Frank Rösch
Pädagogische Hochschule Ludwigsburg

Externe Validität ökologischer Experimente einschätzen können. Befunde und Überlegungen zur Messung und Förderung

Knowing How to Judge the External Validity of Ecological Experiments. Empirical Findings and Considerations for Measuring and Promoting

Lebende Systeme angemessen zu verstehen, setzt Wissen über Ursache-Wirkungsbeziehungen zwischen deren Elementen und mit deren Umgebung voraus. Entsprechende Erkenntnisse zu gewinnen, erfordert Kompetenzen für die eigenständige Planung, Durchführung und Auswertung von Experimenten sowie deren kritische Beurteilung. In der Ökologie sind angesichts der Systemkomplexität in besonderem Maß auch Überlegungen zur Tragweite experimenteller Befunde nötig. Zur Erfassung eines niederschwelligen Verständnisses für zwei Aspekte der externen Validität wurde ein schriftliches Testinstrument entwickelt und validiert sowie die dimensionale Struktur des Konstrukts analysiert. Lernende der 6. (n = 79) und der 9. Klassenstufe (n = 50) (Gymnasien) weisen eine ähnlich niedrige Performanz auf. In Klassenstufe 6 (N = 472, Realschulen) erweist sich das operationalisierte Konstrukt als eindimensional; in einer Interventionsstudie (n = 320) wurden dort keine Effekte des untersuchten Treatments festgestellt. Die Befunde legen Maßnahmen zur Steigerung der Sensitivität, Validität und Reliabilität des vorgestellten Messinstruments nahe. Überdies erscheint es lohnenswert zu klären, ob die Kombination aus ökologischen Realexperimenten, dem Training systemischen Denkens und PC-Simulationen eine spätere Förderung der anspruchsvollen Fähigkeit begünstigen könnte.

Schlüsselwörter: Experimentieren, Kompetenzen, Validität, Ökologie, Systemisches Denken.

In order to understand a biological system, knowledge of cause-and-effect relationships between its elements and with their surroundings is necessary. Gaining this knowledge requires competencies in planning and executing experiments, data analysis, and critical assessment of their design. The complexity of ecological systems places a particularly high demand on the ability to determine the validity of experimental findings. We have developed and validated a written instrument that probes the understanding of two basic aspects of external validity and analyzed the dimensional structure of the construct. We could observe that students (grammar school) in grades 6 (n = 79) and 9 (n = 50) show similar, low, performance on the test. In grade 6 (N = 472; average-performing students), we found that the construct is one-dimensional; an intervention study (n = 320) showed no training effect. Obviously, the sensitivity, the validity, and the reliability of the reported test instrument have to be improved. Besides, the question is raised whether promotion of this cognitively challenging skill could be fostered at a later development stage by a combination of real experiments, training in systems thinking, and computer simulations.

Keywords: experimentation, competencies, validity, ecology, systems thinking.

doi:10.2390/zdb-v19-i1-315

1 Einleitung

Anthropogene Eingriffe verändern auf kurze oder lange Sicht Elemente, Strukturen, Funktionsweisen und Eigenschaften natürlicher Systeme (Smith & Smith, 2009, S. 554, 756 ff., 778 ff.). Da systemisches Denken das Verständnis von Ökosystemen und den verantwortungsvollen, nachhaltigen Umgang damit begünstigen kann (Bräutigam, 2014, S. 24; Riess & Mischo, 2010), stellt es ein wichtiges Bildungsziel von Biologieunterricht dar (KMK, 2005, S. 8 ff.). Dazu gehört u. a. Wissen über Strukturen von Systemen, Interaktionen zwischen deren Elementen sowie Ursache-Wirkungsbeziehungen mit Faktoren aus der Systemumgebung (Riess & Mischo, 2010). Als Instrument kausalanalytischer Erkenntnisgewinnung (Schulz, Wirtz & Starauschek, 2012) kommt dem Experiment somit auch in der Ökologie (Smith & Smith, 2009, S. 17 ff.) und im Hinblick auf ökologische Grundbildung und Umweltbildung große Bedeutung zu (McBride, Brewer, Berkowitz & Borrie, 2013).

Die Förderung von Kompetenzen zur eigenständigen Planung, Durchführung und Auswertung von Experimenten ist in den Bildungsstandards aller naturwissenschaftlichen Fächer(verbünde) verankert (KMK, 2005, S. 10, 14): Lernende sollen dazu befähigt werden, sowohl aussagekräftige Experimente selbst zu planen und durchzuführen, als auch Sicherheit, Belastbarkeit und Grenzen vorgegebener Untersuchungen einzuschätzen (Arnold, Kremer & Mayer, 2013). Die Fähigkeiten, externe Validität zu beurteilen bzw. zu beachten, wurden in der naturwissenschaftsdidaktischen Forschung bislang wenig berücksichtigt. Angesichts der fehlenden Präzisierung input- und prozessorientierter Standards in aktuellen Bildungsplänen ergeben sich für die biologiedidaktische Forschung wichtige Fragen: (a) Welche Teilfähigkeiten umfassen Standards wie „erörtern Tragweite und Grenzen“ und „beurteilen die Aussagekraft“ von Experimenten (KMK, 2005, S. 14)? (b) Wie lassen sich diese Kompetenzen valide operationalisieren sowie adressatengerecht, ressourcenadäquat und reliabel zur Diagnostik bzw. Evaluation von Unterricht erfassen? (c) In welcher Altersstufe ist eine systematische Förderung möglich bzw. effizient? (d) Welche Domäne eignet sich als Lernkontext besonders? (e) Welche Lernumgebungen begünstigen den Kompetenzerwerb?

Im Mittelpunkt dieses Beitrags stehen Teilstudien eines Projekts, das in der biologiedidaktischen Unterrichtsforschung angesiedelt ist. Mit der Entwicklung und Erprobung eines Testinstruments für Klassenstufe 6 zur Messung eines niederschwelligen Verständnisses für Aspekte externer Validität wurden dabei neue Wege beschritten. Darüber hinaus wird von Erfahrungen mit dieser Operationalisierung bei der Evaluation eines kontextbasierten sowie problem- und kompetenzorientierten Trainingskonzepts in der komplexen Domäne Ökologie berichtet. Theoretisch-konzeptionelle Überlegungen widmen sich abschließend der Frage, inwiefern systemisches Denken und computerbasierte Simulationen den Kompetenzaufbau in höheren Klassen unterstützen könnten.

doi:10.2390/zdb-v19-i1-315

2 Theoretischer Hintergrund und Stand der Forschung

2.1 Aussagekraft von Experimenten

Echte Experimente lassen sich methodologisch und epistemologisch von Versuchen und anderen s. g. experimentellen Arbeitsformen abgrenzen (Schulz et al., 2012). So erfordert die Prüfung von Hypothesen zu Ursache-Wirkungsbeziehungen die planmäßige, systematische Manipulation von Naturvorgängen unter Beachtung bestimmter Regeln: Eine als Ursache in Frage kommende Systemgröße (unabhängige Variable, Testgröße) wird von allen anderen Faktoren isoliert und als einzige variiert. Jene müssen für einen aussagekräftigen Vergleich in Kontrollansätzen bestmöglich kontrolliert werden (Bortz & Döring, 2006, S. 57, 526 ff.). Je eindeutiger die Wirkung auf eine Effektgröße (abhängige Variable) auf eine bestimmte Ursache zurückgeführt werden kann, desto höher ist die *interne Validität*. Diese ist für die Generalisierbarkeit der in einem Experiment gewonnenen Erkenntnisse zwar notwendig, jedoch nicht hinreichend. Die Übertragbarkeit experimenteller Befunde ergibt sich vielmehr aus deren *externen Validität*. Ob von einem Experiment auf andere verallgemeinert werden kann, hängt von diversen Aspekten ab (ebd., S. 504): von der Repräsentativität der untersuchten Stichprobe für die Population, dem Untersuchungszeitpunkt und der Beobachtungsdauer, der Ähnlichkeit zwischen spezifischen situationalen und örtlichen Umständen mit natürlichen Settings (man spricht hier von *ökologischer Validität*), den Untersuchungsmethoden und ihren Auswirkungen auf die Forschungsobjekte sowie bei reduktionistischem Vorgehen von den Eigenschaften und der Komplexität des fokussierten (Sub-)Systems mit Blick auf das interessierende Originalsystem.

2.2 Experimentieren als naturwissenschaftlicher Problemlöseprozess

Eigenständiges Experimentieren stellt aus kognitionspsychologischer Sicht einen komplexen Problemlöseprozess dar: Aus einem naturwissenschaftlichen Phänomen Fragen abzuleiten, auf der Basis kognitiver Modelle begründete Hypothesen zu formulieren, ein geeignetes unkonfundiertes Experiment zu planen, es durchzuführen und erhobene Daten angemessen zu interpretieren, erfordert zahlreiche Kenntnisse, Fertigkeiten und mentale Leistungen wie etwa metakognitive Fähigkeiten zur Selbstregulation (Eckhardt, 2010, S. 22) sowie wissenschaftliches Denken (Arnold et al., 2013). Dazu gehören auch deklaratives Methoden- und prozedurales Wissen (Mayer, Grube & Möller, 2008). Kenntnisse und Umsetzungsvermögen zur Steigerung von Reliabilität und Validität sind z. B. bei Überlegungen zu Einfluss- und Störgrößen, speziellen Systemeigenschaften und multiplen oder interagierend wirkenden Ursachen (Keselman, 2003) von Bedeutung: bei der Planung, kritisch reflektierten Durchführung sowie Diskussion und Einschätzung der Übertragbarkeit von Befunden. Verständnis für externe Validität stellt somit eine anspruchsvolle Komponente experimenteller Problemlösefähigkeit dar und tangiert alle Dimensionen des Kompetenzbereichs *Erkenntnisgewinnung*: „Arbeitstechniken“ (z. B. Messgenauigkeit, -wiederholungen), „Wissenschaftliche Erkenntnismethoden“ (z. B. experimentelle Denkweise und Strategien zum Erwerb belastbarer Befunde) sowie „Wissenschaftsverständnis“ (z. B. Sicherheit und Begrenztheit von Laborbefunden) (Mayer & Ziemek, 2006, S. 5).

doi:10.2390/zdb-v19-i1-315

2.3 Förderung experimenteller Problemlösefähigkeit

Im Rahmen forschenden Lernens (Mayer & Ziemek, 2006) können schon Grundschüler effektiv an die eigenständige Nutzung des Kontrollansatzes und der Variablenkontrollstrategie (Klahr & Nigam, 2004) sowie 6.-Klässler an ein Verständnis multivariater Kausalität (Keselman, 2003) herangeführt werden. Zur Förderung eines Bewusstseins für externe Validität fehlen nach unserem Wissen bislang Erkenntnisse. Angesichts eines Anstiegs der Performanz diverser wissenschaftsmethodischer Kompetenzen über die Schuljahre (Mayer et al., 2008) ist anzunehmen, dass neben unterrichtlicher Förderung auch die kognitive Entwicklung eine Rolle spielt. Erkenntnisse über das *Verständnis interner Validität* legen das nahe: Diesbezüglich ist bei Kindern früh ein niederschwelliges Bewusstsein zu beobachten, das dazu befähigt, korrekte Vorgehensweisen zu identifizieren, während die spontane eigenständige Planung aussagekräftiger Experimente noch nicht gelingt (Schneider, Bullock & Sodian, 1998). Sollten – mit Blick auf vermutlich höhere kognitive Anforderungen des *Verständnisses für externe Validität* (s. Abschn. 2.2, 2.4) – Fördermaßnahmen erst bei älteren Lernenden ergriffen werden? Ohne dies auf empirische Studien zurückzuführen, schlug z. B. Meyer (1978, S. 24 f.) vor, den Vergleich von Befunden verschiedener Experimente bzw. die kritische Beurteilung der Versuchsdurchführung erst ab Klassenstufe 8 bzw. 9 zu trainieren. Vor diesem Hintergrund stellen Erkenntnisse zum möglichen Zeitpunkt für den Beginn des gezielten Kompetenzaufbaus ein wichtiges Forschungsdesiderat der Biologiedidaktik dar.

Als besonders wirkungsvoll und ohne Einbußen für Problemlösefähigkeit oder Transferleistung (Klahr & Nigam, 2004) haben sich bei anderen experimentellen Kompetenzen explizites Training, Reflexion und instruktionale Unterstützung i. S. des moderaten Konstruktivismus‘ erwiesen (Ehmer, 2008; Kirschner, Sweller & Clark, 2006; Urhahne & Harms, 2006), z. B. gemäß des *Cognitive Apprenticeship*-Ansatzes (Keselman, 2003): Neben der direkten Vermittlung von Kompetenzen tragen hier instruktionale Maßnahmen wie Hilfen zur Prozessregulation zur Reduktion der Belastung des Gedächtnis-Arbeitsspeichers bei (Kirschner et al., 2006), was die Performanz begünstigt.

Unverzichtbar sind bei der Förderung experimenteller Problemlösefähigkeit zweifelsohne Realexperimente (Baumann, Simon, Wonisch & Guttenberger, 2013): In Echtzeit werden dabei Primärerfahrungen in den Originalsystemen gesammelt. Im Umgang mit Organismen, Geräten und Techniken können methodologische Planungen unmittelbar realisiert und optimiert sowie Fehler anschaulich analysiert werden.

Eine Alternative bzw. Ergänzung hierzu stellt computergestütztes forschendes Lernen dar, welches mittlerweile in der didaktischen Forschung sämtlicher MINT-Fächer eingesetzt wird: Anhand computerbasierter interaktiver Experimentierumgebungen erwerben Lernende in Modellsystemen durch dynamisches Problemlösen (Leutner, Wirth, Klieme & Funke, 2005) Wissen über Ursache-Wirkungsbeziehungen (Eckhardt, 2010, S. 43, 205 f.; Keselman, 2003, Leutner et al., 2005; Riess & Mischo, 2010). Simulationen ermöglichen, selbst in komplexeren Systemen Erkenntnisse zu gewinnen, was in Realexperimenten nur schwer realisierbar wäre (Baumann et al., 2013; Smith & Smith, 2009, S. 864; Urhahne & Harms, 2006). Dies ist nicht nur gefahrlos und ethisch unbedenklich, sondern auch weniger zeit- und ressourcenaufwändig und reduziert sowohl die intrinsische kognitive Belastung als auch die

doi:10.2390/zdb-v19-i1-315

mit lebenden Organismen verbundene Ablenkung sowie die Komplexität der betrachteten Systeme. Adaptive Prompts können dabei individualisiert und prozessbegleitend die Selbstregulation, Strategienutzung und Wissensaneignung unterstützen (Eckhardt, 2010, S. 66 ff., 208; Keselman, 2003), solange sie keine zu hohe kognitive Belastung verursachen (Eckhardt, 2010, S. 206 ff.; Urhahne & Harms, 2006). Baumann et al. (2013) zufolge sind mit computerbasierten Experimenten keine Nachteile hinsichtlich Wissenszuwachs und Behaltensleistung, dafür aber Vorteile hinsichtlich der Binnendifferenzierung verbunden.

Bislang wurde die Förderung experimenteller Kompetenzen (v. a. bezüglich der Variablenkontrollstrategie) isoliert untersucht (z. B. Ehmer, 2008; Klahr & Nigam, 2004). Mit Blick auf hochkomplexe, schwer untersuchbare lebende Systeme sowie auf deren internen und externen Störgrößen, zeitliche Dynamik und Variabilität (Bräutigam, 2014, S. 23 f.), die beim Experimentieren eine Rolle spielen, ergibt sich die Frage, ob für die Förderung eines Verständnisses externer Validität im integrativen bzw. vorausgehenden Training systemischen Denkens in der spezifischen Domäne sowie im Einbezug von computerbasierten Simulationen (Riess & Mischo, 2010) eine besondere Chance liegen könnte (s. Abschn. 2.4, 2.6, 5).

Domänen- bzw. kontextspezifische Herausforderungen können den Lernprozess mit Experimenten entweder begünstigen oder beeinträchtigen (Keselman, 2003). Die Forschung zur Förderung experimenteller Problemlösefähigkeit widmete sich bislang kaum der externen Validität und klammerte herausforderndere Experimente eher aus (Klahr & Nigam, 2004). Viele Studien nutzen weniger komplexe Domänen als Ökologie (Rösch, in Vorbereitung; Rösch, Rieß & Nerb, 2012). Insofern ist zu klären, ob sich diese als Lernkontext zum Aufbau eines Verständnisses für externe Validität eignet.

2.4 Externe Validität bei ökologischen Experimenten

Ökologische Experimente zu planen, durchzuführen und deren Design bzw. Befunde bezüglich Belastbarkeit und Generalisierbarkeit zu beurteilen, ist besonders anspruchsvoll: Kausalzusammenhänge sind in hochkomplexen, nicht linearen zeitlich dynamischen Systemen mit emergenten Eigenschaften, vielen interagierenden Faktoren und unterschiedlichsten Typen von Wechselbeziehungen verortet. Phänomene sind aufgrund intra- und interindividueller Variabilität sowie diverser anderer externer und interner Ursachen i. d. R. nur begrenzt kalkulierbar (Bräutigam, 2014, S. 13 ff.; Riess & Mischo, 2010; Schulz et al., 2012; Smith & Smith, 2009). Einzelbefunde, v. a. auf Basis von geringmächtigen oder Klumpenstichproben, und Kurzzeitbeobachtungen weisen eine geringe Reliabilität und relativ hohe Wahrscheinlichkeit für Stichprobenfehler auf (Bortz & Döring, 2006, S. 435 f.), was die externe Validität begrenzt. Bei Untersuchungen in unnatürlichen Zusammenhängen (z. B. bei Labor-, Mikro- oder Mesokosmos-Experimenten) ist auch die ökologische Validität reduziert. Die Erkenntnis, dass all diese Aspekte der Erkenntnisgewinnung Grenzen setzen, ist ein wichtiger Beitrag zum Wissenschaftsverständnis (Mayer & Ziemek, 2006).

In der Domäne Ökologie lassen sich ergo zahlreiche Aspekte der externen, v. a. auch ökologischen Validität ansprechen und deren Bedeutung für die Sicherheit, Belastbarkeit und Übertragbarkeit experimenteller Befunde sowie besondere Charakteristika biologischer Untersuchungen erarbeiten. Für den schulischen Kontext ist zu klären, wie sich ein

doi:10.2390/zdb-v19-i1-315

entsprechendes Verständnis operationalisieren lässt, welche Dimensionen es aufweist und welche davon wann unterrichtlich förderbar sind.

2.5 Messung des Verständnisses für Kriterien externer Validität und Befunde

In welchem Maß Lernende Aspekte der externen Validität bzw. damit verbundener Reliabilität tatsächlich erkennen bzw. berücksichtigen, lässt sich mithilfe verschiedener *Assessmentmethoden* erfassen. Diese unterscheiden sich u. a. bezüglich Prozessorientierung und Realitätsnähe, Situationsmerkmalen der Messung, Kontextbezug, fokussierten Phasen bzw. Vollständigkeit des experimentellen Problemlöseprozesses sowie Problemtyp (analytische bzw. dynamische Aspekte; Leutner et al., 2005). Die Analyse von *Realexperimenten* (Beobachtung, Videographie oder Lernprozessgrafiken als Grundlage), Logfile-Daten in *computerbasierten interaktiven Simulationen* oder *prozessorientierten Protokollen* erfasst auch selbstregulative Aspekte des experimentellen, dynamischen Problemlöseprozesses (ebd.), was die Validität dieser Messmethoden erhöht (Emden & Sumfleth, 2012; Hammann, Phan, Ehmer & Grimm, 2008). Des Weiteren können *Produkte* von Experimenten, *Conceptmaps* sowie *Interviews* zur Diagnose herangezogen werden (Klahr & Nigam, 2004). Bei großen Stichproben werden jedoch häufig *paper and pencil-Tests* als eine ökonomischere Methode eingesetzt (Ehmer, 2008; Emden & Sumfleth, 2012) – so auch in den Untersuchungen dieses Beitrags. Problematisch ist dabei die geringere Validität des Messverfahrens infolge reduzierter Komplexität, Authentizität, Vollständigkeit und fehlender Interaktion (Hammann et al., 2008). In vielen Studien findet das geschlossene Single- bzw. Multiple-Choice-Format Verwendung (z. B. ebd.; Ehmer, 2008), welches v. a. kognitiv weniger anspruchsvolle Aktivitäten wie Reproduzieren und Selegieren erfasst (Kauertz, Fischer, Mayer, Sumfleth & Walpuski, 2010). Überdies verzerrt dabei Rateverhalten die Befunde. Freitext-Items (Arnold et al., 2013; Ehmer, 2008; Hammann et al., 2008; Mayer et al., 2008) sind ebenfalls nur begrenzt valide, erfordern aber zumindest die kontextangemessene Reorganisation oder gar den Transfer und die Synthese von Problemlöseoperatoren (Kauertz et al., 2010). Verständnis lässt sich hier über die Begründung einer eigenständig erdachten oder aus Vorschlägen ausgewählten Handlungsoption erfassen.
Bei Freitext-Items zur Experimentplanung zeigen Probanden der gymnasialen Oberstufe (Arnold et al., 2013) ähnlich Lernenden der Sekundarstufe I (Mayer et al., 2008) wenig Bewusstsein für Stichprobengröße, Versuchsdauer, Messzeiten und -wiederholung, welche sich auf die externe Validität auswirken. Dies könnte entweder mit dem hohen kognitiven Anspruch entsprechender Kompetenzen zusammenhängen oder auf ungenügende Förderung hinweisen. Ob Lernende im Biologieunterricht Gelegenheit erhalten, ein Bewusstsein für die besondere Komplexität und für wichtige Aspekte der externen Validität ökologischer Experimente aufzubauen, und darin gezielt unterstützt werden, beleuchtet der folgende Abschnitt.

2.6 Systematischer Aufbau von Kompetenzen zur externen Validität im Unterricht?

Einer Befragung von Lehrkräften (Düppers, 1975) zufolge spielten Langzeit- und Freilandexperimente vor vier Jahrzehnten im Biologieunterricht nur eine untergeordnete

doi:10.2390/zdb-v19-i1-315

Rolle. Wenngleich keine aktuelle Analyse der realen unterrichtlichen Praxis vorliegt, deutet Manches darauf hin, dass sich bis heute nicht viel geändert haben dürfte: So zeigt eine Analyse zeitgenössischer Schulbücher und Lehrerhandreichungen der 5. und 6. Klassenstufe in Baden-Württemberg (Rösch, 2013), dass bei den dort vorgeschlagenen Schülerexperimenten praktisch weder die Beurteilung der Aussagekraft, Sicherheit und Übertragbarkeit eine Rolle spielt, noch Stichprobenumfang und Beobachtungsdauer als Kriterien externer Validität thematisiert werden. In vielen Anregungen für Schülerexperimente (z. B. Eckebrecht, Eckebrecht & Kluge, 2006; Freytag, 2007) finden sich auch in höheren Klassenstufen v. a. Kurzzeit- bzw. gut kontrollierbare Labor- oder Mehrspezies-Mikrokosmos-Experimente. Externe und speziell ökologische Validität, Messwiederholungen, zeitliche Dynamik, längerfristige Effekte sowie intra- bzw. interindividuelle Variabilität und Replikationen werden kaum berücksichtigt, das Vorhandensein von Störgrößen oft ausgeblendet. Insgesamt erhalten Lernende selten die Gelegenheit, eigenständig zu experimentieren (Prenzel, Artelt, Baumert, Blum, Hammann, Klieme & Pekrun, 2007).

3 (*Ein*) Verständnis für Aspekte externer Validität in Klassenstufe 6 fördern?

Um Fragen zur Kompetenzstruktur und Förderbarkeit eines niederschwelligen Verständnisses für die Bedeutung von ausgewählten Kriterien externer Validität zu klären, wurde 2010 in 6. Realschulklassen in Baden-Württemberg im Fächerverbund *Naturwissenschaftliches Arbeiten* eine quasiexperimentelle Feldstudie mit Klumpenstichprobe durchgeführt (Roesch, Nerb & Riess, 2015; Rösch et al., 2012). Mit der *Beobachtungsdauer* und der *Stichprobengröße* wurden zwei Kriterien für externe Validität ausgewählt, die bei ökologischen Experimenten bereits in der Orientierungsstufe phänomenologisch-anschaulich erlebt und bei der Planung sowie Auswertung von Experimenten bedacht und konkret reflektiert werden können – andere Aspekte hingegen erachteten wir für diese Klassenstufe als zu abstrakt. Folgende *Forschungsfragen* aus dem Fragenkatalog eines umfangreicheren Forschungsprojekts stehen im Fokus dieses Beitrags:

Erstens wurde die *Dimensionalität* des Konstrukts untersucht: Im Hinblick auf eine valide Messung und gezielte künftige Förderung sollte geklärt werden, ob das Konstrukt in der konkreten Operationalisierung ein- oder zweidimensional ist (s. Abschn. 3.3). Theoriegeleitet wurde angenommen, dass das Verständnis für die Bedeutung der *Beobachtungsdauer* bzw. *Stichprobengröße* zwei empirisch unterscheidbare Teilkonstrukte darstellt: Es ist z. B. denkbar, dass ein Proband um die zeitliche Dynamik und Veränderlichkeit lebender Systeme sowie um die Existenz von Langzeiteffekten weiß, jedoch kein Bewusstsein dafür hat, dass Effekte auf bestimmte Einflüsse interindividuell variabel sind, wodurch die gewählte Stichprobe(ngröße) die Generalisierbarkeit der experimentellen Befunde beeinflusst. Von Interesse war zweitens, in welcher *Ausprägung* ein basales Verständnis in der 6. Klassenstufe bereits vorliegt (s. Abschn. 3.4), und drittens, ob dort eine effektive *Förderung* anhand eines spezifischen Treatments möglich ist (s. ebd.). Mit Blick auf die Erkenntnisse zum Verständnis

doi:10.2390/zdb-v19-i1-315

für domänenübergreifende Grundlagen des Experimentierens (s. Abschn. 2.3) wurde angenommen, dass auch für Aspekte der externen Validität bereits ein basales Verständnis trainiert werden könne. Für die im Folgenden berichteten Teilstudien gibt Tabelle 1 einen Überblick über die jeweiligen (Teil-)Stichproben und deren Eigenschaften bzw. Zusammenhang.

Tabelle 1
Zusammensetzung der berichteten (Teil-)Stichproben

						Alter (J.)		
Phase	**Intention(en)**	**Eigenschaften**	**SchA**	**KlSt**	***n***	***M***	***SD***	***h*(♀)**
Vor-studie	Item-, Skalenana-lyse, Itemauswahl, Überprüfung der Test-Sensitivität	1 Klasse	HS	6	18	11.6	0.7	61 %
		3 Klassen	GY	6	79	11.5	0.6	67 %
		2 Klassen	GY	9	50	14.4	0.6	46 %
Haupt-studie	Untersuchung der Dimensionalität, Konstrukt-validierung	zusammen-gesetzt aus mehreren Experimental-gruppen[*x]	RS	6	431 bzw. 472[#]	11.9	0.5	44 %
	Evaluation der Treatmentwirkung (Intervention)	[x] davon EXP[t]	RS	6	109	11.8	0.5	49 %
		[x] davon SYS[t]	RS	6	127	11.8	0.5	49 %
		[x] davon $KG_{Öko}$[t]	RS	6	84	11.8	0.6	39 %

Anmerkungen. SchA: Schulart. KlSt: Klassenstufe. *n*: Umfang der (Teil-)Stichprobe. J.: Jahre. *M*: Mittelwert. *SD*: Standardabweichung. *h*(♀): Anteil der Mädchen. [*]: auch aus solchen, die im Rahmen dieses Beitrags nicht berichtet werden; vgl. Bräutigam (2014), Roesch et al. (2015), Rösch et al. (2012), Vogel et al. (2011). HS: Hauptschule. GY: Gymnasium. RS: Realschule. EXP: Treatmentgruppe mit Konzept zur Förderung experimenteller Problemlösefähigkeit. SYS: Treatmentgruppe mit Konzept zur Förderung systemischen Denkens. [t]: Teilmenge der jeweiligen Gruppe innerhalb der Gesamtstichprobe ([*]) aufgrund von Missing Data zu einem der Messzeitpunkte. $KG_{Öko}$: Kontrollgruppe (gleiche Ökologiethemen). [#]: Probandenzahl zum Pretest- / Posttest-Zeitpunkt.

3.1 Entwicklung und Erprobung des Messinstruments

3.1.1 Operationalisierung

Auf Basis der Klassischen Testtheorie wurde in mehreren Vorstudien ein multidimensionaler *paper-and-pencil*-Leistungstest für experimentelle Problemlösefähigkeit entwickelt (Rösch et al., 2012). Dieser enthielt u. a. auch Items zum Verständnis für die beiden o. g. Kriterien der externen Validität von Experimenten in komplexen lebenden Systemen. In den ersten Vorstudien wurden diverse Freitext-Formate erprobt, darunter Items mit vorgegebenen Experimenten, deren externe Validität gut bzw. schlecht ausgeprägt war und von den Probanden ausführlich beurteilt werden sollte. Bodeneffekte zeigten, dass die meisten 6.-

doi:10.2390/zdb-v19-i1-315

Klässler an Realschulen Schwierigkeiten haben, solche Arbeitsaufträge zu verstehen bzw. sie angemessen zu beantworten, was die Validität der Datenerhebung beeinträchtigte. Vermutlich ist bei dieser Zielgruppe die Fähigkeit zur Identifikation angemessener Forschungsdesigns bzw. adäquater Begründungen eher ausgeprägt bzw. förderbar (s. Abschn. 2.3). Für eine alternative Operationalisierung sprachen zudem der insgesamt sehr umfangreiche im Projekt eingesetzte Leistungstest (Bräutigam, 2014, S. 189 ff.; Rösch et al., 2012) sowie die begrenzte Konzentrationsfähigkeit der Probanden.

Daher fiel die Wahl auf das Single-Choice-Antwortformat (Bühner, 2011, S. 117). In der von uns vorgenommenen Operationalisierung galt es einzuschätzen, ob in konkret beschriebenen ökologischen Experimenten z. B. eine kurze *Beobachtungsdauer* bzw. eine kleine *Stichprobe* für belastbare und übertragbare Schlussfolgerungen angemessen sind (vgl. Beispiel-Items in Abb. 1, (a) bzw. (b)). Bei den Experiment-Beschreibungen mancher Items sind hingegen längere Beobachtungszeiträume bzw. größere Stichproben angesprochen, und es wird gefragt, ob für die angemessene Interpretation des Experiments auch kurze Zeiträume bzw. kleine Stichproben ausgereicht hätten (s. Tab. 7, Anhang). Die Items gliedern sich in einen strukturgleichen Stamm mit einem Text-Stimulus und einer Abbildung zur Veranschaulichung sowie ein Antwortfeld. Von drei Antwortoptionen ist die korrekte anzukreuzen.

Manche Vogelarten werden bei uns immer seltener. Wissenschaftler möchten untersuchen, ob sich das Abholzen von Wäldern und damit die Schaffung von naturnahen Freiflächen gut auf die Wiederansiedlung bestimmter Vogelarten auswirken. Dazu vergleicht man zwei gleich große Waldstücke in derselben Gegend, auf denen ursprünglich die gleichen Arten vorkamen. Auf einer Fläche wird der Wald abgeholzt, auf der anderen Fläche bleibt der Wald stehen. Nach zwei Jahren zählt man jeweils wieder die Anzahl von Vogelarten auf den beiden Flächen und vergleicht sie.

Wurde die Untersuchung so durchgeführt, dass man die Ergebnisse ohne Zweifel deuten kann?

Kreuze bei (a) und bei (b) jeweils die Meinung an, die richtig ist!

(a)

□ „Da immer Vogelarten dazukommen oder abwandern, ist es egal, wie lange man insgesamt beobachtet. Hauptsache, man vergleicht die Flächen gleichzeitig."

□ „Zwei Jahre lang zu beobachten, reicht vollkommen aus. Schließlich haben bis dahin alle Veränderungen stattgefunden."

□ „Zwei Jahre sind eine zu kurze Zeit – man müsste die Flächen viel länger beobachten."

(b)

□ „Es ist gut, diese Untersuchung auf zwei Flächen durchzuführen, weil das ja ausreicht."

□ „Auch wenn es aufwändiger ist, müsste man dafür mehr Flächen untersuchen."

□ „Wie viele Flächen untersucht werden, ist nicht so wichtig, wenn richtig gearbeitet wird."

Abbildung 1. Beispielaufgabe mit Items zu (a) Beobachtungsdauer und (b) Stichprobengröße (Zeichnung: Verfasser)

doi:10.2390/zdb-v19-i1-315

Neben der eigentlichen Lösung gibt es zwei Distraktoren, die in Vorstudien beobachteten Fehlkonzepten nachempfunden sind. Anhand von Begründungen wurde versucht, deren Plausibilität und damit Attraktivität zu erhöhen (vgl. Abb. 1; Bühner, 2011, S. 119). Korrekt bearbeiteten Items wurde der Wert 1 zugeordnet, falsch beantworteten Items der Wert 0. Bei der statistischen Auswertung wurde das arithmetische Mittel der Items berechnet. Eine Ergänzung dieses Antwortformats um Freitext-Items zur Begründung des Ankreuzverhaltens hätte die Validität der Kompetenzmessung erhöht, konnte jedoch aufgrund der begrenzten Bearbeitungszeit nicht vorgenommen werden.

3.1.2 Erprobung des Messinstruments

Bei der kriteriengeleiteten Optimierung des Tests wurde das Antwortformat besser an die Zielgruppe adaptiert und der Umfang des Subtests im Hinblick auf die begrenzte Bearbeitungszeit reduziert. In der letzten Vorstudie (s. Tab. 1) wurden für die Aspekte *Beobachtungsdauer* bzw. *Stichprobengröße* je zwölf Items erprobt. Tabelle 7 (s. Anhang) gibt einen Überblick über die Kontexte dieser Items und zeigt, auf welche Weise die Kriterien *Beobachtungsdauer* bzw. *Stichprobengröße* in den Items beachtet wurden. Anhand einer Item- und Skalenanalyse (Bühner, 2011, S. 216 ff.) wurden aus den Itempools jeweils die vier Items ausgewählt, die aufgrund ihrer Kennwerte in der Stichprobe der Vorstudie am geeignetsten erschienen. Anders als in der Hauptstudie standen für diese Vorstudie keine Realschulklassen zur Verfügung. Um für die o. g. Analysen die statistische Varianz zu erhöhen, wurden eine vermutlich etwas leistungsschwächere Hauptschul- und mehrere Gymnasialklassen miteinbezogen (s. Tab. 1). Die *Skalenreliabilität* für die acht binären Items wurde mit MPlus 6.1 (Muthén & Muthén, 2010; s. Abschn. 3.3) anhand des Verfahrens von Raykov, Dimitrov und Asparouhov (2010) unter Nutzung der ML (Maximum Likelihood)-Methode geschätzt. Da zu diesem Zeitpunkt noch nicht bekannt war, ob die postulierten Kompetenzdimensionen *Beobachtungsdauer* und *Stichprobengröße* in der Hauptstudie zur Untersuchung der Treatment-Wirksamkeit in einer einzigen Skala oder getrennt in zwei Subskalen behandelt würden, wurde die Skalenreliabilität für dreierlei Item-Batterien ermittelt: Sie beträgt in der Gesamtstichprobe (N = 147) .74 für die Gesamtskala (acht Items) sowie .52 für die Subskalen *Beobachtungsdauer* (vier Items) bzw. .69 für *Stichprobengröße* (vier Items). Diese Werte sind z. T. nicht zufriedenstellend, genügen jedoch für Gruppenvergleiche (Lienert & Raatz, 1998, S. 14). Wie sich in der Hauptstudie zeigte, führte der moderate Stichprobenumfang der Vorstudie (N = 147) zu Stichprobenfehlern (Bühner, 2011, S. 169) hinsichtlich der Skaleneigenschaften und Eignung der Items.

Zur Überprüfung der Sensitivität des Instruments wurden die Mittelwerte der jeweils gepoolten 6. und 9. Gymnasialklassen anhand von t-Tests für unabhängige Stichproben mithilfe der Software IBM SPSS Statistics 22 verglichen: Es wurde angenommen, dass die Performanz infolge fortgeschrittener kognitiver Entwicklung, höheren Domänenwissens und größeren Erfahrungsschatzes bei 9.-Klässlern ausgeprägter sei. In dieser schularthomogenen Stichprobe (n = 129) beträgt die Skalenreliabilität für die Gesamtskala .75 und für die Subskalen *Beobachtungsdauer* und *Stichprobengröße* .49 bzw. .70. Das Instrument erwies sich hier sowohl bezüglich der Gesamtskala als auch der beiden Subskalen als nicht

doi:10.2390/zdb-v19-i1-315

ausreichend sensitiv (s. Tab. 2.). Deskriptivstatistisch sind die Mittelwerte der 9. Klassenstufe zwar gleich groß bzw. hypothesenkonform sogar größer als in der 6. Klasse. Die Durchschnittswerte des Verständnisses für die *Stichprobengröße* unterscheiden sich marginal signifikant – zu erwarten wären jedoch jeweils signifikante Unterschiede.

Tabelle 2
Ergebnisse des Mittelwertvergleichs zwischen 6. und 9. Klassenstufe (Vorstudie, Gymnasium)

	Klassenstufe							
	6		**9**		***t*-Test (unabhängige Gruppen)**			
Itembatterie	***M*[a]**	***SD***	***M*[a]**	***SD***	***t***	***df***	***p***	***d***
Gesamtskala[*]	.47	.30	.50	.25	-0.71	127	.24	0.11
Beobachtungsdauer[x]	.56	.32	.55	.24	0.20	124.2	.42	-0.03
Stichprobengröße[x]	.37	.35	.46	.35	-1.29	127	.10	0.26

Anmerkung. [*]: Gesamtskala zur externen Validität mit 8 Items. [x]: Subskala mit je 4 Items. M^a: arithmetisches Mittel der jeweils gültig bearbeiteten Items, *SD*: Standardabweichung. *t*: t-Wert (Prüfgröße). *df*: Freiheitsgrade. *p*: Irrtumswahrscheinlichkeit (einseitig). *d*: (entspr. der untersch. *n*) korrigierte Effektstärke (Cohens *d*).

Wird bei der Hypothesenprüfung die 6. Hauptschulklasse miteinbezogen, so indiziert das Messinstrument zumindest bezüglich der Subskala *Stichprobengröße* hypothesenkonform einen signifikanten kleinen Mittelwertunterschied zwischen 6. (M = .35, SD = .34) und 9. Klassenstufe (M = .46, SD = .35), $t(145) = -1.71$, $p < .05$, d = .34. Zusammenfassend betrachtet ist zu vermuten, dass die realisierte Operationalisierung zwar die Indikation von größeren Performanzunterschieden gestattet, die Sensitivität jedoch begrenzt ist. Der durch kognitive Reifung oder Förderung erklärbare Unterschied zwischen den Klassenstufen hält sich womöglich in Grenzen. Das Verständnis, dass bei Experimenten mit lebenden Systemen längere Beobachtungsdauern zu höherer Aussagekraft führen, ist in beiden Klassenstufen stärker ausgeprägt als das Verständnis für die Bedeutung der Stichprobengröße (s. auch Tab. 4). Denkbare Ursachen für dieses Phänomen könnten unterschiedlich hohe kognitive Anforderungen oder eine unterschiedlich explizite Thematisierung dieser Kriterien für externe Validität im Alltag sein.

3.2 Überblick über die Interventionsstudie

Die Interventionsstudie zur Förderung experimenteller Problemlösefähigkeit, darunter auch des *Verständnisses für die Bedeutung von Beobachtungsdauer und Stichprobengröße*, war in einem größeren, mehrere Teilstudien umfassenden Forschungsprojekt angesiedelt. In diesem sollte die Effektivität von Unterrichtskonzepten ohne computerbasierte Lernumgebung zur Förderung (a) experimenteller Problemlösefähigkeit (Roesch et al., 2015; Rösch et al., 2012) bzw. (b) systemischen Denkens (Bräutigam, 2014, S. 133; Vogel, Rieß & Nerb, 2011) in einem Pretest-Posttest-Design untersucht werden. Dazu wurden unabhängig voneinander spezifische Treatmentgruppen mit einer adäquaten Kontrollgruppe hinsichtlich der in den Studien jeweils fokussierten Kompetenzen verglichen. Der Pretest fand vor der Intervention

doi:10.2390/zdb-v19-i1-315

statt, der Posttest zwei Wochen danach. Der Unterricht wurde in allen Klassen durch die Fachlehrkräfte erteilt, wobei für die Treatmentgruppen EXP und SYS Fortbildungen durchgeführt wurden und detaillierte Handreichungen vorlagen. Für das Gesamtprojekt bot sich der Lernkontext „Waldökologie“ an: Einerseits können Fragen interner *und* externer Validität bei Experimenten gut thematisiert werden (s. Abschn. 2.4, 2.6). Anhand intraökosystemarer Phänomene und der Verzahnung von Öko- mit Human-Systemen lässt sich andererseits auch systemisches Denken einführen (Bräutigam, 2014, S. 12 ff.; Riess & Mischo, 2010). Überdies können alle vier Kompetenzbereiche der Bildungsstandards i. S. einer Bildung für nachhaltige Entwicklung miteinbezogen werden (KMK, 2005, S. 7, 13 ff.; Roesch et al., 2015).

Im Mittelpunkt dieses Beitrags steht die Treatmentgruppe EXP (s. Tab. 1). Im Verlauf der insgesamt 13 Stunden an der Schule und 2 Tage am Naturschutzzentrum „Ruhestein“ (Nationalpark Schwarzwald) umfassenden Intervention (Roesch et al., 2015; Rösch et al., 2012; Vogel et al., 2011) beschäftigten sich die Lernenden zunächst 8 Unterrichtsstunden an der Schule bzw. einen Tag lang im „Waldklassenzimmer“ des Naturschutzzentrums mit domänenübergreifenden Aspekten und Strategien experimenteller Erkenntnisgewinnung; diese betrafen u. a. die Schritte des hypothetisch-deduktiven Verfahrens, die Formulierung epistemischer Fragen, die systematische Suche im Hypothesenraum, die Bedeutung von Kontrollansätzen sowie die Variablenkontrollstrategie. Die Lernenden befassten sich dabei mit relativ gut kontrollierbaren ein- und zweifaktoriellen Experimenten. Anschließend setzten sie sich mit Aspekten wie Störgrößen, Variabilität und Dynamik in lebenden Systemen auseinander, welche sich auf die Sicherheit, Aussagekraft und Übertragbarkeit experimenteller Befunde auswirken. Dazu befassten sich diese Klassen in den fünf folgenden Schulstunden sowie am zweiten Naturschutzzentrum-Tag auch mit der Steigerung der externen Validität (z. B. Messwiederholungen, längere Beobachtungsdauer, größere Stichproben) sowie mit den Grenzen der Übertragbarkeit von Befunden aus Labor- bzw. Modellexperimenten (s. Abb. 2).

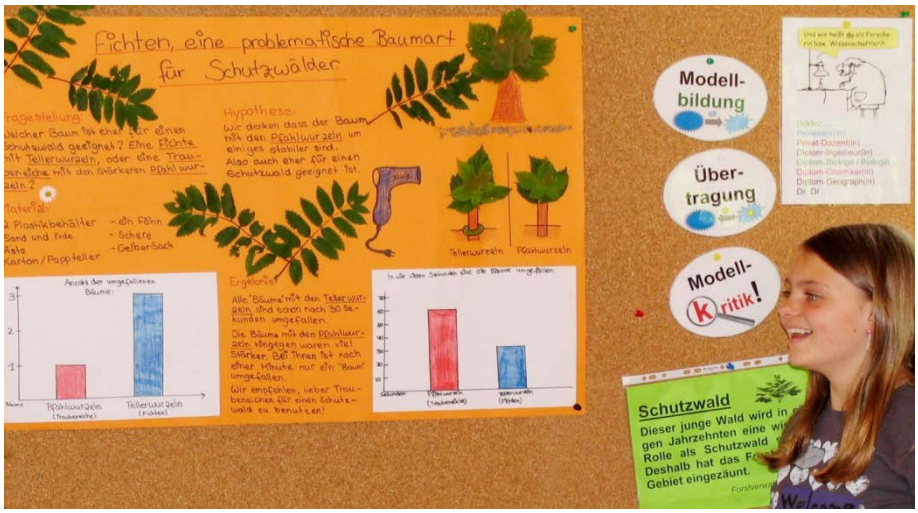

Abbildung 2. Lernende präsentieren Design, Befunde und Interpretation eigener Modellexperimente zu Funktionen von Schutzwäldern und diskutieren Aspekte der Validität (Foto: Verfasser)

doi:10.2390/zdb-v19-i1-315

Zu den Lernaktivitäten gehörten die hypothesenprüfende, kritisch reflektierende Auswertung vorgegebener Langzeituntersuchungen (Populationsschwankungen bei Borkenkäfern und Bekämpfungsstrategien), Untersuchungsaufträge zur Zersetzung, die handlungsorientierte Analyse von Sukzessionsstadien sowie die Durchführung eigener Langzeit-, ethologischer und Modell-Experimente (Einflüsse auf Samenkeimung und Pflanzenwachstum, Ortspräferenz bei Asseln, Schutzfunktionen von Wäldern). Anhand expliziter Reflexion (Arbeitsblätter, Unterrichtsgespräche, Diskussionen) wurden die Erkenntnisse gefestigt, angewandt und vertieft. Instruktionale Unterstützung sollte die Selbstregulation begünstigen, z. B. in Form von Karten mit gestuften Hilfen. Zu Beginn wurde jeweils auch mit Lösungsbeispielen gearbeitet.

Das Treatment einer zweiten Experimentalgruppe SYS (s. Tab. 1) zur Förderung systemischen Denkens war dazu inhaltlich bestmöglich parallelisiert (Bräutigam, 2014, S. 110, 133). Dort wurde an den Tagen am Naturschutzzentrum z. T. auch handlungsorientiert gearbeitet, jedoch weder experimentiert, noch wurden Aspekte der Reliabilität oder Validität angesprochen. Die beteiligten Schulklassen befassten sich mit der Definition und allgemeinen Eigenschaften von Systemen. Anschließend lernten die Probanden Elemente von Waldökosystemen und deren Interaktionen kennen, modellierten System-Ausschnitte mithilfe symbolischer Darstellungsformen und entwickelten Erklärungen sowie Prognosen für das dynamische Systemverhalten, indem sie anhand der Modelle mögliche Dynamiken simulierten (ebd.; Vogel et al., 2011). Dabei erfuhren sie, wie systemisches Denken helfen kann, authentische Problemsituationen zu bearbeiten und Erkenntnisse zu gewinnen.

Die Kontrollgruppe $KG_{Öko}$ (s. Tab. 1) erhielt Unterricht ohne ein spezifisches Konzept zur Förderung experimenteller Problemlösefähigkeit oder systemischen Denkens. Gemäß der verbindlichen Bildungsstandards spielten aber auch dort naturwissenschaftliche Erkenntnismethoden eine Rolle (KMK, 2005, S. 10 f., 14). Durch die inhaltliche Parallelisierung wurden die gleichen ökologischen Inhalte wie in EXP und SYS behandelt (mit minimalen Abstrichen aus pragmatischen Gründen). Neben dieser Kontrollgruppe gab es im Gesamtprojekt noch weitere, sodass den folgenden Ausführungen z. T. eine größere Gesamtstichprobe zugrunde liegt (vgl. Tab. 1).

3.3 Untersuchung der Dimensionalität, Konstruktvalidität und Skalenreliabilität

Die Hypothese zur Dimensionalität (s. Abschn. 3) wurde in mehreren Schritten überprüft. Die vier Items BD_1 bis BD_4, welche die Beobachtungsdauer thematisierten, wurden dazu in einem ersten, zweifaktoriellen Modell (A) dem postulierten Faktor *Verständnis für die Bedeutung der Beobachtungsdauer* (BD) zugeordnet, die vier Items ST_1 bis ST_4 zur Stichprobengröße dem Faktor *Verständnis für die Bedeutung der Stichprobengröße* (ST) (s. Abb. 3). Dem wurde ein zweites, einfaktorielles Modell (B) gegenübergestellt, in dem alle acht Items dem gemeinsamen Faktor *Verständnis für die Bedeutung von Beobachtungsdauer und Stichprobengröße* zuordnet sind (s. Abb. 4).

Aufgrund der Item-Codierung wurde die hypothesenprüfende Konstruktvalidierung in Form von Parameterschätzungen sowie Modellanalysen und -vergleichen mit der Software MPlus 6.1 (Muthén & Muthén, 2010) anhand der Posttest-Werte von 472 Versuchspersonen aus

doi:10.2390/zdb-v19-i1-315

verschiedenen Experimentalgruppen des Projekts durchgeführt (s. o.; Tab. 1) – diese über die hier berichteten Gruppen EXP, SYS und $KG_{Öko}$ (s. Abschn. 3.2, 3.4) hinausreichende Stichprobe ist repräsentativer. Die Daten werden dabei als dichotome kategorial skalierte Indikatoren aufgefasst (Muthén & Muthén, 2010, S. 488). MPlus 6.1 verwendet das WLSMV-Schätzverfahren (Weighted Least Squares Means and Variance adjusted estimation) (ebd., S. 58), welches zu den ADF-(Asymptotically Distribution-Free-) Methoden zählt (Schermelleh-Engel, Moosbrugger & Müller, 2003) und auf Basis der Probabilistischen Testtheorie das 2PL (Zwei-Parameter-Logistische)-Modell (Birnbaum, 1968) nutzt (Bühner, 2011, S. 503 ff.). Für binäre und kategoriale sowie nicht multivariat normalverteilte Daten stellt es im Gegensatz zur ML-(Maximum-Likelihood-) Methode i. A. das angemessenere Verfahren dar (Schermelleh-Engel et al., 2003; Beauducel & Herzberg, 2006).

Zunächst wurde in einer Konfirmatorischen Faktorenanalyse erster Ordnung die Alternativhypothese getestet, es handele sich um zwei empirisch differente Faktoren (Muthén & Muthén, 2010, S. 399 f.) (s. Abb. 3). Zwecks Modellidentifikation wurde in Modell (A) als Spezifikation die Varianz beider Faktoren auf 1.00 fixiert (Bühner, 2011, S. 398).

Im Anschluss wurde die Nullhypothese getestet, die postulierten Faktoren (BD) und (ST) seien empirisch nicht unterscheidbar. Das entsprechende Modell (B) ist ein restriktiveres (Muthén & Muthén, 2010, S. 670), genestetes „Untermodell“ (Bühner, 2011, S. 542; Schermelleh-Engel et al., 2003), welches als weitere theoretische Annahme zwischen den beiden Faktoren (BD) und (ST) eine perfekte Interkorrelation von 1.00 verwendet, diese also künstlich gleichsetzt (s. Abb. 4).

Neben den Abbildungen 3 und 4 ermöglicht Tabelle 3 eine Synopse wichtiger Modell-Fit-Indikatoren dieser Modelle. Beide erfüllen die Voraussetzung signifikanter Faktorladungen (Korrelation des Faktors mit der jeweiligen manifesten Variablen), die jedoch nicht alle über .60 liegen. Niedrige Indikatorreliabilitäten (durch den Faktor aufgeklärte systematische Varianzanteile; Bühner, 2011, S. 310) der manifesten Variablen sind in der sozialempirischen Forschung nicht selten (Beauducel & Herzberg, 2006). Sie liegen hier z. T. unter .40 und indizieren eine gewisse Heterogenität der Faktoren (Bühner, 2011, S. 453): Vermutlich wird ein (unterschiedlich großer) Teil der Varianz auch durch andere Faktoren beeinflusst (ebd., S. 119), was die Testvalidität und die Reliabilitätswerte reduziert.

Tabelle 3
Fit-Indizes der Modelle (A) und (B) (N = 472)

Modell	***df***	χ^2	χ^2/df	***RMSEA***	***CI90***	***CFI***	***TLI***
zweifaktoriell (A)	19	39.056**	2.056	0.047	[0.026; 0.068]	0.969	0.954
einfaktoriell (B)	20	42.993**	2.150	0.049	[0.029; 0.070]	0.964	0.950

Anmerkungen. *df*: Freiheitsgrade. **: $p < .01$. χ^2/df: normed χ^2. *RMSEA*: Root Mean Square Error of Approximation. *CI90*: 90-prozentiges Konfidenzintervall des RMSEA. *CFI*: Comparative Fit Index. *TLI*: Tucker-Lewis-Index.

doi:10.2390/zdb-v19-i1-315

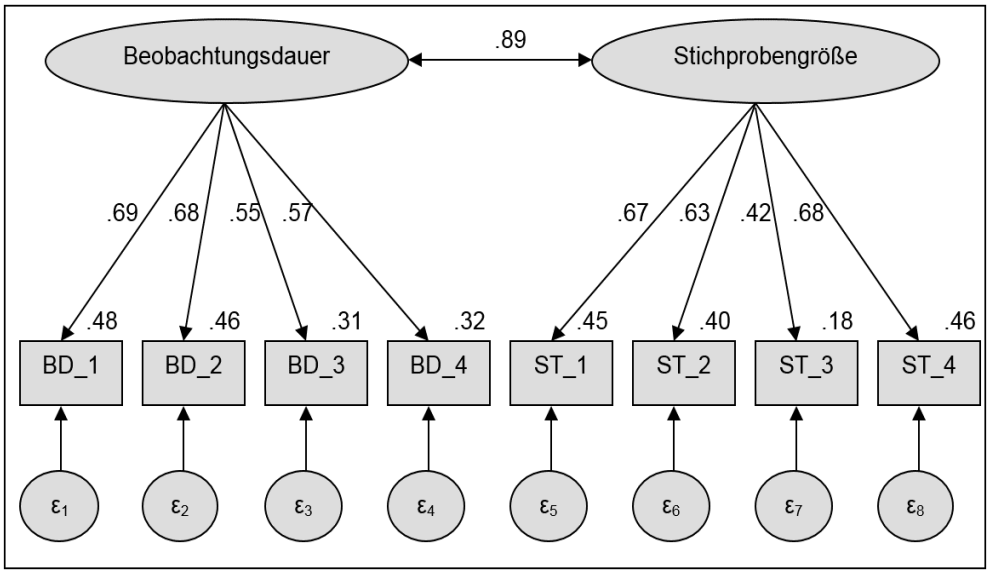

Abbildung 3. Strukturgleichungsmodell des zweifaktoriellen Modells (A) (*N* = 472)

Anmerkungen. BD_i: Item i für *Verständnis für die Bedeutung der Beobachtungsdauer*, ST_j: Item j für *Verständnis für die Bedeutung der Stichprobengröße*. ε_k: Fehlervariable der manifesten Variable k. Die Werte an den Pfeilen geben die geschätzten standardisierten Faktorladungen der Items wieder, die Werte an den Items (Rechtecke) deren jeweilige geschätzte Indikatorreliabilität. $p < .001$.

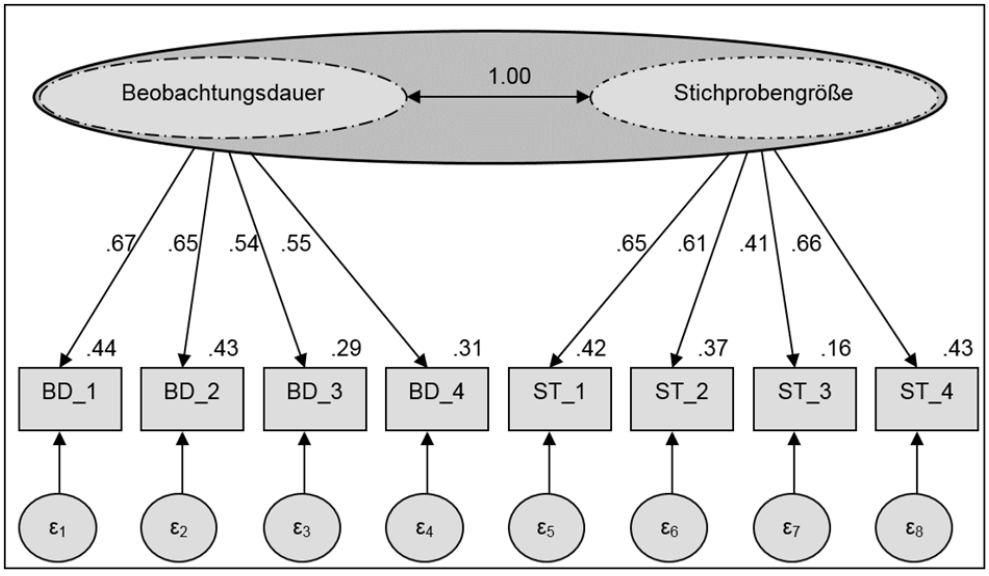

Abbildung 4. Strukturgleichungsmodell des genesteten „einfaktoriellen“ Modells (B) (*N* = 472)

Anmerkungen. S. Abb. 3. Der durch Fixierung der Interkorrelation der Faktoren auf 1.00 künstlich geschaffene Einzel-Faktor ist dunkel schattiert angedeutet (großes Oval).

doi:10.2390/zdb-v19-i1-315

Der Vergleich der Kennwerte für die Gesamtpassung der theoretisch fundierten Modelle (A) bzw. (B) mit den Daten (Globaler Signifikanztest χ^2 bzw. *normed* χ^2) bzw. anderer Fit-Indizes zeigt, dass diese nah beieinander und zum Großteil noch in einem akzeptablen Gütebereich liegen (ebd., S. 418 ff.; Schermelleh-Engel et al., 2003). Das zweifaktorielle Modell (A) scheint dem einfaktoriellen Modell (B) minimal überlegen zu sein. Vor diesem Hintergrund wurde anhand eines χ^2-Differenztests für hierarchisch geschachtelte Modelle (Bühner, 2011, S. 542; Muthén & Muthén, 2010, S. 434 f.) untersucht, ob die empirische Passung des restriktiveren Modells (B) signifikant von Modell (A) abweicht, mit Modell (B) also eine Verschlechterung der Schätzung verbunden ist. Der geringe Unterschied, $\chi^2(1) = 3.904$, $p = .048$, spricht jedoch nicht für eine besonders gute Differenzierbarkeit der beiden Modelle: angesichts der relativ großen Stichprobe, die mit einer höheren Teststärke einhergeht (Bühner, 2011, S. 419 f.), ergibt sich ein größeres α-Fehler-Risiko. Um dieses zu reduzieren, legten wir überdies das Signifikanzniveau bei $\alpha = .01$ fest.

Auch die sehr hohe Interkorrelation der Faktoren ($r = .89$, $p < .001$) spricht gegen diskriminante Validität der postulierten Faktoren und für eine starke Redundanz der jeweils erklärten Informationen. Somit fassten wir bei der Evaluation der Intervention i. d. R. alle acht Items zu *einer* Itembatterie zusammen und betrachteten sie als Indikatoren *eines* zugrundeliegenden Konstrukts *Verständnis für die Bedeutung von Beobachtungsdauer und Stichprobengröße*.

Die *Augenscheinvalidität* (Bühner, 2011, S. 62) ist differenziert zu betrachten: Durch das Itemformat kann – sofern kein Rateverhalten vorliegt – die Reorganisation von Kenntnissen überprüft werden, nicht jedoch divergentes oder gar dynamisches Problemlösen. Ohne Begründung des Antwortverhaltens ist zwar keine Aussage über das zugrunde liegende Verständnis ableitbar, andererseits spielt so die verbale Ausdrucksfähigkeit keine Rolle. Weitere Einschränkungen der Validität werden in Kapitel 4 diskutiert. Das in diesem Beitrag fokussierte Konstrukt lässt sich i. S. *diskriminanter Validität* von allgemeinen kognitiven Fähigkeiten und der Schulnote in Naturwissenschaftlichem Arbeiten abgrenzen (Rösch et al., 2012).

3.4 Untersuchung von Grundperformanz und Treatmentwirkung

Wie Tabelle 4 zeigt, variiert der Schwierigkeitsindex P_i der Items – also der Prozentsatz bzw. hier die Prozentzahl der jeweils korrekten Antworten (Bühner, 2011, S. 222) – stark mit dem Aufgabenkontext, $21.8 < P_i < 66.6$. Dies lässt darauf schließen, dass bereichsspezifisches Vorwissen bzw. kontextbezogene Präkonzepte das Antwortverhalten merklich beeinflussen. In der 6. Klassenstufe wählt durchschnittlich lediglich knapp jede zweite Versuchsperson die korrekte Antwort beim Aspekt *Beobachtungsdauer*. Das Verständnis für eine angemessene *Stichprobengröße* und somit auch für intra- bzw. interspezifische Variabilität ist unter diesen Probanden deutlich geringer ausgeprägt (vgl. auch Tab. 2).

Um die Hypothese zu überprüfen, die Lernumgebung der Experimentalbedingung EXP könne das fokussierte Merkmal fördern, führten wir mithilfe der Software IBM SPSS Statistics 22 anhand der Posttest-Werte Kovarianzanalysen (ANCOVAs) durch. Dabei wurde der Pretest-Wert als Kovariate herangezogen. Mit .58 im Pretest ($N = 431$; s. Tab. 1) und .71 im Posttest ($N = 472$) der Hauptstudie reicht die Reliabilität der Gesamtskala (acht Items) für

doi:10.2390/zdb-v19-i1-315

Gruppenvergleiche aus. Während *andere* Komponenten experimenteller Problemlösefähigkeit durch das Treatment EXP stärker als in der Kontrollgruppe $KG_{Öko}$ gefördert wurden (Rösch et al., 2012), überstiegen die Mittelwerte des *Verständnisses für die Bedeutung von Beobachtungsdauer und Stichprobengröße* in EXP zwar die Mittelwerte von $KG_{Öko}$ deskriptiv (s. Tab. 5), jedoch nicht signifikant, $F(1,190) = 0.09$, $p = .39$ (einseitig).

Tabelle 4
Schwierigkeitsindizes der Items zur externen Validität (Hauptstudie, Pretest, Gruppen EXP, $KG_{Öko}$ und SYS, N = 320)

Subskala *Beobachtungsdauer*					**Subskala *Stichprobengröße***					**Skala**
P_{BD_1}	P_{BD_2}	P_{BD_3}	P_{BD_4}	$\boldsymbol{P_{M\_BD}}$	P_{ST_1}	P_{ST_2}	P_{ST_3}	P_{ST_4}	$\boldsymbol{P_{M\_ST}}$	$\boldsymbol{P_{M\_ges}}$
35.4	66.5	56.6	36.4	**48.7**	39.8	51.7	21.9	35.6	**37.3**	**43.0**

Anmerkungen. P_i: Schwierigkeitsindex der dichotomen Items, gültige Prozent-Sätze (s. Text). BD: *Beobachtungsdauer*. *M*: arithmetisches Mittel der (Sub-)Skalen-Items. ST: *Stichprobengröße*. ges: Gesamtskala.

Tabelle 5
Deskriptive Statistiken und Parameterschätzungen der Gesamtskala zur externen Validität (Hauptstudie)

		Messzeitpunkt				**ANCOVA Parameterschätzung**			
		Pretest		**Posttest**		EXP vs. $KG_{Öko}$		SYS vs. $KG_{Öko}$	
Gruppe	***n***	***M***	***SD***	***M***	***SD***	$\boldsymbol{M_b}$	***SE***	$\boldsymbol{M_b}$	***SE***
EXP	109	.44	.23	.46	.28	$.45^{*}$	.02		
$KG_{Öko}$	84	.38	.20	.43	.24	$.44^{*}$	.03	$.46^{x}$	.03
SYS	127	.45	.24	.50	.29			$.49^{x}$	.02

Anmerkungen. *n*: Anzahl der Versuchspersonen. *M*: Mittelwert der arithmetischen Mittel der bearbeiteten Items. *SD*: Standardabweichung. M_b: geschätzter bereinigter Mittelwert (korrigiert, adjustiert). *: Kovariate werden im Modell für den Pretestwert .41 ausgewertet. ^{x}: dito für den Pretestwert .42. *SE*: Standardfehler.

Obwohl das Treatment SYS primär für die Untersuchung der Förderbarkeit systemischen Denkens entwickelt wurde (Bräutigam, 2014; Vogel et al., 2011), war es infolge der Parallelisierung der Experimentalbedingungen möglich, auch dessen Effekte auf Komponenten experimenteller Problemlösefähigkeit zu betrachten. Da eine systemische Perspektive u. a. komplexe Wechselwirkungen zwischen Systemelementen und damit verbundene Störgrößen und zeitliche bzw. interindividuelle Variabilität sowie die Systemdynamik in den Blick nimmt, interessierte uns, ob sich auch allein das Training systemischen Denkens positiv auf das *Verständnis für die Bedeutung von Beobachtungsdauer und Stichprobengröße* auswirken würde, ohne dass die externe Validität von Experimenten behandelt wurde. Die Gruppe SYS erreichte gegenüber $KG_{Öko}$ auf deskriptiver Ebene höhere Mittelwerte, die jedoch nicht überzufällig groß ausfielen, $F(1,208) = 0.86$, $p = .18$ (einseitig).
In einer differenzierteren Analyse wurden die theoretisch postulierten zwei Komponenten des Konstrukts auch getrennt betrachtet. Bezüglich der Subskala *Beobachtungsdauer* wurden sowohl für das Treatment EXP als auch für SYS positive Effekte auf den Kompetenzaufbau

doi:10.2390/zdb-v19-i1-315

erwartet. Die Skalenreliabilität der Subskala *Beobachtungsdauer* (4 Items) betrug im Pretest des gesamten Forschungsprojekts .50 (N = 431) und im Posttest .58 (N = 472), was für Gruppenvergleiche genügt. Die Befunde zeigen einen interessanten Sachverhalt (s. Tab. 6): Während zwischen den Gruppen EXP und $KG_{Öko}$ kein Mittelwertunterschied beobachtet wurde, $F(1,190) < .001$, $p = .50$ (einseitig), liegen die Mittelwerte von SYS nach der Intervention signifikant über jenen von $KG_{Öko}$: $F(1,208) = 3.29$, $p < .05$ (einseitig), der Effekt ist klein, part. $\eta^2 = .02$.

Tabelle 6
Deskriptive Statistiken und Parameterschätzungen der Subskala Beobachtungsdauer (Hauptstudie)

		Messzeitpunkt				**ANCOVA Parameterschätzung**			
		Pretest		**Posttest**		EXP vs. $KG_{Öko}$		SYS vs. $KG_{Öko}$	
Gruppe	***n***	***M***	***SD***	***M***	***SD***	M_b	***SE***	M_b	***SE***
EXP	109	.49	.28	.48	.33	$.47^{*}$	.03		
$KG_{Öko}$	84	.46	.27	.47	.31	$.47^{*}$	.03	$.48^{x}$	.03
SYS	127	.51	.31	.56	.32			$.56^{x}$	.03

Anmerkungen. *n*: Anzahl der Versuchspersonen. *M*: Mittelwert der arithmetischen Mittel der bearbeiteten Items. *SD*: Standardabweichung. M_b: geschätzter bereinigter Mittelwert (korrigiert, adjustiert). *SE*: Standardfehler. *: Kovariate werden im Modell für den Pretestwert .48 ausgewertet. ^{x}: dito für den Pretestwert .49.

4 Zusammenfassung und Diskussion

Im Mittelpunkt dieses Beitrags stand das Grundverständnis für zwei Kriterien externer Validität beim Experimentieren, welches in der Domäne Ökologie untersucht wurde. Die Operationalisierung des *Verständnisses für die Bedeutung von Beobachtungsdauer und Stichprobengröße* ist eingeschränkt valide, jedoch gut an die Zielgruppe adaptiert (s. Abschn. 2.5, 3.2) und ermöglicht eine objektive, ökonomische Auswertung. In der untersuchten Stichprobe der 6. Klassenstufe sind die zwei theoretisch postulierten Teildimensionen des Konstrukts empirisch praktisch nicht zu unterscheiden, was u. a. durch die relativ niedrigen Faktorladungen und die Inhomogenität der Items bedingt sein könnte. Theoretisch denkbar wäre auch, dass die beiden vermuteten Faktoren u. U. auf ein „gemeinsames übergeordnetes Konstrukt“ zurückzuführen sind (Mayer et al., 2008, S. 71).

Die niedrige Performanz in Klassenstufe 6 überrascht angesichts der Befunde von Arnold et al. (2013) nicht gänzlich; sie stützt überdies die Ergebnisse von Mayer et al. (2008) (s. Abschn. 2.5). Erstaunlich erscheint sie jedoch vor dem Hintergrund, dass das vorliegende Antwortformat eine geringere Herausforderung darstellt als ein offenes (s. Abschn. 2.3, 2.5) und die Aspekte der externen Validität im Verlauf des Treatments anschaulich erfahren und explizit reflektiert wurden. Für dieses Phänomen sowie für den geringen relativen Leistungszuwachs in der Treatmentgruppe EXP kommen verschiedene Ursachen infrage – vermutlich in Kombination: (a) hoher Abstraktions- und somit kognitiver Anforderungsgrad der Kompetenz mit Blick auf die Altersstufe; (b) geringes vorhandenes Domänenwissen

doi:10.2390/zdb-v19-i1-315

sowie relativ weiter Transfer zwischen Unterrichtsinhalten und Testitems; (c) suboptimale Gestaltung der Lernumgebung – z. B. bezüglich Dauer, Übungsmöglichkeiten, extrinsischer kognitiver Belastung durch Art von Lernaktivitäten, -kontext oder Medien (Kirschner et al., 2006). Angesichts mehr oder weniger großer Effekte bei anderen kognitiven experimentellen Kompetenzen in dieser Klassenstufe (Ehmer, 2008; Klahr & Nigam, 2004; Rösch et al., 2012), expliziter Trainingsmaßnahmen (Kirschner et al., 2006) sowie geringer Unterschiede zwischen Unter-, Mittel- und Oberstufe weiterführender Schulen (s. o.; Abschn. 2.5) ergibt sich der Befund vermutlich v. a. aus dem kognitivem Entwicklungsstand, hohen mentalen Anforderungen dieser Kompetenz (s. Abschn. 2.2) und dem anspruchsvollen Lernkontext (Roesch et al., 2015). Sicher spielen auch eingeschränkte ökologische Kenntnisse eine Rolle. Die Annahme, der ausbleibende Effekt resultiere alleine aus fehlendem Domänenwissen, wird allerdings u. a. dadurch relativiert, dass sich in der Vorstudie lediglich geringe bzw. inferenzstatistisch keine Mittelwertunterschiede zwischen 6. und 9. Klassenstufe zeigten, obwohl das Domänenwissen der 9.-Klässler bezüglich der Itemkontexte laut curricularer Vorgaben mittlerweile größer sein müsste. Bereichsspezifische Kenntnisse sind für ein Verständnis externer Validität vermutlich notwendig, jedoch nicht unbedingt hinreichend. Als weitere, möglicherweise ausschlaggebende Faktoren sind die begrenzte Sensitivität und Validität des Instruments zu diskutieren: So ist zu vermuten, dass kleinere Performanzunterschiede nicht erfasst werden können, die zwischen Klassenstufen bzw. zwischen der Treatmentgruppe EXP und der Kontrollgruppe $KG_{Öko}$ zu erwarten waren (s. Abschn. 3.1.2). In der vorliegenden Testversion bezieht die Formulierung mancher Distraktoren per se sachlogisch korrekte Argumente mit ein, um deren Plausibilität zu erhöhen (s. Abschn. 3.1.1, Abb. 1) – dies könnte die Nutzung des Verständnisses für die Bedeutung von Beobachtungsdauer und Stichprobengröße v. a. bei Versuchspersonen mit gering ausgeprägtem Verständnis verfälschend beeinflussen. Angesichts von Performanzunterschieden zwischen Schularten in diversen Bereichen experimenteller Problemlösefähigkeit (Mayer et al., 2008) lässt sich von den Realschülern unserer Studie nicht automatisch auf andere Lernende generalisieren.

Erwartungskonform positiv wirkte sich hingegen domänenspezifisches systemisches Denken auf das Verständnis für die Bedeutung der Beobachtungsdauer aus. In der untersuchten Stichprobe scheint die durch das Treatment SYS nachweislich geförderte Fähigkeit zum systemischen Denken (Bräutigam, 2014, S. 134 ff.) in größerem Maße zur korrekten Beantwortung mancher Items zu befähigen als das Treatment EXP, in dem die Versuchspersonen exemplarisch einzelne kontextualisierte Beispiele kennengelernt haben, bei denen die Beobachtungsdauer bei der Herstellung bzw. Beurteilung von externer Validität eine Rolle spielt.

doi:10.2390/zdb-v19-i1-315

5 Pädagogische Schlussfolgerungen und weiterführende Überlegungen

Falls die Befunde in der Treatmentgruppe EXP nicht in erster Linie auf die eingeschränkte Sensitivität des Testinstrument sondern auf einen geringen Kompetenzzuwachs infolge des kognitiven Entwicklungsstands zurückzuführen wären, würde dies die Überlegungen von Meyer (1978) sowie Mayer und Ziemek (2006) unterstützen, den Aufbau unterschiedlich abstrakter experimenteller Kompetenzen innerhalb der Dimensionen naturwissenschaftlicher Erkenntnisgewinnung (ebd.) gestuft und längerfristig anzugehen. Fördermaßnahmen zum Verständnis externer Validität wären vor diesem Hintergrund in höheren Klassen anzusiedeln, zumal die bedeutsame Domäne Ökologie curricular wieder aufgegriffen und vertieft wird, wodurch umfangreicheres Domänenwissen zur Verfügung stehen würde. Angesichts der Tücken von Überfrachtung durch den Lernkontext sollte die extrinsische kognitive Belastung reduziert werden (Kirschner et al., 2006; Roesch et al., 2015). Möglichkeiten, wie ein Spiralcurriculum für die systematische Förderung dieses Grundverständnisses für externe Validität gestaltet werden könnte, werden an anderer Stelle ausgeführt (Rösch, in Vorbereitung).

Angesichts der in diesem Beitrag vorgestellten Befunde sowie der in Abschnitt 2.3 und der nachfolgend erläuterten Erkenntnisse zum computergestützten Erlernen experimenteller Kompetenzen rücken zwei innovative Ansätze als vielversprechender Gegenstand weiterführender biologiedidaktischer Forschung ins Blickfeld: erstens die Förderung systemischen Denkens zur Unterstützung des Verständnisses für die Beobachtungsdauer als Kriterium für externe Validität. Zweitens erscheint der Einbezug computerbasierter interaktiver Simulationen als zielführend. Systemisches Denken kann bereits ohne den Einsatz computerbasierter Simulationen wirkungsvoll gefördert werden (Bräutigam, 2014, S. 134 ff.; Vogel et al., 2011) – dies wirkt sich positiv auf das Verständnis für die Bedeutung von Langzeitbeobachtungen bei ökologischen Experimenten aus (s. Abschn. 3.4). Wenn aber die Kombination solchen Unterrichts mit computergestützten Lernumgebungen das Training systemischen Denkens noch stärker begünstigt (Riess & Mischo, 2010) und computerbasierte Experimentierumgebungen auch die Performanz experimenteller Kompetenzen steigern helfen (Eckhardt, 2010; Urhahne & Harms, 2006), könnte die Konzeption spezieller Simulationen in ökologischen Kontexten wertvolle Synergieeffekte für ein Grundverständnis von Kriterien externer Validität hervorbringen: Nicht nur die Variablenkontrollstrategie und systematische Vorgehensweise beim Experimentieren stünden im Mittelpunkt – in Modellsystemen könnten auch Versuchsreihen mit variierbar vielen Individuen mit differenten Merkmalsausprägungen und bei zeitlicher Systemdynamik durchgeführt werden. Im Zeitraffer ließen sich Stichproben- und Langzeiteffekte sowie Vorteile von Mehrfachmessungen, Replikation und längerfristiger Beobachtung in beliebig vielen Durchgängen vor Augen führen. Logfile-Daten könnten dabei für Feedback und Diagnose sowie adaptive, prozessbegleitende Prompts zur Unterstützung der Prozessregulation genutzt werden (ebd.).

doi:10.2390/zdb-v19-i1-315

6 Ausblick

Die in diesem Beitrag vorgestellten Untersuchungen stellen erste Schritte in einem bislang wenig untersuchten Feld dar. In Anbetracht der empirischen Befunde und konzeptionellen Überlegungen ergeben sich für künftige biologiedidaktische Studien diverse Forschungsdesiderate: Um die Validität, Sensitivität und Homogenität des Messinstruments zu steigern, sollten sich die Antwortoptionen ausschließlich auf die Aspekte *Beobachtungsdauer* bzw. *Stichprobengröße* beziehen und keine anderen sachlogischen Aspekte ansprechen (s. Kap. 4). Mit Blick auf das kontextabhängige Antwortverhalten gilt es, den Item-Pool zu erweitern und in anderen Klassenstufen zu testen. Dabei sollten Zusammenhänge mit kognitiver Entwicklung, die Korrelation der Performanzwerte mit die Itemkontexte betreffendem Domänenwissen sowie die Weite des Transfers zwischen Lern- und Itemkontexten in den Blick genommen werden. Um die dimensionale Struktur und die Konstruktvalidität besser untersuchen zu können, macht es Sinn, das Testinstrument in höheren Klassenstufen zu erproben – dort sind größere interindividuelle Varianz bzw. Homogenität im individuellen Antwortverhalten zu erwarten. Um einen Einblick in den frühestmöglichen Zeitpunkt effektiver Förderung zu erhalten, empfiehlt es sich, die kognitive Belastung der Lernumgebung zu reduzieren, welche in unserer Studie u. a. durch Kontextmerkmale, Maß der Problemorientierung und Offenheitsgrad der Lernaktivitäten beeinflusst wurde.

Ob eine Lernumgebung, die ein Training systemischen Denkens *und* experimenteller Problemlösefähigkeit sowie gegebenenfalls computerbasierte Simulationen kombiniert, positive Synergieeffekte auf das Verständnis für externe Validität bewirkt oder eher zu hoher kognitiver Belastung führt, ist des Weiteren zu untersuchen. Wertvolle Einsichten zur optimalen Förderung entsprechender Kompetenzen könnten Studien liefern, die mit sensitiveren und valideren Erhebungsmethoden arbeiten und Probanden anderer Schularten und höherer Klassenstufen miteinbeziehen.

Dank

Herzlicher Dank gilt Werner Rieß für die Projektinitiative und – wie auch Josef Nerb – für die engagierte Projektbetreuung, Jana C. Gäde für die Mitarbeit bei der statistischen Auswertung, Karin Schermelleh-Engel, Janina Strohmer, Andreas Schulz und Tenko Raykov für hilfreiche Hinweise und Anregungen sowie Anna-Maria Pils für einen wichtigen Impuls zur richtigen Zeit.

Literatur

Arnold, J., Kremer, K. & Mayer, J. (2013). Wissenschaftliches Denken beim Experimentieren – Kompetenzdiagnose in der Sekundarstufe II. *Erkenntnisweg Biologiedidaktik, 11*, 7-20. http://www.bcp.fu-berlin.de/biologie/arbeitsgruppen/didaktik/Erkenntnisweg/2012/Arnold.pdf?1362740309 (22.08.2015).

doi:10.2390/zdb-v19-i1-315

Baumann, M., Simon, U., Wonisch, A. & Guttenberger, H. (2013). Computersimulation versus Experiment. Gibt es Unterschiede im Erzeugen nachhaltigen Wissens und in der Attraktivität für die Schüler? *Der mathematische und naturwissenschaftliche Unterricht* (*MNU*), *66* (5), 305-310.

Beauducel, A. & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13* (2), 186-203.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bortz, J. & Döring, N. ([4]2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer.

Bräutigam, J. I. (2014). *Systemisches Denken im Kontext einer Bildung für nachhaltige Entwicklung. Konstruktion und Validierung eines Messinstruments zur Evaluation einer Unterrichtseinheit* (Dissertation, Pädagogische Hochschule Freiburg). http://opus.bsz-bw.de/phfr/volltexte/2014/438/pdf/DissertationBraeutigamJulia2014.pdf (22.08.2015)

Bühner, M. ([3]2011). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium.

Düppers, W. (1975). Wieweit ist der Biologieunterricht experimenteller Unterricht? *Der mathematische und naturwissenschaftliche Unterricht* (*MNU*), *28* (4), 197-199.

Eckebrecht, H., Eckebrecht, D. & Kluge, S. (2006). *Experimentesammlung Sekundarstufe I. Natura: Biologie für Gymnasien*. Stuttgart: Klett.

Eckhardt, M. (2010). *Instruktionale Unterstützung beim Lernen mit Computersimulationen im Fach Biologie* (Dissertation, Universität Kiel). http://eldiss.uni-kiel.de/macau/receive/dissertation_diss_00004981 (22.08.2015).

Ehmer, M. (2008). *Förderung von kognitiven Fähigkeiten beim Experimentieren im Biologieunterricht der 6. Klasse: eine Untersuchung zur Wirksamkeit von methodischem, epistemologischem und negativem Wissen* (Dissertation, Universität Kiel). http://eldiss.uni-kiel.de/macau/receive/dissertation_diss_00003034 (22.08.2015)

Emden, M. & Sumfleth, E. (2012). Prozessorientierte Leistungsbewertung. Zur Eignung einer Protokollmethode für die Bewertung von Experimentierprozessen. *Der mathematische und naturwissenschaftliche Unterricht* (*MNU*), *65* (2), 68-75.

Freytag, K. (2007). *Biologische Kurzversuche*. 2 Bde. Köln: Aulis Verlag Deubner.

Hammann, M., Phan, T. T. H., Ehmer, M. & Grimm, T. (2008). Assessing pupils' skills in experimentation. *Journal of Biological Education*, *42* (2), 66-72.

Kauertz, A., Fischer, H. E., Mayer, J., Sumfleth, E. & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften* (*ZfDN*), *16*, 135-153.

Keselman, A. (2003). Supporting Inquiry Learning by Promoting Normative Understanding of Multivariable Causality. *Journal of Research in Science Teaching, 40* (9), 898–921.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41* (2), 75-86.

Klahr, D. & Nigam, M. (2004). The Equivalence of Learning Paths in Early Science Instruction. Effects of Direct Instruction and Discovery Learning. *Psychological Science*, *15* (10), 661-667.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (Hrsg.). (2005), *Beschlüsse der Kultusministerkonferenz. Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004*. München, Neuwied: Wolters Kluwer.

Leutner, D., Wirth, J., Klieme, E. & Funke, J. (2005). Ansätze zur Operationalisierung und deren Erprobung im Feldtest zu PISA 2000. In E. Klieme, D. Leutner & J. Wirth (Hrsg.), *Problemlösekompetenz von Schülerinnen und Schülern. Diagnostische Ansätze, theoretische Grundlagen und empirische Befunde der deutschen PISA-2000-Studie* (S. 21-36). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lienert, G. A. & Raatz, U. ([6]1998). *Testaufbau und Testanalyse*. Weinheim: Beltz Psychologie-Verlags-Union.

doi:10.2390/zdb-v19-i1-315

Mayer, J., Grube, C. & Möller, A. (2008). Kompetenzmodell naturwissenschaftlicher Erkenntnisgewinnung. In U. Harms & A. Sandmann (Hrsg.), *Ausbildung und Professionalisierung von Lehrkräften. Internationale Tagung der Fachsektion Didaktik der Biologie im VBiO, Essen 2007* (S. 63-79). Innsbruck: StudienVerlag.

Mayer, J. & Ziemek, H.-P. (2006). Offenes Experimentieren. Forschendes Lernen im Biologieunterricht. *Unterricht Biologie, 30* (317), 4-12.

McBride, B. B., Brewer, C. A., Berkowitz, A. R. & Borrie, W. T. (2013). Environmental literacy, ecological literacy, ecoliteracy: What do we mean and how did we get here? *Ecosphere, 4* (5), art. 67, 1-20. http://dx.doi.org/10.1890/ES13-00075.1 (22.08.2015)

Meyer, G. (1978). *Die Bedeutung des Experiments für den modernen Biologieunterricht*. Arbeitshilfe 1.12, hrsgg. von der Landeshauptstadt Hannover, Schulamt / Schulbiologiezentrum.

Muthén, L. K. & Muthén, B. O. (2010). Mplus User's Guide. 6th ed. Los Angeles, CA, USA: Muthén & Muthén. http://www.statmodel.com/ugexcerpts.shtml (12.05.2011).

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E. & Pekrun, R. (Hrsg.). (2007), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie. Zusammenfassung*. http://pisa.ipn.uni-kiel.de/zusammenfassung_PISA2006.pdf (26.02.2014).

Raykov, T., Dimitrov, D., & Asparouhov, T. (2010). Evaluation of Scale Reliability with Binary Measures Using Latent Variable Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 17* (2), 265-279.

Riess, W. & Mischo, C. (2010). Promoting Systems Thinking through Biology Lessons. *International Journal of Science Education, 32* (6), 705-725.

Rösch, F. (2013). Förderung experimenteller Problemlösefähigkeit!? Welche Rolle spielen aktuelle Schulbücher und Lehrerhandreichungen? *Der mathematische und naturwissenschaftliche Unterricht* (*MNU*), *66* (5), 299-305.

Rösch, F. (in Vorbereitung). *Förderung und Messung experimenteller Problemlösefähigkeit in der Orientierungsstufe. Befunde und Überlegungen unter besonderer Berücksichtigung der Domäne Ökologie und kontextbasierten Lernens* (Arbeitstitel). Online-Dokument.

Roesch, F., Nerb, J. & Riess, W. (2015). Promoting Experimental Problem-solving Ability in Sixth-grade Students Through Problem-oriented Teaching of Ecology: Findings of an intervention study in a complex domain. *International Journal of Science Education, 37* (4), 577-598.

Rösch, F., Rieß, W. & Nerb, J. (2012). Förderung „experimenteller Problemlösefähigkeit“ im problemorientierten Ökologieunterricht der 6. Klassenstufe? In W. Rieß, M. Wirtz, B. Barzel & A. Schulz (Hrsg.), *Experimentieren im mathematisch-naturwissenschaftlichen Unterricht. Schüler lernen wissenschaftlich denken und arbeiten* (S. 183-198). Münster: Waxmann.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Test of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online, 8* (2), 23-74.

Schneider, W., Bullock, M. & Sodian, B. (1998). Die Entwicklung des Denkens und der Intelligenzunterschiede zwischen Kindern. In F. E. Weinert (Hrsg.), *Entwicklung im Kindesalter* (S. 53-74). Weinheim: Beltz Psychologie-Verl.-Union.

Schulz, A., Wirtz, M. & Starauschek, E. (2012). Das Experiment in den Naturwissenschaften. In W. Rieß, M. Wirtz, B. Barzel & A. Schulz (Hrsg.), *Experimentieren im mathematisch-naturwissenschaftlichen Unterricht. Schüler lernen wissenschaftlich denken und arbeiten* (S. 15-38). Münster: Waxmann.

Smith, T. M. & Smith, R. L. (62009). *Ökologie*. München: Pearson Studium.

Urhahne, D. & Harms, U. (2006). Instruktionale Unterstützung beim Lernen mit Computersimulationen. *Unterrichtswissenschaft, 34* (4), 358-377.

Vogel, A., Rieß, W. & Nerb, J. (2011). Systemisches Denken im Umgang mit Natur. Welche subjektiven Theorien entwickeln SchülerInnen im Umgang mit komplexen Systemen? In S. Holzheu (Red.), *Didaktik der Biologie – Standortbestimmung und Perspektiven. Tagung der Fachsektion Didaktik der Biologie (FDdB) im VBiO 12.09. – 16.09.2011* (S. 66-67). Universität Bayreuth.

doi:10.2390/zdb-v19-i1-315

Kontakt
Frank Rösch
Pädagogische Hochschule Ludwigsburg, Abteilung Biologie und ihre Didaktik, Reuteallee 46, D – 71634 Ludwigsburg, roesch@ph-ludwigsburg.de

doi:10.2390/zdb-v19-i1-315

Anhang

Tabelle 7

Kontexte der in den Items vorgegebenen Experimente (Vorstudie bzw. Hauptstudie)

Experimentkontext und betrachtete Variablen			**Ext. Validität**		**Hauptstudie**
System / Organ.	**Unabhängige V.**	**Abhängige V.**	**BD**	**ST**	**(Item-Nr.)**
Wildschweine (naturnaher Lebensraum)	Umgebungslärm (Nähe von Verkehrswegen)	Schlafverhalten	+	-	ST_3
Bach	Stoffeintrag	Artenvielfalt	-	+	n. v.
Wald	Rodung	Mineralsalzkonzentration im Boden	+	+	BD_3
Hirsche (Mesokosmosexperiment)	Wolfsgeheul	Gewichtsveränderung (als Stressreaktion)	+	-	n. v.
Hauskatzen	Annäherungsrichtung eines Balls	Pfoten-‚Präferenz‘	-	-	BD_4
Flusskrebse im Aquarium	Mikrobielle Pathogene (unter manueller Herausnahme durch Aufpinseln eingebracht)	Verhaltensänderung	-	+	BD_2, ST_2
Haushühner	Stalltemperatur	Dauer der Bebrütungsphase(n)	-	-	ST_4
Würmer in limnischem Habitat	Menge sich zersetzender Biomasse	Reproduktionsrate	-	+	n. v.
Waldökosystem	Schaffung naturbelassener Freiflächen	Anzahl der Vogelarten	-	-	BD_1, ST_1 (Abb. 1)
Weizen in ‚Biosphäre‘ (Experiment im Weltall / auf der Erde)	Gravitation vs. Schwerelosigkeit	Keimung und Wachstum (Geschwindigkeit)	-	-	n. v.
Bienen (Laborexperiment)	Luftdruck	Flugverhalten	-	+	n. v.
Afrikanische Savanne	Ausbringen von Löwen-Kot entlang von Gleisen	Anzahl der Wildunfälle mit Zügen	+	-	n. v.

Anmerkungen. Organ.: Organismen. Ext. Validität: Beachtung bzw. angemessene Umsetzung von Beobachtungsdauer (BD) bzw. Stichprobengröße (ST) als Kriterien für externe Validität. +: gut beachtet / umgesetzt. -: schlecht beachtet. n. v.: in der Hauptstudie nicht verwendet. BD_i bzw. ST_j: Items für die Kriterien Beobachtungsdauer bzw. Stichprobengröße.

doi:10.2390/zdb-v19-i1-315