Höttecke, Feser, Heine, & Ehmke, 2018). So konnten Cassels und Johnstone (1984) einen Effekt auf die Lösungshäufigkeit zeigen, wenn Aufgaben sprachlich vereinfacht wurden (bezüglich Textlänge, Wortschatz und Verwendung von Negationen). Auch verschiedene weitere Untersuchungen zeigten eine verringerte Aufgabenschwierigkeit durch sprachliche Vereinfachungen der Aufgabenstellung (Übersicht siehe Höttecke, Feser, Heine, & Ehmke, 2018). Allerdings gab es auch

[2] Natürlich gibt es auch sprachliche Merkmale, die an den fachlichen Inhalt geknüpft sind, wie zum Beispiel physikalische Fachwörter oder Kollokationen. Diese werden hier aber als Fachwissen verstanden und sind mit den beschriebenen sprachlichen Merkmalen nicht gemeint.

widersprüchliche Ergebnisse: Haag et al. (2015) konnten keine Effekte sprachlicher Vereinfachungen nachweisen. Auch Kieffer et al (2009) gelang dies nicht, sondern lediglich der Beleg eines positiven Effekts der Nutzung von Nachschlagewerken wie Wörterbüchern bei der Bearbeitung der Aufgabe.

An dieser uneindeutigen Forschungslage setzt das Forschungsprojekt „Variation von Aufgaben – Mathematik, Physik, Sprache“, kurz VAMPS, an, mit dem Ziel, verschiedene sprachliche Merkmale des Aufgabentextes und ihren Einfluss auf die Aufgabenschwierigkeit von Fachaufgaben in Mathematik und Physik genauer zu analysieren. Die in dieser Arbeit vorgestellte Untersuchung wurde als Präpilotierung im Kontext des VAMPS-Projektes durchgeführt. Im nächsten Abschnitt wird deswegen zur Einordnung der Untersuchung ein Überblick über dieses Forschungsprojekt und die dafür gezielt konstruierten Leistungsaufgaben gegeben.

3. Einbettung der Untersuchung in das VAMPS-Projekt

Das Projekt „Variation von Aufgaben – Mathematik, Physik, Sprache“, kurz VAMPS, untersucht den Einfluss kognitiv-fachlicher und sprachlicher Merkmale auf die Aufgabenschwierigkeit von Mathematik und Physikaufgaben. Das Projekt wird seit seinem Beginn 2019 von der DFG gefördert und ist an den Universitäten Lüneburg (AG Leiß und Ehmke), Hamburg (AG Schwippert und Höttecke) und Bochum (AG Heine) angesiedelt. Diese Masterarbeit entstand im Kontext dieses Projektes zur ersten Präpilotierung der als Testinstrument entwickelten physikalischen Leistungsaufgaben. Um die Ziele dieser Masterarbeit besser einordnen zu können, wird in diesem Abschnitt das Projekt als Rahmen für die Untersuchung näher vorgestellt. Der Schwerpunkt liegt dabei auf der Beschreibung der Entwicklung von physikalischen Leistungsaufgaben als Testinstrument und des theoretischen Modells, mit dessen Hilfe die Schwierigkeit dieser Aufgaben durch Merkmale der kognitiv-fachlichen Anforderung der Leistungsaufgaben systematisch variiert wurde.

3.1. VAMPS Projekt: Voruntersuchungen und Ziele

Im Rahmen des VAMPS-Projekts sollen in einer experimentellen Studie Leistungsaufgaben aus den Fächern Mathematik und Physik hinsichtlich schwierigkeitserzeugender Merkmale untersucht werden. Ein Fokus liegt dabei auf der systematischen Untersuchung des Einflusses sprachlicher Aufgabenmerkmale auf die Schwierigkeit der Fachaufgaben.

Im vorigen Abschnitt wurden sprachliche Merkmale des Aufgabentextes als Merkmale, die einen Einfluss auf die Aufgabenschwierigkeit haben können, bereits thematisiert. In den Fächern Mathematik und Physik sollen in Leistungsaufgaben keine sprachlichen Fähigkeiten, sondern fachliche Kompetenzen diagnostiziert werden. So sollten Texte in Leistungsaufgaben möglichst für alle Schüler*innen verständlich sein, damit die Messung der fachlichen Leistung nicht durch Einflusseffekte der sprachlichen Fähigkeiten verzerrt wird (Höttecke, Feser, Heine, & Ehmke, 2018). Eine hohe konstruktferne Aufgabenschwierigkeit, erzeugt durch sprachliche Aufgabenmerkmale, würde systematisch Schüler*innen mit weniger ausgeprägten sprachlichen Fähigkeiten (zum Beispiel Schüler*innen mit Deutsch als Zweitsprache) in Prüfungssituationen benachteiligen. Da ein Einfluss sprachlicher Merkmale nicht zu verhindern ist (die Aufgabe muss zwangsläufig im Medium der Sprache übertragen werden), ist es umso wichtiger, die Effekte möglichst genau zu verstehen, um sie in Testsituationen möglichst gut kontrollieren zu können. Es ist also zu untersuchen, ob sich ein Effekt sprachlicher Merkmale auf die Aufgabenschwierigkeit zeigt und unter welchen Bedingungen sein Einfluss wie groß ist. Dazu gibt es bereits Voruntersuchungen, an die das VAMPS-Projekt anknüpft:

In einer Untersuchung von Höttecke, Feser, Heine, & Ehmke (2018) wurden sprachliche Einflüsse auf die Aufgabenschwierigkeit von Physikaufgaben detaillierter untersucht. Dafür wurden sechs verschiedene Physikaufgaben entwickelt, die je aus einem Text (Itemstamm) und einer Multiple-choice-single-select Aufgabe bestanden. Der Itemstamm wurde anhand verschiedener sprachlicher Merkmale, die sich in der Forschung als relevant für die Schwierigkeit eines Textes gezeigt hatten, systematisch in drei verschiedenen sprachlichen Anforderungsniveaus variiert. Diese Aufgaben wurden

anschließend von gut 1300 Schüler*innen in je einem der sprachlichen Niveaus bearbeitet. Zusätzlich absolvierten alle Schüler*innen eine reduzierte Variante des gängigen C-Tests zur Feststellung von Sprachfähigkeiten. Es konnte gezeigt werden, dass die sprachlichen Fähigkeiten der Schüler*innen, welche vorher mit Hilfe des C-Tests erfasst wurden, bis zu einem gewissen Grad geeignet waren, um die Lösungswahrscheinlichkeit der Aufgabe vorauszusagen. Fast alle verwendeten Items wurden mit höherer Wahrscheinlichkeit gelöst, wenn der Schüler oder die Schülerin höhere sprachliche Fähigkeiten besaß (Höttecke, Feser, Heine, & Ehmke, 2018). Allerdings traf die Annahme, dass ein höheres sprachliches Anforderungsniveau zu einer geringeren Lösungswahrscheinlichkeit der Aufgabe führe, nur bei zwei der sechs verwendeten Aufgaben zu. Für eine der Aufgaben konnte sogar ein gegenteiliger Effekt festgestellt werden: Die beiden Versionen der Aufgabe mit höheren sprachlichen Anforderungsniveaus wurden mit größerer Wahrscheinlichkeit korrekt gelöst als das leichteste sprachliche Anforderungsniveau. Allgemein kommt die Studie zu dem Schluss, dass der Effekt sprachlicher Oberflächenmerkmale auf die Itemschwierigkeit, wenn er existiert, eher gering ist (Höttecke, Feser, Heine, & Ehmke, 2018, S. 5), weitere Forschung aber nötig ist.

In Anlehnung an diese Forschungsergebnisse wurden für das VAMPS-Projekt neue Leistungsaufgaben entwickelt, in denen die Variation der sprachlichen Anforderung deutlich systematischer geschehen sollte als bei der Voruntersuchung. Da Aufgabenschwierigkeit ein komplexes Produkt verschiedener miteinander in Zusammenhang stehender Faktoren ist, sollen mit dem neuen Aufgabenformat zusätzlich Interaktionseffekte zwischen kognitiv-fachlichen (also den Fachinhalt betreffenden) und sprachlichen Merkmalen einer Aufgabe in Bezug auf ihre Schwierigkeit untersucht werden. Deswegen wurden die entwickelten Leistungsaufgaben so konstruiert, dass sie zusätzlich zu der Variation in drei sprachlichen Anforderungsniveaus auch in drei kognitiv-fachlichen Anforderungsniveaus systematisch variieren. Damit sollen die Effekte dieser Merkmale einzeln ebenso wie Interaktionseffekte untersucht werden können. Ziel ist es, anhand der Ergebnisse praxisrelevante Aussagen darüber abzuleiten, wie Leistungsaufgaben

gestaltet werden müssen, um eine möglichst objektive, reliable und valide Leistungsdiagnostik und damit auch Leistungsbewertung zu erreichen (Projektbeschreibung VAMPS, 2020). Das Aufgabenformat wird im folgenden Abschnitt detaillierter vorgestellt.

3.2. Aufgabenformat und -entwicklung für das VAMPS-Projekt

Die für das VAMPS-Projekt entwickelten Physikaufgaben sind alle nach dem gleichen Schema aufgebaut, um möglichst viele Aufgabenmerkmale, außer den gezielt variierten sprachlichen und kognitiv-fachlichen Merkmalen, konstant zu halten. So kann in der Auswertung nach der Testbearbeitung durch eine ausreichend hohe Teilnehmerzahl analysiert werden, ob sich ein Teil der empirisch festgestellten Varianz in der Aufgabenschwierigkeit durch diese Merkmale erklären lässt.

Dabei wurden die Aufgaben gezielt so entwickelt, dass der Aufgabentext eine wichtige Rolle bei der Bearbeitung der Aufgabe spielt. So ist zu erwarten, dass in diesen Aufgaben sprachliche Merkmale eine größere Rolle spielen als in den meisten „herkömmlichen" Physikaufgaben, deren Aufgabentexte oft recht kurz gehalten sind. Dadurch soll der Einfluss der Sprache auf die Aufgabenschwierigkeit gezielt verstärkt werden, um ihn zunächst deutlicher messbar zu machen. Eine Aufgabe, die einer Schülerin oder einem Schüler in der Untersuchung vorgelegt wird, besteht deswegen immer aus einem Aufgabentext, der „Aufgabenstamm" genannt wird. Zu dem Text gibt es eine Multiple-Choice Aufgabe, die im Folgenden als „Item" bezeichnet wird. Diese besteht wiederum aus einer Frage bzw. Handlungsaufforderung („Prompt") und fünf verschiedenen Antwortmöglichkeiten, von denen immer genau eine richtig ist. Jeder Aufgabenstamm existiert in drei verschiedenen Versionen, die inhaltlich identisch sind, aber nach sprachlichen Merkmalen drei verschiedenen Anforderungsniveaus entsprechen. Ebenso existieren je drei verschiedene Items, die nach kognitiv-fachlichen Merkmalen ebenfalls in drei verschiedenen Anforderungsniveaus konstruiert wurden. Insgesamt ergibt sich somit eine 3x3 Matrix, die „Unit" genannt wird (siehe Abbildung 3). Eine Unit besteht also aus drei verschiedenen Aufgabenstämmen, differenziert nach sprachlichem Niveau und drei verschiedenen Items, differenziert nach kognitiv-fachlichem Niveau. So kann

einem Schüler oder einer Schülerin eine beliebige Kombination aus sprachlichem und kognitiv-fachlichem Anforderungsniveau vorgelegt werden.

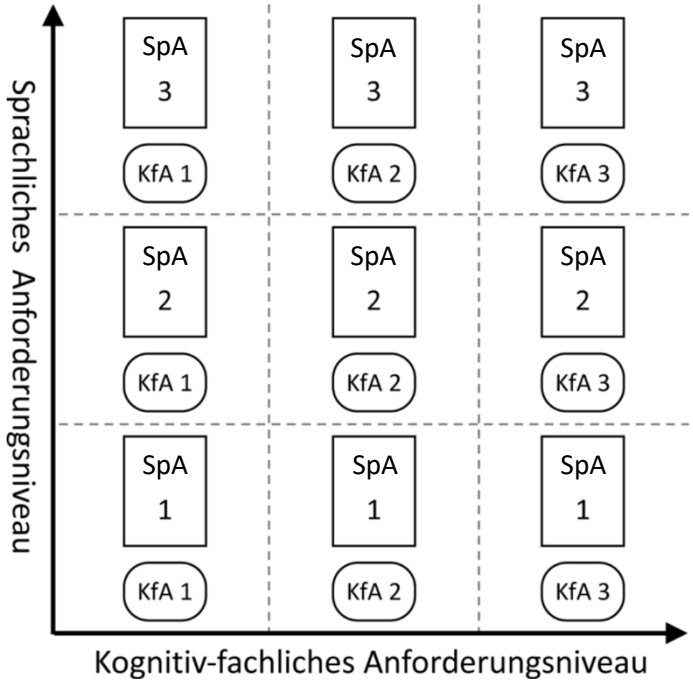


Abbildung 3: schematische Darstellung einer Unitmatrix

Die Items wurden stets so konstruiert, dass sie nicht ohne Informationen aus dem Text lösbar sind. Dies stellt sicher, dass die Schüler*innen den Text auch tatsächlich aufmerksam gelesen haben. Zusätzlich wurde zum korrekten Lösen eines Items immer physikalisches Vorwissen benötigt, damit auch kognitiv-fachliche Aspekte immer relevant sind und nicht einfach nur Lesefähigkeiten ausreichen, um ein Item zu lösen.

Nachdem 35 Aufgabenstämme mit je drei Items zunächst grob entwickelt wurden, wurden die kognitiv-fachlichen Merkmale der Items systematisch erfasst und untersucht, um die Items einem Modell entsprechend den drei Anforderungsbereichen zuzuordnen. Dies geschah diskursiv mit verschiedenen Experten für Physikdidaktik. Die Items wurden dabei teilweise verändert und überarbeitet, damit anschließend 35 modellkonforme Units in die Präpilotierung starten konnten. Zum Zeitpunkt der Präpilotierung lagen die Units nur in den kognitiv-fachlichen Variationen vor, die sprachliche Variation des Aufgabenstammes erfolgte erst später. Dadurch liegt der Fokus dieser Arbeit auf den kognitiv-fachlichen Anforderungsniveaus der Items. Das zur

Konstruktion verwendete und gezielt für dieses Projekt entwickelte Modell der kognitiv-fachlichen Anforderungen wird im nächsten Abschnitt vorgestellt.

3.2.1. Modell: kognitiv-fachliche Anforderung

Alle Units behandeln inhaltlich möglichst lebensweltlichen Problemkontext zum Themengebiet aus dem Bereich der Elektrizitätslehre. Für jede Aufgaben-Unit wurden drei verschiedene Items in den drei kognitiv-fachlichen Anforderungsniveau-Stufen I, II und III entwickelt. „Kognitiv-fachlich“ meint dabei, dass diese Merkmale den physikalischen, also fachlichen Inhalt einer Aufgabe betreffen und damit den kognitiven Verarbeitungsprozess beinflussen. Dazu wurde ein theoretisches Modell der kognitiv-fachlichen Anforderungen genutzt, mit dem die Items gezielt anhand von Merkmalen in drei verschiedene Anforderungsniveaustufen klassifiziert werden konnten. Dabei nehmen die kognitiv-fachlichen Anforderungen (kurz kfA) laut Theorie mit steigendem Anforderungsniveau zu, weswegen erwartet wird, dass ebenso die Aufgabenschwierigkeit mit steigendem Anforderungsniveau zunimmt. Das theoretische Modell entstand auf Grundlage bisheriger Forschung zu schwierigkeitserzeugenden Merkmalen von (Physik-)Aufgaben und wurde diskursiv mit verschiedenen Experten entwickelt. Die genaue Entstehung dieses Modells im Rahmen des VAMPS Projektes ist nicht Gegenstand dieser Arbeit. Eine besondere Herausforderung an das Modell war allerdings, aus der Vielzahl der in der bisherigen Forschung betrachteten Dimensionen einer Aufgabe besonders relevante Merkmale für die kognitiv-fachlichen Anforderungen herauszuarbeiten und diese in ein Modell zu übersetzen, was trotz der Komplexität dieses Themengebietes möglichst schlicht und damit für die Aufgabenkonstruktion gut handhabbar ist. So wurden beispielsweise einige der in anderen Modellen beschriebenen Merkmale nicht berücksichtigt, sondern konstant gehalten. Dadurch konnte Variation in den drei kognitiv-fachlichen Anforderungsniveaus auf eine überschaubare Anzahl verschiedener Dimensionen beschränkt werden. Lebensweltbezug, Offenheit und Repräsentationsformen (genannt im Modell von Maier et al. (2010)) wurden beispielsweise nicht variiert.

In dem Modell wurden sechs verschiedene Dimensionen formuliert, die Merkmale einer Aufgabe betreffen, die ihre kognitive Anforderung beeinflussen. Diese Dimensionen sind: *Anzahl der Prozeduren, Tätigkeitsart, Anzahl der expliziten aus dem Stamm zu identifizierenden Informationen, Anzahl der impliziten aus dem Stamm zu identifizierenden Informationen, Curriculare Gängigkeit* und *Abstraktheit der Darstellungsebene*. Des Weiteren wurden bei der Analyse der Aufgaben die Dimensionen *Anzahl der nötigen fachlichen Vorwissenselemente* und *Anzahl der nötigen alltäglichen Vorwissenselemente* zur Bestimmung des kfA-Niveaus ergänzt. An dieser Stelle sei betont, dass es sich bei der Untersuchung im Rahmen dieser Masterarbeit um eine Präpilotierung der Items in einem sehr frühen Stadium handelte. Das hier vorgestellte Modell zur Beschreibung der kognitiv-fachlichen Anforderungsniveaus ist eine erste Version des Modells, welches im weiteren Verlauf des Projektes noch deutlich ausgeschärft werden wird.

In Tabelle 1 sind die Kriterien dargestellt, anhand derer für jede Unit gezielt drei verschiedene Items in drei verschiedenen kognitiv-fachlichen Anforderungsniveaus konstruiert wurden. Es ist zu erwarten, dass die Gesamtschwierigkeit eines Items für Schüler*innen stark von der Unit abhängig ist und zu einem großen Teil durch den Kontext der Aufgabe und das Vorwissen des Schülers bzw. der Schülerin bestimmt wird. In jedem Stamm wird ein anderer lebensweltlicher Kontext genannt, muss ein anderes Situationsmodell von den Bearbeitenden gebildet werden und ist anderes fachliches Vorwissen nötig. All diese Dinge haben einen großen Einfluss auf die kognitiv-fachliche Anforderung und auf die Schwierigkeit des Items. So wird durchaus erwartet, dass einige Units möglicherweise eine höhere Schwierigkeit haben als andere Units. Ziel ist also nicht, die empirische Schwierigkeit aller Items im selben kfA-Niveau aneinander anzugleichen. Die drei kfA-Niveaustufen sollen lediglich eine Abstufung der Schwierigkeit relativ zueinander, also innerhalb einer Unit erzeugen. Zur Veranschaulichung ist ein Beispiel für eine Variation der Schwierigkeit dreier Units und deren Items relativ zueinander in Abbildung 4 schematisch dargestellt.

Abbildung 4: Beispiel für eine mögliche Variation der Schwierigkeit dreier Units und deren Items relativ zueinander: Die kfA Niveaustufen erzeugen eine relative Abstufung der Schwierigkeit der Items innerhalb einer Unit, aber nicht absolut. So kann beispielsweise das Item im kfA-Nivea III aus Unit 1 eine geringere empirische Schwierigkeit haben als das Item im kfA-Niveau I aus Unit 3.

Aus diesem Grund wurden bei der Konstruktion und Analyse der Items die meisten Merkmale der Aufgabe in den drei kfA-Niveaus relativ zueinander betrachtet. Die Aspekte *Anzahl der Prozeduren* und *Tätigkeitsart* sollten sich dabei zwischen den drei Niveaus immer verändern. Alle anderen Aspekte können, müssen sich aber nicht ändern. Während der Analyse wurde diese Vorgabe dahingehend modifiziert, dass zumindest in einem der beiden genannten Merkmale zwischen zwei kfA-Stufen eine Veränderung stattfinden musste. Die einzelnen Aufgabenmerkmale, anhand derer laut Modell die kognitiv-fachliche Anforderung und damit auch die Schwierigkeit der Items systematisch variiert werden soll, sind im Folgenden kurz dargestellt.

Das Merkmal *Anzahl der Prozeduren* wirkt sich auf die kognitive Verarbeitung aus und stimmt größtenteils mit dem Aspekt der *Komplexität* aus dem Inhaltsstrukturmodell von Kauertz überein (siehe Abschnitt 2.3.3). Um die Anzahl der Prozeduren eines Items zu bestimmen, wurde analysiert, welche Gedankenschritte zur korrekten Lösung notwendig sind. Dabei wurden die Zusammenhänge gezählt, die gebildet werden mussten. Als *einschrittig* (kfA-Niveau I) wurden Items eingestuft, wenn ein Zusammenhang aus zwei Wissenselementen bzw. Informationen gebildet werden musste. Als *mehrschrittig einfach* oder *zweischrittig* (kfA-Niveau II) wurde ein Item eingestuft, wenn zwei Zusammenhänge gebildet werden mussten. Diese konnten entweder unabhängig voneinander getroffen werden, oder logisch aufeinander aufbauen. Als *mehrschrittig komplex* (kfA-Niveau III) wurden Items eingestuft, wenn mehr als zwei Zusammenhänge gebildet werden mussten, die zusätzlich miteinander verknüpft werden mussten. Um die Beziehung der Zusammenhänge untereinander zu

berücksichtigen, wurde bei der Analyse systematisch erfasst, ob die Zusammenhänge in bestimmter Reihenfolge gebildet werden müssen, da sie logisch aufeinander aufbauen. Als *mehrschrittig einfach* gelten Zusammenhänge, die einzeln, ohne feste Reihenfolge gebildet werden können, als *mehrschrittig komplex* gelten dagegen Zusammenhänge, die untereinander verknüpft sind, wo also das Finden eines Zusammenhangs die Voraussetzung dafür ist, dass der zweite Zusammenhang gefunden werden kann. Die Anzahl dieser Prozeduren hat sich über die drei kfA-Niveaus immer verändert, um eine deutliche Veränderung der kognitiven Anforderung zu bewirken.

Ein weiteres Merkmal, welches auch unter anderem bei Kauertz und bei Maier et al. als relevant für die kognitive Komplexität befunden wurde und somit relativ gut theoretisch fundiert ist, ist die Anzahl der Wissenselemente bzw. Informationen, die zur korrekten Lösung der Aufgabe benötigt werden. Je höher diese Anzahl ist, desto höher die benötigte kognitive Verarbeitungskapazität und damit auch das Anforderungsniveau. So wurden bei der Aufgabenanalyse die Anzahlen von folgenden Elementen erfasst: *explizite aus dem Stamm zu identifizierende Informationen, implizite aus dem Stamm zu identifizierende Informationen, benötigte fachliche Vorwissenselemente, benötigte alltägliche Vorwissenselemente*. Explizite Informationen dem Stamm zu entnehmen ist dabei laut Modell mit geringerer kognitiver Aktivität verbunden als implizite Informationen zu erschließen. Die Anzahl jedes dieser Elemente sollte sich von einer niedrigeren zur höheren kfA-Stufe erhöhen oder konstant bleiben. Eine Vorgabe für die absoluten Anzahlen gab es nicht.

Mit dem Merkmal *Tätigkeitsart* wurde erfasst, welche Anforderungen an die Schüler*innen bezüglich der Handlungsaufforderung der Aufgabe gestellt wurden. Die für die einzelnen Niveaus definierten Tätigkeitsarten finden sich im Wesentlichen auch in den drei Anforderungsniveaus der Kultusminister*innenkonferenz wieder und erhalten dadurch ihre theoretische Begründung. Bei der diskursiven Analyse der Aufgaben wurden einige Tätigkeitsarten ergänzt, so zum Beispiel die Differenzierung zwischen einfachem und komplexem Bewerten sowie zwischen nahem und fernem Transfer, wie es auch in dem Modell von Anderson et al. vorgeschlagen wird (siehe oben).

Aufgabenmerkmale	Niveau I	Niveau II	Niveau II
Anzahl der Prozeduren	einschrittig	mehrschrittig einfach/ zweischrittig	mehrschrittig komplex
Tätigkeitsart	Reproduktion, Zusammenhang herstellen,	Anwenden, Zusammenhänge herstellen, einfaches Bewerten, naher Transfer	Generalisieren, Abstrahieren, Reflektieren, Fern-Transfer, komplexes Bewerten
curriculare Gängigkeit (Anwendungshäufigkeit, Intensität)	≤ besonders gängiger bzw. häufiger Regelinhalt, gelegentlicher Standard, seltener Unterrichtsinhalt	≤	≤
Anzahl der expliziten aus dem Stamm zu identifizierenden Informationen (Fakten, Texturen) x_e	$x_{\_exI}$	$x_{\_exI} \leq x_{\_exII}$	$x_{\_exII} \leq x_{\_exIII}$
Anzahl der impliziten aus dem Stamm zu identifizierenden Informationen (Fakten, Texturen) x_i	$x_{\_imI}$	$x_{\_imI} \leq x_{\_imII}$	$x_{\_imII} \leq x_{\_imIII}$
Benötigtes fachliches Vorwissen	Anzahl der Fakten/Konzepte	≤	≤
Benötigtes alltägliches Vorwissen	Anzahl der Fakten	≤	≤
Abstraktheit der Darstellungsebene	≤ gegenständlich, konkret; einfache Abstraktion; komplexe Abstraktion	≤	≤

Tabelle 1: Aufgabenmerkmale zur Bestimmung des kognitiv-fachlichen Anforderungsniveaus

Ein weiteres Aufgabenmerkmal im Modell zur Bestimmung des kfA-Niveaus ist die *curriculare Gängigkeit*. Dafür wurde durch eine Analyse der verschiedenen Bildungspläne der Bundesländer festgestellt, ob der in der Unit behandelte und im Item konkretisierte Kontext mit den dafür benötigten physikalischen Wissenselementen in den Curricula auftaucht. Dabei wurde sowohl die Anwendungshäufigkeit als auch die Intensität beachtet. Die curriculare Gängigkeit sollte zwischen den drei Anforderungsniveaus konstant bleiben oder abnehmen.

Auch die *Abstraktheit der Darstellungsebene* wurde im Modell beachtet, mit dem Ziel, diese möglichst innerhalb einer Unit über die einzelnen kfA-Niveaus hinweg konstant zu halten, um hier ungewollte Einflüsse zu vermeiden. Die meisten Items waren sprachlich-verbal formuliert, einige Items bestanden allerdings aus bildlichen Darstellungen wie z. B. Schaltplänen oder Skizzen.

Zunächst wurden 105 Items für 35 Units in einer ersten Version entwickelt. Dabei wurde systematisch erfasst, welche fachlichen zentralen Konzepte inhaltlich in den Items vorkamen. Ziel war es, möglichst breit über alle curricularen Konzepte der Elektrizitätslehre zu streuen, um ungewollte Einflüsse auf die Aufgabenschwierigkeit durch ein Ungleichgewicht der unterschiedlichen Konzepte zu vermeiden, da bspw. Kauertz einen Einfluss der Dimension *Leitidee* (siehe oben) festgestellt hatte. Anschließend wurden die Items in der Diskussion mit mehreren Physikdidaktik-Expert*innen auf die Passung hinsichtlich des kfA-Modells analysiert. Dabei wurden die Items teilweise leicht angepasst, teilweise grundlegend überarbeitet. Zunächst wurden einige Items gemeinsam in der Gruppe analysiert, damit ein gemeinsames Verständnis des Modells aufgebaut und die Anwendung des Modells zur Analyse der Aufgaben gemeinsam trainiert werden konnte. Vor allem die Aspekte *Anzahl der Prozeduren* und *Tätigkeitsart* sind als interferent anzusehen, da sie stark vom individuellen Lösungsweg abhängig sind, und somit nur mit einem gewissen Interpretationsspielraum eingeschätzt werden können. Dieser Effekt wurde zwar verringert, da die Anzahl der möglichen Lösungswege durch das Aufgabenformat (multiple choice, single select) begrenzt war, trotzdem können die betreffenden Merkmale nicht mit allgemeiner Objektivität bestimmt werden. Somit war es zunächst nötig, sich auf eine möglichst

einheitliche Vorgehens- und Interpretationsweise zu einigen. Dies geschah in mehreren Workshops, in denen Items gemeinsam in der Gruppe der beteiligten Expert*innen diskursiv analysiert wurden. Dazu wurden alle für das kfA-Niveau relevanten Merkmale der Items systematisch erfasst und bei Bedarf angepasst. Anschließend wurden die Items von einzelnen Expert*innen analysiert. Uneindeutige Fälle wurden stets nochmal in der Gruppe zur Diskussion gestellt und Items dann evtl. gemeinsam überarbeitet. Die Entscheidung fiel für diese Vorgehensweise und dagegen, die Items durch Expert*innen unabhängig voneinander analysieren zu lassen und dann die Analyseergebnisse zu vergleichen. Dies hätte zwar Aussagen über die Modellqualität anhand der Interraterreliabilität zugelassen, allerdings hätten die Items für diese Vorgehensweise bei der Analyse nicht mehr verändert werden dürfen. Da zu diesem Zeitpunkt sich sowohl das kfA-Modell als auch die Items in einem allerersten, teilweise unausgereiften Stadium befanden, erschien es deutlich sinnvoller, die Analyse diskursiv in der Gruppe vorzunehmen, um die Items direkt überarbeiten zu können. Somit konnte schon für die Präpilotierung eine bessere Itemqualität erreicht werden.

Da die Itemstämme für die hier vorgestellte Präpilotierung, wie bereits erwähnt, noch nicht in den sprachlichen Anforderungsniveaus variiert wurden, sondern alle in einem mittleren sprachlichen Niveau vorlagen und die systematische und kriteriengeleitete sprachliche Variation erst nach der Präpilotierung vorgenommen wurde, ist das Modell zu den sprachlichen Anforderungsniveaus im folgenden Abschnitt nur sehr knapp dargestellt.

3.2.2. Sprachliche Anforderung

Auch zur systematischen Variation der sprachlichen Anforderung in drei unterschiedlichen Niveaustufen wurde ein theoretisches Modell auf der Grundlage bisheriger Forschung erstellt. Durch die sprachlichen Rahmenbedingungen ergaben sich einige Besonderheiten für das Aufgabenformat: Jeder Aufgabenstamm einer Unit sollte aus genau 24 Propositionen bestehen. Diese Vorgabe führte dazu, dass die entwickelten Aufgaben einen deutlich längeren Aufgabentext hatten, als die meisten Schüler*innen es von bisher im Physikunterricht verwendeten Aufgaben gewohnt waren. Ziel dieses

Aufgabenformates war vor allem, den Einfluss der sprachlichen Merkmale auf die Aufgabenschwierigkeit einer Leistungsaufgabe (siehe Abschnitt 2.3.3) zu verstärken, um ihn besser beobachten und analysieren zu können.

Die Informationen, die zur Lösung der Aufgabe benötigt werden, sollten dabei möglichst gut im gesamten Text des Aufgabenstamms verteilt sein, um sicherzustellen, dass der Text von den Bearbeitenden sorgfältig und vollständig gelesen werden muss. Wesentliche Kriterien bei der sprachlichen Variation in den drei Anforderungsniveaustufen sind dabei sprachliche Komplexität (Flexion, Kasus, Syntax, Morphologie), Eindeutigkeit der Form-Bedeutungsebene (Pronomen, Bezugsnomen) sowie die Frequenz (Häufigkeit der Wörter im Sprachgebrauch). Im sprachlichen Anforderungsniveau I bestand beispielsweise jeder Satz aus nur einer Proposition, im verwendeten Anforderungsniveau II waren es zwei Propositionen pro Satz und im Anforderungsniveau III wurden je vier Propositionen zu einem Satz zusammengefasst. Der Prompt und das Item sollen immer auf dem sprachlichen Anforderungsniveau II formuliert sein, sodass nur der Stamm sprachlich variiert wird.

4. Untersuchungsvorstellung

4.1. Ziele

Die in diesem Abschnitt vorgestellte Erhebung verfolgte im Wesentlichen folgende drei Ziele:

1. Das Sammeln von Hinweisen zur Überarbeitung der Items im Hinblick auf die Pilotierungsstudie und den weiteren Verlauf des VAMPS-Projektes
2. Das Sammeln erster empirischer Evidenz für die Qualität des kfA-Niveaustufen-Modells durch die Schwierigkeitseinschätzungen von Schüler*innen
3. Exploratives Sammeln von Informationen über den Bearbeitungsprozess durch Schüler*innen, vor allem in Hinblick auf Schwierigkeit und schwierigkeitserzeugende Merkmale

Das erste wichtige Ziel lag darin, in der Untersuchung als Präpilotierung der Items für den weiteren Verlauf des VAMPS-Projektes Anhaltspunkte für die Überarbeitung der Aufgaben zu sammeln, bevor diese in die eigentliche Pilotierungsstudie starten. Insbesondere, da das Aufgabenformat für Physikaufgaben untypisch ist und somit den Schüler*innen weitestgehend unbekannt sein sollte, war es wichtig zu testen, wie Schüler*innen mit den Aufgaben umgehen. Dabei sollten etwaige Verständnisprobleme oder sonstige nicht intendierte Hürden bei der Bearbeitung der Aufgaben durch Schüler*innen aufgedeckt werden. So sollten zum Beispiel semantische Missverständnisse vermieden werden. Es sollte überprüft werden, ob die Schüler*innen den Aufgabentext sowie die Aufgabenstellung richtig verstehen und ob sie in der Lage sind, das mentale Modell der in der Aufgabenstellung beschriebenen Situation zu bilden. Außerdem sollte sichergestellt werden, dass die Distraktoren, also die falschen Antwortmöglichkeiten, für die Schüler*innen eine gewisse Attraktivität besitzen. Die Distraktoren sollten bei fehlendem Fachwissen von den Schüler*innen tatsächlich als mögliche richtige Antworten angesehen werden, damit der Test zwischen Wissenden und Unwissenden unterscheiden kann (Schnell, 2016, S. 27f.). Dabei war das Ziel der Präpilotierung nur eine erste, grobe Orientierung. Da die Stichprobe relativ klein war,

konnte es kein Ziel sein, statistische Aussagen darüber zu treffen, wie häufig Distraktoren angewählt werden o. Ä.. Diese Ziele werden mit der anschließenden Pilotierungsstudie verfolgt. Allerdings sollte die Präpilotierung einen Blick aus Schüler*innenperspektive auf die Aufgaben ermöglichen, die ja zuvor ausschließlich von Expert*innen konstruiert und analysiert wurden. Somit sollten möglicherweise unvorhergesehene Probleme aufgedeckt und korrigiert werden, damit die Units schon in verbesserter Qualität in die Pilotierungsstudie starten können.

Das zweite Ziel dieser Untersuchung, ebenso in ihrer Funktion als Präpilotierungsstudie, lag darin, erste Informationen über die Qualität des Modells der kognitiv-fachlichen Anforderungen zu sammeln. Dafür wurde untersucht, ob die Struktur der kognitiv-fachlichen Anforderungen der Items innerhalb der entwickelten Unit-Matrizen durch Schüler*innen, die die Schwierigkeit der Items beurteilen sollen, reproduziert wird.

Ein drittes Ziel lag darin, generelle Informationen über den Bearbeitungsprozess der Aufgaben durch Schüler*innen zu erhalten und offen nach auffälligen Mustern zu suchen. Dieses Ziel wurde vor der Untersuchung mit Absicht wenig konkret formuliert, um induktiv am Material arbeiten zu können und die Untersuchung auf Dinge ausrichten zu können, die bei der Analyse auffällig erscheinen. Als mögliche Forschungsgegenstände wurden hier beispielsweise folgenden Fragen formuliert: Nach welcher Logik und mit welchen Strategien gehen Schüler*innen vor, um zur richtigen Lösung zu kommen? Führen tatsächlich die im Modell vorgesagten Prozesse (abgezählt als Prozeduren) zur Lösung, oder gibt es unerwartete, für die Untersuchung irrelevante Prozesse, die ebenfalls zur Lösung führen können? Lassen sich am Bearbeitungsprozess der Aufgaben eventuell noch weitere Faktoren erkennen, die Schwierigkeit auslösen, im Modell aber nicht enthalten sind? Interessant ist ebenso die Frage, was Schüler*innen aus ihrer Perspektive als schwierigkeitserzeugende Faktoren einer Aufgabe wahrnehmen. Die Entscheidung für einen dieser Aspekte wurde bei der Auswertung der Ergebnisse getroffen und wird in dem zugehörigen Abschnitt (5.3) erläutert.

4.2. Methode

Bei der in dieser Arbeit vorgestellten Untersuchung handelt es sich um eine so genannte „Think-Aloud“-Studie, die als qualitative Forschungsmethode eine Abwandlung der Methode des „Lauten Denkens“ benutzt. Dazu haben die Schüler*innen die entwickelten Physikaufgaben in Kleingruppen bearbeitet, während ihre Gespräche in Audioaufnahmen aufgezeichnet wurden. Zusätzlich wurden bestimmte Merkmale des Bearbeitungsprozesses durch eine*n Testleiter*in mit Hilfe eines Beobachtungsbogens erfasst. Dieser Beobachtungsbogen diente gleichzeitig als eine Art Interviewleitfaden, mit dem die Schüler*innen zu verschiedenen Zeitpunkten um eine begründete Einschätzung der Schwierigkeit gebeten wurden. Der genaue Ablauf der Untersuchung und der Beobachtungsbogen sind im Abschnitt 4.2.3 genauer vorgestellt. Im Folgenden wird eine kurze theoretische Begründung der Methode „Lautes Denken“ sowie der Entscheidung, Schüler*innen nach ihrer Einschätzung der Aufgabenschwierigkeit zu befragen, gegeben.

4.2.1. Lautes Denken

Unter der Methode „Lautes Denken“ wird „[...] das gleichzeitige laute Aussprechen von Gedanken bei der Bearbeitung einer Aufgabe [...]“ (Knoblich & Öllinger, 2006, S. 692) verstanden. Es handelt sich dabei um eine qualitative Forschungsmethode, mit deren Hilfe Gedanken und Strategien erhoben werden können, die von Schüler*innen zur Auswahl der Antwort genutzt werden (Schnell, 2016). Ebenso erhält man Informationen über kognitive Prozesse und mentale Operationen, die während der Bearbeitung der Aufgabe ablaufen (Knoblich & Öllinger, 2006, S. 692). Auch lässt sich mittels der Methode des lauten Denkens die inhaltliche Nachvollziehbarkeit der Item-Formulierungen und die mögliche (In-)Kongruenz zwischen den Intentionen der Testkonstrukteur*innen und den Vorstellungen der Schüler*innen untersuchen (Schnell, 2016, S. 28). So können die geäußerten Gedankengänge Aufschluss darüber geben, ob auch unerwartete oder irrelevante Prozesse zu Lösungen führen können (Hartig, Frey, & Jude, 2012). „Protokolle des lauten Denkens liefern [...] sehr konkrete, inhaltliche Hinweise für die Item-Optimierung.“ (ebd). Deswegen findet diese Methode auch

häufiger bei der Überprüfung der Qualität von Testaufgaben in der Kompetenzmodellierung Anwendung, mit der außerdem die Aufgabenqualität und Item-Validierung geprüft werden kann (Sandmann, 2014, S. 182). Damit ist diese Methode auch im Hinblick auf die für diese Untersuchung verfolgten Ziele geeignet.
Die „Laut-Denk-Protokolle“ wurden bei der Durchführung allerdings (meist) nicht zu individuellen Lösungsprozessen einzelner Schüler*innen erhoben, sondern überwiegend in Kleingruppen von zwei bis drei Schüler*innen. Dadurch wurde die unauthentische und für Schüler*innen ungewohnte Kommunikationssituation des Redens bei der alleinigen Bearbeitung etwas aufgelockert. Die Annahme war, dass in dieser Form den Schüler*innen die Verbalisierung ihrer Denkprozesse etwas erleichtert wird, da der zeitliche Rahmen der Untersuchung es nicht zuließ, vorher Übungen zum Lauten Denken mit den Schüler*innen durchzuführen. Außerdem sind die Schüler*innen durch die Kleingruppen bei kontroversen Meinungen möglicherweise mehr dazu gezwungen, ihre Gedanken zu begründen, was zusätzlich Rückschlüsse auf logische Strukturen im Denkprozess zulässt.

4.2.2. Schüler*innen als Expert*innen für Aufgabenschwierigkeit?

Im Rahmen der Untersuchung wurde die Entscheidung getroffen, Schüler*innen nacheinander jeweils alle drei Items einer Unit vorzulegen (also drei Aufgaben zu dem gleichen Stamm in drei verschiedenen kognitiv-fachlichen Anforderungsniveaus) und diese zu bitten, die Schwierigkeit der Aufgaben auf einer Skala einzuschätzen. Es wurde erwartet, dass die Schüler*innen in ihren Schwierigkeitseinschätzungen der drei Items relativ zueinander zumindest in Teilen die kognitiv-fachlichen Anforderungsniveaus des Modells reproduzieren und somit bestätigen. Dies sollte eine erste Evidenz für die Qualität des Modells der kognitiv-fachlichen Anforderungen liefern.
Dieser Entscheidung gingen theoretische Überlegungen zu der Frage voraus, ob Schüler*innen bei der Einschätzung der Schwierigkeit von Leistungsaufgaben treffende Urteile fällen können. Eine andere in Betracht gezogene Möglichkeit wäre gewesen, Lehrkräfte als Expert*innen für die Einschätzung von Aufgabenschwierigkeit anzusehen, da Lehrkräfte in ihrem Arbeitsalltag ständig mit Leistungsaufgaben und deren

Schwierigkeiten zu tun haben. Allerdings konnten verschiedene Untersuchungen belegen, dass „Lehrer*innen die Schwierigkeit von Aufgaben, die sie im Unterricht verwendeten, nur ungenau einschätzen konnten“ (Astleitner, 2008, S. 70). Möglicherweise sind Schwierigkeitseinschätzungen von Lehrkräften aufgrund ihrer Expertise im Fachgebiet verzerrt. Aus diesem Grund wurde sich dagegen entschieden, die Leistungsaufgaben Lehrkräften zur Einschätzung der Schwierigkeit vorzulegen.

Zur Einschätzung von Aufgabenschwierigkeit durch Schüler*innen gibt es bisher keine nennenswerte Forschung. Bei der Untersuchung von diagnostischen Kompetenzen im Hinblick auf Unterrichtsqualität haben sich allerdings die Urteile von Schüler*innen als (zumindest teilweise) valide gezeigt. Grewe, Strietholt, & Schwippert (2007) kamen nach einer Untersuchung, in der Schüler*innen zur Unterrichtsqualität befragt wurden, zu dem Ergebnis, dass Schüler*innen ein zuverlässiges und reliables Feedback zu Unterricht geben können (Grewe, Strietholt, & Schwippert, 2007). Dies spricht für eine diagnostische Kompetenz von Schüler*innen, die sich möglicherweise auch auf die Beurteilung von Aufgabenschwierigkeiten anwenden lässt. Außerdem sind Schüler*innen in ihrem Schulalltag sehr häufig mit Aufgaben konfrontiert und erhalten auch regelmäßig eine Rückmeldung zu ihrer Performanzleistung. In Klassenarbeiten, Klausuren und Prüfungen sind sie oft der Situation ausgesetzt, entscheiden zu müssen, welche der gestellten Aufgaben für sie am wenigsten schwierig ist, um diese bei Zeitdruck gezielt zuerst zu bearbeiten. Dadurch, so die Vermutung, haben Schüler*innen einen geschulten Blick auf Aufgaben und auf das eigene Vermögen, eine bestimmte Aufgabe richtig lösen zu können. In diesem Sinne wurde die Entscheidung getroffen, Schüler*innen als Expert*innen für Aufgaben und für Aufgabenschwierigkeit anzusehen. Diese Annahme wurde also bewusst trotz geringer theoretischer oder empirischer Fundierung im Rahmen dieser Untersuchung als ein „Versuch“ getroffen.

4.2.3. Beobachtungsbogen und Ablauf der Untersuchung

In diesem Abschnitt wird der Ablauf der Untersuchung und die erhobenen Daten anhand des Beobachtungsbogens vorgestellt, der bei der Beobachtung und Befragung der Schüler*innen verwendet wurde. Dazu wurde jeder Kleingruppe aus Schüler*innen ein*e

Testleitende*r zugeordnet, welche*r durch die Untersuchung führte und dabei den Beobachtungsbogen ausfüllte. Es wurde die Entscheidung getroffen, zusätzlich zu den Audioaufnahmen einige Kriterien auf einem Beobachtungsbogen zu erfassen, um die Daten anschließend möglichst effizient auswerten zu können, auch in Hinblick auf den zeitlichen Druck im weiteren Verlauf des Forschungsprojektes. So sollten auf dem Beobachtungbogen direkt Hinweise auf nötige Überarbeitungen einzelner Items sowie die Schwierigkeitseinschätzungen der Schüler*innen festgehalten werden. Dies sollte auch erlauben, die Audiodateien bei der Auswertung nur gezielt in bestimmten Teilen zu analysieren. Die Beobachtungskriterien sind im folgenden Abschnitt beschrieben und begründet. Zusätzlich enthielt der Bogen an mehreren Stellen freie Felder für Bemerkungen und Kommentare der Testleitenden. Dadurch sollten Hinweise auf unvorhergesehene Probleme gesammelt werden können. Der vollständige Beobachtungbogen findet sich im Anhang dieser Arbeit. Die Untersuchung war in drei Phasen unterteilt, die jede Kleingruppe für jede Unit nacheinander durchlief. Diese sind im Folgenden dargestellt.

4.2.3.1. Phase 1: Lesen und Paraphrasieren

In der ersten Phase der Untersuchung wurden Daten gesammelt, die Rückschlüsse auf die Verständlichkeit des Aufgabenstamms zulassen und somit Hinweise für eine eventuelle Überarbeitung geben sollten. Damit sollte verhindert werden, dass der Text für die Schüler*innen missverständlich formuliert ist oder wichtige Informationen unklar ausgedrückt sind. Dazu bekam die Kleingruppe aus bis zu 3 Schüler*innen den Aufgabenstamm, also den Text mit 24 Propositionen, vorgelegt. Eine*r der Schüler*innen las diesen zunächst laut vor. Dabei markierte die Testleitung in einer eigenen Kopie Stellen, an denen Unsicherheiten beim Lesen (Versprecher, Stocken, etc.) auftauchten. Dadurch konnten Wörter oder Satzstrukturen identifiziert werden, die den Schüler*innen eventuell Probleme bereiten. Den Schüler*innen wurde anschließend angekündigt, dass sie im nächsten Schritt, ohne dabei den Text vorliegen zu haben, diesen zusammenfassen sollten. Sie erhielten deshalb vorher die Möglichkeit, Stellen im Text still nachzulesen. Sie wurden aufgefordert „fertig“ zu sagen, sobald sie der Ansicht

seien, den Text gut verstanden zu haben. Dadurch kann im Nachhinein anhand der Audioaufnahme die Zeit erfasst werden, die die Schüler*innen benötigen, bis sie meinen, den Text verstanden zu haben. Dies kann Hinweise darauf geben, wie gut die Schüler*innen mit dem Text zurechtkommen. Anschließend wurde der Text verdeckt und die Schüler*innen wurden aufgefordert, diesen zusammenzufassen. Dies sollte Aufschluss über die Frage geben, ob den Schüler*innen die Bildung eines mentalen Modells der im Text beschriebenen Situation gelingt, was Voraussetzung für eine erfolgreiche Bearbeitung der Items ist. Die Testleitung notierte im Beobachtungsbogen, ob es bei der Paraphrasierung große Probleme gab oder Unverständnis geäußert wurde. Dies sollte als Hinweis für eine notwendige Überarbeitung des Aufgabenstamms dienen. Auf eine Beurteilung, wie gut die Paraphrasierung gelinge, wurde an dieser Stelle verzichtet, da nicht alle Testleitenden mit den Stämmen und deren Inhalten vertraut waren. Anschließend wurde der Text wieder aufgedeckt und die Schüler*innen bekamen die Möglichkeit, mit Hilfe des Textes noch Informationen zu ergänzen, die sie in der ersten Zusammenfassung noch nicht genannt hatten. Dies geschah aus dem Grund, dass viele der Aufgabenstämme detaillierte Informationen enthalten, wie mehrere Namen oder Zahlen, die normalerweise nur schwer vollständig im Gedächtnis behalten werden können. Das Gelingen der Paraphrasierung sollte aber nicht von dem Memorierungsvermögen der Schüler*innen abhängen. Eine Ergänzung wurde auf dem Beobachtungbogen von der Testleitung notiert. Am Schluss dieser ersten Phase sollten die Schüler*innen die Schwierigkeit des Textes auf einer Skala von 1 bis 10 bewerten und ihre Bewertung begründen. Zur Visualisierung lag eine ausgedruckte Skala vor, die beschriftet war mit „sehr einfach“ (Wert 1) und „sehr schwierig“ (Wert 10). Da alle Aufgabenstämme auf dem mittleren sprachlichen Anforderungsniveau formuliert waren, wurde hier erwartet, dass die Schüler*innen mittlere bis niedrige Werte nennen. Ein sehr hoher Wert wäre ebenfalls ein Hinweis für eine notwendige Überarbeitung gewesen.

4.2.3.2. Phase 2: Aufgabenstellung

In der zweiten Phase der Untersuchung wurde den Schüler*innen zunächst nur eines der drei dem Stamm zugehörigen Items vorgelegt. Sie wurden dazu aufgefordert, sich über die Aufgabe und die Antwortmöglichkeiten zu unterhalten und sich dann möglichst auf eine gemeinsame richtige Lösung zu einigen. Durch die Vorgabe der Einigung sollte die Diskussion der Schüler*innen untereinander bestärkt werden, da diese so bei unterschiedlichen Meinungen mehr dazu gezwungen wurden, ihre Ansicht zu begründen und dafür zu argumentieren, was wiederum Einblicke in die logische Denkstruktur und die damit verbundenen kognitiven Prozesse zulässt. Bewusst wurde den Kleingruppen dabei nur ein gemeinsames Aufgabenblatt gegeben, um die Interaktion der Schüler*innen untereinander zu stärken. Die Testleitung notierte auf dem Beobachtungsbogen, ob die Aufgabenstellung für die Schüler*innen klar war, oder ob diese Probleme bereitete, was wiederum ein Hinweis für eine notwendige Überarbeitung gewesen wäre. Außerdem wurde notiert, ob und wie stark der Aufgabenstamm bei der Bearbeitung berücksichtigt wurde. Dafür konnten die drei Kategorien „intensiv“, „etwas“ oder „gar nicht“ auf dem Beobachtungsbogen angekreuzt werden. Dabei wurde bewusst auf eine fünf- oder mehrstufige Skala verzichtet, um die Beobachtung für die Testleitenden möglichst nicht zu kompliziert zu gestalten, insbesondere da die Testleitenden in dieser Phase der Erhebung viele Merkmale gleichzeitig zu beobachten hatten. Bei dieser dreistufigen Skala wurde bewusst in Kauf genommen, dass es sich um subjektive Einschätzungen der Testleitenden handelt, die lediglich eine grobe Einordnung erlauben. Dies wurde für die verfolgten Ziele als ausreichend angesehen: Mit diesem Beobachtungskriterium sollte untersucht werden, ob die Items wie intendiert so konstruiert sind, dass der Text zwingend notwendig ist, um sie zu lösen. Bei einem „guten Item“ sollte die Aufmerksamkeit der Schüler*innen bei der Bearbeitung zwischen Aufgabenstamm und Item wechseln, was den Einfluss des Aufgabentextes sicherstellt. Wenn dieser also gar nicht beachtet werden würde, könnte dies ein Hinweis auf die Bearbeitungsbedürftigkeit des Items sein. Außerdem wurde notiert, über welche Antwortmöglichkeiten inhaltlich diskutiert wurde, über welche Antwortmöglichkeiten

formal argumentiert wurde (z.B. „Diese Möglichkeit ist länger als alle anderen, deswegen ist das bestimmt die richtige!“), welche Antwortmöglichkeiten ignoriert oder sofort ausgeschlossen wurden und welche Antwortmöglichkeit schließlich als richtig gewählt wurde. Bei einem „guten“ Item sollte über alle Antwortmöglichkeiten inhaltlich diskutiert werden, und keine Möglichkeit aus formalen Argumenten gewählt oder von vornherein ausgeschlossen werden. Diese Daten sollten bei der Überarbeitung der Items dazu dienen, Anhaltspunkte zu sammeln, welche Distraktoren für Schüler*innen unattraktiv sind. Distraktoren, die sofort ausgeschlossen werden können, müssen überarbeitet werden. Auch beim Auftreten formaler Argumente, die einen Hinweis darauf geben, dass man die Aufgabe möglicherweise einfach nur durch Testwissen (test-wiseness) lösen kann, sollte das Item überarbeitet werden. In einem Bemerkungsfeld konnten die Testleitenden zusätzlich Auffälligkeiten beim Bearbeitungsprozess notieren.

4.2.3.3. Phase 3: Schwierigkeitseinschätzung

Damit die Schüler*innen ihre Einschätzung der Schwierigkeit der Aufgabe möglichst unbeeinflusst abgeben konnten, wurden die drei kfA-Niveaus anstatt mit Zahlen ohne erkennbare Hierarchie mit #, & und § bezeichnet, wobei die Bezeichnung unter den verschiedenen Units systematisch permutiert wurde. Nachdem das erste Item in einer Gruppendiskussion gemeinsam gelöst wurde, wurden die Schüler*innen gebeten, die Schwierigkeit der Aufgabe auf einer Skala von 1 bis 10 einzuschätzen und ihre Einschätzung zu begründen. Durch die Begründung sollten Daten darüber gesammelt werden, welche Merkmale Schüler*innen zur Beurteilung der Aufgabenschwierigkeit heranziehen.

Anschließend wurden den Schüler*innen die Aufgabenitems der beiden verbleibenden kognitiv-fachlichen Anforderungsniveaus vorgelegt. Diese beiden Aufgaben sollten nun nicht mehr durch die Schüler*innen gelöst werden, sondern nach dem sorgfältigen Lesen der Aufgabenstellung begründet hinsichtlich ihres Schwierigkeitsgrades eingeschätzt werden. Dies sollte Informationen über die Einschätzung der Schwierigkeit der drei Niveaus relativ zueinander liefern. Diese Werte wurden durch die Testleitung auf dem

Beobachtungbogen notiert. Dabei wurde darauf geachtet, dass die Werte der unterschiedlichen kfA-Niveaus (anonym) den Schüler*innen eindeutig zugeordnet werden können. Dadurch lässt sich erkennen, wie jede*r einzelne Schüler*in die Schwierigkeitsgrade der kfA-Niveaus relativ zueinander einschätzt. Man erhält also Informationen darüber, ob ein*e Schüler*in ein bestimmtes kfA-Niveau als schwieriger einschätzt als ein anderes. Diese Information war das Ziel dieses Teils der Untersuchung. Die Skala dient dazu als Hilfsmittel zur Einordnung und als Hinweis darauf, ob der Abstand zwischen den Schwierigkeiten der drei Items eher als groß oder eher als klein wahrgenommen wird. Damit den Schüler*innen dieser Vergleich der Schwierigkeit der Items untereinander besser gelingt, wurden von den Testleitenden stets noch einmal die Einschätzung des ersten Items der Schüler*innen wiederholt, bevor sie aufgefordert wurden, die anderen beiden Items einzuschätzen.

4.3. Sample und Durchführung

Im Februar 2020 wurde die Untersuchung mit insgesamt 58 Schüler*innen der neunten Jahrgangsstufe an drei Hamburger Schulen mit unterschiedlichen Sozialindizes durchgeführt. Die Schüler*innen kamen je Schule aus der gleichen Klasse. In jeder Klasse wurden neun Kleingruppen zu je 1-3 Schüler*innen gebildet. Je Kleingruppe wurden innerhalb von 90 Minuten drei bis fünf Units bearbeitet. Dadurch konnte jede der 35 Units an jeder der drei Schulen getestet werden, es ergaben sich demzufolge 105 Testsituationen. Jedes Item wurde also genau einmal von einer Kleingruppe intensiv bearbeitet und drei Mal von unterschiedlichen Gruppen hinsichtlich seiner Schwierigkeit eingeschätzt.

Laut dem Hamburger Bildungsplan haben alle Schüler*innen zu dem Zeitpunkt das Themengebiet Elektrizitätslehre bereits in Physik behandelt, was durch die jeweiligen Physiklehrkräfte der Klassen bestätigt wurde. Die Erhebung fand an drei verschiedenen Tagen je innerhalb einer Doppelstunde statt. Die Teilnahme war dabei für alle Schüler*innen freiwillig und fand während der regulären Unterrichtszeit statt. Die gesamte Bearbeitung und Befragung wurde durch Audioaufnahmen dokumentiert. Jede

Kleingruppe wurde durch eine*n Testleiter*in betreut, der/die parallel die genannten Merkmale des Bearbeitungsprozesses auf dem Beobachtungsbogen erfasste.

Zu Beginn jeder Erhebung erhielten alle Schüler*innen gemeinsam eine kurze Einführung über die Ziele und den Ablauf der Untersuchung. Dabei wurde die Methode „Lautes Denken“ erklärt sowie ausdrücklich darauf hingewiesen, dass es bei der Untersuchung darum geht, die Aufgaben zu testen, und nicht das Physikwissen der Schüler*innen. Damit sollte vermieden werden, dass die Schüler*innen Hemmungen haben, bestimmte Gedanken auszusprechen, aus Angst davor, etwas Falsches zu sagen. Die Schüler*innen wurden ebenso über die Audioaufnahmen und Maßnahmen des Datenschutzes aufgeklärt. Danach wurden die Schüler*innen in per Los zugewiesene Kleingruppen eingeteilt. Bei einer der Schulen war am Tag der Durchführung aufgrund eines Sturms nur ein Teil der Schüler*innen anwesend, weswegen dort einige Kleingruppen aus nur einem Schüler bzw. einer Schülerin bestanden. Die Kleingruppen wurden je von einem Testleiter oder einer Testleiterin betreut. Diese erhielten vorher eine Schulung, in der ausgiebig der Beobachtungbogen und die verwendeten Materialien besprochen wurden und es die Möglichkeit gab, Fragen zum Ablauf und Inhalt der Untersuchung zu stellen. Damit sollte möglichst vermieden werden, dass die Untersuchungsergebnisse durch unterschiedliche Interpretationsweisen der Testleitenden oder missverständliche Anweisungen auf dem Beobachtungbogen verfälscht werden. Die Ergebnisse der Erhebung sind im folgenden Kapitel dargestellt.

5. Analyse und Auswertung

Da die Untersuchung mehrere Ziele verfolgt, ist die Analyse und Auswertung der erhobenen Daten ebenso in verschiedene Schritte eingeteilt, in denen diese unterschiedlichen Ziele verfolgt werden. Im Rahmen dieser Masterarbeit wurden nicht alle erhobenen Daten detailliert ausgewertet, da die Untersuchung bewusst einen recht explorativen Charakter hatte, mit dem hauptsächlichen Ziel, mögliche Probleme der Units aufzudecken. Vor der Untersuchung gab es nur Vermutungen, an welchen Stellen noch Probleme liegen könnten (Textverständnis, Verständnis der Aufgabenstellung, Attraktivität der Distraktoren, formaler Ausschluss von Möglichkeiten etc.). Diese im Vorhinein vermuteten Bereiche wurden gezielt auf den Beobachtungsbögen erfasst. Um darüber hinaus weitere, unvorhergesehene Probleme erfassen zu können, wurden möglichst viele verschiedene und umfangreiche Daten erhoben, mit dem Ziel, dann anhand des Materials zu entscheiden, welche Teile intensiver ausgewertet werden.

5.1. Nutzung der Daten zur Überarbeitung der Units

Die 35 in die Präpilotierung gestarteten Units mit insgesamt 105 Items wurden anhand der im Beobachtungsbogen erfassten Daten überarbeitet. Ein wichtiges Ziel der Untersuchung war, anhand der Daten die Units auszuwählen, die vor dem Start in die Pilotierungsarbeit noch einmal überarbeitet werden sollten, sowie die Anzahl der Units im Sinne einer Bestenauslese auf 30 zu reduzieren. Dazu flossen die Merkmale, die dem Beobachtungsbogen zu entnehmen sind, ein. Im Folgenden wird beschrieben, wie die Daten der Beobachtungsbögen zur Überarbeitung der Units genutzt wurden.

5.1.1. Aufgabenstämme

Da jedes Item in der Untersuchung genau von einer Schüler*innen-Kleingruppe bearbeitet bzw. gelöst wurde, wurde jeder Aufgabenstamm von drei verschiedenen Kleingruppen gelesen. Dabei wurde insgesamt nur in 5 der 105 Beobachtungsbögen notiert, dass es Probleme beim Paraphrasieren des Stammes gab. Diese verteilten sich

auf fünf verschiedene Stämme, bei keinem Aufgabenstamm traten in mehreren Kleingruppen unabhängig voneinander Probleme auf.

Der Mittelwert aller Schwierigkeitseinschätzungen der Aufgabentexte liegt bei 3,67 bei einer Standardabweichung von 2,01. Die meisten Aufgabenstämme wurden also als eher leicht eingeschätzt, was den Erwartungen entspricht. Bei allen Schwierigkeitseinschätzungen, auch im Folgenden, ist zu beachten, dass die Werte innerhalb einer Kleingruppe nicht unabhängig voneinander sind. Es ist davon auszugehen, dass Schüler*innen in ihrem Urteil beeinflusst sind, wenn sie schon das Urteil eines Mitschülers oder einer Mitschülerin gehört haben. So weichen bei fast allen Units die Werte innerhalb einer Gruppe deutlich weniger voneinander ab als über die drei Gruppen hinweg, die die gleiche Unit bearbeitet haben. Vor diesem Hintergrund ist es aber ein besonderer Hinweis auf Probleme mit dem Text, wenn für den gleichen Aufgabenstamm Schüler*innen aus unterschiedlichen Kleingruppen, also unabhängig voneinander, die Schwierigkeit des Textes als hoch einschätzen. Dies war, wie man in Abbildung 5 erkennen kann, bei drei Units der Fall: 55_02, 56_01, 64_02, was als Hinweis genommen wurde, den Aufgabenstamm dieser Items noch einmal kritisch zu überarbeiten. Das Diagramm findet sich in höherer Auflösung im Anhang. Bei den beiden erst genannten wurden auch je in einer Kleingruppe Probleme bei der Paraphrasierung im Beobachtungsbogen angegeben.

Generell lässt sich aber festhalten, dass sich beim Verständnis der Aufgabenstämme wenig Probleme ergaben.

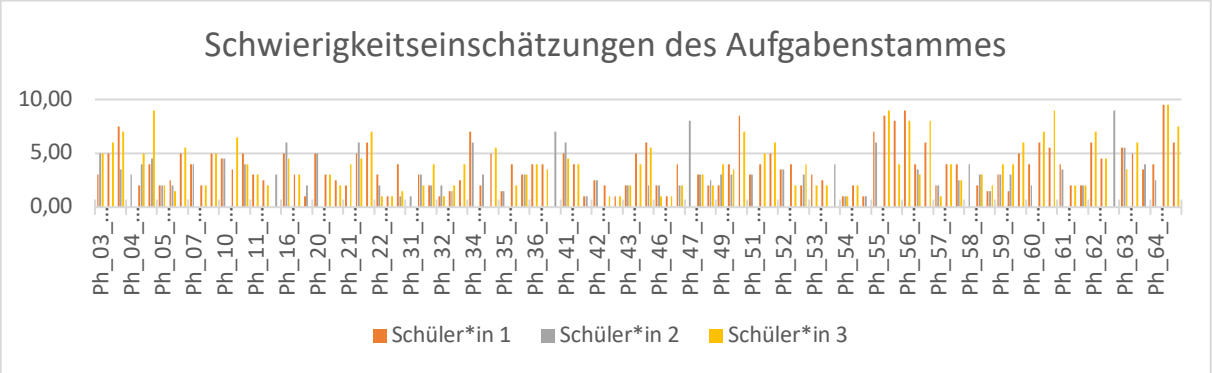
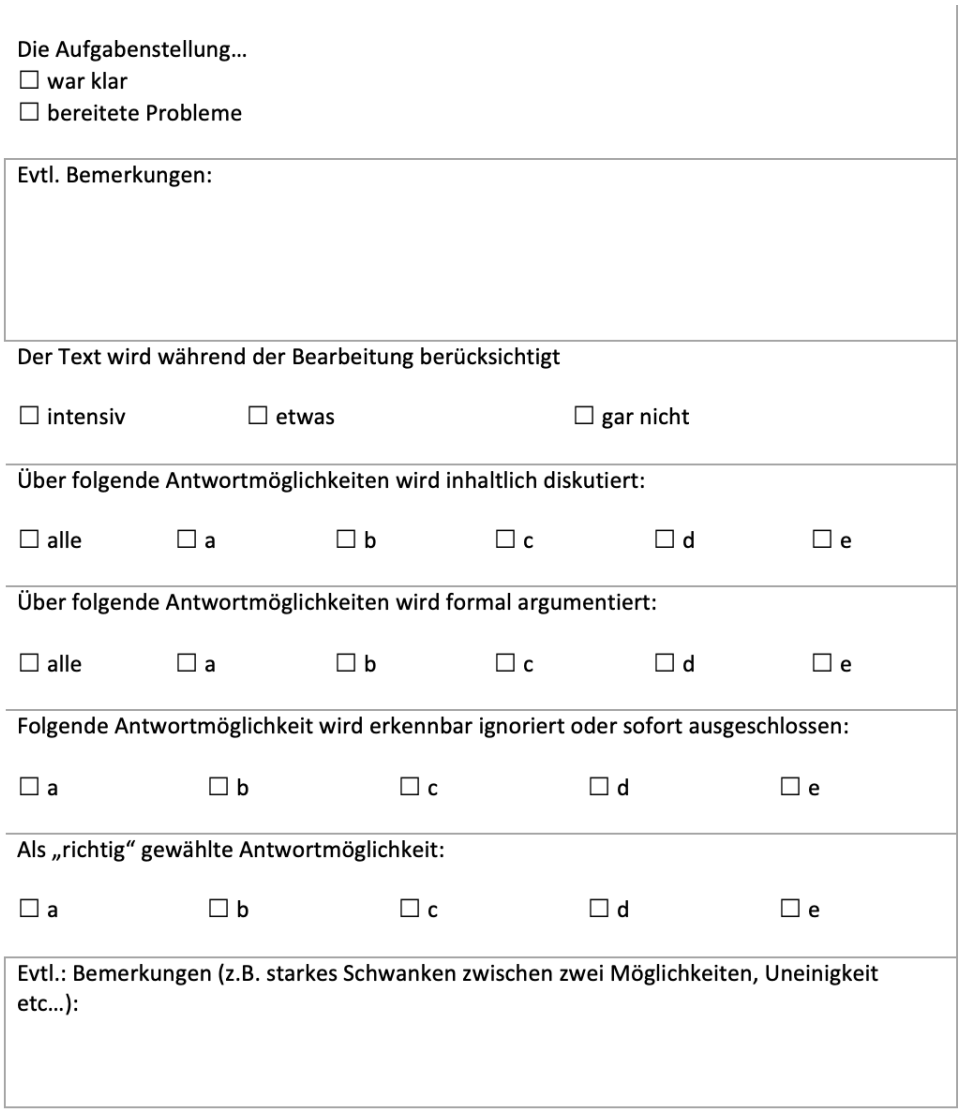


Abbildung 5: Schwierigkeitseinschätzungen des Aufgabenstammes

5.1.2. Items

Anhand der Dokumentationen der Gruppendiskussionen wurden einige Items mehr oder weniger intensiv überarbeitet. Der dafür verwendete Abschnitt des Beobachtungsbogens ist zur besseren Übersicht in Abbildung 6 dargestellt, die Auswertung der Kriterien erfolgt in der Reihenfolge des Beobachtungsbogens. Es zeigte sich allerdings im Nachhinein, dass sich die Dokumentation der Gruppengespräche anhand der Kriterien des Beobachtungsbogens (möglicherweise aufgrund der Vielzahl der Aspekte) für die Testleitenden schwierig gestaltete. Viele der Beobachtungsbögen waren an dieser Stelle unvollständig ausgefüllt. Da diese Daten aber ohnehin nur als Hinweise zu möglichen Überarbeitungen angesehen wurden, stellt dies kein gravierendes Hindernis dar.



Die Aufgabenstellung…
☐ war klar
☐ bereitete Probleme

Evtl. Bemerkungen:

Der Text wird während der Bearbeitung berücksichtigt

☐ intensiv ☐ etwas ☐ gar nicht

Über folgende Antwortmöglichkeiten wird inhaltlich diskutiert:

☐ alle ☐ a ☐ b ☐ c ☐ d ☐ e

Über folgende Antwortmöglichkeiten wird formal argumentiert:

☐ alle ☐ a ☐ b ☐ c ☐ d ☐ e

Folgende Antwortmöglichkeit wird erkennbar ignoriert oder sofort ausgeschlossen:

☐ a ☐ b ☐ c ☐ d ☐ e

Als „richtig" gewählte Antwortmöglichkeit:

☐ a ☐ b ☐ c ☐ d ☐ e

Evtl.: Bemerkungen (z.B. starkes Schwanken zwischen zwei Möglichkeiten, Uneinigkeit etc…):

Abbildung 6: Ausschnitt aus dem Beobachtungsbogen, Gruppendiskussion

In 13 der 105 erhobenen Bearbeitungen wurde durch die Testleitung angekreuzt, dass die Aufgabenstellung Probleme bereitete. Bei den betreffenden Items wurden daraufhin Formulierungen überarbeitet oder angepasst, um die Aufgabenstellung klarer und eindeutiger zu gestalten.

Ein weiteres Beobachtungskriterum war, ob der Aufgabenstamm bei der Bearbeitung des Items erneut zu Rate gezogen wird. In 31 Fällen wurde der Aufgabenstamm bei der Lösung des Items „gar nicht" erneut beachtet, in 33 Fällen „etwas" und in 24 Fällen „intensiv". Allerdings zeigte sich in den Audioaufnahmen, dass eine Nicht-Beachtung in vielen Fällen daran lag, dass der Text gut verstanden worden war und das mentale Situationsmodell im Kopf der Schüler*innen klar war, ohne dass der Text erneut konsultiert werden musste. Ein gesetztes Kreuz bei „gar nicht" wurde aufgrund dessen als Kriterium für eine zwingende Überarbeitung verworfen.

Weitere im Vorhinein festgelegte Kriterien auf Hinweise zur Überarbeitungsbedürftigkeit von Items waren *formale Argumentation* sowie *sofortiger Ausschluss* bei einzelnen Distraktoren. *Formale Argumentation* bezeichnet dabei alle Argumentationen zur Auswahl der richtigen Lösungen, die sich nicht auf inhaltliche, sondern formale Merkmale des Distraktors bezogen (z.B. das Vorkommen bestimmter Wörter, die Länge der Antwortmöglichkeit etc.). Dieses Merkmal wurde nur in 10 der 105 Fälle angekreuzt. Deutlich häufiger (63 Mal) wurde das Kriterium *sofortiger Ausschluss* durch die Testleitenden markiert. Allerdings zeigte sich in den Audioaufnahmen, dass hier ein gewisser Interpretationsspielraum der Bezeichnung „sofort" durch die Testleitenden eine nicht zu vernachlässigende Rolle spielte. Teilweise wurden hier Distraktoren angekreuzt, obwohl sie erst nach der inhaltlichen Diskussion ausgeschlossen wurden. Dies stellt kein Kriterium für eine notwendige Überarbeitung dar. Vielmehr sollten hier Hinweise auf sehr unattraktive Disktraktoren gesammelt werden. So wurden für jedes Item die hier angekreuzten Distraktoren nochmals kritisch in Augenschein genommen und gegebenenfalls überarbeitet, mit dem Ziel, sie attraktiver zu machen. Dabei konnte über Kriterien, die die Attraktivität eines Distraktors für Schüler*innen erhöhen aber nur Vermutungen angestellt werden. Genauere Analysen werden Teil der Pilotierungsstudie sein.

Bei der Auswahl der „richtigen“ Lösung lässt sich feststellen, dass exakt ein Drittel der Items durch die Schüler*innen korrekt gelöst wurde. Von den 75 verbleibenden Items wurden vier weitere teilweise richtig gelöst, was bedeutet, dass die Kleingruppe sich nicht einigen konnte und am Ende zwischen der richtigen und mindestens einer falschen Antwort schwankte. 71 Items wurden falsch gelöst. Eine weitere Betrachtung der Lösungshäufigkeiten ist an dieser Stelle wenig sinnvoll, es können keine Aussagen differenziert nach unterschiedlichen kfA-Niveaus abgeleitet werden, da jedes Item nur von einer einzigen Kleingruppe bearbeitet wurde und somit die Tatsache, ob ein Item richtig oder falsch gelöst wurde, stark von den Lernenden abhängt. Allerdings lassen die genannten Zahlen vermuten, dass die Items tendenziell eher zu schwierig als zu leicht konstruiert wurden. Bei der anschließenden Überarbeitung der Aufgaben wurden deswegen bei einigen Items die kognitiv-fachlichen Anforderungen verringert, so dass teilweise das Item im kfA-Niveau III verworfen wurde, die Niveaus I und II aufrückten und ein neues Item im kfA-Niveau I entworfen wurde, welches geringere kognitive Anforderungen an die Schüler*innen stellt.

5.2. Einschätzung der Aufgabenschwierigkeit durch Schüler*innen

5.2.1. Hypothese

Das zweite Ziel der Untersuchung war es, eine erste empirische Bestätigung des zur Aufgabenkonstruktion und -analyse verwendeten Modells der kognitiv-fachlichen Anforderungsniveaus zu erhalten. Dazu wurden die Angaben, die die Schüler*innen zur Einschätzung der Schwierigkeit der Items gemacht haben, ausgewertet. Zu jedem Item wurde in jeder der drei Schulen die Schwierigkeitseinschätzung einer Kleingruppe erfasst. Es existieren also zu jedem Item drei Datensätze von je 1-3 Schüler*innen. Diese Werte wurden in Excel anhand von verschiedenen Parametern der deskriptiven Statistik ausgewertet. Es wurde davon ausgegangen, dass Schüler*innen aufgrund der in Abschnitt 4.2.2 dargestellten Argumente die Schwierigkeit der Aufgaben innerhalb gewisser Grenzen valide einschätzen können. Daraus ergab sich die Hypothese, dass die Schüler*innen bei der Einordung der Schwierigkeit dreier Items einer Unit zumindest in Teilen die Struktur der kognitiv-fachlichen Anforderungsniveaus laut Modell

reproduzieren, wenn die Werte untereinander verglichen werden. Dies würde bedeuten, dass die von einem bestimmten Schüler oder einer bestimmten Schülerin vergebenen Werte der Schwierigkeitseinschätzung Folgendes erfüllen: kfA I < kfA II < kfA III.

5.2.2. Einschränkungen

Bei der Analyse und Interpretation der Daten ist es zunächst von Bedeutung, einige einschränkende Eigenschaften dieser Daten zu betrachten.

1. Die Schüler*innen legen unterschiedliche Maßstäbe zur Einschätzung der Schwierigkeit an. So vergibt die eine Schülerin für eine leichte Aufgabe vielleicht eine zwei, während ein anderer Schüler eine für ihn als genauso leicht empfundene Aufgabe mit einer vier bewertet. Die Skalen sind in keiner Weise „geeicht". Es wird also nicht erwartet, dass die Werte für kfA I um beispielsweise einen Wert von 2 streuen und die Werte für kfA II und kfA III um Werte wie 5 und 7. Die Gesamtschwierigkeit des Items ist sehr stark abhängig vom Kontext der Unit, also davon, ob die Konstruktion des mentalen Modells gelingt sowie von den Lernenden, also beispielsweise davon, ob die fachlichen Wissensinhalte bekannt sind. Diese beiden Faktoren sind für eine*n Schüler*in allerdings innerhalb einer Unit konstant, weswegen die relativen Unterschiede zwischen den Werten, die ein einziger Schüler den verschiedenen Items einer Unit gibt, von deutlich größerem Interesse sind als die absoluten Werte. Aus diesem Grund war es besonders wichtig, die Werte eindeutig den Schüler*innen einer Kleingruppe zuzuordnen, da ja jede*r Schüler*in seine Schwierigkeitsskala individuell für sich „eicht". Somit kann überprüft werden, ob Schüler*innen öfter als man laut Zufall erwarten würde, tatsächlich kfA III als schwieriger einschätzen als kfA II und kfA I sowie kfA II als schwieriger als kfA I.

2. Die Schüler*innen beurteilen die Schwierigkeit der Items nach unterschiedlich intensiver Bearbeitung. Für eine Unit wurde in jeder Schule ein anderes kfA-Niveau intensiv bearbeitet (d. h. von der Kleingruppe gelöst), während zu den anderen beiden Niveaus je nur eine Einschätzung der Schwierigkeit nach dem Lesen des Items abgegeben wurde. Somit entstand die erste Schwierigkeitseinschätzung der Schüler*innen nach der intensiven Auseinandersetzung mit dem Item. Die

Schwierigkeitseinschätzungen der beiden anderen Items der Unit wurden dagegen lediglich auf der Grundlage (mehr oder weniger) sorgfältigen Lesens getroffen. Dies könnte ebenfalls eventuell zu einer Verzerrung führen. So zeigte sich beispielsweise in einer Untersuchung, in der Student*innen die Schwierigkeit von Biologieaufgaben einschätzen sollten, dass ihre Einschätzungen der Schwierigkeit vor und nach ihrer Bearbeitung voneinander abwichen (Dübbelde, 2013).

3. Auch bei den Schwierigkeitseinschätzungen wird die Aussagekraft der Daten dadurch gemindert, dass die Urteile der Schüler*innen innerhalb einer Kleingruppe oft nicht unabhängig voneinander sind, da die Schüler*innen innerhalb einer Kleingruppe sich gegenseitig beeinflussen. So weichen auch hier die Werte zur Einschätzung eines Items innerhalb einer Gruppe im Mittel deutlich weniger voneinander ab, als die Werte zu diesem Item der verschiedenen Gruppen untereinander.

5.2.3. Ergebnisse

Zur Überprüfung, ob die Schüler*innen die kfA-Niveaus in ihrer Einschätzung der Aufgabenschwierigkeit reproduzieren, wurden die von jedem Schüler und jeder Schülerin vergebenen Werte für die drei Items einer Unit miteinander verglichen. Es gibt sechs verschiedene Möglichkeiten, die drei Items einer Unit in eine Reihenfolge zu bringen. Bei einer rein zufälligen Verteilung würde man also erwarten, dass bei einem Sechstel der Schülerurteile zu den Items einer Unit die Werte eine modellkonforme Sortierung der Items ergeben, also die Bedingung kfA I < kfA II < kfA III erfüllen. Bei der Auswertung wurde analysiert, ob die vergebenen Werte der Schüler*innen mit größerer Wahrscheinlichkeit, also überzufällig, diese Bedingung erfüllen. Es wurden dabei immer nur die Werte eines bestimmten Schülers oder einer bestimmten Schülerin miteinander verglichen und aus bereits genannten Gründen nicht die Werte verschiedener Schüler*innen untereinander. Insgesamt lagen für diese Vergleiche 235 Werte-Triplets vor, bestehend aus den drei verschiedenen Werten, die ein*e Schüler*in den drei Items einer Unit auf der Schwierigkeitsskala zugeordnet hat. 29 dieser Tripletts erfüllten die genannte Bedingung, was einem Anteil von 12,34 % entspricht. Damit liegt der Anteil der modellkonform sortierten Items leicht unter dem Anteil von 16,37 % (also einem

Sechstel), der bei zufälliger Sortierung zu erwarten wäre. Die Schüler*innen haben also die Abstufung der Schwierigkeit laut Modell nicht reproduziert.

Um genauere Aussagen über die Sortierung treffen zu können, wurden zusätzlich statt Item-Tripletts nur Itempaare in jeder möglichen Kombination aus dem Triplett einer Unit miteinander verglichen. So sollte untersucht werden, ob Schüler*innen überzufällig möglicherweise ein oder zwei der drei möglichen Itempaare modellkonform eingeschätzt haben. Wenn ein*e Schüler*in beispielsweise die Werte auf der Schwierigkeitsskala so vergibt, dass sich in aufsteigender Reihenfolge die Sortierung kfA I, kfA III, kfA II ergibt, dann wurde, obwohl das gesamte Triplett nicht modellkonform sortiert wurde, trotzdem das Item im kfA-Niveau III modellkonform als schwerer eingeschätzt als das Item im kfA-Niveau I und das Item im kfA-Niveau II ebenso modellkonform als schwerer als das Item im kfA-Niveau I. In diesem Fall wurde lediglich die Sortierung der kfA-Niveaus II und III vertauscht, sodass trotzdem zwei von drei möglichen Entscheidungen modellkonform getroffen wurden. Um diese Fälle differenzierter betrachten und von Fällen unterscheiden zu können, in denen nur eine (z. B. kfA II, kfA III, kfA I) oder keine einzige Entscheidung (kfA III, kfA II, kfA I) modellkonform getroffen wurde, wurden für jedes Werte-Triplett je die drei möglichen Itempaare gebildet und anschließend überprüft, ob sich für diese Paare eine modellkonforme Sortierung ergibt, also ob (je nach Itempaar) die Bedingung kfA I < kfA II, kfA II < kfA III oder kfA I < kfA III erfüllt ist. Insgesamt wurden so 707 Itempaare miteinander verglichen [3]. Da hier immer nur zwei Werte miteinander verglichen werden, würde man bei einer rein zufälligen Vergabe der Werte auf der Schwierigkeitsskala erwarten, dass die Hälfte der Itempaare modellkonform sortiert sind. Die Schüler*innen reproduzieren also überzufällig das Modell, wenn mehr als die Hälfte der Itempaare die Bedingung erfüllen. Die Ergebnisse dieser Analyse sind in Tabelle 2 dargestellt.

[3] Der Vergleich kfA 2 < kfA 3 wurde zwei Mal häufiger durchgeführt als die anderen beiden Vergleiche, da eine Kleingruppe (bestehend aus zwei Schüler*innen) keine Angaben zur Einschätzung der Schwierigkeit des kfA-Niveaus 1 gemacht hat.

Sortierung laut Modell	kfA I < kfA II	kfA II < kfA III	kfA I < kfA III	gesamt
$n_{konform}$	96	122	113	331
$n_{nonkonform}$	139	115	122	376
Summe	235	237	235	707
Anteil konform	0,41	0,51	0,48	0,47
Anteil nonkonform	0,59	0,49	0,52	0,53

*Tabelle 2: Überprüfung der Schwierigkeitseinschätzungen der Schüler*innen auf Modellkonformität*

Man kann der Tabelle entnehmen, dass auch in dieser Analyse die Schüler*innen in ihrer Einschätzung der Schwierigkeit der Aufgaben die Struktur des Modelles nicht reproduzieren. Der Anteil der modellkonform relativ zueinander eingeschätzten Itempaare beträgt in allen Fällen nahezu 50%. Die Einschätzungen der Schüler*innen verhalten sich relativ zueinander also genau so, wie eine rein zufällige Verteilung es ergeben würde. Das kognitiv-fachliche Anforderungsniveau II wird sogar leicht überzufällig in fast 60 % der Fälle entgegen dem Modell als leichter eingeschätzt als das kognitiv-fachliche Anforderungsniveau I.

Bei der ersten Durchsicht der Beobachtungsbögen entstand jedoch der Eindruck, dass die Schüler*innen möglicherweise das Item, was sie intensiv bearbeitet haben, tendenziell als schwieriger einschätzen, als die beiden anderen Items der Unit, die sie nur gelesen, aber nicht bearbeitet haben. Dies könnte ein Ergebnis der oben beschriebenen Verzerrung durch die unterschiedlichen Beurteilungsgrundlagen sein und eine Erklärung für die dargestellten Zahlen sein. Um dieses Problem zu eliminieren, wurden im nächsten Schritt für jede*n Schüler*in nur die beiden Werte der Items einer Unit miteinander verglichen, die diese*r Schüler*in nicht intensiv bearbeitet hatte. Für diese beiden Items ist die Grundlage, auf der die Beurteilung stattfindet, also gleich, was mögliche Verzerrungen eliminieren sollte. Die Ergebnisse dieser Analysen sind in Tabelle 3 dargestellt.

Sortierung laut Modell	
$n_{konform}$	112
$n_{nonkonform}$	123
Summe	235
Anteil konform	0,48
Anteil nonkonform	0,52

*Tabelle 3: Überprüfung der Schwierigkeitseinschätzungen der Schüler*innen auf Modellkonformität ohne intensiv bearbeitetes Item*

Doch auch hier zeigt sich, dass die Struktur des Modells der kognitiv-fachlichen Anforderungen durch die Schüler*innen nicht reproduziert wird. Wieder liegt der Anteil der modellkonform eingeschätzten Itempaare sehr nahe der Hälfte, was ebenso bei einer rein zufälligen Zuordnung beliebiger Werte zu erwarten wäre. Wie diese nicht erwartungsgemäßen Ergebnisse zu deuten sind, wird im folgenden Abschnitt diskutiert.

5.2.4. Diskussion der Ergebnisse

Abweichend von den Erwartungen reproduzieren Schüler*innen in ihren Einschätzungen der Aufgabenschwierigkeit die Struktur des Modells der kognitiv-fachlichen Anforderungen nicht. Da ein formuliertes Ziel dieser Erhebung darin bestand, Evidenz für die Qualität des Modells zu sammeln, stellt sich nun also die Frage, ob aus den Ergebnissen der Schluss gezogen werden muss, dass die Qualität des Modells ungenügend ist.

Das zur Aufgabenkonstruktion verwendete Modell der kognitiv-fachlichen Anforderungsniveaus befand sich für die Präpilotierung, wie bereits beschrieben, in einem ersten „Rohstadium“. Die Validität des Modells wird in der anschließenden Pilotierungsstudie durch die Ermittlung der empirischen Aufgabenschwierigkeit anhand größerer Fallzahlen empirisch überprüft. Da sich die Kriterien, die im Modell verwendet wurden, allerdings auf bisherige Forschung beziehen, ist es als unwahrscheinlich anzusehen, dass durch die Abstufungen des Modells überhaupt kein Unterschied in der Aufgabenschwierigkeit entsteht. Eine mögliche Fehlerquelle stellt allerdings die Anwendung des Modells zur konkreten Aufgabenkonstruktion dar. Dies wird durch eine

weitere detaillierte Analyse der kognitiv-fachlichen Anforderungen der Aufgaben erneut überprüft und ist nicht Gegenstand dieser Arbeit.

Allerdings gibt es noch eine weitere, unter Umständen wahrscheinlichere Erklärung für die Abweichung der Ergebnisse von den Erwartungen: Die Annahme, dass Schüler*innen als Expert*innen für Aufgabenschwierigkeit anzusehen sind, muss verworfen werden. Möglicherweise konnten die Schüler*innen (zumindest in diesem Untersuchungsdesign) keine treffende Einschätzung der Aufgabenschwierigkeit vornehmen und die vergebenen Werte auf der Schwierigkeitsskala bilden somit die Aufgabenschwierigkeit nicht ab.

Die Annahme, dass Schüler*innen als Expert*innen für Aufgabenschwierigkeit angesehen werden können, wurde wie in Abschnitt 4.2.2 beschrieben, ohne empirisch abgesicherte Grundlage getroffen, sondern eher als Versuch aufgrund von theoretischen Vermutungen. Dabei wurde zwar nicht unbedingt erwartet, dass ein großer Teil der Schüler*innen die kognitiv-fachlichen Anforderungsniveaus bei der Einschätzung der Schwierigkeit reproduziert, da davon ausgegangen wurde, dass Schüler*innen eine fundierte Eischätzung nur innerhalb bestimmter Grenzen vornehmen können. Allerdings wurde durchaus erwartet, dass sich aus den Einschätzungen der Schüler*innen öfter eine modellkonforme Sortierung der Items ergibt, als bei einer rein zufälligen Zuordnung beliebiger Werte zu den Items. Dies ist jedoch nicht der Fall: bei allen Vergleichen der Werte von Itempaaren miteinander zeigte sich stets ungefähr die Hälfte als modellkonform und die andere Hälfte als nonkonform. Dies spricht für eine zufällige Verteilung und gegen ein statistisches Muster und legt den Schluss nahe, dass Schüler*innen im gewählten Untersuchungsdesign keine geeigneten Expert*innen für die Beurteilung der Aufgabenschwierigkeit sind. Das Untersuchungsdesign spielt dabei möglicherweise eine wichtige Rolle. Zwei Drittel der Items wurden nur nach (evtl. oberflächlichem) Lesen beurteilt, ein Drittel nach intensiver Bearbeitung. Einerseits ergibt sich das bereits angesprochene Problem der begrenzten Vergleichbarkeit dieser Einschätzungen untereinander. Will man dieses Problem umgehen und lässt deshalb bei der Analyse die intensiv bearbeiteten Items außen vor, wie in der Auswertung beschrieben, bleiben allerdings nur Einschätzungen von Items, die lediglich aufgrund

von Lesen beurteilt wurden. Dies könnte dazu führen, dass die Schüler*innen sich fast ausschließlich an Oberflächenmerkmalen der Aufgabe orientieren, um ihre Einschätzung der Schwierigkeit zu treffen. Solche formalen Merkmale, wie Satz- oder Wortlängen, Fachwörter, o. Ä., machen aber laut der bisherigen Forschung nur einen kleinen Teil der Aufgabenschwierigkeit aus (siehe Abschnitt 2.3.3), was bedeutet, dass eine anhand dieser Merkmale getroffene Bewertung deutlich weniger aussagekräftig ist. So würden Schüler*innen eventuell öfter die Struktur des Modells reproduzieren, wenn sie alle drei Items einer Unit intensiv bearbeitet und gelöst hätten. Dies war in der durchgeführten Untersuchung aufgrund der begrenzten Zeit nicht möglich. Des Weiteren hätte dabei möglicherweise ein Lerneffekt die Werte verzerrt, da die Items oft aufeinander aufbauen und Schüler*innen so eventuell ein Item einer Unit besser lösen können, wenn sie vorher schon ein anderes Item der gleichen Unit bearbeitet haben.

Aber auch unabhängig davon, ob die Schüler*innen eine Aufgabe nur gelesen oder auch bearbeitet haben, bevor sie diese einschätzen sollen, legen die Ergebnisse die Vermutung nahe, dass für eine Beurteilung der Aufgabenschwierigkeit, die auch die Tiefenstruktur der Aufgabe berücksichtigt, Expertenwissen notwendig ist, über welches Schüler*innen im Allgemeinen offenbar nicht verfügen: einerseits über das fachliche Themengebiet, z.B. die Komplexität der benötigten Wissenselemente, andererseits auch über die notwendigen kognitiven Prozesse und möglichen Lösungswege. Dies würde bedeuten, dass die Erfahrungen, die Schüler*innen in ihrem Alltag mit dem Lösen von Aufgaben und den damit verbundenen Rückmeldungen über richtige und falsche Lösungen machen, nicht ausreichen, um eine diagnostische Kompetenz zu erreichen, mit der (teilweise) die Schwierigkeit von Aufgaben treffend beurteilt werden kann.

Aufgrund dieser überraschenden und ernüchternden Ergebnisse wurde die Entscheidung getroffen, den von der Zielsetzung her vorher bewusst offen gelassenen, qualitativen Teil dieser Arbeit auf die Frage auszurichten, wie (un)zutreffend die Einschätzungen der Aufgabenschwierigkeit der Schüler*innen sind. Dafür wurden in einer qualitativen Auswertung der Audioaufnahmen die Begründungen der Schüler*innen zu ihren Schwierigkeitseinschätzungen untersucht. Dabei wurde analysiert, auf welcher Grundlage diese Urteile gefällt wurden und welche

Überlegungen der Schüler*innen damit verbunden waren, um daraus Rückschlüsse auf die Aussagekraft der Urteile zu ziehen.

5.3. Qualitative Analyse der Begründungen

Die qualitative Auswertung orientierte sich an folgender Forschungsfrage:
Welche Hinweise lassen sich in den Begründungen der Schüler*innen zu ihren Schwierigkeitseinschätzungen finden, die für oder gegen eine zutreffende Beurteilung sprechen und welche möglichen Erklärungen lassen sich daraus für die Diskrepanz zwischen der Schwierigkeitseinschätzung der Schüler*innen und dem Modell ableiten?
Zur Beantwortung dieser Frage wurden die Begründungen der Schüler*innen zu ihren Schwierigkeitseinschätzungen analysiert und nach inhaltlichen Merkmalen klassifiziert. Da im Rahmen dieser Masterarbeit nicht alle Begründungen der Schüler*innen ausgewertet werden konnten, mussten in einem ersten Schritt einige der 105 Testsituationen zur Auswertung ausgewählt werden, die als möglichst exemplarisch für alle Testsituationen angesehen werden können und somit ein möglichst breites Spektrum verschiedener relevanter Phänomene beinhalten. Dazu wurden anhand der Daten der Beobachtungsbögen 12 Testsituationen nach verschiedenen Kriterien ausgewählt, die im folgenden Abschnitt dargelegt sind.

5.3.1. Auswahl der Testsituationen

Die Auswahl der zu analysierenden Testsituationen zielte darauf ab, dass sich diese in den Merkmalen, von denen ein Einfluss auf die Schwierigkeitseinschätzung erwartet wird, möglichst voneinander unterscheiden, um dadurch ein möglichst breites Spektrum verschiedener Einschätzungen und Phänomene zu erhalten.
Astleiter (2008, S. 70) nennt als Faktoren, die die subjektive Aufgabenschwierigkeit, also das Schätzurteil einer Person, beeinflussen, unter anderem „Expertise zu einem Sachverhalt“. Es ist also denkbar, dass Schüler*innen, die ein besseres Verständnis für den in der Unit behandelten physikalischen Sachverhalt haben, eine treffendere Einschätzung der Schwierigkeit der Items abgeben können, da sie die Komplexität der Wissenselemente möglicherweise besser überblicken können. Da Schüler*innen mit

einer guten Expertise für den physikalischen Sachverhalt ein Item mit größerer Wahrscheinlichkeit korrekt lösen, wurden die Testsituationen zunächst in zwei Gruppen eingeteilt, je nachdem ob das bearbeitete Item von der Kleingruppe korrekt gelöst wurde oder nicht.

Um sechs Fälle aus den Testsituationen mit korrekter Lösung des Items und sechs Fälle mit falscher Lösung auszuwählen, wurden drei verschiedene Kriterien herangezogen: Das erste Kriterium bezog sich auf die absoluten Werte der Schwierigkeitseinschätzungen einer Unit durch die Kleingruppe. Innerhalb der betrachteten Fälle sollte sich möglichst ein breites Spektrum der subjektiven Schwierigkeit abbilden. Möglicherweise gelingt es Schüler*innen besser, die Items einer Unit, die insgesamt eher als leicht empfunden wird, einzuschätzen als die Items einer als schwer empfundenen Unit. Um einen groben Eindruck zu erhalten, als wie schwierig eine Unit von einer Kleingruppe insgesamt empfunden wurde, wurden für jede Testsituation die Mittelwerte der Schwierigkeitseinschätzungen für jedes der drei Items gebildet und diese Mittelwerte aufsummiert. Die so erhaltenen Summen befanden sich für alle Testsituationen im Bereich zwischen 3 und 28. Anhand dieser Zahlen wurde sowohl innerhalb der Gruppe der richtig gelösten Items als auch unter den falsch gelösten Items, je eine Testsituation, in der eine Unit als sehr schwer (also nahe 28) und eine Testsituation, in der eine Unit als sehr leicht (nahe 3) eingeschätzt wurde, ausgewählt. Die verbleibenden vier Units wurden so gewählt, dass sie je verschiedene Werte aus dem mittleren Spektrum aufweisen.

Das zweite herangezogene Kriterium zur Auswahl der Testsituationen bezog sich auf den relativen Unterschied der drei vergebenen Werte eines Schülers oder einer Schülerin zueinander. Dahinter steckt die Überlegung, dass, wenn Schüler*innen den Schwierigkeitsunterschied zwischen zwei Items als groß empfinden, sie möglicherweise mehr mit Aufgabenmerkmalen argumentieren, bzw. ihre Einschätzungen genauer begründen, als wenn sie keinen oder kaum einen Unterschied zwischen den Items wahrnehmen. Aus diesem Grund wurde in jeder Testsituation für jede*n Schüler*in die Differenz der abgegebenen Schwierigkeitseinschätzungen gebildet. Die Differenzen jeder der drei möglichen Itempaare wurden in absoluten Werten summiert und diese

Summe als Anhaltspunkt zur Auswahl der Testsituationen genutzt. Hier bewegten sich die Werte in einem Spektrum zwischen 0 und 16. Dabei wurden vorrangig Items mit hohen Differenzen genommen, wobei darauf geachtet wurde, dass in beiden Gruppen auch Einschätzungen mit geringer Differenz enthalten waren.

Das dritte Kriterium zur Auswahl der Testsituationen war die Modellkonformität der Einschätzungen bezüglich der Abstufung der kfA-Niveaus. Dafür wurden in beiden Gruppen sowohl Testsituationen ausgewählt, in denen die Items der Unit (nahezu) vollständig nach Modell eingeschätzt wurden, als auch Testsituationen, in denen die Items genau entgegen oder nur teilweise gemäß dem Modell sortiert wurden. Dazu wurde für jede Testsituation abgezählt, wie viele der drei möglichen Itempaare relativ zueinander modellkonform eingeschätzt wurden. Bei drei Schüler*innen in einer Kleingruppe konnte diese Anzahl also maximal 9 betragen, was bedeuten würde, dass die Unit vollständig modellkonform eingeschätzt wurde. In den Fällen, in denen eine Kleingruppe aus weniger als drei Schüler*innen bestand, wurde die Anzahl entsprechend hochgerechnet, sodass sich für eine bessere Vergleichbarkeit der Werte immer 9 als Maximum ergab.

Außerdem wurde bei der Auswahl der Testsituationen darauf geachtet, dass keine Schüler*innengruppe mehrfach vertreten war. Dies geschah, um die Begründungen von möglichst vielen verschiedenen Schüler*innen zu analysieren, da zu erwarten ist, dass die Herangehensweise bzw. die verwendeten Kriterien zur Einschätzung der Schwierigkeit stark vom Lernenden und dessen diagnostischer Kompetenz abhängen. Da das Vorwissen zu kognitiven Prozessen ebenso relevant ist (also beispielsweise, ob eine Klasse sich intensiv mit Operatoren beschäftigt hat), wurde zusätzlich darauf geachtet, dass alle drei Schulen in beiden Gruppen vertreten waren. Die Tabellen 4 und 5 geben einen Überblick über die ausgewählten Testsituationen und ihre beschriebenen Merkmale, nach denen sie ausgewählt wurden.

Testsituation: Unit und Gruppe	Summe Schwierigkeitswerte	Differenzen	Modellkonformität
63_01: HG	21	10, 9, 12	4
57_01: HA	16	12, 6, 8	1
32_01: HC	16	10, 9, 9	8
42_01: HH	3,5	1, 0	3
21_02: UJ	12	12, 6	6
53_03: PF	24	12, 6	3

*Tabelle 4: Ausgewählte Testsituationen aus der Gruppe der richtig gelösten Items. In der linken Spalte bezeichnet jeweils die Zahl die Unit, und die beiden Buchstaben die Kleingruppe. Der erste Buchstabe der Gruppenbezeichnung bezieht sich dabei auf die Schule, in der die zugehörigen Daten erhoben wurden. Die drei verschiedenen Schulen sind mit H, U und P bezeichnet. Für die Differenzen ergeben sich 2 oder 3 Werte, je nachdem, ob die Kleingruppe aus 2 oder 3 Schüler*innen bestand.*

Testsituation: Unit und Gruppe	Summe Schwierigkeitswerte	Differenzen	Modellkonformität
58_01: HB	18,5	12, 12, 13	9
64_02: UH	28	0, 1	3
42_02: PH	10	12, 7	7,5
22_03: UA	12	12, 2	1,5
31_02: PB	6	2, 2, 0	3
10_03: UE	16	2,1	3

Tabelle 5: Ausgewählte Testsituationen aus der Gruppe der falsch gelösten Items. Weitere Erläuterungen siehe Tabelle 4.

5.3.2. Methode

Für die gewählten Testsituationen wurden die bei der Erhebung aufgenommenen Audiodateien mittels einer inhaltlich-strukturierenden nach Mayring (2010) ausgewertet. Aus Effizienzgründen wurden ausschließlich die Begründungen der Schüler*innen transkribiert. Da das Interesse dem Inhalt gilt, wurden dabei Füllwörter („äh“ o.ä.) weggelassen und Dialektfärbungen usw. der deutschen Schriftsprache angepasst. Besondere Betonungen wurden durch Unterstreichungen deutlich gemacht,

Pausen wurden nur bei einer Länge von 3 Sekunden oder länger durch drei Punkte dargestellt. Bei Bezug auf die Symbole #, & und §, mit denen die unterschiedlichen Items in der Untersuchung gekennzeichnet waren, wurde zur besseren Einordnung in eckigen Klammern das kfA-Niveau ins Transkript eingefügt. Die Transkripte wurden in der Datenverarbeitungssoftware MaxQDA analysiert und codiert. Die Kategorien für die Codierung wurden dabei nach inhaltlichen Gesichtspunkten induktiv am Material gebildet. Dazu wurde das komplette Material in einem ersten Durchgang analysiert und den Begründungen der Schüler*innen Codes zugeordnet. Durch die induktive Kategorienbildung entstanden während der Arbeit am Material laufend neue Kategorien und Subkategorien. In einem zweiten Schritt wurde dann das gesamte Material mit dem so entstandenen Codesystem erneut codiert. Der folgende Abschnitt gibt einen kurzen Überblick über das Codesystem.

5.3.2.1. Vorstellung des Codesystems

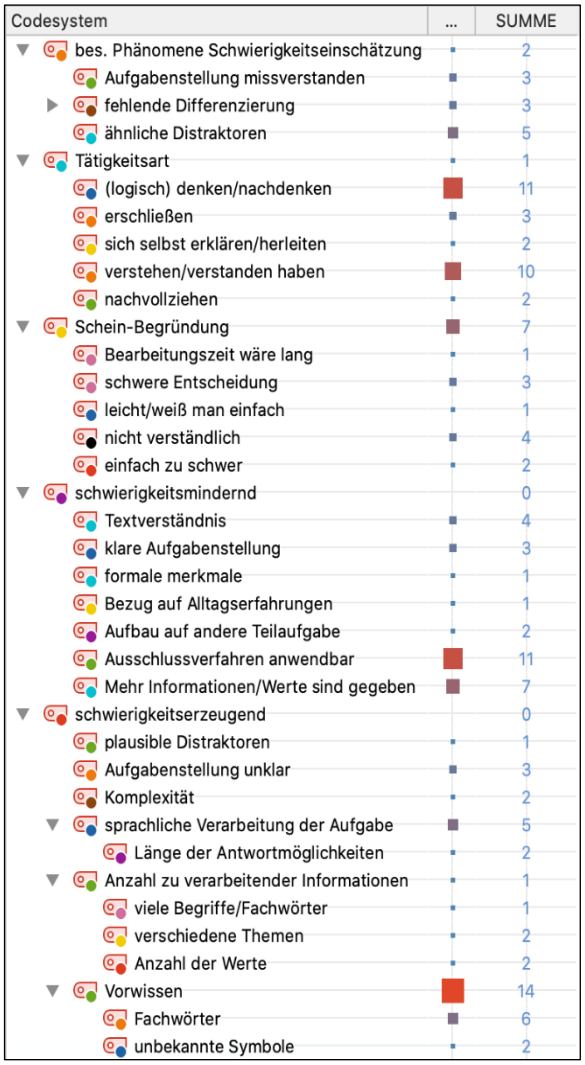


Abbildung 7: Übersicht über das Codesystem und die Summen, wie oft der jeweilige Code vergeben wurde. Die Größe des Quadrates stellt eine Visualisierung dieser Summen dar.

Die induktiv am Material vergebenen Codes wurden in folgende Hauptkategorien eingeteilt: *schwierigkeitsmindernd, schwierigkeitserzeugend, Tätigkeitsart, Schein-Begründung* und *besondere Phänomene*. Alle Kategorien enthalten mehrere Subkategorien, welche Abbildung 7 entnommen werden können. Eine vollständige Code-Matrix, die die Vergabe der Codes in den unterschiedlichen Testsituationen darstellt, findet sich im Anhang dieser Arbeit. Die Hauptkategorien und ihre wichtigsten Subkategorien werden im Folgenden mit Hilfe einiger Ankerbeispiele knapp vorgestellt. Auf eine vollständige Beschreibung aller Subkategorien wird an dieser Stelle verzichtet.

In die Kategorien *schwierigkeitserzeugend* und *schwierigkeitsmindernd* wurden Aussagen von Schüler*innen einsortiert, in denen Gründe genannt wurden, warum eine Aufgabe schwierig bzw. leicht sei. Dabei fanden sich in der Kategorie *schwierigkeitserzeugend* teilweise Merkmale des Modells der kognitiv-fachlichen Anforderungsniveaus wieder. Ein erstes Beispiel dafür ist die *Anzahl zu verarbeitender Informationen,* die die Subkategorien *viele Begriffe/Fachwörter, verschiedene Themen* und *Anzahl der Werte* enthält. Dort wurden beispielsweise folgenden Aussagen zugeordnet.

> „Ich würde 7 bis 8 sagen, weil es sehr viele Zahlen auf einmal sind. Es ist schwer für mich zu verstehen.“ (Ph_63_01 HG, Pos. 24-25)

> „Und ich hatte jetzt nicht das Gefühl, dass es sich jetzt um ein Thema handelt [...], wenn es jetzt um so viele verschiedene Pole, Elektronen, Reibungs- das ist jetzt alles so viel Unterschiedliches, was ich jetzt nicht so einfach fand das zu unterscheiden, deswegen." (Ph_57_01 HA, Pos. 7-12)

Eine weitere, häufig vergebene Subkategorie der genannten schwierigkeitserzeugenden Merkmale ist die Kategorie *Vorwissen.* Beispiele hierfür sind folgende Aussagen:

> „wir hatten ja jetzt den Text vorher quasi, aber da wird dann doch viel mehr Fachwissen verlangt, als man durch den Text jetzt erfahren würde." (Ph_57_01 HA, Pos. 47-49)

> „Also wenn man jetzt was über das Thema wüsste, dann wäre die Aufgabe an sich schon leichter, also dann wäre es halt ne 4 ca. aber so ohne das, auch 8." (Ph_57_01 HA, Pos. 17-18)

Aussagen, die sich konkreter auf bestimmtes Vorwissen bezogen, wurden den Subkategorien *Fachwörter* und *unbekannte Symbole* zugeordnet. Weitere genannte schwierigkeitserzeugende Merkmale waren *plausible Disktraktoren*, *unklare Aufgabenstellung*, *Komplexität* und *sprachliche Verarbeitung der Aufgabenstellung* mit der Subkategorie *lange Antwortmöglichkeiten*.

In der Kategorie der *schwierigkeitsmindernden* Aufgabenmerkmale wurde mit Abstand am häufigsten genannt, dass bei der Lösung der Aufgabe ein Ausschlussverfahren gut anwendbar sei.

> „beim zweiten § [kfA 1] würde ich so 3-4 sagen, weil ich finde es baut halt auf die erste Aufgabe auf und die Antwort ist eigentlich relativ leicht, man kanns wieder nach Ausschlussverfahren machen." (Ph_32_01 HC, Pos. 28-32)

> „ja, das # [kfA 1] find ich ... einfach. Also ich würd so 1 oder 2 sagen, eher 2. Weil es ist, also ich könnte hier drei Sachen ausschließen, aber eine Sache weiß ich nicht ganz genau." (Ph_22_03 UA, Pos. 9-11)

Die weiteren Subkategorien der schwierigkeitsmindernden Merkmale können Abbildung 7 entnommen werden.

Mit dem Code *Tätigkeitsart* wurden alle Passagen markiert, in denen Schüler*innen Aussagen darüber gemacht haben, was man tun müsse, um die Aufgabe richtig zu lösen. Hier wurde oft *denken oder nachdenken* bzw. *verstehen* genannt. In selteneren Fällen wurden Tätigkeiten wie *erschließe*n oder *sich selbst erklären* genannt.

> „Die § [kfA 1] ist so ne 5 würde ich sagen. ja ... ist, man muss halt auch ein bisschen nachdenken" (Ph_32_01 HC, Pos. 25-26)

> „Also beim § [kfA 1], ich würds auch auf 2 schätzen, weil das würde man eigentlich verstehen. Also ich würds jetzt schon verstehen, also wenn man jetzt auch den Text dazu bekommen hätte und es mehrmals auch lesen würde, würde mans verstehen. Also nicht beim ersten Mal, aber beim zweiten Mal liest man dann und denkt nochmal nach, deswegen ist es eigentlich auch einfach." (Ph_21_02 UJ, Pos. 16-20)

In die Kategorie *Schein-Begründung* wurden alle Begründungen von Schüler*innen einsortiert, die keine konkreten Aussagen enthielten. Dazu zählen folgende Beispiele:

> „Ich find zwischen 4 und 5. Weil, keine Ahnung, kann ich nicht begründen." (Ph_42_02 PH, Pos. 7-8)

> „Ich würde 9 nehmen. [...] es war halt schwerer zu verstehen. Die Frage war auch bisschen schwerer." (Ph_63_01 HG, Pos. 7-8)

Mit der Kategorie *Phänomene* wurden Passagen gekennzeichnet, in denen Besonderheiten in den Begründungen der Schüler*innen aufgefallen waren. Diese werden im Abschnitt 5.3.4 separat vorgestellt.

5.3.3. Ergebnisse

Welche Hinweise lassen sich nun in diesen inhaltlich kategorisierten Begründungen der Schüler*innen finden, die für oder gegen eine zutreffende Beurteilung sprechen? Welche Anhaltspunkte gibt es für mögliche Erklärungen der Diskrepanz zwischen Schüler*innenurteilen und Modell?

Der Code, der in den analysierten Testsituationen am häufigsten vergeben wurde, ist *Vorwissen*. Dies ist ein erster Hinweis, der für eine fundierte und damit zutreffende Beurteilung spricht. Die Schüler*innen verbalisieren, dass es vom eigenen Vorwissen abhängig ist, ob sie eine Aufgabe lösen können oder nicht und beziehen dies in ihre Beurteilungen mit ein. Vorwissen ist unumstritten ein wichtiges Merkmal mit Einfluss auf die Aufgabenschwierigkeit. Auch bei der Konstruktion der Aufgaben wurde aus theoretischer Perspektive davon ausgegangen, dass die Gesamtschwierigkeit eines Items vor allem durch das Vorwissen des Bearbeitenden in Zusammenhang mit dem in der Unit beschriebenen physikalischen Kontext erzeugt wird (siehe Abschnitt 3.2.1). Allerdings ist Vorwissen nicht das Merkmal, an dem sich die Abstufung der kfA-Niveaus ergibt. So argumentieren auch die Schüler*innen eher absolut mit dem Vorwissen und ziehen es nicht als Begründung für Schwierigkeitsunterschiede zwischen den Items

heran. Es zeigte sich, dass wenn Schüler*innen also aufgrund von fehlendem Vorwissen eine Aufgabe nicht lösen können, sie dieses Problem bei allen drei Items haben. So wird eine Aufgabe dann oft nur noch aus dem Bauch heraus beurteilt, ohne Differenzierung an anderen Merkmalen der Aufgabe, wie Menge der Informationen oder Tätigkeitsart. Die Unterschiede zwischen den drei kfA-Niveaustufen entstehen laut Modell hauptsächlich durch die beiden Merkmale *Anzahl der Prozeduren* und *Tätigkeitsart*. In diesen beiden Merkmalen wird der kognitive Prozess beschrieben, der zur korrekten Lösung einer Aufgabe durchlaufen werden muss. Diese Reflexion auf metakognitiver Ebene findet sich allerdings in kaum einer der Schüler*innen-Begründungen wieder. Relativ häufig wurde in Bezug auf den kognitiven Prozess einfach „man muss nachdenken" o.Ä. genannt. Dies deutet darauf hin, dass die Schüler*innen schon ein Gefühl dafür haben, dass ein kognitiver Prozess nötig ist, der mehr als Reproduktion bedeutet. Allerdings wird dieser Prozess kaum differenziert, es wird meist nicht darüber nachgedacht, was genau getan werden muss. Nur drei Mal wurde im Unterschied zu „nachdenken" genannt, dass man sich etwas „erschließen" muss. Lediglich ein Schüler bezeichnete seinen Denkprozess genauer als „selber um die Ecke denken" und sortierte dabei die Items modellkonform:

> „Ich würde eher 8 sagen, weil das waren, also hier die Fragen, bisschen kompliziert, so. Also, das steht ja nicht ganz genauso wie im Text, da muss man selbst nochmal um die Ecke selber denken. Nachdenken, was richtig ist, so, logisch." (Ph_21_02 UJ, Pos. 5-8)

Deutlich häufiger wurde genannt, dass man „verstehen" muss. Dies bezog sich aber in den meisten Fällen auf die Aufgabenstellung und in selteneren Fällen (und hier ergibt sich eine Nähe zur Kategorie Vorwissen) darauf, dass man das physikalische Thema verstanden haben muss. Nur in einem Fall wurde *verstehen* als Argument für höhere Anforderungen genannt:

> „Also ich find auch, das ist nen bisschen kompliziert. Also ich find § [kfA 2] ist eigentlich schon schwer. Also man muss halt, wie sie gesagt hat, das Thema verstehen, man muss wirklich wissen, was man da macht, deswegen, ich finds eigentlich schwer, ich würds auf 9 vielleicht einstufen." (Ph_64_02 UH, Pos. 16-19)

Höhere kognitive Tätigkeitsarten wie Bewerten oder sogar Transfer, also Übertragen, wurden nicht genannt. Dies ist ein Hinweis darauf, dass die Schüler*innenurteile aus

dieser Perspektive nur als eher „grob zutreffend“ angesehen werden können, da nur sehr selten zwischen verschieden anspruchsvollen kognitiven Prozessen differenziert wird.
Sehr häufig wurden in den Begründungen der Schüler*innen außerdem Argumente genannt, die sich auf ein Ausschlussverfahren zur Lösung der Aufgabe bezogen. Dies könnte mitunter auch eine mögliche Erklärung dafür sein, dass die kfA-Niveaustufen von den Schüler*innen nicht reproduziert wurden, da so die Aufgabe aus einer anderen Perspektive beurteilt wird. Die kfA-Niveaustufen wurden im Modell immer aufgrund des kognitiven Prozesses, der nötig ist, um den Attraktor als richtig zu identifizieren, bestimmt. Der kognitive Prozess, der nötig ist, um Distraktoren auszuschließen, ist möglicherweise mit anderen Anforderungen verbunden. Schüler*innen scheinen aber oft diese Lösungsstrategie verfolgt zu haben: Die meist genannte Begründung, warum eine Aufgabe leicht(er) sei, ist, dass ein Ausschlussverfahren gut funktioniert. Auch als Begründung für die Schwierigkeit einer Aufgabe wurden teilweise Aussagen getroffen, die auf diese Lösungsstrategie hindeuten, wie z. B. „sich nicht entscheiden können“ oder „plausible Distraktoren“. Dies zeigt sich am Beispiel folgender Aussage:

> „Ich würde sagen 4 § [kfA 3]. Weil ... weil viele, viele Lösungen auch
> Sinn ergeben. Man müsste sich für eine entscheiden.“ (Ph_22_03 UA, Pos. 12-13)

Die Argumentation von Schüler*innen, dass eine Aufgabe umso leichter sei, je besser sich möglichst viele Distraktoren ausschließen lassen, ist zunächst plausibel. In einigen Testsituationen führte das Ausschlussverfahren zur richtigen Lösung, sodass die Aufgabe zutreffend als leicht eingeschätzt wurde. Allerdings gab es auch Fälle, in denen Aufgaben mit dieser Argumentation als sehr leicht eingeschätzt wurden, die falsch gelöst worden waren, da die richtige Antwortmöglichkeit vorschnell ausgeschlossen worden war. Hier lässt sich also kein eindeutiger Hinweis darauf ableiten, ob eine Schwierigkeitseinschätzung mit dieser Begründung zutreffend ist oder nicht. Es zeigt sich aber, wie wichtig bei der Aufgabenkonstruktion eine sorgfältige Analyse der Distraktoren ist, um zu vermeiden, dass diese die kognitive Anforderung der Aufgabe bei einer Lösung nach Ausschlussverfahren vermindern.
Weitere Gründe, die von den Schüler*innen zu ihrer Schwierigkeitseinschätzung genannt wurden, beziehen sich auf Oberflächenmerkmale einer Aufgabe, die laut Theorie zwar ebenso einen Einfluss auf die Aufgabenschwierigkeit haben können,

welcher aber als deutlich geringer anzusehen ist, als der Einfluss von Merkmalen, die Inhaltstruktur und kognitive Prozesse betreffen. So wurde eine Aufgabe als leichter empfunden, wenn weniger Zahlenwerte gegeben waren, z.B.:

> „Hier & [kfA 1] fand ich relativ leicht, da hätte ich 2-3 gegeben. Einfach weil da weniger Werte da waren. Das war nicht so verwirrend." (Ph_63_01 HG, Pos. 34-36)

Ebenso gab es einige Aussagen, die sich auf die sprachliche Verarbeitung der Aufgabenstellung und der Antwortmöglichkeiten bezogen, beispielsweise auf die Wortwahl oder die Länge der Distraktoren.

> „Ich würde auch so 4 sagen. Weil, man musste schon öfter lesen, um es richtig zu verstehen, weil ähnliche Begriffe drin vorkamen." (Ph_32_01 HC, Pos. 10-11)

> „Und die sind auch zum Beispiel länger, die Antworten und somit auch komplexer und nicht so verständlich wie konkrete Fakten." (Ph_32_01 HC, Pos. 15-16)

Insgesamt lässt sich an dieser Stelle das vorläufige Fazit ziehen, dass es durchaus einige Hinweise darauf gibt, dass die Schwierigkeitseinschätzungen der Schüler*innen in Teilen zutreffend sind, da sie sich auf Merkmale beziehen, deren Einfluss auf die Aufgabenschwierigkeit theoretisch begründet und empirisch abgesichert ist.
Allerdings finden sich auch deutliche Hinweise, die gegen eine zutreffende Einschätzung der Schüler*innen sprechen oder zumindest die starke Begrenztheit nahelegen: Ein genauerer Blick auf die Begründungen der Schüler*innen zeigt, dass es ihnen oft schwerfällt, die Unterschiede zwischen den kfA-Niveaus wahrzunehmen. Die Aufgabenmerkmale, die variiert wurden, erzeugen einen vergleichsweise (mehr oder weniger) kleinen Unterschied in der Gesamtschwierigkeit zwischen den kfA-Niveaus, der von den Schüler*innen sehr oft nicht erkannt wird. So wurde in einigen der untersuchten Fälle kein relevanter Unterschied zwischen den drei Items wahrgenommen. Dies zeigt sich in folgenden Zitaten:

> „Also die Aufgaben ähneln sich ja. Und hier würde ich auch ne 10 geben, bei beiden, weil die beiden sind ja so ähnliche Aufgaben." (Ph_64_02 UH, Pos. 11-12)

> „Und # [kfA 3] würde ich auch, das ist ja im Prinzip dasselbe, deswegen würde ichs auch so auf 9 einstufen." (Ph_64_02 UH, Pos. 19-21)

Außerdem findet offenbar teilweise überhaupt keine reflektierte Schwierigkeitseinschätzung statt. So werden in manchen Fällen Werte und

Begründungen mitunter wortgleich von Mitschüler*innen übernommen, die vorher gesprochen haben (z. B. 58_01 HB, Pos. 4 und 6) oder Schein-Begründungen aufgeführt wie „die Frage ist schwerer“. Nur sehr selten werden konkrete Aufgabeneigenschaften genannt und oft findet kein direkter Vergleich der Items miteinander statt. So wurden beispielsweise sehr selten konkrete Unterschiede zwischen den Items benannt. Teilweise wurde auch die Aufgabenstellung grundsätzlich missverstanden, sodass Aussagen in den Distraktoren, die eigentlich auf ihre Richtigkeit bewertet werden sollten, fälschlicherweise als Erklärung gedeutet wurden und das Item deshalb als besonders leicht eingeschätzt wurde. In einem Fall wurde auch nicht erkannt, dass eine Aufgabe mehr als Lesefähigkeiten erfordert:

> „Das # [kfA 2] ist eigentlich eine einfache Aufgabe. Da muss man sich
> den Text einfach nur an einer bestimmten Stelle halt gründlich durchlesen.“ (Ph_21_02 UJ, Pos. 9-10)

So lässt sich die Forschungsfrage für den qualitativen Teil der Untersuchung dahingehend beantworten, dass sich einige Hinweise finden lassen, die dafür sprechen, dass Schüler*innen die Aufgabenschwierigkeit treffend einschätzen können, beispielsweile die Nennung von Kriterien die auch laut bisheriger Forschung als relevant für die Schwierigkeit eingeschätzt wurden. Allerdings überwiegen hier die Kriterien mit eher geringem Einfluss auf die Aufgabenschwierigkeit. Metakognitive Betrachtungen werden von den Schüler*innen nicht durchgeführt. Dies deutet darauf hin, dass die Schüler*innen in diesem Untersuchungsdesign die Aufgabenschwierigkeit nur relativ grob zutreffend einschätzen können. Vielen Schüler*innen gelingt generell eine reflektierte Betrachtung der Aufgabenmerkmale und eine Reflexion über deren Schwierigkeit kaum.

Die Abstufung der kfA-Niveaus zu erkennen, erfordert also offenbar einen recht differenzierten Blick auf Aufgabenschwierigkeit sowie auf Aufgabenmerkmale, da es hier nötig ist, Unterschiede in Details wahrzunehmen. Offensichtlich reicht die diagnostische Kompetenz und das Reflexionsvermögen über eigene kognitive Prozesse der Schüler*innen, die an dieser Untersuchung teilgenommen haben, in vielen Fällen nicht aus, um die Tiefenstruktur der Aufgabe wahrzunehmen und dadurch die Abstufungen der kfA-Niveaus zu erkennen.

5.3.4. Weitere Phänomene

Bei der qualitativen Auswertung der Daten sind zwei weitere Phänomene aufgefallen, die abschließend Erwähnung finden sollen. Dabei handelt es sich um Begründungen für die Schwierigkeitseinschätzung von Schüler*innen, die als plausibel und reflektiert erschienen[4], aber trotzdem zu einer Einschätzung entgegen dem Modell führten, oder Kriterien beinhalteten, die im Modell nicht erfasst wurden.

Insgesamt fünf Mal, also relativ häufig und teilweise unabhängig voneinander, wurde als schwierigkeitserzeugendes Merkmal genannt, dass sich die Distraktoren sehr ähneln und deswegen die sprachliche Verarbeitung, also das Lesen und Verstehen, und deren Unterscheidung voneinander deutlich anspruchsvoller sind.

> „da wurden halt oft so die ähnlichen Wörter benutzt, weswegen man erstmal die Unterschiede zwischen den verschiedenen Aufgaben richtig verstehen muss. Also was da wirklich der große Unterschied ist." (Ph_32_01 HC, Pos. 4-7)

> „Also ich würd beim ersten so 8-9 & [kfA 3] sagen, weil es halt alles gleich klingt und man es dadurch sehr oft lesen muss, um es zu verstehen." (Ph_32_01 HC, Pos. 28-29)

> „Ich würde hier wieder 7 nehmen, weil ein paar von denen sind einfach gleich, man kann die nicht so leicht unterscheiden. Und deshalb ist es schwer, sich für eine zu entscheiden." (Ph_22_03 UA, Pos. 3-5)

Dazu passt auch, dass, entgegen dem Modell, Schüler*innen eine Erhöhung der Anzahl der Elemente einer Aufgabe nicht immer als schwieriger dargestellt haben. Teilweise wurde es als leichter empfunden, wenn mehr Informationen in den Distraktoren vorkamen. Laut Modell sind mehr Informationen eigentlich mit höherem kognitiven Verarbeitungsaufwand verbunden, auch da dann mehr Informationen in den Distraktoren auf ihre Richtigkeit bewertet werden müssen, was auch mehr Fachwissen erfordert. Laut einigen Schüler*innen liegen durch mehr Informationen allerdings mehr Anhaltspunkte zur Auswahl einer Antwortmöglichkeit vor bzw. höhere Chancen, Vorwissen zu aktivieren. Dies zeigt sich in folgenden Beispielen. Im ersten Beispiel empfand ein Schüler eine Aufgabe, in der er zusätzlich zu Schaltplänen auch noch

[4] Hierbei handelt es sich um meine subjektive Einschätzung.

Messergebnisse zuordnen musste, leichter als die Aufgabe in der nur die Schaltpläne zugeordnet werden mussten:

> „Bei mir ist es auch so 4-5. Es stehen ja die ganzen Werte schon drin.
> Man könnte es halt auch schon rechnen."
> (Ph_63_01 HG, Pos. 19-20)

In einer anderen Testsituation schätzte eine Schülerin die Items der kfA-Niveaus II und III mit einer ähnlichen Argumentation als leichter ein, als das Item im kfA-Niveau I:

> „Ich muss sagen, ich find beides bedeutend einfacher als das davor,
> weils überhaupt mal erklärt wurde. Einfach mal so einen zweiten Begriff dazu zu
> haben, den man dann vielleicht mal gehört hat, ist die Wahrscheinlichkeit, dass
> ich mich daran erinnere, weil ich schon mal irgendwo gehört habe, einfach viel
> viel höher bei mir." (Ph_57_01 HA, Pos. 22-26)

Diese beiden Phänomene, ebenso wie die häufige Verwendung des Ausschlussverfahrens als Lösungsstrategie, deuten darauf hin, dass der Einfluss der Distraktoren auf die kognitiv-fachliche Anforderung keinesfalls unterschätzt werden sollte. Sowohl der Inhalt als auch die sprachliche Gestaltung der Distraktoren hat laut Schüler*innen einen Einfluss auf die Aufgabenschwierigkeit. Für die weitere Forschung kann es also sinnvoll sein, diesen noch gründlicher zu analysieren und ebenso Richtlinien für die Distraktoren der Items nach kognitiv-fachlichen und sprachlichen Gesichtspunkten zu formulieren.

6. Diskussion und Grenzen der Methode

In der qualitativen Auswertung ließen sich Hinweise darauf finden, dass die Schüler*inneneinschätzungen der Aufgabenschwierigkeit durchaus in Teilen zutreffend sein können und nicht völlig zufällig, wie sich zunächst in der quantitativen Analyse der Werte angedeutet hatte. Dadurch stellt sich die Frage, ob möglicherweise ein modifiziertes Untersuchungsdesign den Schüler*innen eine reflektiertere Einschätzung der Aufgabenschwierigkeit ermöglicht hätte, in welcher die Abstufungen der kfA-Niveaus beobachtbar gewesen wären. So wären gezieltere Nachfragen der Testleitenden eventuell hilfreich gewesen, um Impulse zur Reflexion über die Aufgabenschwierigkeit zu setzen, beispielsweise „Was genau findest du an dieser Aufgabe schwieriger als an der anderen?“. Hier hat sich im Nachhinein gezeigt, dass das Vorgehen der Testleitenden relativ unterschiedlich war. Einige haben solche oder ähnliche Impulse gesetzt, andere haben bei fehlender Begründung oder Schein-Begründungen nicht genauer nachgefragt. Unterschiedliche Vorgehensweisen, die sich trotz Testleiter*innenschulungen und Beobachtungsbogen mit Interviewleitfaden nicht vermeiden ließen, stellen einen grundsätzlichen Nachteil einer hohen Anzahl an Testleitenden dar. Dieser musste allerdings aus zeitökonomischen Gründen in Kauf genommen werden.

Auch die unterschiedlichen Grundlagen, auf denen die Schüler*innen ihre Einschätzungen vornahmen, also teilweise intensive Bearbeitung, teilweise nur Lesen, stellt eine Grenze der Methode dar. Möglicherweise gelingt es Schüler*innen besser, sich auch über die Tiefenstruktur einer Aufgabe bewusst zu werden, wenn sie diese intensiv bearbeiten. Bei der qualitativen Auswertung hat sich gezeigt, dass einige Aufgabenstellungen beim reinen Lesen nur unzureichend erfasst wurden. Um dieses Problem zu umgehen, hätte man alle drei Items von den Kleingruppen intensiv bearbeiten lassen können. Allerdings hätte dies einerseits wieder zeitökonomische Probleme mit sich gebracht, andererseits wäre dann das Problem der Verzerrung durch Lerneffekte entstanden. Je nach Unit sind die meisten Items deutlich leichter zu lösen, wenn man ein anderes Item der Aufgabe schon gelöst hat, da sie aufeinander aufbauen.

Generell muss auch beachtet werden, dass durch Abstufung der kfA-Niveaus die objektive Aufgabenschwierigkeit modifiziert wurde. Hier wurde sich demnach an Merkmalen orientiert, die sich für viele Schüler*innen als schwierigkeitserzeugend gezeigt haben. Die objektive Aufgabenschwierigkeit muss allerdings nicht mit der subjektiven Aufgabenschwierigkeit übereinstimmen. Da die untersuchte Stichprobe relativ klein war, müssen die hohen Unstimmigkeiten zwischen Modell und Schüler*inneneinschätzungen weder bedeuten, dass Schüler*innen generell unzutreffend urteilen, noch dass das Modell gar nicht zutrifft. Auch dies stellt eine Begrenztheit der Aussagekraft dieser Daten dar.

7. Fazit und Ausblick

Auch wenn in dieser Untersuchung das kfA-Niveaustufenmodell durch die Schüler*innen nicht bestätigt werden konnte, haben sich dennoch wichtige Punkte gezeigt, die zu einer Überarbeitung der Items und damit zu einer höheren Itemqualität für den weiteren Verlauf des VAMPS-Projektes geführt haben. Aus der Untersuchung lässt sich abschätzen, dass die Grenzen, in denen Schüler*innen die Schwierigkeit einer Aufgabe zutreffend beurteilen können, enger sind, als zunächst angenommen.

Im nächsten Schritt gehen die Units in die Pilotierungsstudie mit deutlich größerer Stichprobe, in der festgestellt wird, ob sich die kfA-Niveaustufen in der empirischen Aufgabenschwierigkeit abbilden. Dann wird sich zeigen, ob und wie sehr das Modell der kognitiv-fachlichen Anforderungsniveaus und vor allem seine konkrete Anwendung bei der Aufgabenkonstruktion noch einmal kritisch überprüft werden müssen. Interessant könnte es dann sein, bei den Items, die nicht oder nur leicht überarbeitet wurden, einen Vergleich der Schüler*inneneinschätzungen mit der empirischen Schwierigkeit vorzunehmen, da dann auf empirischer Grundlage beurteilt werden kann, wie zutreffend die Einschätzungen der Schüler*innen tatsächlich sind.

Durch die Beobachtung der Schüler*innen bei der Lösung eines Aufgabenitems wurden Daten über den Bearbeitungsprozess gewonnen, bei denen sich eventuell zusätzliche Hinweise auf schwierigkeitserzeugende Merkmale finden lassen, die von den Schüler*innen nicht explizit genannt wurden, weil sie nicht bewusst wahrgenommen wurden. Eine weitere Auswertung könnte darüber Aufschluss geben, ebenso über Techniken der Lösung von Aufgaben und vieles mehr.

Allgemein wird mit dem VAMPS-Projekt, zu dem diese Arbeit einen kleinen Beitrag leistet, ein wichtiger Schritt hin zum besseren Verständnis des komplexen Konstruktes der Aufgabenschwierigkeit gegangen, der dabei helfen kann, langfristig die Qualität von Leistungstests und von Unterricht generell zu verbessern.

8. Literatur- und Quellenverzeichnis

Abraham, U., & Müller, A. (2009). Aus Leistungsaufgaben lernen. *Praxis Deutsch, 36* (214), S. 4-12.

Aebli, H. (1980). *Denken. Das Ordnen des Tuns.* Stuttgart: Klett-Cotta.

Anderson, L., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching and assessing. A revision of Bloom's taxonomy of educational objectives.* New York: Addison Wesley Longman.

Astleitner, H. (2008). Die lernrelevante Ordnung von Aufgaben nach der Aufgabenschwierigkeit. In J. Thonhauser, *Aufgaben als Katalysatoren von Lernprozessen. Eine zentrale Komponente organisierten Lehrens und Lernens aus Sicht von Lernforschung, Allgemeiner Didaktik und Fachdidaktik* (S. 65-80). Münster/New York/München/Berlin: Waxmann.

Aufschnaiter, S., & Welzel. (1997). Wissensvermittlung durch Wissensentwicklung. Das Bremer Komplexitätsmodell zur quantitativen Beschreibung von Bedeutungsentwicklung und Lernen. *Zeitschrift für Didaktik der Naturwissenschaften, 3*(2), S. 43-58.

Becker-Mrotzek, M., Schramm, K., Vollmer, H., & Thürmann, E. (2013). *Sprache im Fach. Sprachlichkeit und fachliches Lernen.* Münster: Waxmann.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *The Taxonomy of EDucational Objectives, The Classification of Educational Goals, Handbook I: Cognitive Domain.* New York: Dacid McKay Company.

Cassels, J. R., & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice in chemistry. *Journal of Chemical Education, 61* (7), S. 613-615.

Dübbelde, G. (2013). *Diagnostische Kompetenzen angehender Biologie-Lehrkräfte im Bereich der naturwissenschaftlichen Erkenntnisgewinnung. Dissertationsschrift an der Uniersität Kassel.*

Duden. (26. 04 2020). Von "Schwierigkeit" auf Duden online: https://www.duden.de/node/161670/revision/161706 abgerufen

Fischer. (1994). Physiklernen: Eine Herausforderung für die Unterrichtsforschung. In D. Nachtigall, *Didaktik und Naturwissenschaften, Band 3.* Frankfurt: Lang.

Fischer, H., & Draxler, D. (2002). Konstruktion und Bewertung von Physikaufgaben. In E. Kirchner, & W. Schneider, *Physikdidaktik in der Praxis.* Berlin Heidelberg: Springer.

Fischer, H., & Draxler, D. (2006). Konstruktion und Bewertung von Physikaufgaben. In E. Kircher, R. Girwidz, & P. Häußler, *Physikdidaktik. Theorie und Praxis* (S. 639-655). Berlin: Springer.

Florian, C., Sandmann, A., & Schmiemann, P. (2014). Modellierung kognitiver Anforderungen schriftlicher Abituraufgaben im Fach Biologie. *Zeitschrift für Didaktik der Naturwissenschaften* (20), S. 175-189.

Grewe, M., Strietholt, R., & Schwippert, K. (2007). Unterrichtsqualität aus Schülersicht. In K. Möller, P. Hanke, C. Beinbrech, A. K. Hein, T. Kleickmann, & R. Schages, *Qualität von Grundschulunterricht. entwickeln, erfassen und bewerten* (S. 179-182). Wiesbaden : VS Verlag für Sozialwissenschaften.

Höttecke, D., Feser, M. S., Heine, L., & Ehmke, T. (2018). *Do linguistic features influence item difficulty in physics assessments?. Science Education Review Letters*. Von https://edoc.hu-berlin.de/handle/18452/19939 abgerufen

Haag, N., Heppt, B., Roppelt, A., & Stanat, P. (2015). Linguistic simplification of mathematics items: effects for language minority students in Germany. *Eur J Psychol Educ, 30*, S. 145-167.

Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger, & A. Kelava, *Testtheorie und Fragebogenkonstruktion. 2., aktualisierte und überarbeitete Auflage* (S. 142-171). Berlin Heidelberg: Springer-Verlag.

Hopf, M., Schecker, H., & Wiesner, H. (2011). *Physikdidaktik kompakt.* Hallbergmoos: Aulis Verlag.

Impara, J., & Plake, B. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff Standard Setting Method. *Journal of Educational Measurement, 35*(1), S. 69-81.

Kauertz, A. (2008). Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben. In H. Niedderer, H. Fischler, & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 79). Essen: Logos.

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretischen Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos, & R. Lehmann, *TIMSS/III: Bd. 2 Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (S. 57-128). Opladen: Lekse + Budrich.

KMK. (16. Dezember 2004). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss.*

KMK. (2004). *Einheitliche Prüfungsanforderungen in der Abiturprüfung Physik. Beschluss der Kultusministerkonferenz vom 01.12.1989 i.d.F. vom 05.02.2004.*

Knoblich, G., & Öllinger, M. (2006). Die Methode des Lauten Denkens, The Method of Thinking Aloud. In J. Funke, & P. Frensch, *Handbuch der Allgemeinen Psychologie - Kognition* (S. 691-696). Göttingen: Hogrefe.

Leutner, D., Fischer, H. E., Kauertz, A., Schabram, N., & Fleischer, J. (2008). Instruktionspyschologische und fachdidaktische Aspekte der Qualität von Lernaufgaben und Testaufgaben im Physikunterricht. In J. Thonhauser, *Aufgaben als Katalysatoren von Lernprozessen* (S. 169-181). Münster: Waxmann.

Luthiger, H. (2012). Lern- und Leistungsaufgaben in einem kompetenzorienten Unterricht. *Haushalt in Bildung und Forschung* (3), S. 3-14.

Maier, U., Kleinknecht, M., Metz, K., & Bohl, T. (2010). Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potentials von. *Beiträge zur Lehrerinnen- und Lehrerbildung, 28*(1), S. 84-96.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), S. 207-218.

Mayring, P. (2010). *Qualitative Inhaltsanalyse. Grundlagen und Techniken. 11., aktualisierte und überarbeitete Auflage.* Weinheim und Basel: Beltz.

Prenzel, M., Häußler, P., Rost, J., & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft, 30*(1), S. 120-135.

Projektbeschreibung VAMPS. (20. 02 2020). Von Didaktik der Physik. Universität Hamburg: https://www.ew.uni-hamburg.de/einrichtungen/ew5/didaktik-physik/projekte/projekte-laufend/2019-2022-vamps-cvdg.html abgerufen

Sandmann, A. (2014). Lautes Denken - die Analyse von Denk-, Lern- und Problemlöseprozessen. In D. Krüger, I. Parchmann, & H. Schecker, *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 179-188). Berlin Heidelberg: Springer Verlag.

Schnell, C. (2016). "Lautes Denken" als qualitative Methode zur Untersuchung der Validität von Testitems. *Zeitschrift für ökonomische Bildung* (5), S. 26-49.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: University Press.

Weinert, F. E. (1999). Die fünf Irrtümer. *Psychologie heute* (6), S. 28-34.

Wuttke, J. (2006). Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung. In T. Jahnke, & W. Meyerhöfer, *PISA & Co - Kritik eines Programms* (S. 101-154). Hildesheim: Franzbecker Verlag.

9. Anhang

Beobachtungsbogen für die Testleitung

Folgende Informationen bitte ausfüllen, anschließend in das Aufnahmegerät vorlesen:

Bezeichnung Schule:	
Kürzel Testleiter*in:	
Bezeichnung Gruppe:	
Nummer des Items:	
kfA-Stufen-Symbol:	☐ # „Hashtag“ ☐ & „Und-Zeichen“ ☐ § „Paragraph“

Hinweis:

Arbeitsanweisungen für die Schülerinnen und Schüler, die vorgelesen werden sollen, sind im Folgenden fett und in Anführungszeichen geschrieben.
Bsp.: **„Ihr bekommt jetzt einen Text!“**

Arbeitsanweisungen für die Testleitungen, die nicht laut vorgelesen werden sollen, sind kursiv gedruckt.
Bsp.: *Text umdrehen.*

Phase 1: Lesen und Paraphrasieren	Zeit
„**Ihr bekommt von mir einen Text. Ich bitte einen von euch, ihn laut vorzulesen. Die anderen beiden lesen bitte mit.“** *Unsicherheiten beim Lesen im Text markieren.*	0 min
„**Vielen Dank. Wir legen den Text gleich zur Seite und ihr sollt mir dann erzählen, was in dem Text drinsteht. Wenn ihr möchtet, könnt ihr ihn euch jetzt nochmal anschauen. Wenn ihr der Meinung seid, dass ihr ihn verstanden habt, dann sagt bitte einmal „fertig“**.“ *Darauf achten, dass alle drei „fertig“ sagen, ggf. nachfragen. Text umdrehen.*	2 min
„Bitte erzählt mir, was in dem Text drinsteht. Worum geht es? Einer beginnt, die anderen beiden ergänzen bitte“ *Bei längerer Pause evtl. kleine Impulse: „Fällt euch noch mehr ein?“ „Haben die anderen etwas zu ergänzen?“*	3 min
„Vielen Dank. Jetzt nehmen wir den Text wieder dazu. Möchtet ihr noch etwas Wichtiges ergänzen?“ *Text wieder zurückdrehen.* ☐ Es wurde etwas ergänzt. ☐ Bei der Paraphrasierung gab es größere Probleme oder Unverständnis wurde geäußert.	4 min
Evtl. Bemerkungen:	
„Dankeschön. Hier habe ich eine Schwierigkeitsskala von 1 bis 10. 1 bedeutet: Der Text ist sehr leicht verständlich. 10 bedeutet: Er ist sehr schwer verständlich. Wo würdet ihr diesen Text einordnen? Ihr müsst euch nicht einigen, aber begründet bitte kurz eure Einschätzung“. *Schwierigkeitsskala hinlegen.* *Evtl.: Bei fehlender Begründung nachfragen.*	5 min

Wert 1 (Schüler*in links):	Wert 2 (Schüler*in Mitte):	Wert 3 (Schüler*in rechts):

Phase 2: Aufgabenstellung

„Vielen Dank. Wir kommen jetzt in die zweite Phase. Ihr bekommt hier eine Aufgabe von mir, die zum Text gehört. Genau eine Antwortmöglichkeit ist richtig. Einer liest bitte zunächst wieder vor. Unterhaltet euch dann darüber, was vielleicht richtig und was vielleicht falsch ist. Versucht euch auf eine Lösung zu einigen. Denkt daran, möglichst viel zu sprechen.“ *Erste Aufgabenstellung hinlegen.* *Evtl. Impulse: „Warum sind die anderen falsch?“ „Warum habt ihr das ausgewählt?“* *<u>Achtung:</u> Darauf achten, dass die Antwortmöglichkeiten in der Diskussion benannt werden! (a, b, c, d, e)*	6 min
Die Aufgabenstellung… ☐ war klar ☐ bereitete Probleme	
Evtl. Bemerkungen:	
Der Text wird während der Bearbeitung berücksichtigt ☐ intensiv ☐ etwas ☐ gar nicht	
Über folgende Antwortmöglichkeiten wird inhaltlich diskutiert: ☐ alle ☐ a ☐ b ☐ c ☐ d ☐ e	
Über folgende Antwortmöglichkeiten wird formal argumentiert: ☐ alle ☐ a ☐ b ☐ c ☐ d ☐ e	
Folgende Antwortmöglichkeit wird erkennbar ignoriert oder sofort ausgeschlossen: ☐ a ☐ b ☐ c ☐ d ☐ e	
Als „richtig“ gewählte Antwortmöglichkeit: ☐ a ☐ b ☐ c ☐ d ☐ e	
Evtl.: Bemerkungen (z.B. starkes Schwanken zwischen zwei Möglichkeiten, Uneinigkeit etc…):	

„**Vielen Dank. Ich habe jetzt hier wieder die Schwierigkeitsskala. Wie schwer fandet ihr diese Aufgabe? Bitte ordnet das mal mit einem Wert auf der Skala ein. Ihr müsst euch wieder nicht einigen, aber eure Antwort bitte kurz begründen**.“ *Schwierigkeitsskala hinlegen.* *Evtl.: Nachfrage bei fehlender Begründung.*	12 min

Wert 1 (Schüler*in links):	Wert 2 (Schüler*in Mitte):	Wert 3 (Schüler*in rechts):

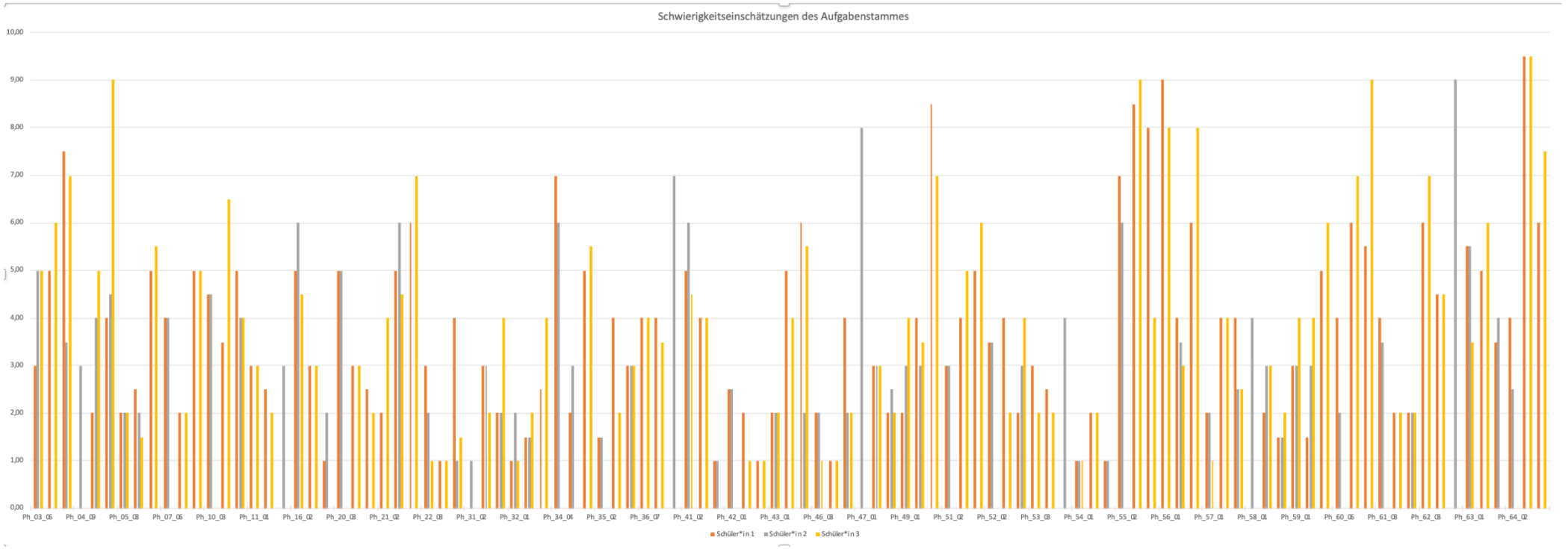
„Jetzt kommen wir zum letzten Schritt. Ich habe hier nun zwei weitere Aufgaben, die auch zu dem Text gehören, den ihr eben gelesen habt. Jeder liest bitte beide Aufgabenstellungen für sich in Ruhe. Die Aufgabe eben habt ihr ja bei *(Werte wiederholen)* **eingeordnet. Ihr sollt diese Aufgaben jetzt nicht lösen, sondern nur ebenfalls auf der Skala einordnen, wie schwer ihr sie findet. Jeder darf sein eigenes Urteil abgeben und bitte wieder kurz begründen.“** *Die weiteren Aufgabenstellungen hinlegen.* *Evtl.: Nachfrage bei fehlender Begründung.*	13 min

Einschätzung Schwierigkeit:

	Wert 3 Schüler*in links	Wert 2 Schüler*in der Mitte	Wert 3 Schüler*in rechts
§			
#			
&			

Ende	15 min

Allgemeine Bemerkungen (besondere Vorkommnisse/ Besonderheiten bei der Mitarbeit etc.):


Schwierigkeitseinschätzungen des Aufgabenstammes
10,00
9,00
8,00
7,00
6,00
5,00
4,00
3,00
2,00
1,00
0,00
Ph_03_05
Ph_04_03
Ph_05_03
Ph_07_05
Ph_10_03
Ph_11_01
Ph_16_02
Ph_20_03
Ph_21_02
Ph_22_03
Ph_31_02
Ph_32_01
Ph_34_04
Ph_35_02
Ph_36_07
Ph_41_02
Ph_42_01
Ph_43_01
Ph_46_03
Ph_47_01
Ph_49_01
Ph_51_02
Ph_52_02
Ph_53_03
Ph_54_01
Ph_55_02
Ph_56_01
Ph_57_01
Ph_58_01
Ph_59_01
Ph_60_05
Ph_61_03
Ph_62_03
Ph_63_01
Ph_64_02
Schüler*in 1
Schüler*in 2
Schüler*in 3

X

Codesystem	Ph_63_...	Ph_57_...	Ph_32_...	Ph_42_01 HH	Ph_21_...	Ph_53...	Ph_58_...	Ph_64_...	Ph_42_02 PH	Ph_22_...	Ph_31_...	Ph_10_...	SUMME
bes. Phänomene Schwierigkeitseinschätzung		1										1	2
Aufgabenstellung missverstanden		2			1								3
fehlende Differenzierung								2				1	3
ähnliche Distraktoren			2						1	2			5
Tätigkeitsart	1												1
(logisch) denken/nachdenken		2	1		3		1	2			1	1	11
erschließen		1	1		1								3
sich selbst erklären/herleiten		2											2
verstehen/verstanden haben		1	3		1	1		2			1	1	10
nachvollziehen							1					1	2
Schein-Begründung				1	2				2		1	1	7
Bearbeitungszeit wäre lang	1												1
schwere Entscheidung							1			2			3
leicht/weiß man einfach				1									1
nicht verständlich	3		1										4
einfach zu schwer	1								1				2
schwierigkeitsmindernd													0
Textverständnis					3							1	4
klare Aufgabenstellung							3						3
formale merkmale											1		1
Bezug auf Alltagserfahrungen											1		1
Aufbau auf andere Teilaufgabe			2										2
Ausschlussverfahren anwendbar		1	2			1			3	2	1	1	11
Mehr Informationen/Werte sind gegeben	2	3				1						1	7
schwierigkeitserzeugend													0
plausible Distraktoren										1			1
Aufgabenstellung unklar							3						3
Komplexität			1		1								2
sprachliche Verarbeitung der Aufgabe		1	4										5
Länge der Antwortmöglichkeiten		1	1										2
Anzahl zu verarbeitender Informationen						1							1
viele Begriffe/Fachwörter		1											1
verschiedene Themen		2											2
Anzahl der Werte	2												2
Vorwissen	1	5		1			2	2			1	2	14
Fachwörter			1			5							6
unbekannte Symbole	1							1					2
SUMME	12	23	19	3	12	9	11	9	7	7	7	11	130

Codematrix, dargestellt mit MAXQDA20

Ph_63_01 HG

[kfA 1]
Person 1: Ich finde, dass diese Aufgabe deutlich schwieriger war, weil mich halt das A oder das V im Kreis bisschen verwirrt haben. Deshalb würde ich diese Aufgabe zwischen 8 und 9 einschätzen. Weil ich mir auch selbst nicht im klaren war, was ich jetzt persönlich wählen würde.

Person 2: Ich würde 9 nehmen. [...] es war halt schwerer zu verstehen. Die Frage war auch bisschen schwerer.

Person 3: Ich würde so zwichen 9 und 10 machen, weil ich auch erstmal die Aufgabenstellung erstmal nicht so gut verstanden habe.

Person 2: Wir haben im Unterricht jetzt nicht durchgenommen, dass das V im Kreis steht, wenn das jetzt die Einheit sein sollte. [...] Das hat uns jetzt am Meisten verwirrt, dass wir das noch nie so gesehen haben.

& [kfA 2]
Person 1: Die würde ich zwischen 4 und 5 einschätzen. Weil es verschiedene Werte gibt. Ich hätte mich für b oder d entschieden, weil das beides in der Aufgabenstellung gegeben ist, muss man sich halt nochmal genau die Aufgabenstellung durchlesen, auf was genau das jetzt passen würde.

Person 2: Bei mir ist es auch so 4-5. Es stehen ja die ganzen Werte schon drin. Man könnte es halt auch schon rechnen.

Person 3: Ich würde sagen 5, weil wir die Einheiten so ein bisschen orientieren müssen, was das eigentlich sein soll, also was für ne Einheit das ist.

§ [kfA 3]
Person 3: Ich würde 7 bis 8 sagen, weil es sehr viele Zahlen auf einmal sind. Es ist schwer für mich zu versehen.

Person 2: Bei mir
ist es 10. Ich habs einfach nicht verstanden. Das wars eigentlich. Ich habe diese Aufgabe mit den Schaltkreisen einfach nicht verstanden.

Person 1: Die liegt im Bereich von 7. Deutlich schwerer als die anderen Aufaben jetzt für mich, weil wenn man sich längere Zeit dafür geben würde, dann würde man das auch verstehen. Wenn das jetzt bezogen ist auf die Aufgabe, könnte man das relativ leicht lösen, denke ich. Aber jetzt nicht so auf Schnelligkeit, deshalb sag ich mal 7.
achso ich dachte das sind Punkte, jetzt bin ich verwirrt. Hier & [kfA 1] fand ich relativ leicht, da hätte ich 2-3 gegeben. Einfach weil da weniger Werte da waren. Das war nicht so verwirrend.

Ph_57_01 HA

[kfa 1]
Person 2: Ich fand die schwer. Also ich fand die WIRKLICH schwer. Weil diese waren einfach so..

Person 1:die wurden halt nicht erklärt, also nicht in der Aufgabe. Das sind halt alles so Dinge, die man sich selbst erklären muss.

Person 2: Und ich hatte jetzt nicht das Gefühl, dass es sich jetzt um ein Thema handelt, also würde man jetzt sagen, gut entweder c) oder d), dann könnte ich noch so denken, gut es geht um das Thema, das weiß ich jetzt, das definiere ich jetzt mal kurz, aber wenn es jetzt um so viele verschiedene Pole, Elektronen, Reibungs- das ist jetzt alles so viel Unterschiedliches, was ich jetzt nicht so einfach fand das zu unterscheiden, deswegen.

Person 1: Ja vor allem, wenn man sich noch nicht davor damit befasst hat und dann halt diesen einen Text liest und dann gehts da nur um dieses Experiment und wie das so aufgebaut ist, dann kann man sich das, find ich, nicht so richtig erklären. Also ich würd sagen so ne 8.

Person 3: Also wenn man jetzt was über das Thema wüsste, dann wäre die Aufgabe an sich schon leichter, also dann wäre es halt ne 4 ca. aber so ohne das, auch 8.

Person 1: ich find das ist wieder so ähnlich wie die andere Aufgabe, das man halt sehr viele Begriffe hat und man muss sich das dann halt so alles selbst herleiten.

Person 2: Ich muss sagen, ich find beides bedeutend einfacher als das davor, weils überhaupt mal erklärt wurde. Einfach mal so einen zweiten Begriff dazu zu haben, den man dann vielleicht mal gehört hat, ist die Wahrscheinlichkeit, dass ich mich daran erinnere, weil ich schon mal irgendwo gehört habe, einfach viel viel höher bei mir.

Person 1: Also ich find halt [...] wenn man nur diesen einen Text hat, ist es schon noch relativ schwierig, also es ist schon einfacher als der andere, weil man da halt irgendwie, da wird mehr erklärt und mehr so beschrieben was passiert, das finde ich deutlich besser, dass man sich das auch so vorstellen kann und dann so auch ausschließen kann, was nicht passiert ist. Aber dennoch ist es so, das man halt, wenn man gar keine Ahnung hat, ja nicht weiß, welche Vorgänge da genau passieren und deswegen ... also wenn man nur diesen einen Text hat, finde ich das schon nicht so einfach. Also wenn man noch nie irgendwas davon gehört hat, was da passiert. [...] ich find die [Aufgaben] sind beide sogar sehr ähnlich im Niveau.

Moderator*in: Kannst du die einstufen auf der Skala?

Person 1: Also ich find, das kann man immer nicht so genau sagen, das hängt halt immer davon ab, was man so davor gemacht hat, ob man...

Moderator*in: und für euch, mit euren Kenntnissen auf dem aktuellen Stand.

Person 1: Achso, dann würd ich sagen, dass es so ne 4 war-

Moderator*in: bei beiden?

Person 1: ja doch, bei beiden.

Person 3: Ich find die Aufgaben werden durch die doch recht unterschiedlichen Lösungen bzw. langen Sätze da ein wenig eher zur Lernaufgabe. Und natürlich ist Physik an sich auch zum Teil ein Lernfach, aber man sollte es ja doch vorher verstanden haben und wir hatten ja jetzt den Text vorher quasi, aber da wird dann doch viel mehr Fachwissen verlangt, als man durch den Text jetzt erfahren würde. Deshalb...und den & [kfA 2] Text finde ich leichter als den § [kfA 3] Text, aber nicht viel. Also würde ich beide so.. also ich würde den & [kfA 2] Text auf ne 5 und den § [kfA 3] Text auf ne 6.

Person 2: Also ich find den § [kfA 3] Text sogar einfacher als den & [kfA 2] Text, weil man halt weniger denken - also ja, man muss natürlich erstmal die Sätze aufnehmen so, das ist natürlich, wenn man müde ist, nicht so einfach, aber ich meine, zum Beispiel jetzt beim § [kfA 3] wo es um Pole geht, muss man sich nichts so richtig erschließen. [...] also wenn da steht 'stoßen sich ab', dann ist es halt relativ logisch, also für uns jetzt würde ich sagen, dass sie sich dann auch anziehen. Aber gleichzeitig ist es auch einfacher, weil man das nicht mal denken muss, sondern es wird einem so hingehalten. Deswegen finde ich § [kfA 3] einfacher, wenn auch nicht so wichtig, diese Information, das heißt ich würde sagen, & [kfA 2] ist so ne 3, weil man halt ein bisschen mehr denken muss und der § [kfA 3] so ne 2.

Ph_32_01 HC

[kfA 2]
Person 1: Also ich würd sagen, dass ist 4. Also man kann eigentlich wieder, nen paar kann man einfach raussuchen, aber ich find die waren jetzt nicht soo ... da wurden halt oft so die ähnlichen Wörter benutzt, weswegen man erstmal die Unterschiede zwischen den verschiedenen Aufgaben richtig verstehen muss. Also was da wirklich der große Unterschied ist. Aber es war jetzt generell bei den Aufgaben so, dass die Antwortmöglichkeiten zum Teil ziemlich ähnlich waren, und deswegen man schauen musste, dass man die gut trennt.

Person 2: Ich würde auch so 4 sagen. Weil man musste schon öfter lesen, um es richtig zu verstehen, weil ähnliche Begriffe drin vorkamen. Aber man versteht es am Ende und kann es auch eigentlich durch Ausschlussverfahren lösen.

Person 3: Ich würde 5-6 sagen, weil ... vielleicht sind mir die Begriffe nicht so bekannt, irgendwie... "Ladungsausgleich" oder so, aber halt, es ist halt wie schon gesagt sehr gleich. Und die sind auch zum Beispiel länger, die Antworten und somit auch komplexer und nicht so verständlich wie konkrete Fakten.

Person 1: Also bei dem & [kfA 3] kann ich sagen, dass ist so ne 8. Die sind ja alle gleich aufgebaut, nur da sind mal nen paar Wörter ausgetauscht, also die Phänomene, bei denen das auftitt und dann halt "nicht" oder "es findet statt", das heißt es hört sich alles ziemlich gleich an, und ja generell finde ich das auch nicht so leicht zu beantworten, weil ich jetzt kein großen - ich wusste nicht, dass es da so einen Unterschied gibt. Wobei, beim Geburtstag könnte man ja denken, dass es sich nicht auf die Socken bezieht. Weil beim Geburtstag findet es ja so normal nicht statt, wenn man jetzt nur die Aufgabe sieht. [...] Die § [kfA 1] ist so ne 5 würde ich sagen. ja ... ist, man muss halt auch ein bisschen nachdenken, aber die Aufgaben sind nicht so gleich, also die Lösungen. Und das lässt sich eigentlich auch erschließen.

Person 2: Also ich würd beim ersten so 8-9 & [kfA 3] sagen, weil es halt alles gleich klingt und man es dadurch sehr oft lesen muss, um es zu verstehen. Und beim zweiten § [kfA 1] würde ich so 3-4 sagen, weil ich finde es baut halt auf die erste Aufgabe auf und die Antwort ist eigentlich relativ leicht, man kanns wieder nach Ausschlussverfahren machen.

Person 3: Ich würde bei dem hier § [kfA 1] sagen, dass es auch so 3 ist. Weil das haben wir ja schon aus der anderen Aufgabe erschlossen und es ist halt b. Und bei & [kfA 3] ist es halt ziemlich gleich. [...] Vielleicht so bei 7. 7-8. Es ist halt nicht soo verständlich.

Ph_42_01 HH

[kfA 3]
Person 1: 1! Also man muss es halt wissen, wenn mans weiß, dann ist es 1.

Person 2: Ich sag 2.

& [kfA 1]
Person 1: Das ist auch wieder ganz einfach. 1. Wenn halt der Wolfram-Faden durchtrennt ist, dann ist halt der Stromkreis unterbrochen.

Person 2: Ja und nicht kurz geschlossen, da bleiben wir bei unserer Meinung. Wenn man die Aufgabe davor gemacht hat, auf jeden Fall. Also würde ich wieder 1 sagen.

§ [kfA 2]
Person 2: 1

Person 1: 1

Ph_21_02 UJ

& [kfA 3]
Person 1: Also wenn man den Text versteht, dann ist es eigentlich gar nicht so schwer, ich würd das bei 4 hintun. Also gar nicht so schwer.

Person 2: ich würde eher 8 sagen, weil das waren, also hier die Fragen, bisschen kompliziert, so. Also, das steht ja nicht ganz genauso wie im Text, da muss man selbst nochmal um die Ecke selber denken. Nachdenken, was richtig ist, so, logisch.

Person 1: Das # [kfA 2] ist eigentlich eine einfache Aufgabe. Da muss man sich den Text einfach nur an einer bestimmten Stelle halt gründlich durchlesen. Also es würde schwieriger sein, wenn man den Text nur einmal zum Lesen bekommt und ihn danach nicht mehr lesen darf, dann würde es schwieriger sein, weil der Text halt nen bisschen kompliziert ist. Aber wenn man den immer vor sich liegen hat, ist es eigentlich ne einfache Aufgabe halt. Und ich würde es bei 1 oder 2 einschätzen.

Person 2: Also beim § [kfA 1], ich würds auch auf 2 schätzen, weil das würde man eigentlich verstehen. Also ich würds jetzt schon verstehen, also wenn man jetzt auch den Text dazu bekommen hätte und es mehrmals auch lesen würde, würde mans verstehen. Also nicht beim ersten mal, aber beim zweiten mal liest man dann und denkt nochmal nach, deswegen ist es eigentlich auch einfach.

Person 1: Also bei dieser Aufgabe § [kfA 1] ist es halt so, das ist so ne Frage, die so zum allgemein-, das muss man sich am Ende halt so erschließen, weil das am Ende des Textes gar nicht richtig gesagt wird, also da muss man halt die einzelnen Schritte so zusammenfassen und dann so gucken, was hat Klaus oder Inge

da halt gemacht, deswegen ist das schon so ein bisschen schwieriger, ich würde das bei 5 hinmachen.

Person 2: ich würde # [kfA 2] auf drei schätzen. Weil, also, das steht nicht im Text, aber man würde schon so sich das selber denken, wenn man selber nachdenkt, so, die Fragen beantworten, halt wenn man selber nachdenkt.

Ph_53_03 PF

§ [kfA 1]
Person 1: 10, weil... oder nee, 9, weil ... man konnte es sich schon ein ganz bisschen denken wegen den Batterien. Aber man weiß, also ich weiß halt nicht was ne Reihenschaltung oder Parallelschaltung ist, also ich weiß es gar nicht, deswegen könnte ich mir von den beiden Sachen eigentlich nichts denken.

Person 2: Ich würd 4 sagen. Ja, weil so wie ich den Text verstanden hab, das mit Batterien war und nur bei d) und e) gibt es Batterien, ich aber nicht den Unterschied weiß von Reihenschaltung oder Parallelschaltung.

[kfA 2]
Person 1: Ich würde sagen wieder 9. Also es ist einfach schwer zu verstehen, vor allem auch wegen jetzt zwei Schaltungen oder einer oder so. Und halt immer wegen Reihenschaltung, Parallelschaltung.

Person 2: Ich würd 10 sagen. Weil das ja irgendwie noch mehr Reihen- und Parallelschaltung hat und ich immer noch nicht weiß, was der Unterschied ist.

& [kfA 3]
Person 1: Also die ist schon ein bisschen leichter, weil, hier steht, also im Text steht, dass hier irgendwas mit 1,5 steht und wenn man dann hier zwei Parallelschaltungen oder Reihenschaltungen halt zusammenrechnet sind das dann 3 wahrscheinlich und deswegen ist es schon ein bisschen leichter, aber man weiß ja immer noch nicht den Unterschied zwischen Reihenschaltung und Parallelschaltung. Ich würde sagen 6.

Person 2: Ich würde wieder 10 sagen. Weil, keine Ahnung. Das mit den 3 Volt ergibt zwar Sinn, aber da steht immer noch Reihenschaltung und Parallelschaltung.

Ph_58_03 HB

[kfA 3]
Person 1: Sehr schwierig. Das war unverständlich, also 10. Weil die Aufgabenstellung war nicht so konkret.

Person 2: Ja, das find ich auch, ich würd sagen, so 9. Also weil die Aufgabenstellung nicht so konkret war, ob es jetzt halt, was am höchstwahrscheinlichsten ist, woran er stirbt, oder was am schlimmsten ist, was passieren kann. Und ... ja ... ich weiß jetzt auch nicht genau, ob jetzt nen Hirnschlag schlimmer ist, oder irgendwie... man kann es schlecht einschätzen. Ich bin mir nicht sicher, was davon jetzt am Schlimmsten ist, bei den meisten.

Person 3: Ja, ich würde auch sagen, 9. War halt ein bisschen unverständlich, was man da halt machen soll.

Person 1: Die & [kfA 1] finde ich leicht, also so 4. Weil die Aufgabe verständlich ist und man auch die einzelnen Situationen einschätzen kann und das § [kfA 2] finde ich schwieriger, das finde ich halt man kann die Konsequenzen nicht so richtig nachvollziehen. So ne 7.

Person 2: Also ich würde diese Aufgabe § [kfA 2] auf ne 6 einschätzen, weil also die Aufgabenstellung ist klar, also die Situation, man kann sie nachvollziehen, man muss aber schon so nen bisschen darüber nachdenken. Und die & [kfA 1] schätze ich so auf ne 3 ein, weil die Situationen sind klar und die Aufgabenstellung auch.

Person 3: Ja hier & [kfA 1] würde ich auch so sagen, 2-3, eigentlich aus den selben Gründen, wie die anderen jetzt gesagt haben. Und die § [kfA 2] würde ich auf ne 5 vielleicht werfen. Also eigentlich kann ich nur wiedergeben, was die anderen schon gesagt haben, also es ist klar gestellt so, und trotzdem sind da so ein paar, keine Ahnung, so Sätze, wo man nicht ganz zu 100 % weiß, was man genau... also ich könnte mich jetzt vielleicht nicht ganz so gut enscheiden, aber ich verstehe die Aufgabe trotzdem gut.

Ph_64_02 UH

& [kfA 1]
Person 1: 10. Weil wir haben das einfach noch nie gesehen, und müssen hier einfach raten. Und ja, wussten auch nicht, was mit den Eingängen gemeint ist

Person 2: Ich find auch eher schwierig. So 8-9. Weil, ich weiß nicht ob das richtig war, was wir gerade gesagt haben, aber man muss halt nen bisschen logisch denen, so. Man muss ein bisschen nachdenken, also auch eher raten, weil wir haben sowas noch nie gesehen und deswegen haben wir jetzt einfach nur geraten. Wie es dann, ob es so möglich sein könnte, wie es halt zu stande gekommen ist, aber, ja, auch eher schwierig.

Person 1: Also die Aufgaben ähneln sich ja. Und hier würde ich auch ne 10 geben, bei beiden, weil die beiden sind ja so ähnliche Aufgaben. Und wie gesagt, wir haben das noch nie gesehen. Und joa, man müsste einfach nachdenken. Logisch denken. In den Text nochmal gucken. Ich glaube man müsste den Text die ganze Zeit durchlesen, um das richtig zu verstehen.

Person 2: Also ich find auch, das ist nen bisschen kompliziert. Also ich find § [kfA 2] ist eigentlich schon schwer. Also man muss halt, wie sie gesagt hat, das Thema verstehen, man muss wirklich wissen, was man da macht, deswegen, ich finds eigentlich schwer, ich würds auf 9 vielleicht einstufen. Und # [kfA 3] würde ich auch, das ist ja im Prinzip das selbe, deswegen würde ichs auch so auf 9 einstufen.

Ph_42_02 PH

§ [kfA 2]
Person 1: Also ich würde sagen ne 3. Weil es waren halt so vier Antworten, also a) bis d) war so, man könnte sich denken, dass alle richtig sind. Aber, weil es ja ein Kurzschluss war und es nur eine Anwort hier mit dem Kurzschluss gibt, habe ich mich dann am Ende halt für b) entschieden. Also deswegen 3.

Person 2: Ich find zwischen 4 und 5. Weil, keine Ahnung, kann ich nicht begründen.

Person 2: Ich find den Zettel, den ich hatte & [kfA 1] einfach. Weil, wir bei der letzten Aufgabe besprochen hatten, uns geeinigt hatten auf den Kurzschluss und es bei dieser Aufgabe auch nur eine Antwort auf den Kurzschluss gibt.

Person 1: Ich würde bei dem Blatt # [kfA 3] eine 7 geben. Weil es ist sehr sehr schwer. Also es gibt sehr, also alle haben sozusagen fast den gleichen Inhalt. Also ich weiß jetzt nicht, welches ich aussuchen sollte. Also es war halt schwer zu verstehen.

Person 1: und bei & [kfA 1] finde ich auch 1. Weil es halt nur einen Satz gibt, der halt was mit dem Kurzschluss zu tun hat. Und die anderen halt passen gar nicht zu dem Text hier, also zu dem, was passiert. Deswegen würde ich sagen 1.

Person 2: Ich würde # [kfA 3] zwischen 3 und 4 einschätzen, weil es da nur zwei Antworten gibt mit dem Kurzschluss und ich würde mich da auf c) einigen.

Ph_22_03 UA

& [kfA 2]
Person 1: Ich würde hier wieder 7 nehmen, weil ein paar von denen sind einfach gleich, man kann die nicht so leicht unterscheiden. Und deshalb ist es schwer, sich für eine zu entscheiden.

Person 2: Ich würde so 5 sagen. Das war schon einfacher, aber trotzdem war es schwer sie zu unterscheiden, weil sich viele oder sich manche Antwortmöglichkeiten auch sehr identisch sind.

Person 1: ja, das # [kfA 1] find ich ... einfach. Also ich würd so 1 oder 2 sagen, eher 2. Weil es ist, also ich könnte hier drei Sachen ausschließen, aber eine Sache weiß ich nicht ganz genau.

Person 2: Ich würde sagen 4 § [kfA 3]. Weil ... weil viele, viele Lösungen auch Sinn ergeben. Man müsste sich für eine entscheiden.

Person 1: Hier § [kfA 3] würde ich ne 1 nehmen. Weil ich find, hier könnte ich die meisten Sachen schon ausschließen. Und ich finde nur eine Antwortmöglichkeit sinnvoll.
Person 2: Ich würde 5 nehmen # [kfA 1]. Weil, ich kann mich nicht wirklich so entscheiden, welches richtig ist.

Ph_31_02 PB

§ [kfA 2]
Person 1: 1. Also zwei Aussagen waren ja nen bisschen dämlich. Da sind schon zwei ausgefallen. Da hatte man ne höhere Chance zu gucken. Und es ist irgendwie auch verständlich, weil auf einem Hocker aus Holz, Holz leitet ja keinen Strom. Und mit dem Schlüssel, der ist ja aus Eisen, und wie ich weiß, wirds ja immer weiter geleitet, dann durch den Körper halt. Deswegen, d) ist halt verständlich.

Person 2: Ne 2.

Person 3: Ich auch ne 2.

Person 3: Ich glaube ich würde dem & [kfA 1] und dem # [kfA 3] eine 2 geben. Weil die Fragen ganz simpel gestellt wurden.

Person 2: Ich würde dem & [kfA 1] eine 2 geben, weil ... darum (lacht). Und dem # [kfA 3] würde ich eine 3 geben. Bei dem & [kfA 1] kann man sich schon glaub ich ein wenig denken, warum das so ist.

Person 1: Ich würde auch sagen, & [kfA 1] und # [kfA 3] ne 2. Ich kenn das ja auch. Manchmal, wenn ich meinen Mantel anziehe, meinen Bademantel, der ist aus Stoff. Wenn ich dann meine Haare daran reibe, dann spüre ich halt so wie es da oben so knallt, bisschen. Wenn ich dann die Türklinke anfasse, dann kommt so nen blauer (unverständlich) so boom. Bisschen. In meiner Erfahrung auch so.

Ph_10_03 UE

& [kfA 1]
Person 1: Also diesmal würde ich schon nen bisschen schwieriger sagen. So 5-6. Ja, also nen bisschen schwerer als der Text, so. Weil man hat ja auch so verschiedene Ansichten, also darauf.

Person 2: Also wenn man sich so alles erstmal durchgelesen hat, dann erschienen einem so mehrere Sachen als richtig und als man dann aber mehr überlegt hat und logisch gedacht hat, dann konnte man sich für eine Antwort entscheiden. Also man konnte dann halt so ne engere Auswahl treffen. Also würde ich auch 5-6 sagen.

Person 1: Also ich find das # [kfA 3] ist jetzt auch so 5-6, weil dafür braucht man auch so nen bisschen mehr Vorwissen. Und wenn das halt nicht so ausreicht, könnte man die Frage halt schwierig beantworten.

Und bei §[kfA 2] ist es eigentlich genauso. Man bräuchte halt Vorwissen dafür, ja. Ansonsten ist es eigentlich nicht so schwer. Also so 4-6.

Person 2: Also ich find die Aufgabe mit dem # [kfA 3], also ich finde die Aufgabe ist gut zu verstehen, aber man muss halt genau wissen, was die im Text dazu gesagt haben. Und das ist halt nen bisschen schwerer das so ... ja wie du schon gesagt hast, Vorwissen. Und deswegen würde ich auch die Aufgabe so 5-6 sagen und die mit dem § [kfA 2] die finde ich irgendwie leichter, weil man kann da halt so sehen, was halt gefragt wird. So die elektrische Abstoßung, dann wird halt immer gesagt, was, und dann kann man das besser nachvollziehen. Und deswegen würde ich da so 4 sagen.

Eigenständigkeitserklärung

Hiermit versichere ich an Eides statt, dass ich die Arbeit eigenständig verfasst habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen und die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium.

Datum, Unterschrift