ZfDN (2018) 24:1–15
https://doi.org/10.1007/s40573-017-0069-0

ORIGINAL PAPER

Schwierigkeitserzeugende Aufgabenmerkmale bei Multiple-Choice-Aufgaben zur Experimentierkompetenz im Biologieunterricht: Eine Replikationsstudie

Moritz Krell[1]

Eingegangen: 1. Juni 2017 / Angenommen: 30. November 2017 / Online publiziert: 11. Dezember 2017

Zusammenfassung

Die Entwicklung von Instrumenten zur Erhebung von Experimentierkompetenz ist ein bedeutsames Aufgabenfeld der Biologiedidaktik. Diese Studie repliziert Befunde einer Vorgängerstudie zur schwierigkeitserzeugenden Wirkung der Merkmale Aufgabenkomplexität (niedrig, hoch), Teilkompetenz (Suche im Hypothesenraum, Testen von Hypothesen, Analyse von Evidenz) und Aufgabenkontext (sechs verschiedene Kontexte) bei Multiple-Choice-Aufgaben zur Experimentierkompetenz im Biologieunterricht. Durch systematische Kombination der drei Merkmale wurden 36 Aufgaben konstruiert. Zur Erklärung der schwierigkeitserzeugenden Wirkung der Aufgabenkontexte wurden deren Bekanntheit, Interessantheit und Relevanz („Kontext-Personen-Valenzen“) erhoben. 708 Schülerinnen und Schüler (8. und 9. Jahrgangsstufe) haben die Aufgaben bearbeitet. Zur Analyse der schwierigkeitserzeugenden Wirkung der Aufgabenmerkmale wurde das Linear Logistische Test-Modell (LLTM) eingesetzt. Zusammenfassend konnten die Befunde der Vorgängerstudie zur schwierigkeitserzeugenden Wirkung der Aufgabenkomplexität und der Teilkompetenzen erfolgreich repliziert werden. Ebenso zeigten sich signifikante Effekte der Kontext-Personen-Valenzen auf die Schwierigkeit der Multiple-Choice-Aufgaben. Insgesamt zeigen die Ergebnisse, dass die Aufgabenkomplexität, die Teilkompetenz sowie die Bekanntheit, Interessantheit und Relevanz von Aufgabenkontexten bei der Konzeption von Tests zur Experimentierkompetenz im Biologieunterricht berücksichtigt werden sollten.

Schlüsselwörter Experimentierkompetenz · Biologieunterricht · Multiple-Choice-Aufgaben · Schwierigkeitserzeugende Aufgabenmerkmale · Linear Logistisches Test-Modell (LLTM) · Replikationsstudie

Abstract

The development of instruments to assess experimental competencies is an important part of biology education research. This study replicates findings of a previous study about the effect of the characteristics task complexity (low, high), competence aspect (forming hypotheses, planning experiments, analyzing data), and task context (six different contexts) on the difficulty of multiple-choice-tasks assessing experimental competencies in biology education. 36 tasks were developed by systematically combining the three characteristics. In order to explain the difficulty generating effect of the task contexts, their familiarity, interestingness, and relevance (“context-person-valences”) were assessed. 708 students (grades 8 and 9) answered the tasks. The Linear Logistic Test-Model (LLTM) was applied to analyze the characteristics’ contribution to task difficulty. Summarizing, the findings of the previous study regarding the difficulty generating effect of task complexity and competence aspect could be replicated successfully. Significant effects of the context-person-valences on the difficulty of the multiple-choice-tasks were found as well. In total, the findings of this study show that the task complexity, the competence aspect, as well as the familiarity, interestingness and relevance of the task contexts should be taken into account when developing tests for the assessment of experimental competencies in biology education.

Keywords Experimental competencies · Biology education · Multiple-choice-tasks · Difficulty generating

Zusatzmaterial online Zusätzliche Informationen sind in der Online-Version dieses Artikels (https://doi.org/10.1007/s40573-017-0069-0) enthalten.

✉ Moritz Krell
moritz.krell@fu-berlin.de

[1] Didaktik der Biologie, Freie Universität Berlin, Schwendenerstraße 1, 14195 Berlin, Deutschland

task characteristics · Linear Logistic Test-Model (LLTM) · Replication study

Einleitung

Experimentelle Designs sind wesentliche methodische Zugänge für die Untersuchung kausaler Zusammenhänge zwischen Variablen in der Biologie (Köchy 2006; Mahner und Bunge 1997; Wellnitz und Mayer 2013). Experimente können als künstliche und meist apparativ vermittelte Eingriffe in die Natur verstanden werden, bei denen durch Eingrenzung und gezielte Variation vorherrschender Bedingungen versucht wird, die funktionellen Verknüpfungen zwischen unabhängigen Variablen (UV) und abhängigen Variablen (AV) unter Berücksichtigung von Kontrollvariablen (KV) zu erkennen (Köchy 2006). In der Forschungspraxis treten allerdings vielfältige Variationen dieses „Idealkonzepts" (Köchy 2006) des Experiments auf (Höttecke und Rieß 2015). Aufgrund dieser Bedeutung von Experimenten für die naturwissenschaftliche Praxis wird Experimentierkompetenz in den Bildungsstandards für den Mittleren Schulabschluss des Fachs Biologie dem Kompetenzbereich Erkenntnisgewinnung zugeordnet; vier von 13 Standards beziehen sich hier auf die Durchführung und das Verständnis experimenteller Untersuchungen (E5 bis E8; KMK 2005). Entsprechend sind die Konzeption von Unterrichtsansätzen zur effektiven Förderung (z. B. Meier und Wellnitz 2013) und die Entwicklung von Instrumenten zur validen Erhebung von Experimentierkompetenz (z. B. Hammann et al. 2007) bedeutsame Aufgabenfelder der Biologiedidaktik.

Bei der Entwicklung von Testinstrumenten muss geprüft werden, inwieweit Testergebnisse valide als Indikatoren für die Ausprägung der Experimentierkompetenz von Schülerinnen und Schülern interpretiert werden dürfen. Insbesondere aufgrund der Komplexität von Kompetenzen (Klieme et al. 2008) und dem Prozesscharakter des Experimentierens (Schreiber et al. 2016) muss untersucht werden, inwiefern die jeweils vorgenommene Operationalisierung in Form eines Testinstruments die Experimentierkompetenz zu diagnostizieren erlaubt (Shavelson 2013). In diesem Zusammenhang kann die Analyse schwierigkeitserzeugender Aufgabenmerkmale dazu beitragen, die Generalisierbarkeit von Testergebnissen zu prüfen. Schwierigkeitserzeugende Aufgabenmerkmale, die nicht auf das zugrunde liegende Konstrukt zurückgeführt werden können (illegitime Quellen von Aufgabenschwierigkeit), stellen die Generalisierbarkeit und damit die valide Interpretation eines Testergebnisses als Indikator für die Kompetenzausprägung einer Person in Frage (Messick 1995; Shavelson 2013; Stiller et al. 2016). Die Analyse schwierigkeitserzeugender Aufgabenmerkmale ist folglich bedeutsamer Teil der Kompetenzmodellierung und Testentwicklung (Hartig und Frey 2012; Prenzel et al. 2002).

Im Zusammenhang mit Experimentierkompetenz liegen bereits Befunde vor, die die jeweils operationalisierte Teilkompetenz und die Aufgabenkomplexität als schwierigkeitserzeugende Aufgabenmerkmale nachweisen (z. B. Hammann et al. 2007; Krell und Vierarm 2016; Mannel et al. 2015; Wellnitz und Mayer 2013). In den bislang vorliegenden Arbeiten wurden die Aufgabenmerkmale allerdings größtenteils nicht systematisch bereits während der Aufgabenbearbeitung berücksichtigt. Auch deshalb gelten schwierigkeitserzeugende Merkmale im Zusammenhang mit Experimentierkompetenz als empirisch noch nicht ausreichend abgesichert (Schecker et al. 2016). Krell und Vierarm (2016) haben daher systematisch die Aufgabenmerkmale Teilkompetenz, Komplexität sowie Aufgabenkontext (im Sinne von Oberflächenmerkmalen der eingesetzten Aufgaben; vgl. Opfer et al. 2012; Schnotz und Baadte 2015) in einem Facettendesign bei der Konstruktion von Multiple-Choice (MC)-Aufgaben zur Experimentierkompetenz im Biologieunterricht berücksichtigt. Die Befunde zeigen, dass sich die drei Merkmale signifikant auf die Schwierigkeit der MC-Aufgaben auswirken (Krell und Vierarm 2016). Die primäre Zielsetzung vorliegender Studie ist die Replikation dieser Befunde.

Die Bedeutsamkeit (d. h. Generalisierbarkeit) von Befunden zu schwierigkeitserzeugenden Aufgabenmerkmalen sollte in Replikationsstudien geprüft werden (Fischer 1995; Hartig et al. 2012). Grundsätzlich gelten Replikationsstudien als essenziell für den wissenschaftlichen Erkenntnisfortschritt (Lamal 1991). Trotz dieser allgemein anerkannten Bedeutung werden Replikationsstudien in sozial- und bildungswissenschaftlichen Disziplinen eher selten durchgeführt und publiziert (Yong 2012). Der Grund hierfür wird vor allem in der herrschenden Veröffentlichungspraxis gesehen (Lamal 1991; Neuliep und Crandall 1993; Yong 2012): Studien, die bestehende Ergebnisse replizieren, haben es wegen des (vermeintlich) fehlenden Neuigkeitswertes oftmals schwer, in angesehenen Fachzeitschriften publiziert zu werden (Fanelli 2012; Yong 2012). Daher wird angenommen, dass allein aufgrund statistischer Logik (*false positives*) ein großer Teil der publizierten Befunde nicht verallgemeinerbar ist: „It can be proven that most claimed research findings are false" (Ioannidis 2005, S. 696).

Die vorliegende Arbeit kommt der Forderung einer stärkeren Berücksichtigung von Replikationsstudien nach und hat das Ziel, die Befunde aus Krell und Vierarm (2016) zur schwierigkeitserzeugenden Wirkung der Aufgabenmerkmale Teilkompetenz, Komplexität und Aufgabenkontext bei MC-Aufgaben zur Experimentierkompetenz von Schülerinnen und Schülern unter Verwendung eines Facettendesigns

Tab. 1 In dieser Arbeit umgesetzte Operationalisierung von Experimentierkompetenz

Teilkompetenzen	Fähigkeiten	Fertigkeiten
Suche im Hypothesenraum	Identifizieren naturwissenschaftlicher Hypothesen	SuS identifizieren die einem Untersuchungsdesign zugrunde liegende Hypothese
Testen von Hypothesen	Planen einer naturwissenschaftlichen Untersuchung	SuS identifizieren das zur Prüfung einer gegebenen Hypothese notwendige Untersuchungsdesign
Analyse von Evidenz	Auswerten naturwissenschaftlicher Untersuchungsergebnisse	SuS identifizieren auf der Grundlage eines gegebenen Untersuchungsdesigns und -ergebnisses die korrekte Schlussfolgerung

SuS Schülerinnen und Schüler

(experimentelles Untersuchungsdesign; Hartig und Frey 2012) zu replizieren.

Direkte Replikationsstudien prüfen die Wiederholbarkeit und Generalisierbarkeit von Befunden, während in konzeptuellen Replikationsstudien Hypothesen geprüft werden und das Design der Vorgängerstudien entsprechend angepasst wird (z. B. durch den Einbezug zusätzlicher Variablen; Schmidt 2009). In dieser Arbeit wird einerseits eine direkte Replikation zur Analyse der schwierigkeitserzeugenden Wirkung von Teilkompetenz und Komplexität umgesetzt, indem das Testinstrument von Krell und Vierarm (2016) weitgehend unverändert eingesetzt wird. Für die Erklärung der schwierigkeitserzeugenden Wirkung von Aufgabenkontexten (Krell und Vierarm 2016) werden „Kontext-Personen-Valenzen“ erhoben (Bekanntheit; Interessantheit, Relevanz; Werner et al. 2014, 2015), was einer konzeptuellen Replikation zur Hypothesenprüfung entspricht (Schmidt 2009).

Grundlagen

Experimentierkompetenz

Aufgrund der Bedeutung experimenteller Praxis für die biologische Forschung ist Experimentierkompetenz in den Bildungsstandards für den Mittleren Schulabschluss des Fachs Biologie im Kompetenzbereich Erkenntnisgewinnung verankert. Schülerinnen und Schüler sollen am Ende der zehnten Jahrgangsstufen in der Lage sein, Untersuchungen mit geeigneten qualifizierenden oder quantifizierenden Verfahren durchzuführen (E5), einfache Experimente zu planen, die Experimente durchzuführen und/oder sie auszuwerten (E6), Schritte aus dem experimentellen Weg der Erkenntnisgewinnung zur Erklärung anzuwenden (E7) sowie Tragweite und Grenzen von Untersuchungsanlage, -schritten und -ergebnissen zu erörtern (E8; KMK 2005, S. 14).

In der fachdidaktischen Forschung wurden bereits unterschiedliche Ansätze umgesetzt, Experimentierkompetenz (oder verwandte Konstrukte wie das Verständnis experimenteller Denk- und Arbeitsweisen; z. B. Mathesius et al. 2014; Vorholzer et al. 2016) zu operationalisieren. Die meisten Ansätze operationalisieren Experimentierkompetenz basierend auf unterschiedlichen Phasen des Experimentierens (z. B. Hammann et al. 2007; Wellnitz und Mayer 2013) und lassen sich dahingehend unterscheiden, ob ihnen eine dreidimensionale (z. B. Scientific Discovery as Dual Search (SDDS)-Modell; Glug 2009; Hammann et al. 2007; Mannel et al. 2015; Vorholzer et al. 2016) oder eine vierdimensionale (z. B. Modell wissenschaftlichen Denkens; Dasgupta et al. 2014; Mathesius et al. 2014; Wellnitz und Mayer 2013) Struktur zugrunde liegt. Darüber hinaus wird Experimentierkompetenz in den meisten Studien produktorientiert in Form von papiergebundenen Testinstrumenten erfasst (insb. MC-Aufgaben; z. B. Hammann et al. 2007; Krell und Vierarm 2016; Mannel et al. 2015; Mathesius et al. 2014), während in einzelnen Studien prozessorientierte hands-on oder computerbasierte Erhebungen umgesetzt werden (z. B. Kambach und Upmeier zu Belzen 2016; Schecker et al. 2016; Schreiber et al. 2016).

In der vorliegenden Arbeit wurde Experimentierkompetenz basierend auf dem SDDS-Modell in die drei Teilkompetenzen *Suche im Hypothesenraum, Testen von Hypothesen* und *Analyse von Evidenz* differenziert. Tab. 1 zeigt, welche Fähigkeiten und Fertigkeiten für das korrekte Beantworten der in dieser Arbeit eingesetzten MC-Aufgaben umgesetzt werden müssen (Krell und Vierarm 2016). Hierbei werden Fähigkeiten als theoretische Konzepte betrachtet, die durch messbare Fertigkeiten operationalisiert sind (Frey 2006; Glug 2009).

Die vorgenommene Operationalisierung von Experimentierkompetenz (dreidimensional, MC-Aufgaben) stellt eine Reduktion des zugrunde liegenden Konstrukts dar, weshalb die Interpretation der Testergebnisse (d. h. messbarer Fertigkeiten) als Indikatoren für die Kompetenzausprägung von Schülerinnen und Schülern ein möglichst gutes Verständnis der zur Aufgabenbearbeitung erforderlichen kognitiven Prozesse voraussetzt (Shavelson 2013). Hierzu kann die Analyse schwierigkeitserzeugender Aufgabenmerkmale beitragen.

Analyse schwierigkeitserzeugender Aufgabenmerkmale

Die Analyse schwierigkeitserzeugender Aufgabenmerkmale ist von erheblicher Relevanz für die empirische Bildungsforschung. Oftmals werden drei Erträge solch einer Analyse unterschieden (z. B. Fleischer et al. 2013; Hartig und Frey 2012; Leucht et al. 2012). *Erstens* können Kenntnisse über schwierigkeitserzeugende Aufgabenmerkmale die systematische Entwicklung von Testaufgaben unterschiedlicher Schwierigkeit anleiten (z. B. Leucht et al. 2012). *Zweitens* kann die Analyse schwierigkeitserzeugender Aufgabenmerkmale a posteriori dabei helfen, eine entstandene Skala zu erklären (z. B. Schecker et al. 2016). *Drittens* kann die Analyse schwierigkeitserzeugender Aufgabenmerkmale zur Konstruktvalidierung beitragen (z. B. Embretson und Daniel 2008), welche hierbei als Konstruktrepräsentation verstanden wird (Embretson 1983). Grundgedanke hierbei ist, dass Annahmen (kognitive Theorien) darüber, welche Anforderungen und kognitiven Prozesse ein Konstrukt konstituieren, durch Testaufgaben operationalisiert und im Sinne von Hypothesen überprüft werden können. Für diesen Ansatz sollten die entsprechenden Anforderungen die Testentwicklung leiten. Im Kontext der Kompetenzmodellierung wird der ursprünglich auf die Prüfung allgemeiner kognitiver Theorien fokussierte Ansatz der Konstruktrepräsentation (Embretson 1983) weiter gefasst (Hartig und Frey 2012). Dabei kann insbesondere durch den Einbezug inhaltsbezogener Aufgabenmerkmale dem kontextualisierten Charakter von Kompetenzen entsprochen werden (Fleischer et al. 2013).

Die Analyse schwierigkeitserzeugender Aufgabenmerkmale kann dazu beitragen, die *valide Interpretation* eines Testergebnisses als Indikator für die Kompetenzausprägung einer Person zu prüfen (Kane 2013; Shavelson 2013): „That is, can one reliably and validly interpret (infer) from a person's performance on a small sample of tasks [...] the level of competence in the full domain?" (Shavelson 2013, S. 80).

Es wird zwischen legitimen und illegitimen Quellen von Aufgabenschwierigkeit unterschieden (Messick 1995; Stiller et al. 2016). Während legitime Quellen von Aufgabenschwierigkeit auf das zugrunde liegende Konstrukt zurückgeführt werden können (z. B. Teilkompetenzen, Niveaustufen), stellen illegitime Quellen von Aufgabenschwierigkeit (z. B. Textlänge, Aufgabenformat) die Generalisierbarkeit und damit die valide Interpretation eines Testergebnisses als Indikator für die Kompetenzausprägung einer Person in Frage (Messick 1995; Shavelson 2013; Stiller et al. 2016). Merkmale, die als illegitime Quellen von Aufgabenschwierigkeit identifiziert werden, sollten daher, etwa durch den Einbezug motivationaler Variablen (z. B. Interesse; van Vorst et al. 2015; Werner et al. 2014, 2015), weitergehend untersucht und können in Form von „merkmalsbezogenen Teilkompetenzen" (Prenzel et al. 2002, S. 124) bei der Kompetenzmodellierung berücksichtigt werden.

Für eine systematische Untersuchung schwierigkeitserzeugender Aufgabenmerkmale sollte die Testkonstruktion einem Facettendesign unterliegen, in dem die zu untersuchenden Aufgabenmerkmale als Dimensionen vollständig gekreuzt sind (Hartig und Frey 2012). Nehm und Kollegen (z. B. Nehm und Ridgway 2011; Opfer et al. 2012) schlagen bei der Erfassung des Verständnisses von Evolution in diesem Sinne vor, den Einfluss von Aufgabenkontexten (*item features*) zu untersuchen, indem in Aufgaben ein konstantes Tiefenmerkmal (*deep item feature*, z. B. der Selektionstyp) durch unterschiedliche Oberflächenmerkmale (*surface features*, z. B. Organismen) kontextualisiert und damit variiert wird (vgl. Schnotz und Baadte 2015). Befunde deuten für Novizen auf einen Einfluss von Aufgabenkontexten auf die Aufgabenschwierigkeit hin, während Experten eher das Tiefenmerkmal einer Aufgabe erkennen und damit über Kontexte hinweg konsistent antworten (Clough und Driver 1986; Nehm und Ridgway 2011).

Schwierigkeitserzeugende Aufgabenmerkmale im Zusammenhang mit Experimentierkompetenz: Teilkompetenz, Komplexität und Aufgabenkontext

Es werden unterschiedliche Systematisierungen schwierigkeitserzeugender Aufgabenmerkmale vorgeschlagen (z. B. Krell und Krüger 2011; Prenzel et al. 2002; Stiller et al. 2016). Anknüpfend an die Unterscheidung zwischen legitimen und illegitimen Quellen von Aufgabenschwierigkeit werden Merkmale, die die Operationalisierung des entsprechenden Konstrukts betreffen (kompetenzrelevante Merkmale bzw. legitime Quellen von Aufgabenschwierigkeit), und Merkmale, die eher die Oberflächenstruktur eines Tests betreffen (kompetenzirrelevante Merkmale bzw. illegitime Quellen von Aufgabenschwierigkeit), unterschieden (z. B. Krell und Krüger 2011; Prenzel et al. 2002). Sowohl die theoretische Einordnung einzelner Merkmale als auch empirische Befunde zu ihrer schwierigkeitserzeugenden Wirkung sind nicht über Konstrukte hinweg generalisierbar (Gut 2012; Messick 1995; Prenzel et al. 2002). Im Zusammenhang mit Experimentierkompetenz können die in einer Aufgabe operationalisierte Teilkompetenz und die Aufgabenkomplexität als legitime Quellen von Aufgabenschwierigkeit betrachtet werden (Gut 2012; Mannel et al. 2015).

Die *Teilkompetenzen Suche im Hypothesenraum, Testen von Hypothesen* und *Analyse von Evidenz* erfordern die Umsetzung spezifischer Fertigkeiten (Glug 2009). Bei den Aufgaben zu *Suche im Hypothesenraum* und *Testen von Hypothesen* ist jeweils ein Aspekt des betrachteten Experiments im Aufgabenstamm vorgegeben (*Suche im Hypothesenraum*: Untersuchungsdesign; *Testen von Hypothesen*: Hypothese), während bei den Aufgaben zu *Analyse von*

Evidenz das Untersuchungsdesign und die Untersuchungsergebnisse vorgegeben und miteinander in Beziehung gesetzt werden müssen (Tab. 1). *Analyse von Evidenz* ist daher kognitiv anspruchsvoller als *Suche im Hypothesenraum* und *Testen von Hypothesen* (Glug 2009). Dementsprechend konnten Krell und Vierarm (2016) in einer Studie mit Schülerinnen und Schülern der Jahrgangsstufen 9 und 10 zeigen, dass Aufgaben zu *Analyse von Evidenz* signifikant schwerer sind als Aufgaben zu *Suche im Hypothesenraum* und *Testen von Hypothesen*. Demgegenüber argumentieren Hammann et al. (2007), dass für die Teilkompetenzen *Suche im Hypothesenraum* und *Analyse von Evidenz* vor allem bereichsspezifisches Wissen erforderlich ist, während *Testen von Hypothesen* stärker auf methodischem Wissen basiert. Entsprechend berichten die Autoren, dass Aufgaben zu *Testen von Hypothesen* für Schülerinnen und Schüler der Jahrgangsstufen 5 und 6 „etwas schwerer zu lösen waren" (Hammann et al. 2007, S. 40) als Aufgaben zu *Suche im Hypothesenraum* und *Analyse von Evidenz*. Insgesamt fallen psychometrische Befunde zur Dimensionalität von Experimentierkompetenz (oder verwandten Konstrukten) nicht einheitlich aus (z. B. Hammann et al. 2007: zweidimensionale Struktur; Vorholzer et al. 2016: eindimensionale Struktur) und erschweren daher ein Urteil über die relative Schwierigkeit der drei Teilkompetenzen.

Die *Aufgabenkomplexität* kann durch die Beschreibung von Experimenten mit einer unterschiedlichen Variablenanzahl variiert werden (Gut 2012; Krell und Vierarm 2016; Mannel et al. 2015). Empirische Befunde belegen, dass die Schwierigkeit von Aufgaben mit zunehmender Variablenanzahl steigt (z. B. Krell und Vierarm 2016; Mannel et al. 2015). Dies kann mit der höheren kognitiven Belastung bei der Bearbeitung komplexerer Aufgaben erklärt werden (Kauertz et al. 2010; Krell 2017). Bezüglich des in der Vorgängerstudie (Krell und Vierarm 2016) eingesetzten Testinstruments konnte bei Schülerinnen und Schülern, die Aufgaben zu Experimenten mit zwei unabhängigen Variablen bearbeitet haben, ein signifikant höherer Cognitive Load nachgewiesen werden als bei Schülerinnen und Schülern, die Aufgaben zu Experimenten mit nur einer unabhängigen Variable bearbeitet haben (Krell 2017).

Die Einordnung von *Aufgabenkontexten* als legitime oder illegitime Quelle von Aufgabenschwierigkeit hängt entscheidend von der Definition und Operationalisierung des Kontextbegriffs ab (Hartig 2008). Bei einer eher weiten Operationalisierung des Kontextbegriffs (z. B. im Sinne einer wissenschaftlichen Disziplin; Krell et al. 2015) können Aufgabenkontexte aufgrund der Definition des Kompetenzbegriffs im Sinne einer kontextspezifischen Leistungsdisposition (Klieme et al. 2008) als legitime Quellen von Aufgabenschwierigkeit und schwierigkeitserzeugende Effekte von Aufgabenkontexten als definitorischer Bestandteil einer Kompetenz betrachtet werden (Krell und Vierarm 2016). Bei einer eher engen Definition des Kontextbegriffs (z. B. im Sinne unterschiedlicher Organismen als Oberflächenmerkmale; Opfer et al. 2012) wären Aufgabenkontexte als illegitime Quellen von Aufgabenschwierigkeit zu betrachten, da mit dem Kompetenzbegriff der Anspruch verbunden ist, die jeweils operationalisierte Fähigkeit (Tab. 1), also das Tiefenmerkmal einer Aufgabe (Opfer et al. 2012), in variablen Situationen anwenden zu können (Klieme et al. 2008; Weinert 2001). Im Zusammenhang mit Experimentierkompetenz fallen Befunde zu Kontexteffekten nicht einheitlich aus. Zum Beispiel berichten Vorholzer et al. (2016, S. 37), dass in einem Test zur Erfassung von experimentellen Denk- und Arbeitsweisen „nicht von einem systematischen Zusammenhang zwischen Aufgabenschwierigkeit und Aufgabenkontext auszugehen ist". Demgegenüber konnten Krell und Vierarm (2016) Aufgabenkontexte als signifikante Prädiktoren der Aufgabenschwierigkeit bei einem Test zur Erfassung der Experimentierkompetenz nachweisen. Hier waren Aufgaben mit Experimenten zur Hefegärung signifikant schwerer als parallele Aufgaben zur Samenkeimung und Schimmelbildung. In beiden Studien (Krell und Vierarm 2016; Vorholzer et al. 2016) wurden Aufgabenkontexte als konkrete experimentelle Settings im Sinne von Oberflächenmerkmalen operationalisiert. Es wird angenommen, dass motivationale Variablen wie die Interessantheit und Relevanz von Kontexten zu einer kontextabhängigen Fokussierung und Ausdauer bei der Aufgabenbearbeitung führen, darüber hinaus fördert eine hohe Bekanntheit von (bzw. ein hohes Vorwissen zu) Kontexten routiniertes Handeln bei der Aufgabenbearbeitung und es werden weniger kognitive Ressourcen für das Erschließen des Oberflächenmerkmals einer Aufgabe benötigt (Kalyuga und Renkl 2010; Roesler et al. 2014, 2016; Schiefele et al. 1993; van Vorst et al. 2015; Werner et al. 2014, 2015). Die schwierigkeitserzeugende Wirkung von Aufgabenkontexten kann demnach als subjektspezifische Interaktion zwischen Kontext- und Personenmerkmalen verstanden werden („Kontext-Personen-Valenzen"; Werner et al. 2014, 2015). In Bezug auf die Interessantheit von Aufgabenkontexten können Roesler et al. (2016) diese Annahme für den Kompetenzbereich Bewertung bestätigen, nicht jedoch für den Kompetenzbereich Fachwissen.

Zusammenfassend liegen somit bereits Befunde zur schwierigkeitserzeugenden Wirkung der Merkmale Teilkompetenz, Komplexität und Aufgabenkontext vor. Allerdings wurden in den vorliegenden Arbeiten größtenteils keine Facettendesigns umgesetzt (z. B. Mannel et al. 2015; Vorholzer et al. 2016; Wellnitz und Mayer 2013), weshalb die schwierigkeitserzeugenden Merkmale methodisch nicht kontrolliert geprüft (Hartig und Frey 2012) und daher empirisch nicht ausreichend abgesichert sind (Schecker et al. 2016).

Zielsetzung der Studie

Die primäre Zielsetzung dieser Studie ist die Replikation bestehender Befunde, hier der Studie von Krell und Vierarm (2016), zur schwierigkeitserzeugenden Wirkung der Aufgabenmerkmale Teilkompetenz, Komplexität und Aufgabenkontext bei MC-Aufgaben zur Experimentierkompetenz im Biologieunterricht unter Verwendung eines Facettendesigns (Hartig und Frey 2012). Hierbei wird die direkte Replizierbarkeit der Befunde zur schwierigkeitserzeugenden Wirkung der Merkmale Teilkompetenz und Komplexität geprüft, während die Befunde zur schwierigkeitserzeugenden Wirkung von Aufgabenkontexten durch Kontext-Personen-Valenzen erklärt werden sollen (konzeptuelle Replikation; Schmidt 2009). Die übergeordnete Fragestellung dieser Studie lautet: Inwieweit wirken sich die Teilkompetenz, die Aufgabenkomplexität und der Aufgabenkontext auf die Schwierigkeit von MC-Aufgaben zur Experimentierkompetenz im Biologieunterricht aus? Die folgenden Hypothesen H1, H2, H3 können aus dem Stand der Forschung abgeleitet werden:

H1 Aufgaben zur Teilkompetenz *Analyse von Evidenz* sind aufgrund der zur Aufgabenbearbeitung notwendigen anspruchsvolleren Fertigkeit schwerer als Aufgaben zu den Teilkompetenzen *Suche im Hypothesenraum* und *Testen von Hypothesen*, welche sich nicht signifikant in ihrer Schwierigkeit unterscheiden (Glug 2009; Krell und Vierarm 2016).

H2 Aufgaben zu Experimenten mit hoher Komplexität (höhere Anzahl unabhängiger Variablen) sind aufgrund der höheren kognitiven Belastung während der Aufgabenbearbeitung schwerer als Aufgaben zu Experimenten mit geringer Komplexität (geringere Anzahl unabhängiger Variablen; Kauertz et al. 2010; Krell und Vierarm 2016; Mannel et al. 2015).

H3 Aufgaben mit Kontexten mit hoher Bekanntheit, Interessantheit und Relevanz sind aufgrund der damit verbundenen Fokussierung und Routine bei der Aufgabenbearbeitung leichter als Aufgaben mit Kontexten mit niedriger Bekanntheit, Interessantheit und Relevanz (Kalyuga und Renkl 2010; Roesler et al. 2014, 2016; Schiefele et al. 1993; van Vorst et al. 2015; Werner et al. 2014, 2015).

Methoden

Testinstrument

Die Aufgabenentwicklung basierte auf einem Facettendesign, in dem die Dimensionen Teilkompetenz (*Suche im Hypothesenraum, Testen von Hypothesen, Analyse von Evidenz*), Komplexität (niedrig, hoch) und Aufgabenkontext (sechs unterschiedliche experimentelle Settings: Beutefangverhalten [BEUTE], Hefegärung [HEFE], Schimmelbildung [SCHIM], Samenkeimung [KEIM], Atmungsfrequenz [ATMNG], Fotosynthese [FOTO]; vgl. Glug 2009; Hammann et al. 2007; Krell und Vierarm 2016) vollständig gekreuzt und dergestalt 36 MC-Aufgaben entwickelt wurden (Tab. 2); jede Aufgabe enthält einen Attraktor und drei Distraktoren. Die Auswahl der Aufgabenkontexte orientierte sich an bereits bestehenden Studien, die ebenfalls auf dem SDDS-Modell basieren (Glug 2009; Hammann et al. 2007). Die Aufgabenkomplexität wurde durch die Beschreibung von Experimenten mit einer (niedrig) beziehungsweise zwei (hoch) unabhängigen Variablen variiert. Aufgabenkontexte werden in dieser Arbeit als Oberflächenmerkmale einer Aufgabe verstanden.

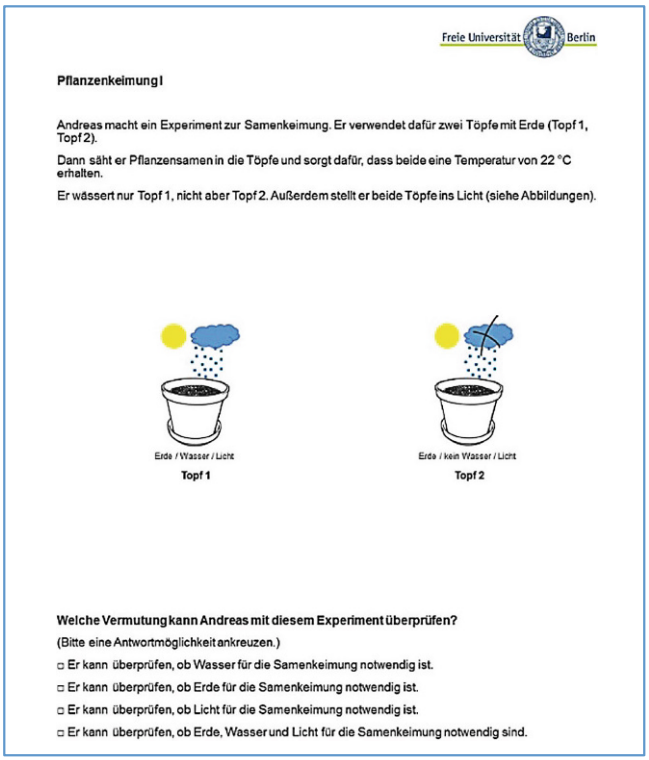
Die Aufgabenentwicklung basierte auf kontextfreien Vorlagen (Tiefenmerkmale), die durch die oben genannten experimentellen Settings (Oberflächenmerkmale) kontextualisiert wurden (vgl. Nehm und Ridgway 2011; Opfer et al. 2012). Im Sinne einer Replikationsstudie (Schmidt 2009) wurde das Testinstrument aus Krell und Vierarm (2016) weitgehend beibehalten und lediglich um drei Kontexte erweitert (FOTO, BEUTE und ATMNG). Abb. 1 zeigt exemplarisch die kontextfreien Vorlagen sowie die unterschiedlich komplexen Aufgaben zur Teilkompetenz *Suche im Hypothesenraum* im Kontext Samenkeimung, die weiteren Aufgaben sind Onlinematerial 1 zu entnehmen.

Tab. 2 In dieser Arbeit umgesetztes Facettendesign mit den drei Dimensionen Teilkompetenz, Komplexität und Aufgabenkontext

		Teilkompetenz		
		Suche im Hypothesenraum	Testen von Hypothesen	Analyse von Evidenz
Komplexität	Niedrig	BEUTE, HEFE*, SCHIM*, KEIM*, ATMNG, FOTO	BEUTE, HEFE*, SCHIM*, KEIM*, ATMNG, FOTO	BEUTE, HEFE*, SCHIM*, KEIM*, ATMNG, FOTO
	Hoch	BEUTE, HEFE*, SCHIM*, KEIM*, ATMNG, FOTO	BEUTE, HEFE*, SCHIM*, KEIM*, ATMNG, FOTO	BEUTE, HEFE*, SCHIM*, KEIM*, ATMNG, FOTO

Mit einem * gekennzeichnet sind diejenigen Kontexte, die im Vergleich zur Vorgängerstudie (Krell und Vierarm 2016) ergänzt wurden

Komplexität niedrig	Komplexität hoch
Aufgabenstamm [Person X] führt ein Experiment zum [biologischen Phänomen Y] durch. Sie nutzt dazu das [Untersuchungsdesign Z].	Aufgabenstamm [Person X] führt ein Experiment zum [biologischen Phänomen Y] durch. Sie nutzt dazu das [Untersuchungsdesign Z].
Ansatz 1: A^+, B^+, C^+ Ansatz 2: A^+, B^-, C^+	Ansatz 1: A^+, B^+, C^+ Ansatz 2: A^+, B^-, C^+ Ansatz 3: A^+, B^+, C^- Ansatz 4: A^+, B^-, C^-
Item Welche Vermutung kann [Person X] mit diesem Experiment überprüfen?	Item Welche Vermutung kann [Person X] mit diesem Experiment überprüfen?
☐ A^+ ist für AV^+ notwendig ☐ B^+ ist für AV^+ notwendig ☐ C^+ ist für AV^+ notwendig ☐ A^+, B^+, und C^+ sind für AV^+ notwendig	☐ A^+ und B^+ sind für AV^+ notwendig ☐ A^+ und C^+ sind für AV^+ notwendig ☐ B^+ und C^+ sind für AV^+ notwendig ☐ A^+, B^+, und C^+ sind für AV^+ notwendig



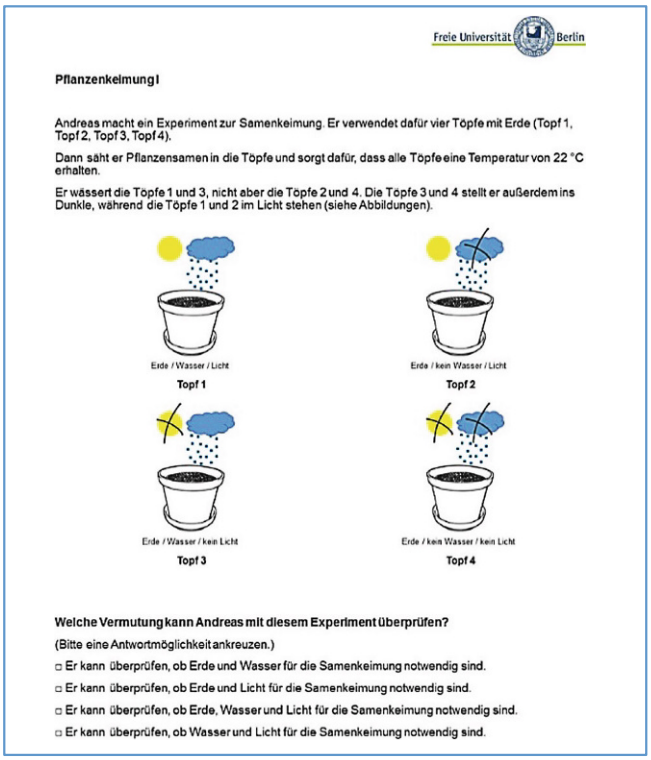


Abb. 1 Die kontextfreien Vorlagen zur Aufgabenentwicklung (*oben*) sowie die unterschiedlich komplexen Aufgaben zur Teilkompetenz *Suche im Hypothesenraum* im Kontext Samenkeimung (*unten*). In den kontextfreien Vorlagen werden drei UV unterschieden (A, B, C). Diese und die AV können entweder ausgeprägt/vorhanden (+) oder nicht ausgeprägt/nicht vorhanden (–) sein. Die Platzhalter [...] sind in den Aufgaben kontextspezifisch ausgeführt

Zur Prüfung von H3 (konzeptuelle Replikation) wurden Skalen zur schülerseitigen Erhebung der Bekanntheit, Interessantheit und (Alltags-) Relevanz („Kontext-Personen-Valenzen“) der sechs Aufgabenkontexte genutzt (jeweils drei Items, 4-stufige Ratingskala; Werner et al. 2014, 2015).

Testadministration

Die Testadministration wurde umgesetzt durch Lehramtsstudierende (z. B. Masterarbeitskandidatinnen) und erfolgte entsprechend der Vorgängerstudie im Rahmen des normalen Unterrichts und geleitet von einem Manual zur Testinstruk-

tion (Durchführungsobjektivität). Um den Arbeitsaufwand für die Probanden in angemessenen Grenzen zu halten (etwa 30 min), wurden 37 Testheftvarianten mit jeweils acht oder neun MC-Aufgaben nach einem *incomplete block design* entwickelt (Cochran und Cox 1957; Frey et al. 2009). Als Grundlage wurde ein *incomplete latin square design* für 37 Aufgaben gewählt (Cochran und Cox 1957, S. 532) und nach folgenden Kriterien adaptiert: 1) Löschen der Aufgabe 37 im Design, um es an die Aufgabenanzahl der Studie anzupassen, 2) parallele Aufgaben mit hoher und niedriger Komplexität nicht direkt aufeinanderfolgend, um eine Beeinflussung der Aufgabenbearbeitung zu vermeiden, 3) parallele Aufgaben der Teilkompetenzen *Suche im Hypothesenraum* und *Testen von Hypothesen* nicht direkt aufeinanderfolgend, um die Bearbeitung der Aufgaben zu *Suche im Hypothesenraum* nicht durch Vorgabe einer Hypothese zu beeinflussen, 4) Aufgaben mit gleichem Kontext nicht direkt aufeinanderfolgend, um die Aufmerksamkeit bei der Bearbeitung durch einen Wechsel der Kontexte zu erhöhen.

Jedem Testheft wurden die Skalen zur schülerseitigen Erhebung der Bekanntheit, Interessantheit und Relevanz (Werner et al. 2014, 2015) in Bezug auf drei der sechs Aufgabenkontexte vorangestellt, welche ebenfalls nach einem *incomplete block design* kombiniert wurden (6 Aufgabenkontexte, 10 Kombinationen, 3 Aufgabenkontexte pro Kombination; Cochran und Cox 1957, S. 471).

Stichprobe

Insgesamt haben $N = 708$ Schülerinnen und Schüler mit einem durchschnittlichen Alter von 14 Jahren freiwillig an der Studie teilgenommen (Gelegenheitsstichprobe, Jahrgangsstufen 8 und 9, Gymnasien und Sekundarschulen, 53 % weiblich). Jede Aufgabe wurde 164 bis 186 Mal bearbeitet ($M = 168$).

Im Sinne einer Replikationsstudie wurde eine mit der Vorgängerstudie vergleichbare Stichprobe gezogen (bei Krell und Vierarm 2016: Gelegenheitsstichprobe, Jahrgangsstufen 9 und 10, Gymnasien und Sekundarschulen, 51 % weiblich). Weil die Aufgaben für Schülerinnen und Schüler der zehnten Jahrgangsstufe relativ leicht zu lösen waren ($M = 0{,}65$; Krell und Vierarm 2016), wurden zur Vermeidung von Deckeneffekten Schülerinnen und Schüler der Jahrgangsstufen 8 und 9 befragt.

Datenauswertung

Für einen ersten Zugang zu den Daten wurde die Lösungswahrscheinlichkeit der Aufgaben nach Aufgabenmerkmal gruppiert und verglichen (deskriptiv). Zusammenhänge zwischen den erhobenen Kontext-Personen-Valenzen und der Lösungswahrscheinlichkeit wurden mittels Korrelationsanalysen geprüft. Nicht erfolgte Bearbeitungen eines Items (missings) wurden durchgängig als fehlende Werte behandelt.

Eine etablierte Vorgehensweise zur Analyse schwierigkeitserzeugender Aufgabenmerkmale besteht in einem zweischrittigen Verfahren: (1) Schätzung von Aufgabenschwierigkeitsparametern; (2) Regressionsanalyse mit den Schwierigkeitsparametern als abhängiger und den betrachteten Merkmalen als Prädiktorvariablen (z. B. Krell und Krüger 2011; Prenzel et al. 2002). Dieses Vorgehen wird allerdings als fehleranfällig kritisiert (Cho et al. 2013; Embretson und Daniel 2008; Wilson und De Boeck 2004), beispielsweise weil aufgrund der Schätzung auf Itemebene die Standardfehler tendenziell groß ausfallen (Embretson und Daniel 2008) und weil Schätzfehler der Schwierigkeitsparameter in der Regel nicht im Regressionsmodell berücksichtigt werden (Cho et al. 2013). Entsprechend zeigen Studien, dass im zweischrittigen Verfahren die Signifikanz von Prädiktoren (d. h. Aufgabenmerkmalen) unterschätzt werden kann (z. B. Embretson und Daniel 2008; Hartig et al. 2012). Wilson und De Boeck (2004) empfehlen daher die Anwendung erklärender IRT-Modelle. Ein solches Modell ist das Linear Logistische Test-Modell (LLTM; Fischer 1995), das als eine restriktivere und sparsamere Variante des Einparametrisch-Logistischen „Rasch“ Modells (1PLM) betrachtet werden kann (Mair und Hatzinger 2007). Im LLTM ist der Aufgabenschwierigkeitsparameter (β_i') als lineare Kombination von merkmalsbezogenen Schwierigkeitsparametern (α_k) konzipiert (Fischer 1995): $\beta_i' = \sum_{k=1}^{N}(\alpha_k \chi_{ik})$. Das LLTM stellt also eine Operationalisierung der Hypothese dar, dass sich die Aufgabenschwierigkeit additiv aus den jeweils spezifizierten Merkmalen ergibt und ist daher für die Analyse der Konstruktrepräsentation etabliert (z. B. Baghaei und Kubinger 2015; Embretson 1983; Embretson und Daniel 2008; Hartig und Frey 2012). Zur Prüfung der jeweils operationalisierten Hypothese muss die Modellgültigkeit des LLTM geprüft werden, wozu zwei Schritte vorgeschlagen werden (Fischer 1995): (1) Nachweis der Gültigkeit des 1PLM für die vorliegenden Daten; (2) Prüfung, inwieweit die im LLTM geschätzten Aufgabenschwierigkeitsparameter (β_i') mit den im 1PLM geschätzten Aufgabenschwierigkeitsparametern (β_i) übereinstimmen. Ergänzend kann ein Modellvergleich zwischen 1PLM und LLTM durchgeführt werden (Fischer 1995; Wilson und De Boeck 2004). Die Idee hinter diesem vergleichenden Vorgehen ist die Bewertung der relativen Passung des LLTM; also der Annahme einer additiven, durch Aufgabenmerkmale zu erklärenden Aufgabenschwierigkeit (LLTM), im Vergleich zur freien Schätzung von Aufgabenschwierigkeiten für jede Aufgabe (1PLM). Das (gültige) 1PLM wird somit als Referenz für die Prüfung des sparsameren LLTM betrachtet. Für die Parameterschätzung wurde in dieser Studie das R-Paket eRm genutzt (Mair und Hatzinger 2007).

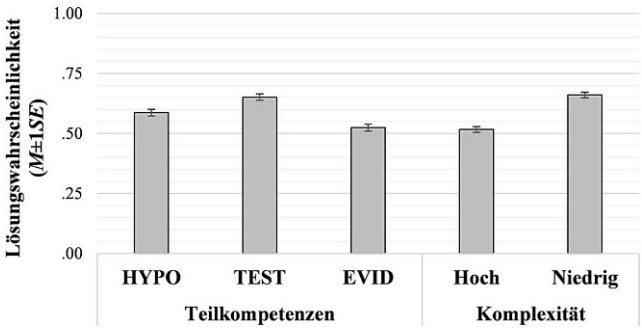


Abb. 2 Lösungswahrscheinlichkeit der MC-Aufgaben, aufgeteilt nach Teilkompetenzen (*links*) und Komplexität (*rechts*); HYPO: *Suche im Hypothesenraum*, TEST: *Testen von Hypothesen*, EVID: *Analyse von Evidenz*

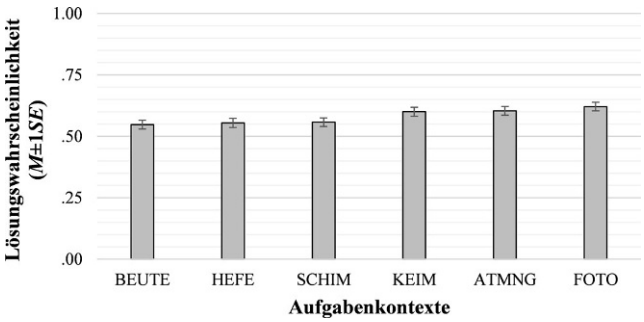


Abb. 3 Lösungswahrscheinlichkeit der MC-Aufgaben, aufgeteilt nach Kontexten

Ergebnisse

Deskriptive Analysen

Die mittlere Lösungswahrscheinlichkeit aller Aufgaben liegt bei $M = 0{,}586$ ($SE = 0{,}009$) und für die einzelnen Aufgaben im Bereich $0{,}288 \leq M \leq 0{,}825$. Die deskriptiven Analysen zeigen, dass die Aufgaben zu *Analyse von Evidenz* ($M = 0{,}524$, $SE = 0{,}014$) schwerer sind als die zu *Suche im Hypothesenraum* ($M = 0{,}587$, $SE = 0{,}014$) und *Testen von Hypothesen* ($M = 0{,}652$, $SE = 0{,}014$; Abb. 2). Außerdem sind die Aufgaben mit hoher Komplexität ($M = 0{,}517$, $SE = 0{,}012$) schwerer als die Aufgaben mit niedriger Komplexität ($M = 0{,}660$, $SE = 0{,}012$; Abb. 2).

Die mittlere Lösungswahrscheinlichkeit der Aufgaben für die Aufgabenkontexte liegen zwischen $M = 0{,}548$ ($SE = 0{,}012$; Kontext BEUTE) und $M = 0{,}621$ ($SE = 0{,}018$; Kontext FOTO; Abb. 3).

Gemäß H3 wurde bei einem Vergleich der Valenz-Werte in Abhängigkeit der Kontexte eine der Abb. 3 ähnliche Reihenfolge erwartet. Diese Erwartung kann nicht bestätigt werden, eine Tendenz ist mit Ausnahme des Aufgabenkontexts ATMNG bezüglich der Bekanntheit erkennbar (Abb. 4, *oben*). Insgesamt ergeben sich aber keine signifikanten Pearson-Korrelationen zwischen der mittleren Lösungswahrscheinlichkeit der Aufgabenkontexte und deren Valenz-Werten (Bekanntheit: $r_P = 0{,}130$, $p = 0{,}450$; Interessantheit: $r_P = 0{,}024$, $p = 0{,}891$; Relevanz: $r_P = 0{,}006$, $p = 0{,}974$; jeweils $N = 36$).

Die bislang dargestellten Befunde bieten einen ersten Zugang zu den Daten, erlauben aber keine Kontrolle der jeweils nicht betrachteten Aufgabenmerkmale. Daher werden die drei betrachteten Merkmale im Folgenden mit Hilfe des LLTM gemeinsam in einer Analyse als potenziell schwierigkeitserzeugend geprüft. Die Matrix der betrachteten Aufgabenmerkmale sind dem Anhang 2 zu entnehmen.

Analyse im LLTM (H1, H2, H3)

Nachweis der Gültigkeit des 1PLM: Zunächst wurde die Dimensionalität der Daten geprüft. Theoriebasiert wurden neben dem eindimensionalen (1D) 1PLM ein zwei- (nach der Komplexität), ein drei- (nach den Teilkompetenzen) und ein sechsdimensionales (nach Komplexität je Teilkompetenz) 1PLM in die Prüfung einbezogen. Auf Basis der Informationskriterien (vgl. Burnham und Anderson 2000) ergab sich keine eindeutige Präferenz für eines der Modelle (Tab. 3). Aus Sparsamkeitsgründen wurde daher das eindimensionale 1PLM umgesetzt (Burnham und Anderson 2000).

Die Modellprüfung des (eindimensionalen) 1PLM ergibt zufriedenstellende Werte. Auf Aufgabenebene liegen die MNSQ-Werte im Bereich $0{,}500 \leq \text{MNSQ} \leq 1{,}500$ (*productive for measurement*; Linacre 2002; Wright und Linacre 1994) und indizieren eine angemessene Passung des 1PLM ($0{,}599 \leq \text{MNSQ}_{\text{infit}} \leq 1{,}290$; $0{,}793 \leq \text{MNSQ}_{\text{outfit}} \leq 1{,}141$). Die *person separation reliability* beträgt 0,643. Der Likelihood-Ratio-Test („Andersen-Test“) fällt weder für die internen Teilungskriterien Median ($\text{LR}(35) = 39{,}701$; $p = 0{,}268$) und Mittelwert ($\text{LR}(35) = 48{,}943$; $p = 0{,}059$) noch für das externe Kriterium Jahrgangsstufe ($\text{LR}(35) = 39{,}701$; $p = 0{,}268$) signifikant aus; dies weist auf Itemhomogenität und damit ebenfalls auf die Gültigkeit des lPLM hin (Mair und Hatzinger 2007).

Zusammenhang zwischen Aufgabenschwierigkeitsparametern: MNSQ-Werte deuten auf eine angemessene Passung des LLTM hin ($0{,}711 \leq \text{MNSQ}_{\text{infit}} \leq 1{,}214$; $0{,}606 \leq \text{MNSQ}_{\text{outfit}} \leq 1{,}313$); die *person separation reliability* beträgt 0,626. Es besteht kein signifikanter Unterschied zwischen den im 1PLM ($M = 0{,}000$, $SD = 0{,}767$) und im LLTM ($M = 0{,}000$, $SD = 0{,}478$, $p = 0{,}997$) geschätzten Aufgabenschwierigkeiten (*t*-Test für verbundene Stichproben), ein hoher Pearson-Korrelationskoeffizient weist auf einen starken Zusammenhang beider Parameter hin ($r_P = 0{,}656$; $p < 0{,}001$; d. h. $R^2 = 0{,}430$).

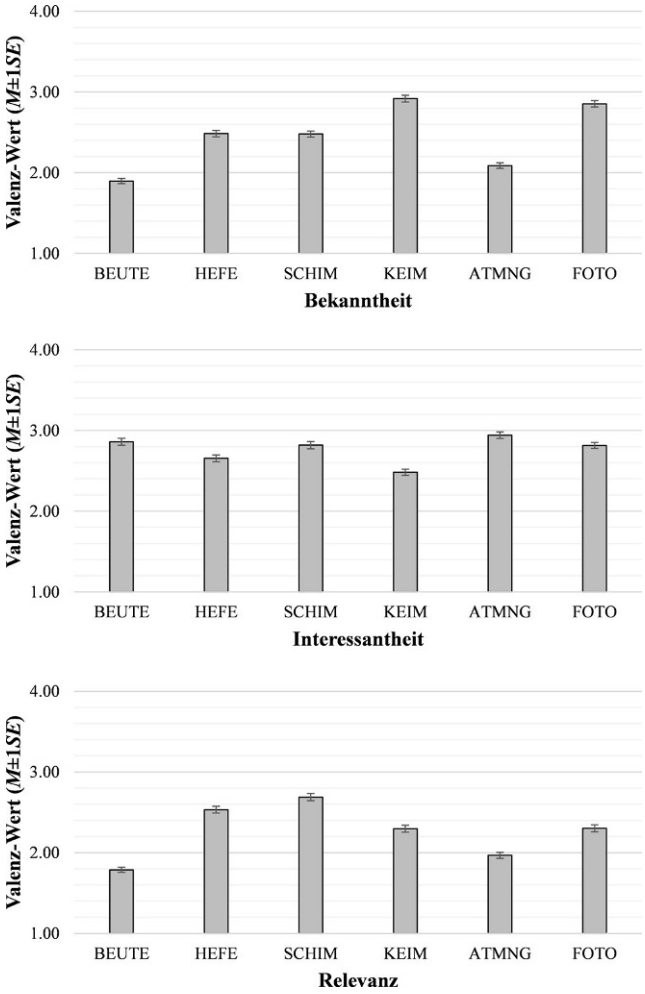


Abb. 4 Valenz-Werte der Kontexte

Modellvergleich: Der auf den Informationskriterien beruhende Modellvergleich zwischen 1PLM und LLTM deutet eine signifikant bessere Passung des 1PLM an (Tab. 4).

Für alle im LLTM geschätzten Schwierigkeitsparameter (α_k) enthält das 95 %-Konfidenzintervall nicht den Wert Null (Tab. 5), die Parameter können also auf dem Niveau $p < 0{,}05$ als signifikant von null verschieden betrachtet werden. Hierbei sind Aufgaben zu *Suche im Hypothesenraum* und *Testen von Hypothesen* signifikant leichter als Aufgaben zu *Analyse von Evidenz*[1] und die Aufgaben mit hoher Komplexität sind signifikant schwerer als diejenigen mit niedriger Komplexität. Die Kontext-Personen-Valenzen (Mediansplit) wirken sich nicht einheitlich aus; während Aufgaben mit Kontexten mit hoher Bekanntheit und Interessantheit erwartungsgemäß signifikant leichter sind als Aufgaben mit Kontexten mit niedriger Bekanntheit und Interessantheit, erhöht eine hohe Relevanz die Aufgabenschwierigkeit.

Nach der Modellgleichung des LLTM (siehe Abschnitt *Datenauswertung*) ergeben sich die im LLTM geschätzten Aufgabenschwierigkeiten durch Summation der α_k-Parameter – zum Beispiel für die Aufgabe zur Teilkompetenz *Suche im Hypothesenraum*, mit hoher Komplexität und im Kontext Samenkeimung (Abb. 1): $\beta_i' = (-0{,}324) + 0{,}720 + (-0{,}366) = 0{,}030$. (Die übrigen Aufgabenschwierigkeiten sind Onlinematerial 2 zu entnehmen.).

Zusammenfassung und Diskussion der Ergebnisse

In dieser Studie wurden die experimentellen Teilkompetenzen *Suche im Hypothesenraum, Testen von Hypothesen* und *Analyse von Evidenz*, die Aufgabenkomplexität sowie die Kontext-Personen-Valenzen Bekanntheit, Interessantheit und Relevanz der sechs Aufgabenkontexte BEUTE, HEFE, SCHIM, KEIM, ATMNG und FOTO als schwierigkeitserzeugende Aufgabenmerkmale eines MC-Tests zur Experimentierkompetenz im Biologieunterricht untersucht. Methodisch wurde die Studie als Replikationsstudie konzipiert. Durch den Einsatz des (erweiterten) Testinstruments von Krell und Vierarm (2016) sollten die Befunde zu den Teilkompetenzen (H1) und der Aufgabenkomplexität (H2) direkt repliziert werden. Mit Hilfe der Skalen zu den Kontext-Personen-Valenzen (Werner et al. 2014, 2015) sollten die von Krell und Vierarm (2016) gefundenen schwierigkeitserzeugenden Effekte der Aufgabenkontexte (H3) konzeptuell repliziert werden. Die Konzeption eines Facettendesigns erlaubt eine kontrollierte Prüfung der formulierten Hypothesen, was in bislang vorliegenden Arbeiten (z. B. Mannel et al. 2015; Vorholzer et al. 2016; Wellnitz und Mayer 2013) aufgrund der jeweils umgesetzten Testentwicklung oftmals nicht möglich war.

Anschließend an vorliegende Befunde (Glug 2009; Krell und Vierarm 2016) wurde erwartet, dass Aufgaben zur Teilkompetenz *Analyse von Evidenz* aufgrund der zur Aufgabenbearbeitung notwendigen anspruchsvolleren Fertigkeit signifikant schwerer sind als Aufgaben zu den Teilkompetenzen *Suche im Hypothesenraum* und *Testen von Hypothesen* (H1). Diese Hypothese kann sowohl basierend auf der deskriptiven Analyse der Rohdaten (Abb. 2) als auch im LLTM bestätigt werden (Tab. 5). Die Befunde von Krell und Vierarm (2016) konnten somit direkt repliziert werden. Insgesamt liegen allerdings uneinheitliche Ergebnisse über die relative Schwierigkeit der Teilkompetenzen *Suche im Hypothesenraum, Testen von Hypothesen* und *Analyse von Evidenz* vor (z. B. Glug 2009; Hammann et al. 2007; Krell und Vierarm 2016; Vorholzer et al. 2016). Ham-

[1] Im LLTM mit der Teilkompetenz *Testen von Hypothesen* als Vergleichsstandard ergeben sich ebenfalls signifikante Unterschiede in der Aufgabenschwierigkeit zwischen *Testen von Hypothesen* und *Suche im Hypothesenraum*; wobei *Suche im Hypothesenraum* signifikant schwerer ist ($\alpha_k = 0{,}388$, $SE\alpha_k = 0{,}076$, $CI_{95\%}$: 0,239/0,536).

Tab. 3 Modellvergleich der 1PLM basierend auf Informationskriterien

	*n*Par	Deviance	AIC	cAIC	BIC	aBIC
1PLM, 1D	37	7193	7267	7473	7436	7318
1PLM, 2D	39	7178	7256	7473	7434	7310
1PLM, 3D	42	7164	7248	7482	7440	7306
1PLM, 6D	57	7124	7238	7555	7498	7317

Die Modellprüfung wurde mit ACER ConQuest (Wu et al. 2007) umgesetzt, also unter Anwendung des *marginal Maximum Likelihood*-Schätzers, die Prüfung des LLTM (Tab. 4) unter Anwendung des *bedingten Maximum Likelihood*-Schätzers (vgl. Rost 2004)

Tab. 4 Modellvergleich basierend auf Informationskriterien sowie Likelihood-Differenz (LD)-Test

	*n*Par	Deviance	AIC	cAIC	BIC	aBIC	LD-Test
1PLM	35	4194	4264	4458	4423	4312	313.702(29); $p < 0{,}001$
LLTM	6	4507	4519	4553	4547	4528	

mann et al. (2007) schlagen alternativ zu H1 die zur Aufgabenbearbeitung notwendige Wissensart als ausschlaggebend für die relative Schwierigkeit der drei Teilkompetenzen vor. Demnach müsste *Testen von Hypothesen* (methodisches Wissen) eine von *Suche im Hypothesenraum* und *Analyse von Evidenz* (bereichsspezifisches Wissen) abweichende Schwierigkeit aufweisen. Diese Hypothese kann basierend auf den Daten nicht sicher falsifiziert werden. Zwar bestehen auch signifikante Unterschiede in der Aufgabenschwierigkeit zwischen *Suche im Hypothesenraum* und *Analyse von Evidenz*, die Effekte fallen aber eher klein aus (Abb. 2; Tab. 5). Im Gegensatz zu den Befunden von Hammann et al. (2007) waren Aufgaben zu *Testen von Hypothesen* in dieser Studie signifikant leichter als Aufgaben zu *Suche im Hypothesenraum* und *Analyse von Evidenz* (Tab. 5; FN 1).

Bezogen auf die Aufgabenkomplexität wurde in Einklang mit bestehenden Befunden (Kauertz et al. 2010; Krell und Vierarm 2016; Mannel et al. 2015) erwartet, dass Aufgaben mit hoher Komplexität aufgrund der höheren kognitiven Belastung während der Aufgabenbearbeitung signifikant schwerer sind, als Aufgaben mit geringer Komplexität (H2). Diese Erwartung kann sowohl basierend auf der deskriptiven Analyse der Rohdaten (Abb. 2) als auch im LLTM bestätigt (Tab. 5) und die Ergebnisse von Krell und Vierarm (2016) somit direkt repliziert werden. Die Aufgabenkomplexität wurde in dieser Studie durch die Variablenanzahl der betrachteten experimentellen Settings variiert. Während Mannel et al. (2015) drei Komplexitätsstufen umsetzen (1 Fakt: 1 AV; 1 Zusammenhang: 1 AV, 1 UV; 2 Zusammenhänge: 1 AV, 1 UV, 1 KV), wurden in dieser Studie Aufgaben mit niedriger (1 AV, 1 UV) und Aufgaben mit hoher (1 AV, 2 UV) Komplexität konstruiert. Die schwierigkeitserzeugende Wirkung der durch die Variablenanzahl variierten Aufgabenkomplexität ist empirisch gut belegt (z. B. Krell und Vierarm 2016; Mannel et al. 2015; diese Studie) und kann kognitionspsychologisch plausibel mit der steigenden kognitiven Belastung erklärt werden (Kauertz et al. 2010; Krell 2017). Durch eine Variation der Variablenanzahl können demnach systematisch parallele Aufgaben unterschiedlicher Schwierigkeit entwickelt werden (vgl. Hartig und Frey 2012; Leucht et al. 2012).

Schließlich wurde erwartet, dass die in Krell und Vierarm (2016) gefundenen schwierigkeitserzeugenden Effekte der Aufgabenkontexte mit Hilfe der Kontext-Personen-Valenzen Bekanntheit, Interessantheit und Relevanz (Werner et al. 2014, 2015) erklärt werden können (H3). Zunächst konnte die relative Schwierigkeit der drei auch in Krell und Vierarm (2016) eingesetzten Aufgabenkontexte HEFE, SCHIM und KEIM repliziert werden (Abb. 3). Es wurde erwartet, dass Kontexte mit hoher Bekanntheit, Interessantheit und Relevanz aufgrund der damit verbundenen Fokussierung und Routine bei der Aufgabenbearbeitung signifikant leichter sind als Aufgaben mit Kontexten mit niedriger Bekanntheit, Interessantheit und Relevanz (Roesler et al. 2014, 2016; Schiefele et al. 1993; van Vorst et al. 2015; Werner et al. 2014, 2015). Basierend auf den Rohdaten kann diese Hypothese nicht bestätigt werden. Im LLTM kann H3 nur bezüglich der Bekanntheit und Interessantheit bestätigt werden, wobei der Effekt für alle drei Valenzen eher klein ausfällt (Tab. 5). Für die Interessantheit und Relevanz liegt das Vertrauensintervall für den α_k-Parameter außerdem sehr dicht bei null (Tab. 5), was einen nur knapp signifikanten Effekt anzeigt. Damit erweitert diese Studie die Befundlage zur schwierigkeitserzeugenden Wirkung von Kontext-Personen-Valenzen (Kompetenzbereiche Fachwissen, Bewertung; Roesler et al. 2016) um den Kompetenzbereich Erkenntnisgewinnung. Der negative Einfluss der Bekanntheit von Aufgabenkontexten auf die Aufgabenschwierigkeit kann damit erklärt werden, dass bei bekannten Kontexten weniger kognitive Ressourcen für das Erschließen der Oberflächenmerkmale einer Aufgabe benötigt werden (Kalyuga und Renkl 2010). Grundsätzlich kann sich dieser förderliche Effekt der Bekanntheit (bzw.

Tab. 5 Parameter im LLTM

		α_k	$SE_{\alpha k}$	$CI_{95\%}$	
Kompetenz	*Suche im Hypothesenraum* (1 = ja)	–0,324	0,076	–0,473	–0,175
	Testen von Hypothesen (1 = ja)	–0,712	0,077	–0,862	–0,561
Komplexität	Hoch (1 = zwei UV)	0,720	0,062	0,598	0,842
Valenzen	Bekanntheit (1 = über median)	–0,366	0,133	–0,627	–0,105
	Interessantheit (1 = über median)	–0,199	0,094	–0,382	–0,015
	Relevanz (1 = über median)	0,239	0,110	0,024	0,454

des Vorwissens) auch umkehren, zum Beispiel wenn fachlich nicht belastbare prä-instruktionale Konzepte vorliegen (Krell und Vierarm 2016) oder eigene und die in Aufgaben angebotenen Schemata parallel verarbeitet werden müssen (*expertise reversal effect*; Kalyuga und Renkl 2010). Eine belastbare Erklärung für den erwartungswidrigen Befund zur Relevanz der Aufgabenkontexte erlauben die vorliegenden Daten nicht. Van Vorst et al. (2015, S. 34) heben hervor, dass die Relevanz „von zentraler Bedeutung", aber gleichzeitig das „am schwierigsten handhabbare und am wenigsten definierte" Kontextmerkmal ist. Die Autoren stellen in Frage, inwiefern Relevanz als eigenständiges Kontextmerkmal oder vielmehr als abhängige Variable betrachtet werden sollte, die durch andere Kontextmerkmale bestimmt werden kann (van Vorst et al. 2015). Es sollte somit geprüft werden, inwiefern die Relevanz weiterhin neben der Bekanntheit und Interessantheit als Kontext-Personen-Valenz zur Erklärung von schwierigkeitserzeugenden Effekten von Aufgabenkontexten herangezogen werden kann (Werner et al. 2014, 2015). Darüber hinaus macht es die (postulierte) Abhängigkeit der Relevanz von anderen Kontextmerkmalen schwierig, die erwartungswidrigen Befunde dieser Studie zu erklären. Hierzu müssten in anschließenden Studien weitere Kontextmerkmale (z. B. Authentizität; van Vorst et al. 2015) erhoben und zur Erklärung herangezogen werden.

Zur Analyse schwierigkeitserzeugender Aufgabenmerkmale werden oftmals zweischrittige Verfahren umgesetzt, welche jedoch als fehleranfällig kritisiert werden (z. B. Cho et al. 2013; Embretson und Daniel 2008). In dieser Studie konnte ein signifikanter Effekt der Kontext-Personen-Valenzen im LLTM nachgewiesen werden. Diese Ergebnisse werfen die Frage auf, inwiefern Befunde zur Kontextunabhängigkeit von Testergebnissen (z. B. Vorholzer et al. 2016) belastbar sind, und legen nahe, in künftigen Studien zur Analyse schwierigkeitserzeugender Aufgabenmerkmale erklärende IRT-Modelle zu verwenden (z. B. LLTM; Wilson und De Boeck 2004).

Zusammenfassend konnten die Befunde von Krell und Vierarm (2016) zur schwierigkeitserzeugenden Wirkung der experimentellen Teilkompetenzen (H1: *Analyse von Evidenz* schwerer als *Suche im Hypothesenraum* und *Testen von Hypothesen*) und der Aufgabenkomplexität (H2: Aufgaben mit hoher Komplexität schwerer als Aufgaben mit geringer Komplexität) direkt repliziert werden. Die Befunde zur schwierigkeitserzeugenden Wirkung der Aufgabenkontexte fielen teilweise erwartungswidrig aus (H3). Da sich allerdings signifikante Effekte der Kontext-Personen-Valenzen zeigten, ist die Interpretierbarkeit der Testergebnisse als Indikator für die Kompetenzausprägung der Schülerinnen und Schüler fraglich (Kane 2013; Shavelson 2013). Wird Kompetenz als (rein) kognitive Leistungsdisposition verstanden (Klieme et al. 2008), stellen die gefundenen Effekte der Kontext-Personen-Valenzen illegitime Quellen von Aufgabenschwierigkeit dar, die für eine gültige Interpretation der Testergebnisse als Indikator für Experimentierkompetenz kontrolliert werden müssen. Insbesondere aufgrund der sehr großen Anzahl möglicher Aufgabenkontexte (*large universe of possible tasks*; Shavelson 2013) sollte daher weitergehend untersucht werden, inwiefern in MC-Tests zur Experimentierkompetenz von Schülerinnen und Schülern unterschiedliche Kontexte berücksichtigt werden müssen, um ein vorliegendes Testergebnis valide interpretieren zu können. In Abhängigkeit der Definition des Kompetenzbegriffs (Klieme et al. 2008; Weinert 2001) können schwierigkeitserzeugende Effekte motivationaler Variablen auch als legitime Quellen von Aufgabenschwierigkeit interpretiert werden (vgl. Roesler et al. 2014). Sofern motivationale, volitionale und soziale Bereitschaften als Teil des Kompetenzbegriffs verstanden werden (Weinert 2001), kann die Erklärung der in Krell und Vierarm (2016) gefundenen schwierigkeitserzeugenden Wirkung von Aufgabenkontexten durch die hier betrachteten Kontext-Personen-Valenzen (Werner et al. 2014, 2015) als Ansatz betrachtet werden, die Aufgabenkontexte als legitime Quellen von Aufgabenschwierigkeit

zu begründen. Hieran zeigt sich, dass die Entscheidung darüber „what constitutes construct-irrelevant variance is a tricky and contentious issue“ (Messick 1995, S. 743). Die Erklärung der schwierigkeitserzeugenden Wirkung von Aufgabenkontexten (Oberflächenmerkmale) durch Kontext-Personen-Valenzen trägt vor diesem Hintergrund dazu bei, nachgewiesene Effekte besser zu verstehen.

Methodisch können sowohl die in dieser Studie vorgenommene Operationalisierung von Experimentierkompetenz als auch das zur Datenanalyse eingesetzte LLTM kritisch diskutiert werden. Köchy (2006) bezeichnet die oben skizzierte Definition von Experiment als „Idealkonzept“ und Höttecke und Rieß (2015) zeigen, dass das Testen von Hypothesen nur eine von vielen experimentellen Strategien ist (neben z. B. explorativem Experimentieren). Die Testergebnisse des in dieser Studie eingesetzten MC-Tests können entsprechend maximal als Indikator für die Experimentierkompetenz im Sinne des SDDS-Modells (Tab. 1) interpretiert werden. Inwiefern das Generalisieren auf ein über das SDDS-Modell hinausgehendes Verständnis von Experimentierkompetenz zulässig ist, muss empirisch geprüft werden. Neben der Operationalisierung des Experimentbegriffs kann die Form des Testinstruments das Testergebnis beeinflussen und damit dessen Generalisierbarkeit in Frage stellen. Schreiber et al. (2016) zeigen, dass die Ergebnisse von produktorientierten und prozessorientierten Erhebungen teilweise (in Abhängigkeit der experimentellen Teilkompetenzen) nur mäßig korrelieren. Die Autoren schreiben der prozessorientieren Erhebung aufgrund des Prozesscharakters des Experimentierens die größere kriteriale Validität zu.

Das LLTM ist etabliert für die Analyse schwierigkeitserzeugender Aufgabenmerkmale im Sinne einer Prüfung der Konstruktrepräsentation (z. B. Baghaei und Kubinger 2015; Embretson 1983; Embretson und Daniel 2008; Hartig und Frey 2012). Trotzdem kann die mit dem LLTM getroffene Annahme eines additiven Zusammenhangs der Schwierigkeit einzelner Merkmale kritisiert werden (Hartig et al. 2012). Des Weiteren wurden in dieser Studie nur Haupteffekte im LLTM geprüft. Es ist allerdings durchaus vorstellbar, dass sich die Teilkompetenzen *Suche im Hypothesenraum* und *Analyse von Evidenz* (erfordern nach Hammann et al. (2007) bereichsspezifisches Wissen) in Abhängigkeit der Bekanntheit des jeweiligen Kontexts schwierigkeitserzeugend auswirken. In weiteren Studien sollten daher, basierend auf einem Test mit einer größeren Anzahl an Aufgaben, solche Interaktionseffekte zwischen den hier betrachteten Aufgabenmerkmalen geprüft werden. Die im LLTM geschätzten Aufgabenschwierigkeiten erklären etwa 43 % der Varianz der im 1PLM geschätzten Schwierigkeiten, was deutlich über dem Grenzwert eines großen Effekts liegt ($R^2 \geq 0{,}26$; Hartig und Frey 2012) und dafür spricht, dass die Annahme eines additiven Zusammenhangs im vorliegenden Fall plausibel ist. Diese sehr gute Varianzaufklärung kann mit der systematischen Aufgabenentwicklung erklärt werden, aufgrund derer sich die Aufgaben primär in den betrachteten Merkmalen unterscheiden. Die Befunde zeigen also, dass die hier betrachteten Merkmale zusammengenommen knapp die Hälfte der Varianz der im 1PLM geschätzten Aufgabenschwierigkeiten erklären können. Umgekehrt bedeutet dies, dass 57 % der Varianz der im 1PLM geschätzten Aufgabenschwierigkeiten nicht mit dem hier spezifizierten LLTM erklärt werden können. Dies deutet darauf hin, dass weitere Faktoren die Schwierigkeit der hier betrachteten Aufgaben bestimmen (z. B. eher formale Aufgabenmerkmale wie die Textlänge; Prenzel et al. 2002; Stiller et al. 2016).

Trotz der guten Erklärung der Aufgabenschwierigkeiten im LLTM ergibt sich eine signifikant bessere Modellpassung für das 1PLM (Tab. 4). Die im Vergleich schlechte Modellpassung des LLTM ist ein häufig auftretender Befund (z. B. Baghaei und Kubinger 2015; Wilson und De Boeck 2004), der mit der strengen Annahme einer vollständigen Erklärung der Aufgabenschwierigkeiten durch die spezifizierten Merkmale erklärt wird (Fischer 1995). Der auf Informationskriterien beruhende Vergleich zweier Modelle erlaubt keine Aussage über die absolute Passung der betrachteten Modelle (Burnham und Anderson 2000). Da eine im Vergleich schlechtere Modellpassung also nicht notwendigerweise eine absolut schlechte Modellpassung bedeutet, wird eine Prüfung der im LLTM geschätzten Schwierigkeitsparameter im Sinne einer prognostischen Validierung durch Replikationsstudien vorgeschlagen (Fischer 1995; Hartig und Frey 2012). Dies wurde in der vorliegenden Studie umgesetzt.

Zusammenfassend belegt die vorliegende Studie die Teilkompetenzen *Suche im Hypothesenraum, Testen von Hypothesen, Analyse von Evidenz*, die Aufgabenkomplexität und die Kontext-Personen-Valenzen Bekanntheit, Interessantheit, Relevanz als schwierigkeitserzeugende Aufgabenmerkmale bei einem MC-Test zur Experimentierkompetenz im Biologieunterricht. Diese Merkmale sollten bei zukünftigen Testkonzeptionen berücksichtigt und mit Hilfe erklärender IRT-Modelle (z. B. LLTM; Fischer 1995) geprüft werden, um die valide Interpretation von Testergebnissen als Indikator für die Experimentierkompetenz von Schülerinnen und Schülern zu gewährleisten (Shavelson 2013).

Literatur

Baghaei, P., & Kubinger, K. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation*, *20*, 1–11.

Burnham, K., & Anderson, D. (2000). *Model selection and inference: a practical information-theoretic approach*. New York: Springer.

Cho, S.-J., Gilbert, J., & Goodwin, A. (2013). Explanatory multidimensional multilevel random item response model. *Psychometrika*, *78*, 830–855.

Clough, E., & Driver, R. (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. *Science Education*, *70*, 473–496.
Cochran, W., & Cox, G. (1957). *Experimental designs*. New York: Wiley.
Dasgupta, A., Anderson, T., & Pelaez, N. (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE – Life Sciences Education*, *13*, 265–284.
Embretson, S. (1983). Construct validity. *Psychological Bulletin*, *93*, 179–197.
Embretson, S., & Daniel, R. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly*, *50*, 328–344.
Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.
Fischer, G. (1995). The linear logistic test model. In G. Fischer & I. Molenaar (Hrsg.), *Rasch models* (S. 131–155). New York: Springer.
Fleischer, J., Koeppen, K., Kenk, M., Klieme, E., & Leutner, D. (2013). Kompetenzmodellierung. *Zeitschrift für Erziehungswissenschaft*, *16*(S1), 5–22.
Frey, A. (2006). Strukturierung und Methoden zur Erfassung von Kompetenz. *Bildung und Erziehung*, *59*, 125–166.
Frey, A., Hartig, J., & Rupp, A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: theory and practice. *Educational Measurement: Issues and Practice*, *28*, 39–53.
Glug, I. (2009). *Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung* (Doctoral dissertation). Christian-Albrechts-Universität, Kiel. Retrieved from http://eldiss.uni-kiel.de/macau/receive/dissertation_diss_00003649
Gut, C. (2012). *Modellierung und Messung experimenteller Kompetenz*. Berlin: Logos.
Hammann, M., Phan, T., & Bayrhuber, H. (2007). Experimentieren als Problemlösen. *Zeitschrift für Erziehungswissenschaft*, *10*(S8), 33–49.
Hartig, J. (2008). Kompetenzen als Ergebnisse von Bildungsprozessen. In N. Jude, J. Hartig & E. Klieme (Hrsg.), *Kompetenzerfassung in pädagogischen Handlungsfeldern* (S. 15–25). Bonn & Berlin: BMBF.
Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau*, *63*, 43–49.
Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, *72*, 665–686.
Höttecke, D., & Rieß, F. (2015). Naturwissenschaftliches Experimentieren im Lichte der jüngeren Wissenschaftsforschung. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*, 127–139.
Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701.
Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications. *Instructional Science*, *38*, 209–215.
Kambach, M., & zu Belzen, U. A. (2016). Wie experimentieren Lehramtsstudierende der Biologie? In M. Hammann & U. Gebhard (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik. Band 7* (S. 229–246). Innsbruck: Studienverlag.
Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.
Kauertz, A., Fischer, H., Mayer, J., Sumfleth, E., & Walpulski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 135–153.
Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 3–22). Göttingen: Hogrefe.
KMK (2005). Sekretariat der Ständigen Konferenz der Kultusminister der Länder. In der BRD] (Hrsg.), *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss*. München & Neuwied: Wolters Kluwer.
Köchy, K. (2006). Lebewesen im Labor. *Philosophia naturalis*, *43*, 74–110.
Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Education*, *4*, 1280256.
Krell, M., & Krüger, D. (2011). Forced Choice-Aufgaben zur Evaluation von Modellkompetenz im Biologieunterricht: Empirische Überprüfung konstrukt- und merkmalsbezogener Teilkompetenzen. *Erkenntnisweg Biologiedidaktik*, *10*, 53–68.
Krell, M., & Vierarm, A. (2016). Analyse schwierigkeitserzeugender Aufgabenmerkmale bei einem Multiple-Choice-Test zum Experimentieren. In M. Hammann & U. Gebhard (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik*. Band 7 (S. 283–298). Innsbruck: Studienverlag.
Krell, M., Reinisch, B., & Krüger, D. (2015). Analyzing students' understanding of models and modeling referring to the disciplines biology, chemistry, and physics. *Research in Science Education*, *45*, 367–393.
Lamal, P. (1991). On the importance of replication. In J. Neuliep (Hrsg.), *Replication research in the social sciences* (S. 31–35). Newbury Park: SAGE.
Leucht, M., Harsch, C., Pant, H., & Köller, O. (2012). Steuerung zukünftiger Aufgabenentwicklung durch Vorhersage der Schwierigkeiten eines Tests für die erste Fremdsprache Englisch durch *Dutch Grid* Merkmale. *Diagnostica*, *58*, 31–44.
Linacre, J. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.
Mahner, M., & Bunge, M. (1997). *Foundations of biophilosophy*. Berlin: Springer.
Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling. *Journal of Statistical Software*, *20*, 1–20.
Mannel, S., Walpuski, M., & Sumfleth, E. (2015). Erkenntnisgewinnung: Schülerkompetenzen zu Beginn der Jahrgangsstufe 5 im naturwissenschaftlichen Anfangsunterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*, 99–110.
Mathesius, S., Upmeier zu Belzen, A., & Krüger, D. (2014). Kompetenzen von Biologiestudierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung. *Erkenntnisweg Biologiedidaktik*, *13*, 73–88.
Meier, M., & Wellnitz, N. (2013). Beobachten, Vergleichen und Experimentieren mit Wasserflöhen. *Praxis der Naturwissenschaften. Biologie in der Schule*, *62*, 4–9.
Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*, 741–749.
Nehm, R. H., & Ridgway, J. (2011). What do experts and novices „see" in evolutionary problems? *Evolution: Education and Outreach*, *4*, 666–679.
Neuliep, J., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality*, *8*, 21–29.
Opfer, J., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design. *Journal of Research in Science Teaching*, *49*, 744–777. https://doi.org/10.1002/tea.21028.
Prenzel, M., Häußler, P., Rost, J., & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft*, *30*, 120–135.
Roesler, M., Wellnitz, N., & Mayer, J. (2014). Motivationale Einflüsse auf schriftliche Testleistungen im Fach Biologie. *Erkenntnisweg Biologiedidaktik*, *13*, 179–195.
Roesler, M., Wellnitz, N., & Mayer, J. (2016). Die Rolle affektiver Variablen bei der Bearbeitung kontextualisierter Testaufgaben. In M.

Hammann & U. Gebhard (Hrsg.), *Lehr- und Lernforschung in der Biologiedidaktik* (S. 265–281). Innsbruck: Studienverlag.
Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.
Schecker, H., Neumann, K., Theyßen, H., Eickhorst, B., & Dickmann, M. (2016). Stufen experimenteller Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, *22*, 197–213.
Schiefele, U., Krapp, A., & Schreyer, I. (1993). Metaanalyse des Zusammenhangs von Interesse und schulischer Leistung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *10*, 120–148.
Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100. https://doi.org/10.1037/a0015108.
Schnotz, W., & Baadte, C. (2015). Surface and deep structures in graphics comprehension. *Memory & Cognition*, *43*, 605–618.
Schreiber, N., Theyßen, H., & Schecker, H. (2016). Process-oriented and product-oriented assessment of experimental skills. In N. Papadouris, A. Hadjigeorgiou & C. Constantinou (Hrsg.), *Insights from research in science teaching and learning. Contributions from science education research 2* (S. 29–43). Cham: Springer.
Shavelson, R. (2013). On an approach to testing and modeling competence. *Educational Psychologist*, *48*, 73–86.
Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning. *Assessment & Evaluation in Higher Education*, *41*, 721–732.
Vorholzer, A., von Aufschnaiter, C., & Kirschner, S. (2016). Entwicklung und Erprobung eines Tests zur Erfassung des Verständnisses experimenteller Denk- und Arbeitsweisen. *Zeitschrift für Didaktik der Naturwissenschaften*, *22*, 25–41.
van Vorst, H., Dorschu, A., Fechner, S., Kauertz, A., Krabbe, H., & Sumfleth, E. (2015). Charakterisierung und Strukturierung von Kontexten im naturwissenschaftlichen Unterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*, 29–39.
Weinert, F. (2001). Vergleichende Leistungsmessungen an Schulen – eine umstrittene Selbstverständlichkeit. In F. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim & Basel: Beltz.
Wellnitz, N., & Mayer, J. (2013). Erkenntnismethoden in der Biologie. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 315–345.
Werner, M., Schwanewedel, J., & Mayer, J. (2014). Does the context make a difference? In C. Constantinou, N. Papadouris & A. Hadjigeorgiou (Hrsg.), *E-Book proceedings of the ESERA 2013 conference* (S. 81–89). Nicosia: European Science Education Research Association. Retrieved from http://www.esera.org/media/eBook_2013/Strand%208/ESERA_eBook_Part_8.pdf.
Werner, M., Schwanewedel, J., & Mayer, J. (2015). Bewertungskompetenz und der Einfluss von Kontexten und Kontext-Personen-Valenzen. In U. Gebhard, M. Hammann & B. Knälmann (Hrsg.), *Bildung durch Biologieunterricht* (S. 58–59). Hamburg: Universität Hamburg.
Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory item response models* (S. 43–74). New York: Springer.
Wright, B., & Linacre, J. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.
Wu, M. L., Adams, R., & Wilson, M. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Camberwell: ACER Press.
Yong, E. (2012). Bad copy: In the wake of high-profile controversies, psychologists are facing up to problems with replication. *Nature*, *485*, 298–300.