Physik und Didaktik in Schule und Hochschule
PhyDid 1/15 (2016), S. 49 - 63

Wie genau beurteilen Schülerinnen und Schüler ihre eigenen experimentellen Fähigkeiten?

– Ein Ansatz zur praktikablen Diagnostik experimenteller Fähigkeiten im Unterrichtsalltag –

How accurate can students assess their own experimental skills?

– An approach to feasibly assess experimental skills in the classroom –

Nico Schreiber*, Heike Theyßen*, Martin Dickmann*

*Universität Duisburg-Essen, Universitätsstraße 2, 45117 Essen, nico.schreiber@uni-due.de, heike.theyssen@uni-due.de und martin.dickmann@uni-due.de

(Eingegangen: 23.05.2016; Angenommen: 01.12.2016)

Kurzfassung

Im vorliegenden Artikel wird ausgehend von der Herausforderung einer formativen, individuellen Diagnostik experimenteller Fähigkeiten im Unterrichtsalltag ein potenziell praktikabler Ansatz zur Unterstützung dieser Diagnostik vorgeschlagen. Der Ansatz beruht darauf, dass Schülerinnen und Schüler ihre experimentelle Performanz mittels Checklisten selbst beurteilen. Solche Selbstbeurteilungen können Lehrkräfte zur Diagnostik nutzen.
Zur theoretischen Fundierung des vorgeschlagenen diagnostischen Ansatzes führen wir zunächst in die Grundlagen zur Modellierung und zur Diagnostik experimenteller Fähigkeiten sowie in das Forschungsfeld zur Genauigkeit von Selbstbeurteilungen ein. Eine zentrale Voraussetzung für die zielführende Nutzung von Schülerselbstbeurteilungen zur Diagnostik experimenteller Fähigkeiten ist, dass hierbei eine hinreichend hohe Genauigkeit erzielt werden kann. Zu dieser Fragestellung stellen wir ein Studiendesign und erste empirische Ergebnisse vor. Die Ergebnisse zeigen, dass die Fähigkeit sich selbst zu beurteilen im Mittel vergleichsweise hoch, aber individuell unterschiedlich stark ausgeprägt ist. Die sich daraus ergebenden Konsequenzen für die Diagnostik experimenteller Fähigkeiten im Unterrichtsalltag sowie weitergehende Forschungsperspektiven werden diskutiert.

Abstract

In this paper we introduce an approach to feasibly assess experimental skills in the classroom. This approach is based on students' self-assessments of their experimental performance. Such self-assessments can be used by teachers to assess experimental skills.
To supply a theoretical basis for the suggested approach, we introduce into the modelling and diagnostics of experimental skills. Further, we give an overview of the research on self-assessment accuracy. For the feasible use of self-assessments as a diagnostic tool of experimental skills, it is essential to make sure that a high accuracy of the self-assessments can be achieved. Concerning self-assessment accuracy, we present a study design and first empirical results. Results indicate that self-assessment accuracy is on average quiet high, but differs from person to person. Consequences for the assessment of experimental skills in the classroom and research perspectives are discussed.

1. Einleitung

Physiklehrkräfte sollen die experimentellen Fähigkeiten ihrer Schülerinnen und Schüler fördern (z. B. [1]; [2]). Dazu werden u. a. Schülerexperimente eingesetzt. Eine Voraussetzung für die geforderte individuelle Förderung (z. B. [3]), stellt eine individuelle Diagnostik experimenteller Fähigkeiten dar.

Ziel dieses Beitrags ist es, die Herausforderung einer praktikablen formativen Individualdiagnostik experimenteller Fähigkeiten im Unterrichtsalltag herauszustellen und einen potenziellen Ansatz zur Bewältigung dieser Herausforderung vorzustellen: die Einbeziehung von Schülerselbstbeurteilungen. Dazu stellen wir im Folgenden zunächst die theoretischen Grundlagen zur Modellierung und Diagnostik experimenteller Fähigkeiten vor und führen in das Forschungsfeld zur Genauigkeit von Selbstbeurteilungen ein. Daraus leiten wir einen potenziell praktikablen Ansatz zur Unterstützung der formativen Diagnostik ab und stellen ihn – gestützt durch die Ergebnisse einer ersten empirischen Studie – zur Diskussion.

1.1. Experimentelle Fähigkeiten

Universelle Aussagen dazu, was naturwissenschaftliches Experimentieren ist, sind kaum möglich [4]. Konsens ist, dass das Experimentieren facettenreich ist. Zu den vielfältigen Facetten des Experimentierens gehören u. a. der Kontext des Experiments, das Herstellen von Versuchsanordnungen sowie Strategien und Fertigkeiten des Experimentierens [5, S. 294f].

Im fachdidaktischen Diskurs wird stark eingeschränkt, was unter dem Begriff des Experimentierens zu verstehen ist [4]. Im Fokus stehen die Planbarkeit, Wiederholbarkeit und das Durchführen der Variablenkontrollstrategie im Rahmen von Hypothesentests (ebd.).

Auf den naturwissenschaftlichen Unterricht bezogene Modelle experimenteller Fähigkeiten sind aus verschiedenen Perspektiven entstanden, z. B. aus der Lehrerperspektive [6], der normativen Perspektive [7] oder der deskriptiven Perspektive [8]. Die Mehrheit lässt sich dem Teilprozessansatz [9] zuordnen, der das Experimentieren als einen Dreischritt von Planung, Durchführung und Auswertung modelliert [10, S. 17f]. Diesem Dreischritt werden – im Detail abhängig vom Modell – experimentelle Fähigkeiten zugeordnet, die man in der Regel sowohl in nationalen (z. B. [2, S. 11]) als auch in internationalen Curricula (z. B. [1]) findet. Einen gemeinsamen Kern dieser Modellierungen stellt die Durchführung dar, die unterschiedlich fein aufgeschlüsselt wird, aber in aller Regel die Komponenten des Aufbaus und der Messung enthält (z. B. bei [7]).

1.2. Experimentelle Fähigkeiten im Unterrichtsalltag diagnostizieren

Vorliegende Methoden zur Diagnostik experimenteller Fähigkeiten aus der empirischen, fachdidaktischen Forschung sind im Unterrichtsalltag eher ungeeignet und bieten wenig Hilfestellung für eine formative Individualdiagnostik: Eine Diagnostik durch schriftliche Tests ist zumindest im Bereich der konkreten Durchführung von Experimenten wenig valide (z. B. [11]; [12]). Zu aufwendig sind Testverfahren wie Experimentiertests mit Realexperimenten oder Simulationen. Insbesondere eine Diagnostik anhand verhaltensbasierter Maße (z. B. [13]; [14]) ist mit Realexperimenten in der Schule nicht praktikabel. Zudem sind die vorliegenden Verfahren für eine summative Diagnostik konzipiert.

Videoaufzeichnungen von Schülerinnen und Schülern beim Experimentieren können Lehrkräften detailreiche Informationen über die experimentelle Performanz der Schülerinnen und Schüler liefern. Sie sind jedoch im Unterrichtsalltag zu aufwendig zu realisieren und zu analysieren.

Um im Unterrichtsalltag möglichst viele Daten beim Experimentieren zu sammeln, bleibt Lehrkräften die Möglichkeit, die Versuchsprotokolle als Datenquelle heranzuziehen. Klassische Versuchsprotokolle geben gute Informationen über die Planung und die Auswertung eines Experiments [15, S. 29f]. Aus den Versuchsprotokollen wird allerdings nicht deutlich, wie die Schülerinnen und Schüler tatsächlich beim praktischen Aufbauen und Messen vorgegangen sind. Um dem abzuhelfen, schlagen Emden und Sumfleth [16] die Nutzung von Prozessprotokollen vor, die während des Experimentierens von den Schülerinnen und Schülern erstellt werden und genauere Auskunft über den Ablauf des Experiments geben können. Zur Auswertung der Prozessprotokolle sind jedoch Prozessgrafiken anzufertigen „was zeitaufwendig ist und im Unterrichtsalltag zu Problemen führen könnte“ [16, S. 74]. Alternativ müssten Lehrkräfte alle Schülerinnen und Schüler während des Experimentierens systematisch und genau beobachten, um deren Fähigkeiten zum praktischen Aufbauen und Messen zu diagnostizieren. In der Regel experimentieren jedoch ca. 30 Schülerinnen und Schüler in Partner- oder Kleingruppenarbeit. Beobachtungen durch die Lehrkraft werden daher im Unterrichtsalltag meist unsystematisch bleiben. Die geringe und lückenhafte Basis für die individuelle Diagnostik lässt erwarten, dass es zu Urteilstendenzen und Urteilsfehlern kommt, wie sie in anderen Zusammenhängen nachgewiesen wurden (z. B. [17, S. 138]). Hier schlagen Maiseyenka und Kollegen [18] vor, die direkte Beobachtung durch ein Diagnoseraster (sog. "Spinnennetzmodell") zu unterstützen. Dieses wird jedoch von Lehrkräften kaum genutzt [18, S. 11]. Hier wünschen Lehrkräfte weitere Konkretisierungen [19, S. 9]. Denn tatsächlich stehen die Lehrkräfte auch mit diesem Diagnoseraster vor der Herausforderung, geeignete Daten über ihre Schülerinnen und Schüler beim Experimentieren zu sammeln. Im Allgemeinen ist die Genauigkeit von Lehrerurteilen zwar einigermaßen hoch (Metaanalyse von [20, S. 755]), es besteht aber Potenzial zur Verbesserung [ebd.].

Somit stellt sich die Frage, wie man im Unterrichtsalltag die Datenbasis für die Diagnostik möglichst praktikabel erweitern kann. Eine Möglichkeit stellen Selbstbeurteilungen der Schülerinnen und Schüler dar. In einer Lernsituation kann man davon ausgehen, dass Schülerinnen und Schüler wertvolle Informationen zu eigenen Stärken und Schwächen geben können [21, S. 23f]. Diese Informationen können von der Lehrkraft als Erweiterung der Datenbasis für ihre Diagnostik praktikabel erhoben und genutzt werden [22]. Aus der breiteren Datenbasis könnten potenziell genauere Diagnosen resultieren. Um diese Informationen zu generieren sind Selbstbeurteilungen notwendig, die mit Checklisten erhoben werden können. Der skizzierte Ansatz zur Erweiterung der Datenbasis beruht allerdings auf der Voraussetzung, dass die Selbstbeurteilungen der Schülerinnen und Schüler möglichst genau sind.

Ferner können die Selbstbeurteilungen im Rahmen des Self-Monitorings beim selbstregulierten Lernen experimenteller Fähigkeiten von den Schülerinnen und Schülern selbst genutzt werden, um den weite-

ren Lernprozess zu steuern. Eine genaue Selbstbeurteilung wirkt positiv auf die Selbstregulation ([23]; [24]). Die effektivere Selbstregulation steht dann im Zusammenhang mit einem erhöhten Lernzuwachs [25]. Kompetente Lernende sollten daher in der Lage sein, das Resultat ihres Lernprozesses möglichst genau selbst beurteilen zu können.

Im Folgenden wird der theoretische Hintergrund zu Selbstbeurteilungen und deren (potenzieller) Genauigkeit ausgeführt.

2. Selbstbeurteilungen und Genauigkeit von Selbstbeurteilungen

Im Allgemeinen unterscheidet man zwischen prospektiven und retrospektiven Selbstbeurteilungen [26, S. 8]. Bei einer prospektiven Selbstbeurteilung beurteilen Lernende auf Basis ihrer Lernerfahrung, wie gut sie bei einer anschließenden Aufgabenbearbeitung in einer Testsituation abschneiden werden. Diese Art der Selbstbeurteilung ist unter dem Begriff *Judgment of Learning (JOL)* bekannt (z. B. [27]; [28]). Im Unterschied zum *JOL* wird beim *self-assessment* (z. B. [29]) eine Selbstbeurteilung nach einer Aufgabenbearbeitung eingeholt (retrospektive Selbstbeurteilung).

Zur Bestimmung der Genauigkeit von Selbstbeurteilungen werden Selbstbeurteilungen mit Urteilen von einem externen Beurteilungsmaßstab verglichen. Je stärker beide Urteile übereinstimmen, desto höher ist die Genauigkeit der Selbstbeurteilung (Details zu Maßen der Urteilsgenauigkeit s. Abschnitt 3.2).

Zur Genauigkeit von Selbstbeurteilungen liegen zahlreiche Befunde aus unterschiedlichen Domänen und Forschungsrichtungen vor (u. a. Classroom Assessment: [30]; Higher Education: [31]; Fremdsprachen: [32]; Calibration-Forschung: u. a. [33]; [34]).

Einen Überblick über den Forschungsstand zur Genauigkeit von Selbstbeurteilungen gibt die Metasynthese von Zell und Krizan [35]. Zell und Krizan analysierten 22 Metaanalysen aus verschiedenen Domänen. Die in den Metaanalysen untersuchten Studien beschäftigen sich mit der Genauigkeit von prospektiven und retrospektiven Selbstbeurteilungen. Die Teilnehmenden waren im Mittel mindestens 13 Jahre alt. Gegenstand der Selbstbeurteilung waren Fähigkeiten, Fertigkeiten, Fachwissen, Selbstkonzept, Selbstwirksamkeit, Selbstachtung und Selbstbewusstsein. Die Selbstbeurteilung wurde mit einem externen Beurteilungsmaßstab (Noten, Tests, Expertenurteile) verglichen und ein Zusammenhang quantitativ angegeben. Als Maß für die Genauigkeit der Selbstbeurteilung wählen Zell und Krizan die Korrelation zwischen Selbstbeurteilung und externem Beurteilungsmaßstab. Sie beträgt über alle Metaanalysen hinweg $\overline{r}$ = .29 (SD = .11). In 21 Studien lagen die Korrelationen zwischen .09 und .39. Nur eine Metaanalyse zur Sprachkompetenz stach mit $\overline{r}$ = .63 [32] heraus. Die Genauigkeit von Selbstbeurteilungen ist damit im Allgemeinen als moderat einzuschätzen.

Zur Genauigkeit bei Selbstbeurteilungen im Kontext des Experimentierens liegen bisher nur wenige Befunde vor (z. B. [36]; [37]). Diese deuten zunächst auf eine zufriedenstellende Genauigkeit von Selbstbeurteilungen hin. Stefani [36] untersuchte die Urteilsgenauigkeit von Studierenden im biologischen Anfängerpraktikum. Zu Beginn des Praktikums entwickelten Praktikumsleitung und Studierende gemeinsam die Beurteilungskriterien, um damit am Ende den Praktikumserfolg zu beurteilen. Am Ende des Semesters beurteilten zunächst die Lehrenden 87 Versuchsprotokolle. Anschließend beurteilten die Studierenden ihre eigenen Versuchsprotokolle ohne Kenntnis der Beurteilung durch die Lehrenden. Stefani berichtet eine Korrelation von r = .93 zwischen den Selbstbeurteilungen der Studierenden und den Fremdbeurteilungen durch die Lehrenden. Die eigene experimentelle Performanz war hier nicht Gegenstand der Selbst- und Fremdbeurteilung.

In einer Studie von van der Jagt et al. [37] wurde untersucht, ob sich die Qualität eines Experiments durch Selbstbeurteilungen mit Checklisten verbessert. 24 Oberstufenschülerinnen und -schüler führten selbst geplante, offene Experimente durch und beurteilten sich selbst anhand einer Checkliste. Die Checkliste bestand aus 19 Beurteilungskriterien zur Genauigkeit, Reliabilität und Validität der Vorgehensweise beim Experimentieren. Für jedes Beurteilungskriterium wurden fünf Qualitätsstufen definiert. Um die Genauigkeit der Selbstbeurteilungen zu bestimmen, wurden die Selbstbeurteilungen mit Expertenurteilen (Fachdidaktiker) verglichen. Die Selbstbeurteilungen basierten auf den Beobachtungen der eigenen Performanz. Die Expertenurteile basierten auf den Versuchsprotokollen. Van der Jagt et al. wählen als Maß für die Urteilsgenauigkeit die prozentuale Übereinstimmung zwischen Schülerselbstbeurteilung und Expertenurteil. Bei sechs Beurteilungskriterien lag die Übereinstimmung bei über 80 %. Zwei Kriterien wiesen Übereinstimmungen von unter 40 % auf. Bei allen anderen elf Kriterien lagen Übereinstimmungen zwischen 49 % und 70 % vor.

3. Erhebung von Selbstbeurteilungen und Maße der Urteilsgenauigkeit

3.1. Checklisten zur Erhebung von Selbstbeurteilungen

Um Selbstbeurteilungen zu erheben, hat es sich etabliert, Checklisten einzusetzen, die das zu diagnostizierende Konstrukt durch Beurteilungskriterien operationalisieren (z. B. [38]; [39]). Dabei führen Checklisten die Beurteilungskriterien explizit auf und fordern dazu konkrete Beurteilungen durch klare Arbeitsaufträge ein ([40]; [41]).

Checklisten kommen häufig zur Beurteilung von Produkten, wie beispielsweise Versuchsprotokollen,

Essays und Werkstücken, zum Einsatz. Die Beurteilung von Prozessen ist seltener Gegenstand der Diagnostik mit Checklisten [40].

Auch zur Diagnostik experimenteller Fähigkeiten liegen veröffentlichte Checklisten vor (z. B. [37]; [42]; [43]). Diese sind in der Regel nicht aufgabenspezifisch formuliert und dienen meist der Fremdbeurteilung. Vogt, Müller und Kuhn [44] analysierten die Konzeption vorliegender Checklisten zum Experimentieren und formulierten als Resultat eine aufgabenunspezifische Checkliste mit 51 Kriterien, die im Hinblick auf Inhaltsvalidität und konvergente Validität sowie Reliabilität und Objektivität überprüft wurde. Allerdings dient diese Checkliste nicht der Selbstbeurteilung, sondern der Beurteilung von Versuchsprotokollen durch Lehrende an der Universität. Für Selbstbeurteilungen beim Experimentieren liegen beispielsweise Checklisten von Heinicke und Bellingrath [45], Schreiber und Nawrath [46], Struck [47] sowie van der Jagt et al. [37] vor.

3.2. Maße der Urteilsgenauigkeit

Um die Genauigkeit von Fremd- bzw. Selbstbeurteilungen zu ermitteln, wird ein externer Beurteilungsmaßstab benötigt. Dazu werden in der Regel Expertenurteile (z. B. von Lehrkräften) oder schriftliche Tests herangezogen. Die Urteile aus dem externen Beurteilungsmaßstab werden mit der Selbst- bzw. Fremdbeurteilung verglichen. Je näher die Selbst- bzw. Fremdbeurteilung am Urteil des externen Beurteilungsmaßstabes liegt, desto genauer gilt sie [48, S. 202].

Zur Bestimmung der Urteilsgenauigkeit ist ein korrelativer Ansatz verbreitet ([48, S. 202]; [49]). Dabei berechnet man eine Korrelation zwischen Selbst- bzw. Fremdbeurteilungen und den Urteilen auf Basis des externen Beurteilungsmaßstabes. Eine hohe Korrelation bedeutet eine hohe Urteilsgenauigkeit. Zu unterscheiden ist, ob eine Korrelation für jeden Urteilenden separat (Maß für die Genauigkeit eines Urteilenden – Individualebene) oder über eine untersuchte Stichprobe (Maß für die Genauigkeit einer Gruppe von Urteilenden – Gruppenebene) berechnet wird. Beispielsweise untersuchten Anders et al. [50] die Genauigkeit von Lehrerurteilen (Fremdbeurteilung), indem sie für jede Lehrkraft eine Rangkorrelation für den Zusammenhang zwischen Schülerleistung und Lehrerurteil berechneten. Im Unterschied dazu berechnete beispielsweise Stefani [36] genau eine Korrelation zwischen Lehrenden und Studierenden über die gesamte Stichprobe hinweg, um die Urteilsgenauigkeit von Studierenden im biologischen Praktikum zu bestimmen (s. Abschnitt 2).

Ein Nachteil am korrelativen Ansatz ist, dass Korrelationen keine Auskunft über Urteilstendenzen geben [48, S. 205]. Urteilstendenzen können durch Differenzen ($\Delta_T = X_{Urteil} - X_{Maßstab}$; Definitionen der Abkürzungen siehe Tab. 1) ausgedrückt werden. Ist $\Delta_T = 0$, dann ist die Urteilsgenauigkeit optimal. Differenzen sind als Über- oder Unterschätzungen zu interpretieren.

Wenn ein Selbst- bzw. Fremdbeurteiler mehrere Beurteilungen durchführt, gibt die durchschnittliche Differenz die Urteilstendenz an (s. Tab. 1), d. h. die mittlere Über- oder Unterschätzung. Allerdings können sich hierbei positive und negative Abweichungen kompensieren. Die einzelnen Über- und Unterschätzungen berücksichtigt der sogenannte Urteilsfehler, der die mittlere betragsmäßige Abweichung anzeigt (s. Tab. 1, Urteilsfehler). Voraussetzung zum Bilden von Differenzen ist, dass die verwendeten Skalen bei Beurteilungsmaßstab und Urteilenden identisch sind [48, S. 203].

Urteilstendenz (Über- oder Unterschätzung)	Urteilsfehler
$\Delta_T = \frac{\sum(x_{Urteil} - x_{Maßstab})}{n}$	$\Delta_F = \frac{\sum \lvert x_{Urteil} - x_{Maßstab} \rvert}{n}$
n: Anzahl der Urteile, x_{Urteil}: z. B. Selbstbeurteilung, Lehrerurteil, $x_{Maßstab}$: externer Beurteilungsmaßstab	

Tab. 1: Urteilstendenz und Urteilsfehler, modifiziert aus [51, S. 646]

4. Fragestellung

Die oben skizzierten Ansätze zur Unterstützung der formativen Individualdiagnostik sowie des selbstregulierten Lernens experimenteller Fähigkeiten im Unterrichtsalltag setzen möglichst genaue Schülerselbstbeurteilungen voraus. Die Frage, ob diese Voraussetzung erfüllt ist, kann jedoch anhand des bisherigen Forschungsstandes nicht beantwortet werden:

Im Allgemeinen deuten die Befunde aus Metaanalysen auf eine moderate Genauigkeit von Selbstbeurteilungen hin. Allerdings zeigen einzelne Studien (z. B. [32]), dass vergleichsweise genaue Selbstbeurteilungen möglich sind.

Zur Genauigkeit von Selbstbeurteilungen beim Experimentieren liegen nur wenige Befunde vor, die sich zudem auf ältere Zielgruppen oder abweichende Zielsetzungen (u. a. summative Beurteilung) beziehen, sodass ihre Übertragbarkeit zu hinterfragen ist. Ferner wurde die Urteilsgenauigkeit nur auf Gruppenebene berechnet. Im Zusammenhang mit dem Experimentieren liegen den Autoren keine Befunde zur Urteilsgenauigkeit auf Individualebene vor.

Daher muss vor einer weiteren Vertiefung des oben skizzierten Ansatzes zunächst die folgende Fragestellung empirisch geklärt werden:

Wie hoch ist die Urteilsgenauigkeit von Schülerinnen und Schülern bei der Selbstbeurteilung der eigenen Performanz beim Experimentieren?

Da sich aus dem Stand der Forschung keine belastbaren Hypothesen zu dieser Fragestellung ableiten lassen, wird die Fragestellung zunächst explorativ, mit der im Folgenden vorgestellten Studie bearbeitet. Da die Problematik der formativen Individualdiagnostik insbesondere für die konkrete Durchführung der Experimente besteht (s. Abschnitt 1), fokussiert diese Studie auf den Aufbau von Experimenten und die Durchführung von Messungen.

5. Design der Studie

Um die in Abschnitt 4 aufgeworfene Fragestellung zu bearbeiten, müssen Schülerinnen und Schüler in Einzelarbeit experimentieren und dabei eine funktionstüchtige Versuchsanordnung aufbauen sowie Messungen durchführen und dokumentieren. Anschließend müssen sie ihre experimentelle Performanz selbst beurteilen. Außerdem muss die Urteilsgenauigkeit auf Individualebene bestimmt werden können. Dies führt zu den folgenden grundlegenden Anforderungen an das Studiendesign (Abschnitt 5.1), die für die vorliegende Studie konkretisiert werden (Abschnitt 5.2).

5.1. Grundlegende Anforderungen

Aufgabenstellungen: Zur Bestimmung der Urteilsgenauigkeit sind Aufgabenstellungen notwendig, bei deren Bearbeitung die Schülerinnen und Schüler experimentelle Performanz in Aufbau und Messung zeigen können. Die inhaltlichen und methodischen Anforderungen müssen curricular valide sein und bei den Schülerinnen und Schülern die intendierten, experimentbezogenen kognitiven Prozesse auslösen. Vor dem Hintergrund des zu diagnostizierenden Konstrukts experimenteller Fähigkeiten scheint es notwendig, mehrere Aufgaben einzusetzen, um eine reliable Beurteilung der Schülerfähigkeiten zu erreichen. Hierzu liegen unterschiedliche Empfehlungen vor, die zwischen 8 und 23 Aufgaben liegen, um eine Reliabilität von .80 zu erreichen (z. B. [52, S. 1051]).

Externer Maßstab: Zur Bestimmung der Urteilsgenauigkeit wird die Selbstbeurteilung mit einem externen Maßstab verglichen. Varianz in der Urteilsgenauigkeit sollte dabei möglichst weitgehend auf die unterschiedlich ausgeprägte Fähigkeit der Schülerinnen und Schüler zur Selbstbeurteilung zurückzuführen sein. Deshalb sollte der externe Maßstab keine zusätzliche "Fehlervarianz" erzeugen, sondern – zusammen mit den Aufgabenstellungen – eine möglichst objektive, reliable und valide Beurteilung der experimentellen Fähigkeiten erlauben. Lehrerurteile stellen in dieser Hinsicht nicht unbedingt einen geeigneten externen Maßstab dar (Details zur Kritik an Expertenurteilen als Beurteilungsmaßstab z B. bei [49, S. 67]).

Selbstbeurteilungen: Für die an der Studie teilnehmenden Schülerinnen und Schüler darf nicht der Eindruck entstehen, dass sie sich in einer Testsituation befinden oder unter Leistungsdruck stehen. Dazu muss transparent kommuniziert werden, dass ihre experimentelle Performanz nicht benotet wird und auch die Selbstbeurteilungen nicht zur Benotung herangezogen werden. Es wäre mit eher ungenauen Selbstbeurteilungen zu rechnen, wenn die Selbstbeurteilung im Rahmen einer summativen Diagnostik erhoben und zur Beurteilung verwendet würde. Denn in einer solchen Situation ist davon auszugehen, dass sich Schülerinnen und Schüler möglichst gut darstellen wollen (z. B. [29]).

Außerdem sollten die Selbstbeurteilungen unmittelbar nach jeder Aufgabenbearbeitung erhoben werden, um die Urteilsgenauigkeit bezogen auf die konkreten Aufgabenstellungen zu erfassen.

Maß für die Urteilsgenauigkeit: Um die Ergebnisse in den Forschungsstand zur Urteilsgenauigkeit einordnen zu können, bietet sich als Maß der Urteilsgenauigkeit eine Korrelation an. Dabei ist vorteilhaft, dass dieses Maß normiert ist und Konventionen zur Beurteilung seiner Ausprägung vorliegen. Wenn als Maß der Urteilsgenauigkeit für jede Schülerin und jeden Schüler eine Korrelation berechnet werden soll (Individualebene), dann sind mehrere Aufgabenstellungen notwendig.

5.2. Umsetzung des Studiendesigns

Als Basis für das Zeigen experimenteller Performanz wurden Aufgabenstellungen aus dem computerbasierten MEK-LSA-Experimentiertest [53] genutzt, da die in 5.1 genannten Anforderungen an Aufgabenstellungen und Bewertungsmaßstab für diesen Test bereits umfangreich empirisch überprüft wurden (siehe 5.2.1). Die Schülerinnen und Schüler bearbeiteten in Einzelarbeit jeweils 24 experimentelle Aufgabenstellungen aus diesem Test. Unmittelbar im Anschluss an jede Aufgabenbearbeitung beurteilen sie selbst die eigene experimentelle Performanz (siehe 5.2.2). Prozessdaten der Aufgabenbearbeitungen dienten als Grundlage für die externe Beurteilung der experimentellen Performanz (siehe 5.2.3).

Die an der Studie freiwillig teilnehmenden Schülerinnen und Schüler erhielten die Information, dass ihre Leistungen beim Experimentieren nicht benotet und auch die Selbstbeurteilungen nicht zur Benotung herangezogen werden. So sollte deutlich werden, dass sich die Schülerinnen und Schüler nicht in einer Testsituation befinden.

Als Maß für die Urteilsgenauigkeit wird eine Rangkorrelation gewählt (siehe 5.2.4). Aufgrund der Einzelarbeit und der hinreichend großen Anzahl an Aufgabenstellungen pro Person kann die Urteilsgenauigkeit sowohl auf Individualebene (Korrelation der Selbst- und Fremdeinschätzung über die 24 Aufgabenbearbeitungen hinweg) als auch auf Gruppenebene (Korrelation der über jeweils 24 Aufgabenstellungen gemittelten Selbst- und Fremdeinschät-

zungen über die gesamte Stichprobe hinweg) berechnet werden (siehe 5.2.3).

5.2.1. Aufgabenstellungen mit Simulationen

Die Aufgabenstellungen des MEK-LSA-Experimentiertests erfordern die Planung, Durchführung oder Auswertung typischer Schülerexperimente zur Mechanik, Optik oder Elektrizitätslehre der Sekundarstufe I. Sie sind zusammengefasst zu Rahmenaufgaben, sogenannten "Units". Jede Unit beginnt mit der Einführung einer experimentell zu klärenden Fragestellung, z. B. ob bei einer Glühlampe Stromstärke und Spannung proportional sind (Abb. 1). Es folgen je zwei Aufgabenstellungen zur Planung, Durchführung und Auswertung eines Experiments, mit dem diese Fragestellung geklärt werden kann. Die Aufgabenstellungen zur Durchführung verlangen den funktionsfähigen Aufbau des entsprechenden Experiments und die Durchführung und Dokumentation der Messungen. Zur Sicherung der lokalen statistischen Unabhängigkeit enthält jede Aufgabe die zur Bearbeitung notwendigen Zwischenlösungen der vorangegangenen Aufgabenstellungen dieser Unit. Beispielsweise wird für die Aufgabe zum funktionsfähigen Aufbau die Planung mit Skizze, Geräteauswahl und Beschreibung der geplanten Vorgehensweise vorgegeben. Für die Aufgabe zur Durchführung der Messung werden der funktionsfähige Aufbau und eine vorbereitete Messtabelle vorgegeben. Aufbau und Messung erfolgen in dem computerbasierten Format anhand interaktiver Simulationen, die in ihren Handlungsmöglichkeiten einer realen Experimentiersituation möglichst nahekommen (Abb. 2).

In mehreren Validierungsstudien konnte gezeigt werden, dass die Aufgabenstellungen (zusammen mit dem in 5.3.3 beschriebenen Beurteilungsmaßstab) trotz des computerbasierten Formats zur objektiven, reliablen und validen Messung experimenteller Fähigkeiten eingesetzt werden können ([54]; [55]).

Für die hier beschriebene Studie, die auf die Durchführung von Experimenten fokussiert, wurden aus allen Units lediglich die beiden Aufgaben zur Durchführung der Experimente (Aufbau und Messung) ausgewählt. Den Schülerinnen und Schülern wurde aus jeder Unit zuerst die Einführung der experimentellen Fragestellung (Abb. 1) präsentiert. Im Anschluss bearbeiteten die Schülerinnen und Schüler die Aufgabenstellungen zu Aufbau und Messung.

Vor Beginn der Datenerhebung wurden die Schülerinnen und Schüler anhand einer Trainingsaufgabe in den Umgang mit den Aufgabenstellungen, insbesondere mit den interaktiven Simulationen, eingeführt.

5.2.2. Selbstbeurteilungen experimenteller Performanz

Nach jeder Aufgabe beurteilten die Schülerinnen und Schüler ihre experimentelle Performanz auf einer siebenstufigen Likert-Skala (Abb. 3 und 4). Durch die erkennbare Zuweisung der Zahlen zu den Ankreuzoptionen wurde eine möglichst metrische Interpretation der Skala durch die Schülerinnen und Schüler angestrebt.

Aus pragmatischen Gründen wurde keine aufgabenspezifische Checkliste mit mehreren Beurteilungskategorien verwendet.

Aufgrund der Vertrautheit der Schülerinnen und Schüler mit Checklisten zur Selbstbeurteilung experimenteller Fähigkeiten (s. Abschnitt 5.3) wurde auf ein Training im Umgang mit den Selbstbeurteilungen vor Beginn der Datenerhebung verzichtet.

5.2.3. MEK-LSA-Experimentiertest Beurteilungsmaßstab für Simulationen

Als externer Maßstab für die Beurteilung der experimentellen Performanz dient der für den Test entwickelte Beurteilungsmaßstab [53]. Während der Aufgabenbearbeitung werden auf dem Server detaillierte Navigationsdaten gespeichert, die die Rekonstruktion aller Eingaben und aller Aktionen innerhalb der interaktiven Simulationen zulassen. Anhand dieser Rekonstruktionen wird die Performanz pro Person und pro Aufgabenstellung auf einer dreistufigen Skala beurteilt: geeignet (Stufe 2), teilweise geeignet (Stufe 1) und ungeeignet (Stufe 0)[1]. In einem ausführlichen Kodiermanual sind für jede Aufgabe inhaltliche Merkmale angegeben, anhand derer die Stufen zugewiesen werden. Beim Aufbau unterscheiden sich teilweise geeignete von ungeeigneten Lösungen z. B. dadurch, dass bei teilweise geeigneten Lösungen ein korrekter Grundaufbau, z. B. ein geschlossener Stromkreis mit Glühlampe, vorliegt. Von geeigneten Lösungen unterscheiden sie sich dadurch, dass bei teilweise geeigneten Lösungen zwar der Grundaufbau korrekt ist, jedoch damit noch keine Messung möglich wäre, z. B. weil Stromstärke- oder Spannungsmessgeräte falsch in den Stromkreis eingebaut sind. Die Zuweisung der Stufen gelingt mit zufriedenstellender Beurteilerübereinstimmung ($.63 < \kappa < 1$, mittleres $\kappa = .81$). Die Auswertung der Aufgaben zur Messung erfolgt anhand der Navigationsdaten automatisiert. Hierbei werden insbesondere der Messbereich und die Messintervalle beurteilt.

5.2.4. Datenauswertung: Berechnung der Urteilsgenauigkeit

Die Urteilsgenauigkeit wird auf Gruppen- und Individualebene berechnet. Datengrundlage zur Berechnung der Urteilsgenauigkeit sind jeweils die siebenstufigen Selbstbeurteilungen der Schülerinnen und Schüler einerseits und andererseits die dreistufigen Beurteilungen aus dem MEK-LSA-Experimentiertest-Beurteilungsmaßstab. Bei acht Schülerinnen und Schülern fehlt jeweils eine Selbstbeurteilung, sodass

[1] Diese Werte werden im Folgenden als Testwerte bezeichnet.

die zugehörigen Testwerte gestrichen wurden. Für diese acht Schülerinnen und Schüler liegen folglich 23 Selbstbeurteilungen und 23 korrespondierende Testwerte vor.

Individualebene: Auf Individualebene wird die Urteilsgenauigkeit für jede Schülerin bzw. jeden Schüler angegeben. Dazu wird die Rangkorrelation (Spearmans ρ) über die vorliegenden Selbstbeurteilungen und die Testwerte berechnet.

Gruppenebene: Zur Bestimmung der Urteilsgenauigkeit auf Gruppenebene wird für jede Schülerin bzw. jeden Schüler jeweils ein Mittelwert für alle vorliegenden Selbstbeurteilungen und ein Mittelwert für alle Testwerte berechnet. Dabei werden beide Skalen metrisch interpretiert (s. Diskussion zum Skalenniveau in [56, S. 181f]). Somit liegen auf Gruppenebene jeweils 28 Mittelwerte für Selbstbeurteilungen und Testwerte vor. Zur Bestimmung der Urteilsgenauigkeit der Gruppe wird die Rangkorrelation (Spearmans ρ) über die (mittleren) Selbstbeurteilungen und die Testwerte berechnet. Diese Vorgehensweise setzt voraus, dass die Selbstbeurteilungen und die Testwerte reliable Skalen bilden. Letzteres wurde für den Test zwar bereits gezeigt, ist aber aufgrund der spezifischen Aufgabenauswahl noch einmal zu prüfen.

Auf eine Berechnung des Urteilsfehlers oder von Urteilstendenzen muss verzichtet werden, weil aufgrund der unterschiedlichen Anzahl an Stufen bei der Selbstbeurteilungsskala und bei dem externen Beurteilungsmaßstab keine interpretierbare Differenz berechnet werden kann.

5.3. Stichprobe

Die Gelegenheitsstichprobe besteht aus 13 Schülerinnen und 15 Schülern, die ca. 16 Jahre (Median) alt sind. Sie besuchen die 10. Jahrgangsstufe einer städtischen, integrierten Gesamtschule in Nordrhein-Westfalen. Die Physiknote der Schülerinnen und Schüler beträgt 3,0 (Median) und die Mathematiknote 2,5 (Median). Insgesamt ist von einer eher leistungsstarken Gesamtschulklasse auszugehen, weil zwei Drittel der Schülerinnen und Schüler am Ende von Jahrgangsstufe 10 die Qualifikation für die gymnasiale Oberstufe erreichten.

Die Selbstkonzepte bezogen auf Physik und Experimentieren (erhoben in Anlehnung an Skalen nach [57]) sind eher hoch ausgeprägt. Zudem geben die Schülerinnen und Schüler an, dass sie alle zwei bis drei Stunden im Physikunterricht experimentieren, was auf Erfahrung mit physikalischen Experimenten hindeutet. Auf Erfahrung mit naturwissenschaftlichen Experimenten deutet auch hin, dass 19 Schülerinnen und Schüler seit der sechsten Jahrgangsstufe Naturwissenschaften als Wahlpflichtfach belegen. Außerdem sind die Schülerinnen und Schüler erfahren mit Selbstbeurteilungen, weil die Klasse im gesamten Schuljahr mit Checklisten zum Experimentieren gearbeitet hat [22].

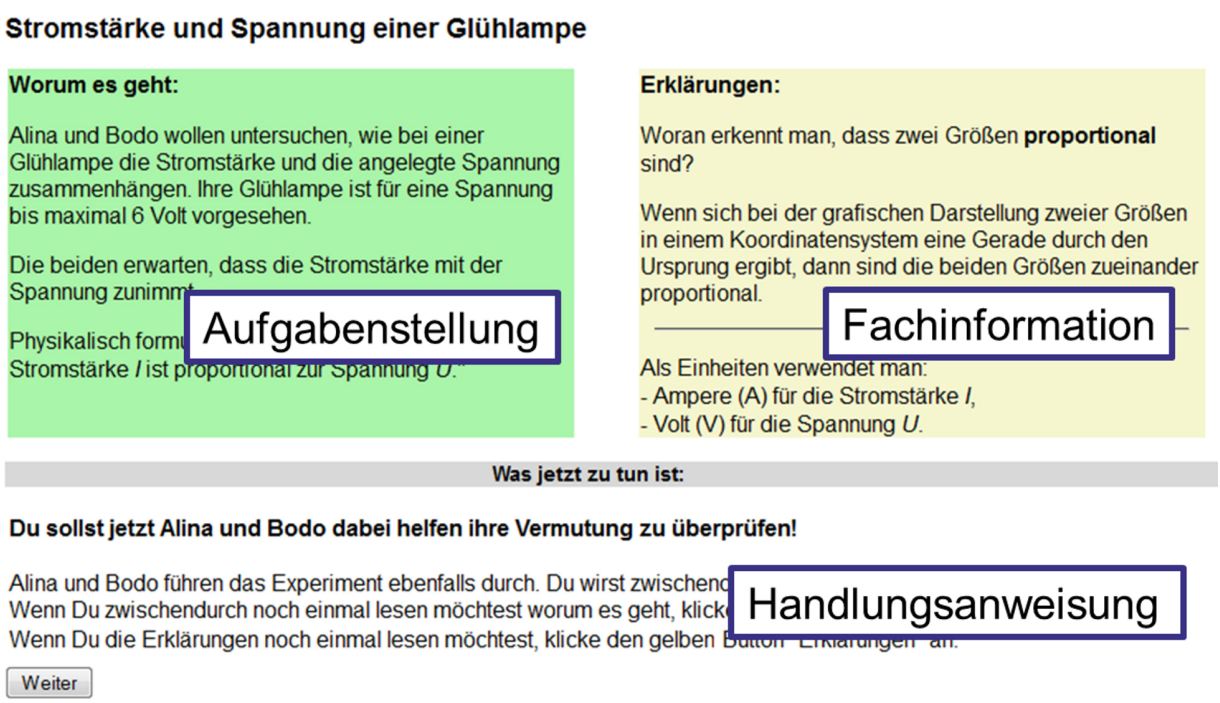

Abb. 1: Aufgabenstamm zur beispielhaften Aufgabenstellung mit dem Thema „Stromstärke und Spannung einer Glühlampe“ (nach [47]).

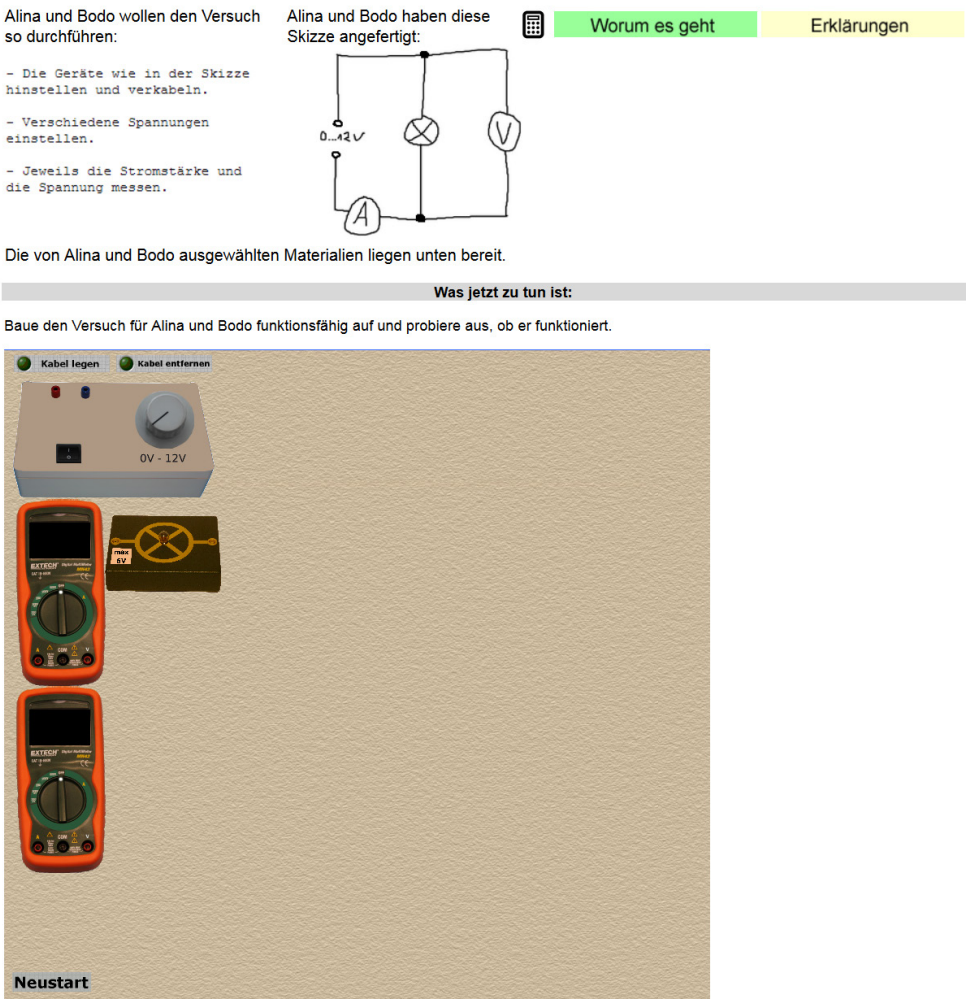

Abb. 2: Aufbauitem zur beispielhaften Aufgabenstellung mit dem Thema „Stromstärke und Spannung einer Glühlampe“ (nach [47]).

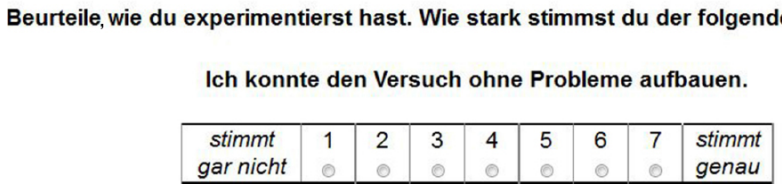

Abb. 3: Selbstbeurteilung nach der Bearbeitung einer Aufgabe zum Aufbauen (Screenshot).

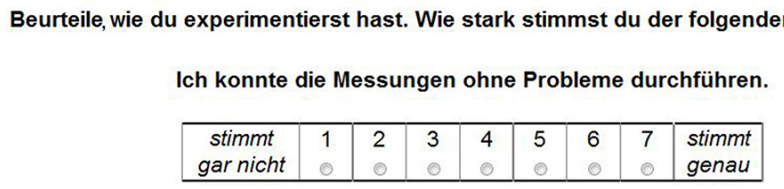

Abb. 4: Selbstbeurteilung nach der Bearbeitung einer Aufgabe zum Messen (Screenshot).

6. Ergebnisse

6.1. Gruppenebene

Als Maß für die interne Konsistenz der Selbstbeurteilungen bzw. der Testwerte wird Cronbachs α berechnet. Für die Selbstbeurteilungen beträgt $\alpha = .93$, für den externen Beurteilungsmaßstab .72. Somit dürften die α-Werte für den Zweck der Studie zufriedenstellend sein (s. Reliabilitätsanforderungen [56, S. 199]).

Die Urteilsgenauigkeit, ausgedrückt durch Spearmans ρ, beträgt .63***.

6.2. Individualebene

Auf Individualebene liegt eine breite Streuung an Korrelationen vor (s. Abb. 5). Die Korrelationen liegen im Bereich zwischen $\rho = -.58^{**}$ und $\rho = .80^{***}$. An den Rändern des Bereichs liegen die insgesamt neun signifikanten Korrelationen. Unabhängig von der Signifikanz der Korrelationen, gibt es nach Cohens Konvention [58] insgesamt zwölf niedrige, sechs mittlere und vier hohe positive Korrelationen. Außerdem liegen sechs Korrelationen im negativen Bereich.

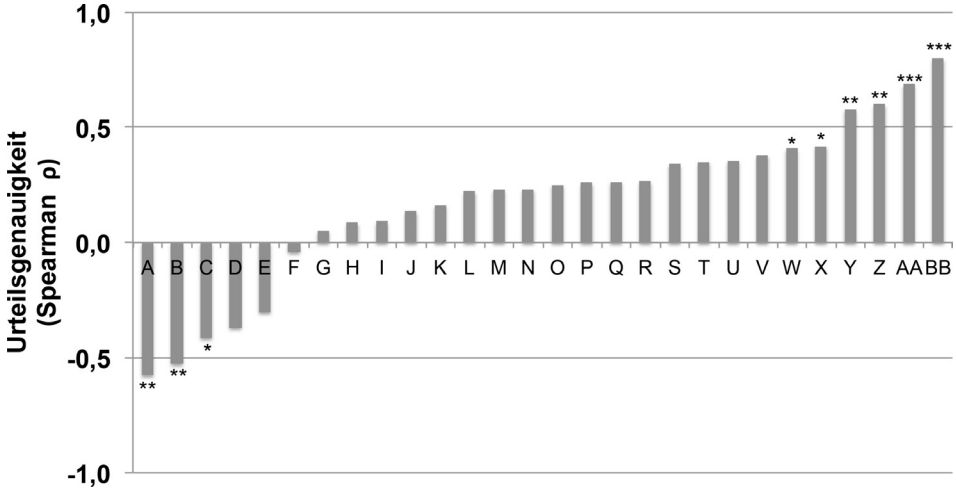

Abb. 5: Korrelationen für die Schülerinnen und Schüler A bis BB (*signifikant $p < 0{,}05$; **sehr signifikant $p < 0{,}01$; hoch signifikant $p < 0{,}001$).

7. Diskussion

Der Ansatz, Schülerselbstbeurteilungen zur Unterstützung der formativen Individualdiagnostik experimenteller Fähigkeiten im Unterrichtsalltag zu nutzen, setzt eine genaue Selbstbeurteilung durch die Schülerinnen und Schüler voraus. Der Frage, wie genau Schülerinnen und Schüler ihre experimentelle Performanz selbst beurteilen können, geht die vorgestellte Studie nach. Im Folgenden werden die Ergebnisse auf Gruppen- und Individualebene diskutiert. Anschließend wird auf die Limitationen der Studie eingegangen.

7.1. Gruppenebene

Die Urteilsgenauigkeit auf Gruppenebene, gemessen als Rangkorrelation, ist im Vergleich zu vorliegenden Studien hoch ($\rho = .63^{***}$). In der Metasynthese zur Urteilsgenauigkeit [36] lag die mittlere Korrelation mit .29 (Standardabweichung .11) deutlich niedriger. Die von Ross [32] berichtete Korrelation von .63 stellt bereits einen Ausreißer nach oben dar. Trotzdem ist Optimierungsbedarf vorhanden.

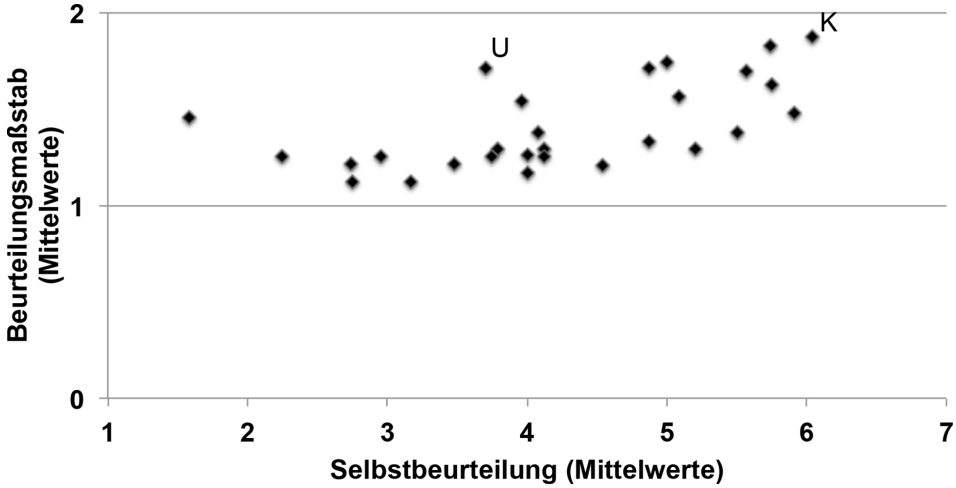

Abb. 6: Streudiagramm zur Urteilsgenauigkeit der Schülerinnen und Schüler auf Gruppenebene mit den ausgewählten Schülern U und K (siehe Text).

Das Streudiagramm (Abb. 6) zeigt, dass die Performanz aller Schülerinnen und Schüler nach dem externen MEK-LSA-Experimentiertest-Beurteilungsmaßstab im Mittel zwischen Stufe 1 und 2 und folglich in der oberen Hälfte des Leistungsspektrums liegt. Im Gegensatz dazu sehen 39% aller Schülerinnen und Schüler ihre Performanz im Mittel in der unteren Hälfte des Leistungsspektrums.

Ferner sind im Streudiagramm individuelle Unterschiede bei der Urteilsgenauigkeit zu erkennen. Beispielsweise kann sich Schüler K im Mittel tendenziell genau beurteilen, während Schüler U seine Leistung im Mittel offenbar stark unterschätzt. Eine detailliertere Analyse auf Individualebene erscheint daher angebracht.

7.2. Individualebene

Auf Individualebene bestätigt sich, dass die Schülerinnen und Schüler sich in ihrer Urteilsgenauigkeit unterscheiden. Zieht man – wie hier geschehen – als Maß der Urteilsgenauigkeit die Rangkorrelation zwischen Selbstbeurteilung und externem Beurteilungsmaßstab heran, sind offenbar nur wenige Schülerinnen und Schüler in der Lage sich genau zu beurteilen: In dieser Stichprobe gibt es nur vier hohe, positive Korrelationen.

Bedenklich erscheint zunächst, dass sechs Korrelationen im negativen Bereich liegen. Allerdings liegt insbesondere bei den signifikanten negativen Korrelationen die Vermutung nahe, dass die Likert-Skala bei der Selbstbeurteilung falsch herum gelesen wurde. Diese Erklärung ist plausibel, weil die Likert-Skala der Selbstbeurteilung, die während des regulären Unterrichts über das Schuljahr hinweg verwendet wurde, genau andersherum beschriftet war.

Insgesamt fällt auf, dass der überwiegende Teil der Korrelationen nicht signifikant ist und im niedrigen Bereich liegt. Hier kann man allerdings nicht ohne Weiteres von einer niedrigen Urteilsgenauigkeit ausgehen. In den Abbildungen 7, 8 und 9 sind für exemplarisch ausgewählte Schülerinnen und Schüler Blasendiagramme dargestellt, in denen die Testwerte über die Selbstbeurteilungen aufgetragen sind.

Die in Abb. 7 nicht signifikante Korrelation ist offenbar darauf zurückzuführen, dass dieser Schüler (F) bei der Selbstbeurteilung häufig die mittleren Stufen drei und vier ankreuzt. Folglich tendiert der Schüler bei der Selbstbeurteilung zur Mitte. Die Tendenz zur Mitte stellt einen typischen Urteilsfehler dar, der beispielsweise auch im Kontext der Urteilsgenauigkeit von Lehrkräften gefunden wurde (z. B. [17, S. 138]).

Auch die hohe Korrelation von $\rho = .80***$ ist durchaus interpretierbar (s. Abb. 8). Tendenziell urteilt dieser Schüler (BB) korrekt. Im Blasendiagramm ist die Varianz der experimentellen Performanz auffällig; auch schwächere Leistungen sind keine Seltenheit und werden vom Schüler selbst entsprechend eingeschätzt.

Ganz anders verhält es sich bei Schüler K (s. Abb. 9; vgl. auch Abb. 5 und 6). In fast 90% aller Aufgaben erreicht dieser Schüler die höchste Stufe nach Kriterien des externen Beurteilungsmaßstabs. In 54% aller Fälle stimmt die höchste Stufe des externen Beurteilungsmaßstabs mit der höchsten Stufe der Selbstbeurteilung überein. Folglich würde man nicht davon ausgehen, dass der Schüler eine so niedrige Urteilsgenauigkeit hat, wie eine nicht signifikante Korrelation von $\rho = .16$ suggeriert. Bei geringen oder nicht signifikanten Korrelationen kann eine durch fehlende Varianz bedingte Unterschätzung der Urteilsgenauigkeit vorliegen. Diese methodische Einschränkung durch die Verwendung von Korrelationen zur Bestimmung der Urteilsgenauigkeit wird im folgenden Abschnitt aufgegriffen.

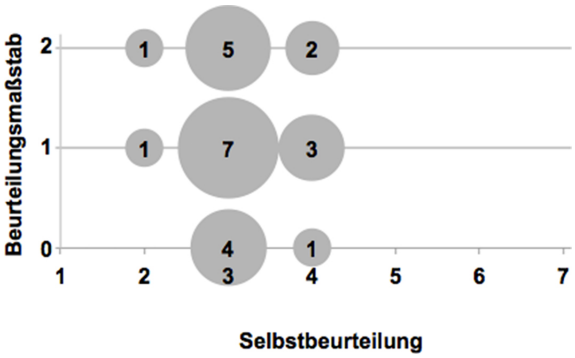

Abb. 7: Tendenz zur Mitte ($\rho = –.04$) bei Schüler F

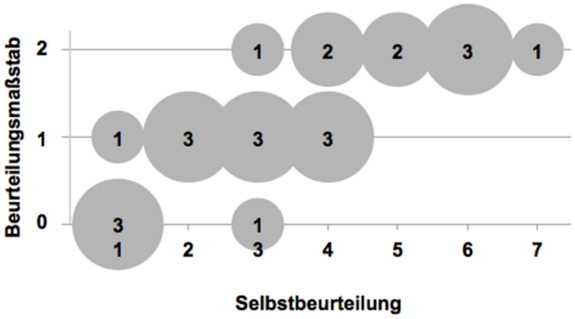

Abb. 8: Hohe Korrelation ($\rho = .80***$) bei Schüler BB

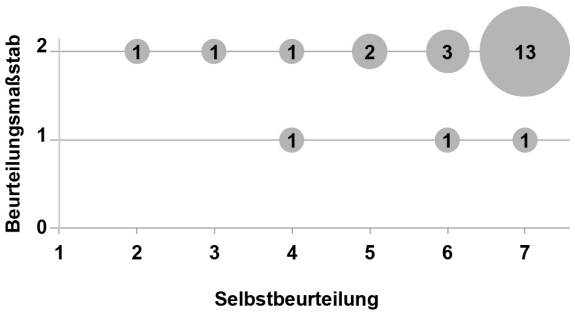

Abb. 9: Deckeneffekt ($\rho = .16$) bei Schüler K

7.3. Limitationen der Studie

Die Limitationen der vorgelegten Studie betreffen insbesondere die Stichprobe sowie die Methodik bei der Datenerhebung und Datenauswertung. Die Aspekte werden im Folgenden genauer erläutert.

Stichprobe: Die Gelegenheitsstichprobe ist sehr klein. Sie besteht nur aus 28 Schülerinnen und Schülern, die einer Klasse angehören. Ferner weist die Stichprobe besondere Stichprobeneigenschaften auf. Zu diesen besonderen Eigenschaften gehören insbesondere die hohen Erfahrungen beim Experimentieren mit Schülerexperimenten sowie mit Selbstbeurteilungen beim Experimentieren. Eine Generalisierbarkeit der Ergebnisse ist somit nicht gegeben. Deut-

lich wird jedoch bereits mit dieser Stichprobe, dass individuelle Unterschiede existieren und eine hohe Urteilsgenauigkeit sowohl auf Gruppen- als auch auf Individualebene möglich ist.

Situation der Datenerhebung: Die Situation während der Datenerhebung unterscheidet sich von einer Lernsituation im Unterricht, in der es vollkommen untypisch wäre, 24 experimentelle Aufgabenstellungen in ca. drei Unterrichtsstunden individuell am Computer bearbeiten zu lassen. Deshalb ist nicht auszuschließen, dass Schülerinnen und Schüler sich in einem selbstregulierten Lernprozess anders selbst beurteilen als in dieser Studie.

Aufgabenunspezifische Beurteilungskriterien: Aufgrund zeitlicher Begrenzung wurde pro Aufgabenbearbeitung nur eine Selbstbeurteilung in einer Beurteilungskategorie erfragt (s. Abschnitt 5.2.2). Das Beurteilungskriterium[2] war in allen Aufgaben identisch, also aufgabenunspezifisch. Wenn die Beurteilungskriterien aufgabenunspezifisch sind, fehlt den Lernenden unter Umständen eine Verbindung zu der aktuellen aufgabenspezifischen Performanz [32]. Diese fehlende Verbindung führt potenziell zu zusätzlicher Varianz bei der Selbstbeurteilung. Beispielsweise besteht die Gefahr, dass Beurteilende als Grundlage für ihre Selbstbeurteilung ihre Anstrengung wählen (z. B. bei [59]). Eine hohe Anstrengung muss allerdings nicht immer mit einer starken Leistung und damit einer hohen Urteilsgenauigkeit einhergehen. Ferner wäre es schlecht, wenn manche Schülerinnen und Schüler sich anhand eines individuellen, internen Beurteilungsmaßstabs beurteilen würden, anstatt sich an inhaltlichen, relevanten Beurteilungskriterien zu orientieren.

Verständlichkeit der Beurteilungskriterien: Die Verständlichkeit der Beurteilungskriterien ist zu gewährleisten (z. B. [60, S. 147]; [61]). In dieser Studie könnten die signifikanten negativen Korrelationen auf ein Verständnisproblem hindeuten (s. Abschnitt 6.1.1).

Maß der Urteilsgenauigkeit: Um die Anschlussfähigkeit an vorliegende Studien herzustellen, wurde die Korrelation als Maß für die Urteilsgenauigkeit gewählt. Wie die Beispiele in Abschnitt 7.2 zeigen, sind in dieser Studie Korrelationen als Maß für die Urteilsgenauigkeit nur eingeschränkt interpretierbar. Dies hängt damit zusammen, dass die Berechnung der Korrelation stark von der Varianz der Beurteilungen auf beiden Skalen abhängt [49, S. 75] und diese bei einigen Schülerinnen und Schülern nicht gegeben ist. Andere Maße, wie der Urteilsfehler oder Urteilstendenzen (vgl. Abschnitt 3.2), lassen sich aufgrund der fehlenden Passung der Skalen für Selbstbeurteilung und externen Beurteilungsmaßstab nicht berechnen. Für weitere Studien bietet sich daher an, für beide Beurteilungen Skalen mit gleicher Anzahl an Stufen zu verwenden. Um eine Varianz auf beiden Skalen zu erreichen, bieten sich Aufgabenstellungen mit mittlerer Aufgabenschwierigkeit an. Solche Aufgabenstellungen erfordern auch eher selbstregulierte Prozesse als zu leichte oder zu schwere Aufgabenstellungen.

[2] Beurteilungskriterium beim Aufbauen: Ich konnte den Versuch ohne Probleme aufbauen. Beurteilungskriterium beim Messen: Ich konnte die Messungen ohne Probleme durchführen.

8. Fazit und Ausblick

In diesem Artikel wird aus dem Stand der Forschung zur Diagnostik experimenteller Fähigkeiten und zur Genauigkeit von Selbstbeurteilungen ein innovativer Ansatz zur praktikablen, formativen Diagnostik experimenteller Fähigkeiten im Unterrichtsalltag abgeleitet (s. Abschnitt 1.2). Dieser Ansatz setzt voraus, dass die Selbstbeurteilungen der Schülerinnen und Schüler möglichst genau sind.

Ferner ist die Genauigkeit der Selbstbeurteilungen auch vor dem Hintergrund eines selbstregulierten Lernprozesses beim Experimentieren relevant (s. Abschnitt 1.2). Kompetent Lernende sollten die eigenen Lernhandlungen möglichst genau selbst beurteilen können.

Dass eine Selbstbeurteilung experimenteller Fähigkeiten durch Schülerinnen und Schüler im Mittel (Gruppenebene) mit vergleichsweise hoher Genauigkeit grundsätzlich möglich ist, konnte trotz einiger Limitationen (vgl. Abschnitt 7.3) bereits mit der vorgelegten Studie gezeigt werden. Es wird jedoch auch deutlich, dass nicht mit einer einheitlich hohen Urteilsgenauigkeit für alle Schülerinnen und Schüler zu rechnen ist (Individualebene). Vielmehr ist davon auszugehen, dass die Fähigkeit, sich genau zu beurteilen, individuell unterschiedlich ausgeprägt ist. Dabei ist festzuhalten, dass es Schülerinnen und Schüler gibt, die die eigene Performanz beim Experimentieren verhältnismäßig genau selbst beurteilen können. Offen ist jedoch noch, wie sich die individuell unterschiedliche Fähigkeit zur Selbstbeurteilung erklären lässt.

Somit erscheint es sowohl aus Forschungsperspektive als auch aus unterrichtspraktischer Perspektive sinnvoll, den Ansatz, Selbstbeurteilungen zur Diagnostik experimenteller Fähigkeiten zu nutzen, weiter zu verfolgen.

Forschungsperspektive: Zu untersuchen ist insbesondere die individuell unterschiedlich ausgeprägte Urteilsgenauigkeit, die sich durch Varianz bei der individuellen Urteilsgenauigkeit ausdrückt. Offen ist, welche Variablen diese Varianz zu einem möglichst großen Teil erklären könnten.

Einen Teil der Varianz in der individuellen Urteilsgenauigkeit können möglicherweise verschiedene personenbezogene Variablen, wie z. B. das Geschlecht, erklären. Beispielsweise zeigt sich in der vorliegenden Studie im Zusammenhang mit dem Geschlecht auf Gruppenebene, dass die Schülerin-

nen ihre Performanz im Mittel tendenziell etwas genauer beurteilen können als ihre Mitschüler[3] ($\rho_w = .77$**; $\rho_m = .54$*). Allerdings wird anhand eines Streudiagramms (Abb. 10) deutlich, dass insbesondere die Schülerinnen ihre Performanz unterschätzen: Nach Beurteilungsmaßstab liegen die Leistungen aller Schülerinnen und Schüler in der oberen Hälfte des Leistungsspektrums. In diesem Bereich liegen auch fast alle Selbstbeurteilungen der Schüler, aber im Wesentlichen nur drei der Schülerinnen. Diese deskriptiven Daten zeigen bereits, dass unterschiedliche Urteilstendenzen bei den Geschlechtern zu erwarten sind. Allerdings war eine Berechnung von Urteilstendenzen in dieser Studie nicht möglich (s. Abschnitt 7.3).

Auf Basis der vorliegenden Literatur ist die Annahme plausibel, dass personenbezogene Variablen die Urteilsgenauigkeit beeinflussen können. Brown und Harris [30] konnten in ihrem Review mit 84 Studien die folgenden potenziellen personenbezogenen Einflussvariablen identifizieren: akademische Fähigkeiten (insbesondere Fachwissen), Geschlecht, Motivation, Erfahrung (im Zusammenhang mit dem Alter), Ethnie sowie Selbstwirksamkeitserwartung und Selbstwert. Diese Variablen sind nach unserem Kenntnisstand a) noch nicht gemeinsam im Hinblick auf ihren Einfluss auf die Urteilsgenauigkeit analysiert worden und b) noch nicht im Kontext der Selbstbeurteilung experimenteller Fähigkeiten. Deshalb sollen in einer Folgestudie die oben genannten personenbezogenen Variablen als potenzielle Prädiktoren für die Urteilsgenauigkeit gemeinsam untersucht werden. Hier wäre beispielsweise eine Hypothese, dass Lernende mit einer hohen, aktuellen Motivation eher genaue Selbstbeurteilungen abgeben als Lernende mit einer niedrigen aktuellen Motivation.

Bei einer solchen Untersuchung potenzieller Prädiktoren sind zusätzliche, methodische Varianzquellen zu reduzieren. Darüber hinaus erscheint es sinnvoll zu sein, den Urteilsfehler und Urteilstendenzen als (zusätzliche) Maße für die Urteilsgenauigkeit heranzuziehen. Beide Maße können auch bei geringer Varianz der experimentellen Performanz sinnvoll interpretiert werden.

In weiteren Schritten wäre zu untersuchen, wie stark die genauen Selbstbeurteilungen der Schülerinnen und Schüler mit den Urteilen der Lehrkräfte übereinstimmen und inwiefern die Urteilsgenauigkeit der Lehrkräfte gesteigert werden kann, wenn diese die (genauen) Selbstbeurteilungen für ihre Diagnosen berücksichtigen.

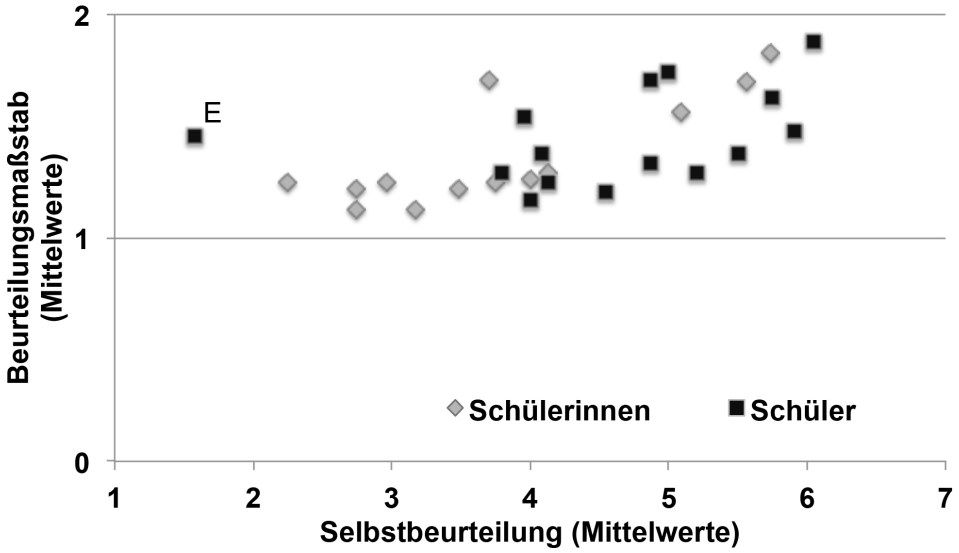

Abb. 10: Streudiagramm zur Urteilsgenauigkeit der Schülerinnen und Schüler auf Gruppenebene – getrennt nach Geschlechtern und ohne Ausreißer (E) bei den Schülern berechnet.

Unterrichtspraktische Perspektive: Auf Gruppenebene (z. B. Klassen) deutet sich an, dass mehrere Selbstbeurteilungen pro Schülerin bzw. pro Schüler eine geeignete zusätzliche Datenquelle darstellen, um experimentelle Fähigkeiten in Lernphasen potenziell genauer als bisher zu diagnostizieren und grundsätzliche – nicht aufgabenspezifische – Fördermaßnahmen zu planen.

Allerdings wird auf Individualebene deutlich, dass die einzelnen Selbstbeurteilungen nicht pauschal zur Erweiterung der Datenbasis einer Lehrkraft herangezogen werden können, um die individuelle Performanz einzelner Schülerinnen und Schüler zu einer Aufgabenstellung zu diagnostizieren. Dafür kommen nur Selbstbeurteilungen von Schülerinnen und Schülern infrage, die ihre Performanz potenziell genau beurteilen können.

[3] Proband E wird als Ausreißer in den folgenden Ausführungen nicht berücksichtigt.

Außerdem wird für das selbstregulierte Lernen experimenteller Fähigkeiten wichtig sein, gerade die Schülerinnen und Schüler mit geringer Urteilsgenauigkeit zu identifizieren. Diese Schülerinnen und Schüler benötigen zunächst Unterstützung durch eine Lehrkraft – kurzfristig bei der Auswahl geeigneter Lernmaterialien, mittelfristig bei der Verbesserung ihrer Fähigkeit zur Selbstbeurteilung.

Grundsätzlich ist zu berücksichtigen, dass sich die Ergebnisse nicht auf den Einsatz von Selbstbeurteilungen zur summativen Diagnostik experimenteller Fähigkeiten und zur Leistungsbewertung übertragen lassen. Bei einem Einsatz zur summativen Diagnostik wäre mit Verzerrungen bei den Selbstbeurteilungen zu rechnen, weil Schülerinnen und Schüler verständlicherweise ihre Leistung möglichst positiv darstellen möchten (s. Abschnitt 5.1).

9. Literatur

[1] National Research Council (NRC) (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Washington, DC: The National Academies Press.

[2] KMK (2005). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss.* München: Luchterhand.

[3] Ministerium für Schule und Weiterbildung des Landes NRW (MSW NRW) (2014). *Schulgesetz für das Land Nordrhein-Westfalen*. Abgerufen am 22.04.2015 unter https://www.schulministerium.nrw.de/docs/Recht/Schulrecht/Schulgesetz/Schulgesetz.pdf

[4] Höttecke, D. & Rieß, F. (2015). Naturwissenschaftliches Experimentieren im Lichte der jüngeren Wissenschaftsforschung – Auf der Suche nach einem authentischen Experimentbegriff der Fachdidaktik. *Zeitschrift für Didaktik der Naturwissenschaften,* 21(1), 127-139.

[5] Höttecke, D. (2008). Fachliche Klärung des Experimentierens. In D. Höttecke (Hrsg.), *Kompetenzen, Kompetenzmodelle, Kompetenzentwicklung* (S. 293-295). Münster: LIT.

[6] Nawrath, D., Maiseyenka, V. & Schecker, H. (2011). Experimentelle Kompetenz – Ein Modell für die Unterrichtspraxis. *Praxis der Naturwissenschaften – Physik in der Schule*, 60(6), 42-49.

[7] Schreiber, N., Theyßen, H. & Schecker, H. (2009). Experimentelle Kompetenz messen? *Physik und Didaktik in Schule und Hochschule,* 5(3), 92-101.

[8] Meier, M. & Mayer, J. (2014). Selbständiges Experimentieren: Entwicklung und Einsatz eines anwendungsbezogenen Aufgabendesigns. *Der mathematische und naturwissenschaftliche Unterricht*, 67(1), 4-10.

[9] Gut, C., Metzger, S., Hild, P. & Tardent, J. (2014b). Problemtypenbasierte Modellierung und Messung experimenteller Kompetenzen von 12- bis 15-jährigen Jugendlichen. *PhyDid B, Didaktik der Physik, Beiträge zur DPG-Frühjahrstagung 2014 in Frankfurt*, 1-9.

[10] Emden, M. (2011). Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens. Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 118). Berlin: Logos.

[11] Schreiber, N., Theyßen, H. & Schecker, H. (2014). Diagnostik experimenteller Kompetenz: Kann man Realexperimente durch Simulationen ersetzen? *Zeitschrift für Didaktik der Naturwissenschaften*, 20(1), 161-173.

[12] Shavelson, R. J., Ruiz-Primo, M. A. & Wiley, E. W. (1999). Note on Sources of Sampling Variability in Science Performance Assessments. *Journal of Educational Measurement,* 36(1), 61-71.

[13] Thillmann, H., Gößling, J., Marschner, J., Wirth, J. & Leutner, D. (2013). Metacognitive knowledge about and metacognitive regulation of strategy use in self-regulated scientific discovery learning: New methods of assessment in computer-based learning environments. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (S. 575-588). New York, NY: Springer.

[14] Chen, Z. & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variable strategy. *Child Development,* 70(5), 1098-1120.

[15] Schreiber, N., Theyßen, H. & Schecker, H. (2015). Process-Oriented and Product-Oriented Assessment of Experimental Skills in Physics: A Comparison. In N. Papadouris, A. Hadjigeorgiou & C. P. Constantinou (Eds.), *Insights from Research in Science Teaching and Learning: Selected Papers from the ESERA 2013 Conference* (S. 29-43). Cham u. a.: Springer.

[16] Emden, M. & Sumfleth, E. (2012). Prozessorientierte Leistungsbewertung des experimentellen Arbeitens. Zur Eignung einer Protokollmethode zur Bewertung von Experimentierprozessen. *Der mathematische und naturwissenschaftliche Unterricht (MNU)*, 65(2), 68-75.

[17] Helmke, A. (2010). *Unterrichtsqualität und Lehrerprofessionalität – Diagnose, Evaluation und Verbesserung des Unterrichts.* Seelze-Velber: Friedrich Verlag.

[18] Maiseyenka, V., Schecker, H. & Nawrath, D. (2013). Kompetenzorientierung des naturwissenschaftlichen Unterrichts – Symbiotische Ko-

operation bei der Entwicklung eines Modells experimenteller Kompetenz. *Physik und Didaktik in Schule und Hochschule*, 1(12), 1-17.

[19] Nawrath, D. & Peters, S. (2014). Experimente für das Lernen nutzen. *Naturwissenschaften im Unterricht – Physik*, 144, 4-9.

[20] Südkamp, A., Kaiser J. & Möller, J. (2012). Accuracy of Teachers' Judgments of Students' Academic Achievement: A Meta-Analysis. *Journal of Educational Psychology*, 104, 743-762.

[21] Winter, F. (2006). Diagnosen im Dienst des Lernens – Diagnostizieren und Fördern gehören zum Unterricht. In G. Becker, M. Horstkemper, E. Risse, L. Stäudel, R. Wernin, & F. Winter (Hrsg.), *Diagnostizieren und Fördern – Stärken entdecken – Können entwickeln. Friedrich Jahresheft XXIV* (S. 22-25). Seelze: Friedrich Verlag.

[22] Schreiber, N. & Theyßen, H. (2016). Sind Selbstbeurteilungen beim Experimentieren praktikabel und nützlich? In C. Maurer (Hrsg.), *Authentizität und Lernen – das Fach in der Fachdidaktik. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Berlin 2015* (S. 164-166). Regensburg: Universität Regensburg.

[23] Mazzoni, G., Cornoldi, C. & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition*, 18(2), 196-204.

[24] Nelson, T. O. & Leonesio, R. J. (1988). Allocation of Self-Paced Study Time and the "Labor-in-Vain Effect". *Journal of Educational Psychology: Learning Memory and Cognition*, 14 (4), 676-686.

[25] Thiede, K. W., Anderson, M. C. M. & Therriault, D. (2003). Accuracy of Metacognitive Monitoring Affects Learning of Texts. *Journal of Educational Psychology*, 95(1), 66-73.

[26] Baars, M. (2014). *Instructional Strategies for Improving Self-Monitoring of Learning to Solve Problems.* Abgerufen am 04.04.2016 unter http://hdl.handle.net/1765/77825

[27] Koriat, A., Ackerman, R., Lockl, K. & Schneider, W. (2009). The easily learned, easily remembered heuristic in children. *Cognitive Development*, 24(2), 169-182.

[28] Nelson, T. O. & Dunlosky, J. (1991). When People's Judgments of Learning (JOLs) are Extremely Accurate at Predicting Subsequent Recall: The "Delayed-JOL Effect". *Psychological Science*, 2(4), 267-270.

[29] Andrade, H. L. (2010). Students as the Definitive Source of Formative Assessment: Academic Self-Assessment and the Self-Regulation of Learning. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of Formative Assessment*. New York, London: Routledge.

[30] Brown, G. T. L. & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.). *The SAGE handbook of research on classroom assessment* (S. 367-393). Thousand Oaks, CA: Sage.

[31] Falchikov, N. & Boud, D. (1989). Student Self-Assessment in Higher Education: A Meta-Analysis. *Review of Educational Research* 59(4), 395-430.

[32] Ross, S. (1998). Self-assessment in second language testing: a meta-analysis and analysis of experimental factors. *Language Testing*, 15(1), 1-20.

[33] Wiley, J., Griffin, T. D. & Thiede, K. W. (2005). Putting the Comprehension in Metacomprehension. *The Journal of General Psychology*, 132, 408-428.

[34] Lin, L.-M. & Zabrucky, K. M. (1998) Calibration of Comprehension: Research and Implications for Education and Instruction. *Contemporary Educational Psychology*, 23, 345-391.

[35] Zell, E. & Krizan, Z. (2014). Do People Have Insight Into Their Abilities? A Metasynthesis. *Perspectives on Psychological Science,* 9(2), 111-125.

[36] Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education*, 19(1), 69-75.

[37] Van der Jagt, S., van Rens, L., Schalk, H., Pilot, A. & Beishuizen, J. (2012). Development of a Student Self-Evaluation Instrument in Inquiries. In *Proceedings of the NARST 2012 conference.* Zugriff am 28.05.2014 unter http://www.dudocprogramma.nl/docs/DUDOC2/saskia-van-der-jagt-paper-narst-2012.pdf

[38] Birri, T. & Smit, R. (2013). Lernen mit Rubrics. Kompetenzen aufbauen und beurteilen. *Pädagogik,* 3/2013, 36-39.

[39] Andrade H. L., Wang X., Du, Y. & Akawi, R. L. (2009). Rubric-Referenced Self-Assessment and Self-Efficacy for Writing. *The Journal of Educational Research*, 102(4), 287-302.

[40] Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review* 2(2), 130-144.

[41] Arter, J. & McTighe, J. (2001). *Scoring Rubrics in the Classroom: Using Performance Criteria for Assessing and Improving Student Performance.* Thousand Oaks, Calif.: Corwin Press.

[42] Rutgers Physics and Astronomy Education Group (RPAEG) (2008). *Scientific Ability Rubrics*. Zugriff am 03.07.2013 unter http://paer.rutgers.edu/ScientificAbilities/Rubrics/default.aspx

[43] Nadji, T., Lach, M. & Blanton, P. (2003). Assessment Strategies for Laboratory Reports. *The Science Teacher*, 41(1), 56-57. Zugriff am 03.07.2013 unter ftp://ftp.aip.org/epaps/phys_teach/E-PHTEAH-41-022301/Rubric.doc

[44] Theyßen, H., Schecker, H., Gut, C., Hopf, M., Kuhn, J., Labudde, P., Müller, A., Schreiber, N. & Vogt, P. (2014). Modelling and Assessing Experimental Competencies in Physics. In C. Bruguière, A. Tiberghien & P. Clément (Eds.), *9th ESERA Conference Selected Contributions: Topics and Trends in Current Education – Contributions from Science Education Research* (S. 321-337). Heidelberg u. a.: Springer.

[45] Heinicke, S. & Bellingrath, M. (2015). Diagnose, Feedback und Feedforward: Methoden-Werkzeuge und Hilfen für eine alltagstaugliche Lernbegleitung. *Naturwissenschaften im Unterricht – Physik,* 147/148, 40-45.

[46] Schreiber, N. & Nawrath, D. (2014). Experimentelle Fähigkeiten mit Schülerselbstbeurteilungen diagnostizieren. *Naturwissenschaften im Unterricht – Physik*, 144, 14-18.

[47] Struck, Y. (2015). Methodenwerkzeug = Diagnostikwerkzeug? Anregungen für den Einsatz geeigneter Methoden-Werkzeuge zur Diagnostik. *Naturwissenschaften im Unterricht – Physik,* 147/148, 24-29.

[48] Jussim, L. (2012). *Social Perception and Social Reality*. Oxford u. a.: Oxford University Press.

[49] Ward, M., Gruppen, L. & Regehr, G. (2002). Measuring Self-assessment: Current State of the Art. *Advances in Health Sciences Education,* 7(1), 63-80.

[50] Anders, Y., Kunter, M., Brunner, M., Krauss, S. & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht,* 57, 175-193.

[51] Fitzgerald, J. T., White, C. B. & Gruppen, L. D. (2003). A longitudinal study of self-assessment accuracy. *Medical Education,* 37(7), 645-649.

[52] Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, 33 (10), 1045-1063.

[53] Theyßen, H., Schecker, H., Neumann, K., Eickhorst, B. & Dickmann, M. (2016). Messung experimenteller Kompetenz – ein computergestützter Experimentiertest. *Physik und Didaktik in Schule und Hochschule*, 15(1), 26-48.

[54] Dickmann, M. (2016). Messung von Experimentierfähigkeiten. *Validierungsstudien zur Qualität eines computerbasierten Testverfahrens.* Berlin: Logos.

[55] Schecker, H., Neumann, K., Theyßen, H., Eickhorst, B. & Dickmann, M. (2016). Stufen experimenteller Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften,* 22*(*1), 197-213.

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

[57] Brell, C. (2008). Lernmedien und Lernerfolg – reale und virtuelle Materialien im Physikunterricht. Empirische Untersuchungen in achten Klassen an Gymnasien (Laborstudie) zum Computereinsatz mit Simulationen und IBE. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 74). Berlin: Logos.

[58] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

[59] Ross, J. A., Rolheiser, C. & Hogaboam-Gray, A. (1998). Skills training versus action research in-service: Impact on student attitudes to self-evaluation. *Teaching and Teacher Education*, 14(5), 463-477.

[60] Keller, S. (2011). Beurteilungsraster und Kompetenzmodelle. In W. Sacher & F. Winter (Hrsg.), *Diagnose und Beurteilung von Schülerleistungen* (S. 143-160). Schorndorf: Schneider.

[61] Reddy, Y. M. & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435-448.