

Aus Fehlern wird man klug

Über die Relevanz eines adäquaten Verständnisses von „Messfehlern“

S. Heinicke, J. Glomski, B. Priemer u. F. Rieß

1 | Einleitung – Warum Fehler klug machen

Messfehler¹ haben gemeinhin eine schlechte Reputation. Sie sind ungeliebte Begleiter experimenteller Messungen. Jeder möchte sie so gering wie möglich halten, sie werden wo es geht vernachlässigt, korrigiert oder wegdiskutiert. Wie Mellor [1] ausdrückt: „Messfehler sind eine ermüdende, aber triviale Wucherung an der sauberen deduktiven Struktur der Naturwissenschaft“. In diesem Artikel möchten wir eine Lanze brechen für diese unliebsamen Störenfriede. Dies hat zwei Gründe: In unseren Augen bieten gerade Messfehler eine wunderbare Gelegenheit, etwas Entscheidendes über die Natur naturwissenschaftlicher Messungen zu lernen. Denn die Abschätzung, wie (un)genau ein Messwert ist und welche Faktoren seine Genauigkeit beeinflussen können, legt auch fest, welches Vertrauen man ihm entgegenbringen und für wie verlässlich man ihn halten kann. Ohne die Betrachtung dieser Verlässlichkeit ist eine Bewertung des Messergebnisses oder ein Vergleich mit Erwartungen oder Daten anderer Wissenschaftler nicht möglich. Und ohne eine Auseinandersetzung mit der kritischen Bewertung von Messdaten ist es wiederum schwierig, Schülerinnen und Schülern Bewertungskompetenzen zu vermitteln, wie sie von den Bildungsstandards [7] eingefordert werden. Zum Zweiten zeigen die Ergebnisse der fachdidaktischen Forschung auf diesem Gebiet, dass sich das Thema entgegen der üblichen Schulpraxis nicht umgehen lässt: Auch ein explizites Nichtthematisieren birgt implizite Aussagen über den Umgang mit Messdaten und ihrer begrenzten Genauigkeit. Wer die Thematisierung der den Mess-

daten innewohnenden Unschärfe umgeht, lässt einen entscheidenden Aspekt bei der Bewertung von Messergebnissen aus dem Spiel und öffnet den Raum für Faustregeln und ad hoc-Einschätzungen.

Warum ist eine solche Bewertungskompetenz von Messdaten wichtig? Mit Sicherheit ist der adäquate Umgang mit Messfehlern für unterschiedliche Menschen auch unterschiedlich wichtig. Bei einem Ingenieur oder einem Naturwissenschaftler gehört ihre Betrachtung zum Handwerk, um die Vergleichbarkeit, Objektivität und Anwendbarkeit ihrer Daten zu gewährleisten. Für einen Arzt, einen Schreiner oder einen Mechaniker haben Messfehler wiederum eine ganz andere Bedeutung. Aber auch über diese eher naturwissenschaftlich-technischen Bereiche hinaus spielen Messfehler oder die Verlässlichkeit von Daten im gesellschaftlichen Alltag eine zwar unbekanntere, aber nicht unbedeutende Rolle. Spektakuläre aktuelle Beispiele für die Frage nach der Verlässlichkeit von Messdaten sind u.a. die verschiedenen und einander widersprechenden Modelle zum Klimawandel, Studien zum möglichen Zusammenhang zwischen Handyutzung und Gehirntumoren oder die Verlässlichkeit der Testverfahren zur Neuen Grippe, aber auch Wetterdaten oder Sonntagsumfragen. Die Forderung nach einem kritischen Umgang mit Daten und der Entmystifizierung der „exakten Naturwissenschaft“ muss daher die Diskussion der Messfehler und Grenzen der messenden Naturwissenschaft mit einschließen. Das Thema besitzt also auch für solche Schülerinnen und Schüler, die keine weitere naturwissenschaftliche Ausbildung erleben werden, beachtliche Relevanz für ihr gesellschaftliches und alltägliches Lebensumfeld.

Bislang nimmt diese Thematik in Unterricht und Forschung wenig Raum ein. Dieser Artikel soll daher zunächst einen Überblick über den Stand nationaler und internationaler Forschung zu Lernvorstellungen und Verständnisschwierigkeiten vom Umgang mit Messdaten und ihrer begrenzten Genauigkeit bieten. Die Ursachenforschung für die Schwierigkeiten der Lernenden steckt allerdings noch in den Anfängen.

Viele der Verständnisschwierigkeiten liegen sicherlich auch darin begründet, dass der Umgang mit Messunsicherheiten auch aus fachlicher Sicht bislang wenig geklärt ist. Erst seit Ende der 1970er Jahre wird ernsthaft um einen international einheitlichen Standard zum Umgang mit Messunsicherheiten gerungen. Mit diesem Beitrag möchten wir Fragen zur derzeitigen unterrichtlichen Praxis im Umgang mit und der Thematisierung von Messfehlern aufwerfen und die Wichtigkeit der expliziten Thematisierung herausstellen. Wir möchten dazu ermutigen, die in diesem Thema verborgenen Reichtümer für den Unterricht zu entdecken.

2 | Ergebnisse nationaler und internationaler fachdidaktischer Forschung

Im naturwissenschaftlichen Unterricht geht es längst nicht nur darum, den Schülerinnen und Schülern fachliche Inhalte der Naturwissenschaft zu vermitteln. Vielmehr sollen sie einerseits in die Lage versetzt werden, ein altersgemäßes Verständnis davon zu entwickeln, wie die Naturwissenschaft die reale Welt beschreibt und erklärt, andererseits, welche Prozesse dieser naturwissenschaftlichen Erkenntnis zugrunde liegen ([8]). Die fachdidaktische Forschung der vergangenen Jahrzehnte hat gezeigt, dass diese beiden Bereiche eng miteinander verknüpft sind.

Darum ist das traditionelle Primat der fachlichen Inhalte in den letzten Jahren einer Gleichberechtigung der so genannten inhaltlichen und handlungsbezogenen Kompetenzen gewichen, die sich auch in den Bildungsstandards der Kultusministerkonferenz [7] widerspiegelt. Zur Stärkung der handlungsbezogenen Kompetenzen, und um den Schülerinnen und Schülern Einblick in die Hintergründe unseres naturwissenschaftlichen Wissens zu vermitteln, dienen auch gerade die praktischen Übungen im Unterricht, wie sie in der naturwissenschaftlichen Ausbildung seit über 100 Jahren eine große Rolle spielen. Aus fachdidaktischer Sicht stellt sich allerdings die Frage, was die Lernenden anhand dieser prak-

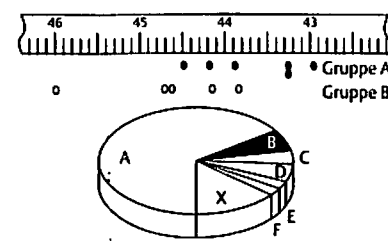


Abb. 1: Antworten der Schüler und Schülerinnen zur Frage nach der Übereinstimmung zweier Messreihen

tischen Übungen aber tatsächlich lernen und verstehen können.

Obwohl die praktischen Anteile in Unterricht und Lehre auf eine lange Tradition zurückblicken, wissen wir von Seiten der fachdidaktischen Forschung bisher nur wenig über die Vorstellungen von Lernenden, wenn es darum geht, wie Lernende naturwissenschaftliche Evidenz nutzen, wie sie Messdaten sammeln und interpretieren, speziell, wie sie mit der begrenzten Genauigkeit experimenteller Messdaten umgehen und sie bewerten. Außer in [9], [10], [11], [12] und [13] ist dieser Frage in der fachdidaktischen Forschung bisher nur wenig Beachtung geschenkt worden. Gerade für den deutschen Bildungsraum lagen hierzu bislang keine Ergebnisse vor. Daher wurden an der Universität Oldenburg im vergangenen Jahr zwei Studien unter 165 Schülerinnen und Schülern der 8. und 9. Klasse der Realschule und unter 160 Studierenden im ersten Jahr ihres Physikstudiums durchgeführt. Die Ergebnisse dieser Befragungen befinden sich in guter Übereinstimmung mit denen der internationalen Studien und werden im Folgenden als Beispiele dienen.

2.1 Die zwei Welten der Messergebnisbetrachtung

Es gehört sicher zur Erfahrung vieler Lehrender, dass Schüler und auch Studierende in ihren Versuchsprotokollen entweder eine Fehlerdiskussion nur dann durchführen, wenn das Ergebnis zu stark vom erwarteten Wert abweicht oder die geforderte Fehlerrechnung durchführen ohne den berechneten Wert anschließend in der Ergebnisdiskussion auch tatsächlich zu benutzen. Messwert und zugehörige Unsicherheit bilden für die meisten Lernenden anscheinend keine Informationseinheit und können in der Ergebnisdiskussion völlig getrennt voneinander stehen.

Antworten der Schülerinnen und Schüler:

Übereinstimmung der Werte (A+B):

A: Nein. Die Messwerte sind zu unterschiedlich. (64%)

B: Ja. Die Werte stimmen alle ungefähr überein. (7%)

Wiederkehrende Werte (C+D):

C: Ja, weil ein paar Werte mehrmals vorkommen. (3%)

D: Nein, dafür müssten alle Werte gleich sein (6%)

Berücksichtigung der Streuung / Unsicherheiten (E):

E: Ja, im Vergleich der Unterschiede und Ungenauigkeit. (2%)

Design der Messreihe (F):

F: Nein. Die Gruppen haben unterschiedlich oft gemessen. (2%)

X: Nicht zuzuordnen. (7%) Keine Angabe. (7%)

Bei der Durchführung eines Experiments zieht die Genauigkeit, mit der die Messungen durchgeführt werden sollen, eine Reihe von wesentlichen Entscheidungen nach sich. Es muss vor der Messung entschieden werden, welche Genauigkeit des Endergebnisses in der Messung angestrebt wird, um das Versuchsdesign, Geräte und Messinstrumente entsprechend auszuwählen. Während der Messung muss entschieden werden, welche und wie viele Daten aufgenommen werden sollen. Dies hängt neben der angestrebten Ergebnisgenauigkeit auch von Art und Größe der Unsicherheit der einzelnen Messdaten ab. Schließlich muss das am Ende erhaltene Ergebnis bewertet werden: Ohne die Angabe seiner Verlässlichkeit (z. B. in Form seiner Messunsicherheit) lässt sich keine Entscheidung darüber treffen, ob er einem erwarteten Wert entspricht oder ob ein gewisser physikalischer Zusammenhang anhand der Messung nachgewiesen werden kann oder nicht.

Masnick u. Klahr [9] stellten in einer Studie unter US-amerikanischen Zweit- und Viertklässlern (ca. 8 und 10 Jahre alt) fest, dass die Schülerinnen und Schüler in beiden Altersgruppen viele verschiedene Fehlerquellen aufzählen konnten, die das Ergebnis eines Versuches beeinflussen könnten. Die älteren Schüler nannten dabei sogar beinahe doppelt so viele Faktoren wie die jüngeren. Für die Entscheidungen während des Messprozesses und die Auswertung der erhaltenen Daten scheint das Wissen um mögliche Einflussfaktoren allerdings in den Hintergrund zu treten. Masnick u. Klahr [9] bemerkten, dass nur wenige der befragten Lernenden bei der Einschätzung der Vertrauenswürdigkeit der Ergebnisse die benannten möglichen Fehlerquellen ins Spiel brachten. Sie schlossen daraus, dass Lernende bereits ab einem sehr frühen Alter reiche, aber unsystematische Vorstellungen von Messfehlern und ih-

ren möglichen Quellen haben, diese Vorstellungen jedoch nicht gut in ein kohärentes Verständnis von Messungen und Dateninterpretationen eingebettet seien.

Beispiele hierfür zeigen auch die Ergebnisse der Oldenburger Studie. Bei der Bewertung zweier Messreihen nannte rund ein Viertel der Befragten mögliche Fehlerquellen als Gründe für die Streuung der Messwerte. Ihre Entscheidung, ob die beiden Messreihen übereinstimmen, machten sie allerdings fast ausnahmslos alleine an den Ergebnissen der Messreihen fest, wie Abb. 1 zu entnehmen ist. Nur in zwei Antworten wird eine Verknüpfung zwischen den Faktoren und der Aussagekraft der einzelnen Messwerte hergestellt: „Die Ergebnisse stimmen überein, weil es meistens nur Unterschiede von ein paar mm sind, und die können durch Messungenauigkeiten auftauchen.“ Auch unter den befragten Studierenden begründeten rund 60% ihre Entscheidung allein mit den Zahlenwerten. Bei der zusätzlichen Angabe des Literaturwertes gaben ebenfalls rund 60% der Studierenden an, dass das entscheidende Kriterium der Messgenauigkeit die Abweichung des Ergebnisses vom Literaturwert sei.

Messwert und Unsicherheit (die zwei „Welten“) bilden in den Vorstellungen der Mehrzahl der Lernenden unseren Befragungen zufolge keine Informationseinheit über die Messung. Eine Messung, soviel auch immer über ihre begrenzte Genauigkeit bekannt und verstanden ist, führt zu einem Ergebnis, das durch einen einfachen Zahlenwert repräsentiert werden kann. Seine Genauigkeit manifestiert sich in der Abweichung vom Zielwert und ergibt sich nicht inhärent aus der Geschichte der Messung. Die maßgebliche Ursache für diese Schwierigkeiten der Lernenden ist in den Vorstellungen der Lernenden von der Natur der Messdaten zu suchen. Dieses Verständnis bewirkt, dass

¹ Allen über die verwendete Terminologie ließen sich viele Seiten hären ([1], [3], [4]). Wir wenden auf die Vielzahl der Begriffe im Hinblick auf die Lernvorstellungen und die internationalen Richtlinien in den nächsten beiden Kapiteln kurz ein. Um aber schon vorher die argsten sprachlichen Klüppel zu umschiffen, werden wir im folgenden zwischen dem traditionell gebräuchlichen Begriff Messfehler (Abweichung des gemessenen Wertes vom Zielwert der Messung) und der im Fachsprachgebrauch auch recht unüblichen Messunsicherheit (Ergebnisunterschied, das dem Ergebnis der Messung realistisch zugeordnet werden kann) unterscheiden (Definitionen nach [5] bzw. [6]).

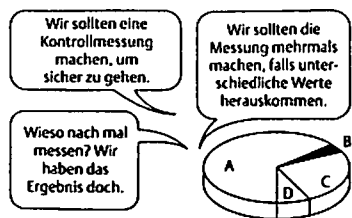
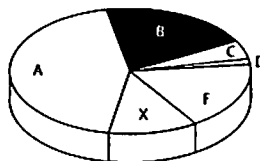


Abb. 2: Antworten der Schülerinnen und Schüler zur Wiederholung einer Messung.

Frage: Gruppe C hat mehrmals gemessen und diese Werte erhalten. Was sollen sie als Ergebnis angeben und warum?

1. Messung	439 mm
2. Messung	430 mm
3. Messung	442 mm
4. Messung	445 mm
5. Messung	433 mm
6. Messung	433 mm



Antworten der Schülerinnen und Schüler:

Man sollte ...

- A: ... nach wiederkehrenden Werten suchen (44%).
- B: ... den Mittelwert bilden (22%).
- C: ... ein Ergebnisintervall angeben (4%).
- D: ... den ersten Messwert nehmen (1%).
- F: ... keine Angabe; 17%
- X: ... nicht zuzuordnen (12%).

Abb. 3: Antworten der Schüler und Schülerinnen zum Ergebnis einer Messreihe.

die üblichen Routinen der Fehlerbehandlung zwar erlernt, aber in ihrer Tiefe nicht verstanden und hinterfragt werden. Diese beiden Punkte werden wir im Folgenden näher beleuchten.

2.2 Gut gelernt ist noch nicht halb verstanden

Lubben und Millar [11] befragten über 600 Schülerinnen und Schüler im Alter von 11, 14 und 16 Jahren über ihre Vorstellungen zum Umgang mit Messdaten. Obwohl es sich hierbei nicht um eine Langzeitstudie mit denselben Schülern über mehrere Jahre hinweg handelt, sind dennoch interessante Unterschiede zwischen den Altersgruppen festzustellen. Sie kamen zu dem Ergebnis, dass viele der Befragten bei der Bewertung ihrer Messdaten auf subjektive Einschätzungen oder Faustregeln zurückgriffen. Die Häufigkeit der Begründung, dass man durch eine mehrfache Wiederholung der Messung die Streuung der Messwerte in den Griff bekommen könne, stieg in den drei Altersgruppen von 18% auf 50%, dann auf 71%. Bei der Ergebnisangabe nahm die Zahl der Lernenden, die den Mittelwert der Messreihe berechnen würden, von 28% auf 56% auf 67% zu, während die Suche nach wiederkehrenden Werten in einer Messreihe als Hinweis auf den „richtigen“ Wert von 46% auf 34% auf 23% abnahm. Typische mathematische Routinen der Datenanalyse spielen nach dieser Studie mit zunehmendem Alter eine immer größere Rolle in der Argu-

mentation. Allerdings stellten Lubben und Millar [11] auch andere Effekte fest. Beispielsweise blieb der Anteil der Schülerinnen und Schüler, die eine Übereinstimmung zwischen zwei Messreihen allein am Abstand ihrer Mittelwerte festmacht, mit 42%, bzw. 38% und 40% über die Altersgruppen hinweg relativ konstant. In Übereinstimmung mit weiteren Studien ([12], [14]) zogen sie außerdem das Resümee, dass die befragten Lernenden zwar in gewissem Maße durch entsprechende Instruktion in die Lage versetzt würden, die Routinen der Messwertanalyse und -bewertung anzuwenden. Es sei ihnen aber offensichtlich unklar, was unter „genauer“ zu verstehen sei oder warum fünf Messungen eine höhere Ergebnisgenauigkeit erzielen können sollten als zwei. Auch in der Oldenburger Studie unter 14-16-jährigen Schülerinnen und Schülern gaben 25% an, dass die Messgenauigkeit durch eine Wiederholung verbessert werden könne, wobei in der Mehrzahl der Fälle offen blieb, warum die Messwiederholung sich positiv auf die Genauigkeit auswirken sollte. 60% der Befragten hingegen wollten die Messung ein- oder mehrmals wiederholen, um den zuvor bestimmten Wert zu überprüfen und ihn durch genügend viele Messungen zu bestätigen.

Die Ergebnisse der vorliegenden Studien lassen vermuten, dass einerseits das Wissen der Lernenden mit zunehmendem Alter (und mit zunehmender Instruktion) wächst, dass das Ergebnis einer Messreihe genauer

ist als das einer Einzelmessung. Es ist aber nur sporadisch Wissen darüber vorhanden, dass dieser Effekt etwas mit der (zufälligen) Streuung der Messdaten zu tun hat. Letztendlich liegt der eigentliche Dreh- und Angelpunkt der Schülerargumentationen lediglich auf dem Wert des Ergebnisses, während zusätzliche Informationen über die Art und Verlässlichkeit dieses Wertes oder Informationen über die Ergebnisproduktion eher eine untergeordnete Rolle spielen.

2.3 Über die Natur eines Messwertes

Wie die oben beschriebenen Studien festgestellt haben, tendiert eine große Anzahl der Lernenden dazu, die Verlässlichkeit einer Messung an der Abweichung des Ergebnisses vom Zielwert festzumachen. Hier geht es letztlich um den Abgleich reiner Zahlenwerte – eine Argumentation, die *Allie et al.* [15] als point Paradigma bezeichnen. In dieser Vorstellung werden die einzelnen Messwerte als voneinander unabhängige (punktgenaue) Zahlenwerte betrachtet. Das Ergebnis einer einzigen Messung kann in diesem Verständnis, sofern sie mit aller Sorgfalt und geeigneten Messgeräten durchgeführt wurde, den wahren Wert einer Messgröße liefern. Eine Messwiederholung dient der Kontrolle und Bestätigung der vorangegangenen Messung und soll eventuelle Fehler ausschließen. Kommen in einer Messreihe Messwerte mehrfach vor, ist dies ein Hinweis darauf, dass es sich hierbei um das richtige Ergeb-

nis handelt ([16]). Als Beispiel enthält Abb. 3 eine Übersicht über Schülerantworten zum Ergebnis einer Messreihe. 44% der Befragten gaben als Ergebnis den Wert an, der in der Messreihe doppelt vorkam. Diesem point-Verständnis steht das sogenannte „set“-Verständnis gegenüber, wonach ein einzelner Messwert nur als eine Näherung an den Wert der Messgröße angesehen wird. Es handelt sich hierbei nicht um einen punktgenauen Zahlenwert, sondern um ein Wertintervall, dessen Schwerpunkt durch den ermittelten Bestwert und dessen Breite durch die Unsicherheit der Messung gegeben sind. Aufgrund dieser Unsicherheit ist es sinnvoll, mehrere Messungen durchzuführen, um verlässliche Informationen über die zu ermittelnde Messgröße zu erhalten. Die Unsicherheit eines Messwertes oder die begrenzte Genauigkeit der Messung sind sowohl entscheidende Faktoren für die Wahl des Versuchsdesigns, die Anzahl und Art aufzunehmender Daten, als auch limitierende Faktoren in der anschließenden Bewertung und Interpretation der Daten. Das set-Verständnis wird als fachlich adäquater und der naturwissenschaftlichen Forschung entsprechender angesehen. Es ist unter Schülerinnen und Schülern deutlich seltener zu finden als Vorstellungen im Sinne der point-Definition. (In der Oldenburger Studie der Schülerinnen und Schüler lag der Anteil der im Sinne des point-Verständnis kodierten Antworten je nach Fragekontext zwischen 70% und 90%.)

Obwohl ungewollt, mag das auch daran liegen, dass durch das weitverbreitete Design der geschlossenen Fragestellung in praktischen Übungen eine point-Vorstellung wenig hinterfragt, wenn nicht sogar bestärkt wird [17]. In solchen Fragestellungen ist der Zielwert von vornherein bekannt, und es ist das Ziel der Messung, diesen Wert zu reproduzieren. Aber auch verschiedene fachliche Aspekte der traditionellen Fehlerrechnung, die zum Teil historisch begründet sind, leisten ihren Beitrag zur Verfestigung der point-Vorstellung oder zumindest zur Verwirrung der Lernenden.

3 | Offizielle Richtlinien zum Umgang mit Messunsicherheiten: Die Entwicklung eines internationalen Standards

Die Schwierigkeiten, die Lernende aber auch Lehrende mit der Diskussion von Messunsicherheiten verbinden, haben ihre Ursache u.a. in einer Reihe fachlicher Unstimmigkeiten. In diesem Kapitel möchten wir

auf einige aufmerksam machen und ihnen eine internationale Norm zum Umgang mit Messunsicherheiten entgegensetzen, die – obwohl bereits 1993 veröffentlicht – im deutschen Bildungsraum bisher wenig Anklang gefunden hat.

3.1 Die Kritik des traditionellen Ansatzes und der Wunsch nach einem einheitlichen internationalen Standard

Das Problem der begrenzten Genauigkeit experimenteller Messungen ist so alt wie die messende Naturwissenschaft selbst. Die verschiedenen Methoden, mit ihr umzugehen, sind vielfältig und spiegeln immer auch das Wissenschaftsverständnis ihrer Zeit wieder. Der quantitative Umgang mit Messunsicherheiten nimmt seinen Anfang in den Arbeiten von *Gauss* und anderen Mathematikern zu Beginn des 19. Jahrhunderts.

Diese unter dem Begriff „Fehlerrechnung“ bekannte Vorgehensweise stellt auch heute im Rahmen des Physikunterrichts oder -studiums das übliche Werkzeug dar, wenn der Prozess und insbesondere das Ergebnis einer Messung kritisch reflektiert und diskutiert werden sollen. Die durch die Fehlerrechnung zu diesem Zweck vorgegebene Routine wird in diesem Artikel auch als „traditioneller“ Ansatz bezeichnet. Dieser Ansatz ist im Wissenschaftsverständnis und den Bedürfnissen der damaligen wissenschaftlichen Forschung seiner Zeit verankert [18]. Aus der Sicht heutiger naturwissenschaftlicher Praxis und Lehre sind mit diesem Ansatz jedoch Probleme verbunden, die zu erheblichen Lernschwierigkeiten führen können. Auf die schwerwiegendsten Unstimmigkeiten soll im Folgenden näher eingegangen werden.

Notation

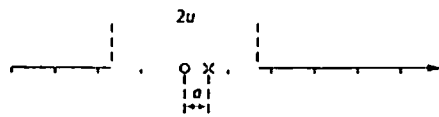
Die Angabe der Unsicherheit der Daten gehört heute zum guten Ton wissenschaftlicher Forschung. Aufgrund der Vielfältigkeit der Ansätze entwickelte sich allerdings auf internationaler oder interdisziplinärer Ebene keine Einigkeit darüber, wie eine solche Unsicherheit anzugeben sei. Betrachtet man beispielsweise die recht trivial anmutende Ergebnisangabe $m = (85 \pm 5)g$. Was besagt dabei die Angabe ± 5 ? Es könnte damit beispielsweise die persönliche Abschätzung des Experimentators gemeint sein, aber auch die Hälfte eines Skalenteils, die Standardabweichung σ , die Standardabweichung des Mittelwertes $\sigma_m = \sigma/\sqrt{n}$, eine erweiterte Standardab-

weichung (z. B. $\pm 2\sigma$ oder $\pm 3\sigma$) oder eine andere Größe. Während in der Physik üblicherweise eine einfache Standardabweichung angegeben wird, sind in der Chemie hingegen doppelte oder dreifache Angaben der Standardabweichung gängig. Fehlt ein entsprechender Standard, ist es nahezu unmöglich, zwei verschiedene Messergebnisse miteinander zu vergleichen.

Terminologie

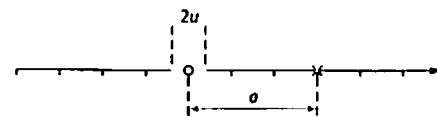
Es ist aber nicht nur die fehlende Einheitlichkeit der Aussagen, die zu einem Unbehagen in der Verwendung der traditionellen Methode führt. Auch die fehlende Eindeutigkeit und Einheitlichkeit der Terminologie bietet Stolpersteine nicht zuletzt auch für die Lernenden. Der Begriff „Fehler“ steht im Rahmen des traditionellen Ansatzes sowohl für Effekte, die für Abweichungen eines Messwerts vom Zielwert führen, als auch für diese Abweichung selbst. Zusätzlich wird der durch die Standardabweichung gegebene Bereich, bei dem es sich um eine Aussage über die Verlässlichkeit und Reproduzierbarkeit eines Ergebnisses handelt, oft als Messfehler bezeichnet. Die Verwendung dieses Fehler-Begriffes ist zudem nicht unproblematisch, da er einerseits im Lernverständnis durchweg sehr negativ belegt ist (Fehlverhalten der Experimentatoren) und darüber hinaus zur Vorstellung verleitet, dass bei fehlerfreier Durchführung des Experimentes eine Abweichung vom Sollwert umgangen werden könne. Um den Begriff „Fehler“ zu umgehen, werden darum in der Schulbuchliteratur häufig weitere Begriffe wie Messunsicherheit, Mess(un)genauigkeit, Abweichung usw. hinzugenommen. Diese werden jedoch oft synonym verwendet, wie das folgende Beispiel eines aktuellen Schulbuchs [19] verdeutlicht: „Hierbei [beim Messen] sind Abschätzungen nötig und es bleiben Messunsicherheiten. Sie werden als Messfehler bezeichnet. [...] Werden Messungen in einer Versuchsreihe wiederholt, treten Schwankungen auf. Die Messwerte streuen um einen Mittelwert. Solche Messungenauigkeiten nennt man statistische Fehler.“ Ob eine solche synonyme Verwendung unterschiedlicher Begriffe vom Standpunkt der Lernenden aus allerdings hilfreich ist, muss aufgrund der Ergebnisse der Oldenburger Studie stark bezweifelt werden. Wie die Ergebnisse zeigen, rufen die einzelnen Begriffe bei den Lernenden sehr unterschiedliche Assoziationen hervor, die sich miteinander nicht in Einklang bringen lassen [4].

Beispiel 1: Große Messunsicherheit und kleine Messabweichung



Der ermittelte Bestwert (gelber Punkt; z. B. der Mittelwert einer Messung) liegt nahe am Zielwert (grauer Kreuz; z. B. ein Literaturwert), d. h. er besitzt eine geringe Messabweichung a , ist jedoch mit einer großen Messunsicherheit u (z. B. die Standardabweichung des Mittelwerts bei Normalverteilung) behaftet. Da sich das Messergebnis prinzipiell an jeder Stelle des von der Unsicherheit festgelegten Intervalls befinden kann (bei einer Normalverteilung liegt ein Messwert mit 68% Wahrscheinlichkeit in dem durch die Standardabweichung u festgelegtem Intervall), ist die Nähe zum wahren Wert zufällig, d. h. es handelt sich um eine zufällig kleine Messabweichung. Deshalb bedeutet eine kleine Messabweichung nicht zwangsläufig, dass das erhaltene Messergebnis „besonders gut“ ist. Erst durch Kenntnis der Messunsicherheit kann die Güte der Messung und somit das Messergebnis selbst bewertet werden.

Beispiel 2: Kleine Messunsicherheit und große Messabweichung



Betrachtet man die Differenz zwischen Bestwert (gelber Punkt; z. B. der Mittelwert einer Messung) und dem Zielwert (grauer Kreuz; z. B. ein Literaturwert), d. h. die Messabweichung a , wirkt das Messergebnis auf den ersten Blick „sehr schlecht“. Die geringe Unsicherheit u zeigt jedoch, dass man sich bezüglich des gemessenen Werts sehr sicher sein kann, d. h. dass die Messung durchaus genau erfolgte (bei einer Normalverteilung liegt ein Messwert mit 68% Wahrscheinlichkeit in dem durch die kleine Standardabweichung u festgelegtem Intervall). Liegt ein Referenzwert vor (z. B. ein Literaturwert), kann in diesem Fall aus der großen Messabweichung gefolgert werden, dass eine systematische Messabweichung vorliegt.

Infokasten 1: Unterschied zwischen Messunsicherheit und Messabweichung anhand zweier Beispiele

Die Wahrscheinlichkeitsdichtefunktion $p(x)$ gibt die Wahrscheinlichkeit dafür an, dass ein bestimmter Wert x gemessen wird. Durch Integration der Wahrscheinlichkeitsdichtefunktion über dem Intervall $x_{best} \pm u$ (bzw. $x_{best} \pm ku$ bei Verwendung einer erweiterten Unsicherheit) erhält man die Wahrscheinlichkeit, Werte für x zu erhalten, die in dieses Intervall fallen. Dies wird als Vertrauensniveau bezeichnet. Die für die Praxis relevantesten Wahrscheinlichkeitsdichtefunktionen sind in den Infokästen 3 bis 5 tabellarisch aufgeführt.

Die auf Grundlage statistischer Betrachtungen oder anderer Erkenntnisse mathematisch modellierte Wahrscheinlichkeitsdichtefunktion enthält sämtliche Informationen über das Messergebnis. So ist ihre durchschnittliche Breite $2u$ ein Maß für die für den Grenzfall angenommene Streuung der Messwerte. Diese ergibt sich aus dem zweiten Moment:

$$(2u)^2 = \int \left(x - \frac{b}{2}\right)^2 p(x) dx.$$

Die Position x_{best} des Maximums von $p(x)$ entspricht dem besten Schätzwert für den Wert der Messgröße.

Infokasten 2: Wahrscheinlichkeitsdichtefunktionen

„Fehlerkategorien“

Ebenfalls schwierig scheint für die Lernenden die Kategorisierung in „zufällige“ und „systematische Fehler“ zu sein. Während systematische Fehler das Messergebnis unter identischen Bedingungen stets auf die gleiche Art und Weise beeinflussen, führen statistische Fehler zu zufälligen Abweichungen vom gesuchten Wert. Diese Kategorisierung, die eines der Fundamente des traditionellen Ansatzes darstellt, erfolgt dementsprechend in Hinblick auf die Wirkung der störenden Einflüsse. Lernende tendieren nach den Ergebnissen der Oldenburger Studie allerdings dazu, die Einflüsse nach ihrer Ursache zu unterscheiden: Einflüsse, die auf den Experimentierenden, die verwendeten

Geräte oder die äußeren Bedingungen zurück zu führen sind. So kommt es mitunter zu einer Vermischung dieser beiden Betrachtungsweisen, sodass beispielsweise die zufälligen Messfehler den Experimentierenden, die systematischen den Geräten zugeordnet werden. Über diese Verwirrung hinaus müssen aber auch fachliche Zweifel an der Berechtigung der Klassifizierung angemeldet werden. Denn einerseits kann aus einer eigentlich zufälligen Abweichung einer Eingangsgröße in der weiteren Rechnung eine systematische werden, falls sie nur ein einziges Mal gemessen und dieser Einzelmesswert in den folgenden Rechnungen verwendet wird. Andererseits sind diese beiden Kategorien nicht trennscharf. Die we-

gnigsten Abweichungen sind vollständig zufälliger oder vollständig systematischer, d. h. konstanter Natur.

An diese fehlende Trennschärfe der Kategorien schließt sich nahtlos die Schwierigkeit an, dass im traditionellen Ansatz nur die zufälligen Unsicherheiten berücksichtigt werden. Dies lässt sich historisch damit begründen, dass die mathematischen Werkzeuge der Korrektur konstanter Einflüsse auch zu Gauss' Zeiten schon sehr ausgefeilt waren. Neu war zu Beginn des 19. Jahrhunderts hingegen die Modellierung solcher Unsicherheiten, die einer gewissen Zufälligkeit unterlagen und daher durch eine Korrekturrechnung nicht beseitigt werden konnten. Dem traditionellen (Gauss'schen) Ansatz liegt daher die Annahme zugrunde, dass im Vorfeld einer Messung alles unternommen wurde, systematische Einflüsse zu korrigieren. In der heutigen Unterrichts- und Lehrpraxis wird allerdings sehr viel weniger Wert auf eine derartige Korrektur systematischer Einflüsse gelegt, als es zur Zeit der Formulierung des Ansatzes üblich war. Dies ist problematisch, da das Ausmaß der Unkenntnis der Effekte, die die Messung systematisch beeinflussen, nicht in die Diskussion der Verlässlichkeit des Messergebnisses einfließt. So bleiben diese unmodellierten Einflüsse als unbekannte Spielverderber abseits der Fehlerrechnung liegen und erhalten zum Einen die Verantwortung für alle Abweichungen vom Sollwert, die am Ende der Ergebnis-

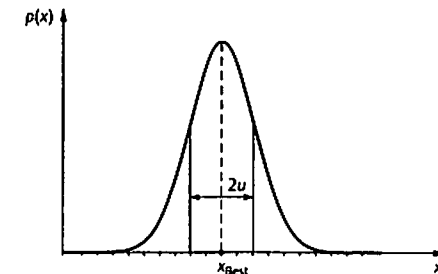
und Fehlerrechnung noch übrig bleiben und bekommen zum Anderen das Label „Nur von erfahrenen Experten zu lösen“ aufgedrückt.

Vermischung von Vertrauensbereichen und Wahrscheinlichkeitsdichtefunktionen

Die traditionelle Fehlerrechnung beschränkt sich also auf die Modellierung von statistischen Streuungen einer (großen, da normalverteilten) Messreihe, während den systematischen Einflüssen durch Korrekturrechnungen begegnet wird. Neben der Problematik, dass solche aufwendigen Korrekturen in Schule und Ausbildung heute nicht mehr adäquate Berücksichtigung finden, stellt sich eine weitere Frage gegenüber der üblichen unterrichtlichen Praxis: Was geschieht in der Umsetzung des traditionellen (statistischen) Ansatzes mit solchen Messgrößen, deren Wert nur einmal bestimmt würde? Die Abschätzung ihrer Unsicherheit ist höchst uneinheitlich und die Verteilung der mit dieser Unsicherheit zusammenhängenden Wahrscheinlichkeit hat nichts mit dem der Normalverteilung der Mehrfachmessung gemein. Die Infokästen 4 und 5 geben hierzu Beispiele. Wenn nun aber die Ergebnisse der Einzel- und der Mehrfachmessung zu einem Ergebniswert verrechnet werden, werden auch oft ihre Unsicherheiten mit Hilfe der Fehlerfortpflanzung oder des Größtfehlergesetzes zu einer Unsicherheit des Ergebniswertes verschmolzen. Dies stellt mathematisch aufgrund der unterschiedlichen Aussagen der berechneten oder abgeschätzten Messunsicherheiten allerdings ein schwerwichtiges Problem dar und führt zu einer Ergebnisunsicherheit, deren Aussage vollkommen undefiniert ist.

3.2 Internationale Einigung: die Erstellung des GUM

Um diese Widersprüche und Uneinheitlichkeiten zu beheben, regte im Jahr 1979 das BIPM (Bureau International des Poids et Mesures) an, eine Arbeitsgruppe einzurichten, die aus den verschiedenen nationalen Standards eine internationale Norm entwickeln sollte. 21 nationale metrologische Staatsinstitute (unter anderem auch das Deutsche Institut für Normung, DIN) reichten auf Anfrage des BIPM ihre bisher verwendeten Verfahren bezüglich des Umgangs mit Messunsicherheiten ein. Eine internationale Arbeitsgruppe entwickelte in den



Mögliche Anwendung	Modelliert in den meisten Fällen die beobachtete Häufigkeitsverteilung der Messwerte bei wiederholten Messungen einer Größe (Ermittlungsmethode A)
Standardunsicherheit	$u = \frac{\sigma}{\sqrt{n}}$ mit σ = Standardabweichung, n = Anzahl der Messungen
Beispiel	Ausschnitt einer Messreihe: T in °C: 28,4; 29,0; 28,9; 25,4; 29,4 ... $x_{best} = 28,2$ °C $u = 0,7$ °C $\rightarrow (28,2 \pm 0,7)$ °C
Vertrauensniveau	Standardunsicherheit: 68% Erweiterte Unsicherheit: 95% mit $k = 2$; 99% mit $k = 3$

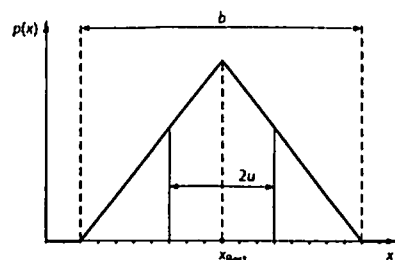
Infokasten 3: Gaußfunktion

folgenden Jahren einen konkreten Leitfa-den, der in der ersten Version 1993 und in der zweiten 2008 als *Guide to the Expression of Uncertainty in Measurement (GUM)* herausgegeben wurde. Neben Definitionen und Erläuterungen der wichtigsten Begriffe des Messwesens beinhaltet der GUM eine konkrete Empfehlung zum Umgang mit und der Erfassung von Messunsicherheiten. Bei einem Blick in die Schulbuchliteratur stellt man fest, dass zum einen das Thema Messunsicherheiten nur in einem Bruchteil der Schulbücher Eingang gefunden hat; in den Büchern der Sekundarstufe I kommt es so gut wie nie vor. Außerdem haben wir bei unseren Analysen bislang kein Schulbuch finden können, das die Empfehlungen des GUM implementierte. Ähnliches lässt sich sogar in der Lehrliteratur feststellen. Der GUM scheint im deutschen Bildungsraum also bislang wenig Resonanz gefunden zu haben. Da offensichtlich an vielen Stellen internationale Kompromisse notwendig wurden, ist der GUM in den vergangenen 15 Jahren nicht frei von Kritik geblieben, vor allem, da er doch noch sehr im Geiste der traditionellen Formulierung steht. Warum es trotzdem lohnenswert ist, sich mit ihm

zu beschäftigen und warum er eine Hilfe sein kann, die Widersprüchlichkeiten des traditionellen Ansatzes zu umgehen, wollen wir im Folgenden aufzeigen.

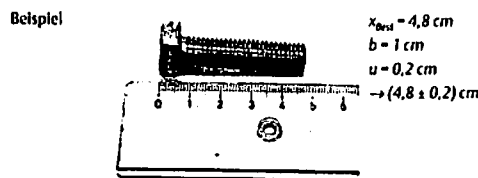
3.3 Chancen des GUM

Zu den oben genannten Kritikpunkten bietet der GUM einige Abhilfe. Eine wichtige Neuerung zum traditionellen Ansatz stellt die Verwendung der Begriffe „Messunsicherheit“ und „Messabweichung“ dar. Die offizielle Definition des Begriffs „Messunsicherheit“ lautet nach GUM [6]: „Dem Messergebnis zugeordneter Parameter, der die Streuung der Werte kennzeichnet, die vernünftiger Weise dem Messergebnis zugeordnet werden können“. Durch die Messunsicherheit wird demnach ein Intervall so definiert, dass man mit einer bestimmten Wahrscheinlichkeit davon ausgehen kann, dass ein weiterer Messwert für die untersuchte Messgröße ebenfalls in dieses Intervall fällt. Die Messunsicherheit ist also ein Maß dafür, wie verlässlich das Messergebnis ist. Die Differenz zwischen Messergebnis und Zielwert (z. B. ein Literaturwert) wird als Messabweichung bezeichnet. Durch diese klare Terminologie lassen sich diver-



Mögliche Anwendung Modelliert die angenommene Häufigkeitsverteilung der Messwerte beim einmaligen Ablesen von einer analogen Anzeige (Ermittlungsmethode B)

Standardunsicherheit $u = \frac{1}{2} \frac{b}{\sqrt{6}}$



Vertrauensniveau Standardunsicherheit: 65%
Erweiterte Unsicherheit: 97% mit $k = 2$ (Mit $k = 3$ wird die Unsicherheit größer als b)

Infokasten 4: Dreiecksfunktion

se Verständnisschwierigkeiten vermeiden, da Messabweichungen eindeutig von Messunsicherheiten abgegrenzt und die damit verbundenen Konzepte nicht durch den Sammelbegriff „Messfehler“ durcheinander gebracht werden (siehe Infokasten 1). Anstelle der traditionellen Kategorien der systematischen und zufälligen Messabweichungen unterscheidet der GUM die Messunsicherheiten anhand ihrer Ermittlungsmethode durch die Kategorien „Typ A“ und „Typ B“. Eine Erfassung von Messunsicherheiten, die durch die statistische Analyse wiederholter Beobachtungen erfolgt, wird vom GUM als Ermittlungsmethode A bezeichnet. Wird die Messunsicherheit auf anderem Wege bestimmt (z.B. durch Literatur- und Geräteherstellangaben oder durch Abschätzung auf Grundlage sämtlicher zur Verfügung stehender Information, schlicht durch alle nicht-statistischen Methoden), so handelt es sich nach dem GUM um die Ermittlungsmethode B. Die Wahl der geeigneten Ermittlungsmethode ist von den Gegebenheiten der jeweiligen Messung abhängig.

Die Ermittlungsmethoden A und B sind mathematisch als gleichwertig anzusehen,

denn beide Ermittlungsmethoden beruhen auf der Modellierung einer Wahrscheinlichkeitsdichtefunktion, die sämtliche zur Verfügung stehenden Informationen über die Messung enthält (z.B. Bestwert und Standardunsicherheit, siehe Infokasten 2). Bei der Ermittlungsmethode A wird diese aus der beobachteten Häufigkeitsverteilung abgeleitet, welche in der Regel der Normalverteilung entspricht (siehe Infokasten 3). Bei der Ermittlungsmethode B wird eine Wahrscheinlichkeitsdichtefunktion a priori angenommen, die nach [6] „auf dem Überzeugungsgrad beruht, dass ein Ereignis eintritt“ (siehe Infokasten 4 und 5). Durch die Ermittlungsmethode B können zudem auch Messunsicherheiten berücksichtigt werden, die sich aus Einzelmessungen einer Größe ergeben. Anstelle subjektiver und uneinheitlicher Abschätzungen lässt sich also nun mithilfe des GUM selbst eine solche Unsicherheit (beispielsweise beim Ablesen eines Zollstocks, eines analogen Amperemeters, einer digitalen Waage) eindeutig und einheitlich angeben.

Mit diesen Einheiten ist es dem GUM gelungen, einen einheitlichen Standard nicht nur zur Erfassung, sondern auch zur Anga-

be der Messunsicherheit zu schaffen. Auf dieser Grundlage kann nun die Verlässlichkeit eines Ergebniswertes eingeschätzt oder das Ergebnis mit anderen Messungen oder einem Referenzwert verglichen werden. Zudem ist es nach dem Ansatz des GUM in jedem Fall mathematisch korrekt, verschiedene Unsicherheiten durch das Fortpflanzungsgesetz miteinander zu verrechnen und somit eine resultierende Wahrscheinlichkeitsdichtefunktion zu modellieren. Durch die klare Terminologie lassen sich außerdem diverse Verständnisschwierigkeiten vermeiden, da Messabweichungen eindeutig von Messunsicherheiten abgegrenzt und die damit verbundenen Konzepte nicht durch den Sammelbegriff „Messfehler“ durcheinander gebracht werden.

Auch wenn momentan der traditionelle Ansatz noch stark verbreitet ist – was sich vielleicht dadurch begründen lässt, dass der Ansatz des GUM eine sinnvolle Alternative zur Fehlerrechnung bieten, die auch eine fruchtbare Bereicherung des Physikunterrichts darstellen könnte.

4.1 Fazit und Ausblick

Der Umgang mit Messfehlern bzw. Messunsicherheiten ist ein komplexes Thema, welches sowohl aus fachlicher als auch aus Sicht der Lernenden mit zahlreichen Schwierigkeiten verbunden ist. Trotzdem ist diese Thematik im Kontext einer Diskussion über die Grenzen der Naturwissenschaft und der Aussagekraft und Verlässlichkeit von Messergebnissen unverzichtbar. Für die Unterstützung der Lernprozesse der Schülerinnen und Schüler ist eine explizite Thematisierung im Unterricht erforderlich. Um die Verständnisschwierigkeiten adäquat begegnen zu können, sollte dabei der Schwerpunkt nicht nur auf der Vermittlung von bestimmten mathematischen Methoden liegen, sondern den Schülerinnen und Schülern die Gelegenheit gegeben werden, die durchzuführenden Routinen nicht nur anzuwenden, sondern auch verstehen zu können.

Um die „zwei Welten“ der Datenbewertung einander anzunähern und ein Verständnis im Sinne des set-Paradigmas zu fördern ist es hilfreich, im Unterricht das Verständnis von Messergebnissen als Wertintervalle anstelle von bloßen Zahlenwerten zu unterstützen. Um ein Gefühl für diese den Messdaten innewohnende Unsicherheit zu entwickeln, könnten die Schülerinnen und Schü-

ler angehalten werden, die Ergebnisunsicherheit auf Basis ihrer Erfahrungen mit dem durchgeführten Versuch quantitativ abzuschätzen. Die Verwendung einer einheitlichen und eindeutigen Terminologie, wie sie im GUM vorgeschlagen ist, kann zudem helfen, Verständnisschwierigkeiten der Lernenden vorzubeugen.

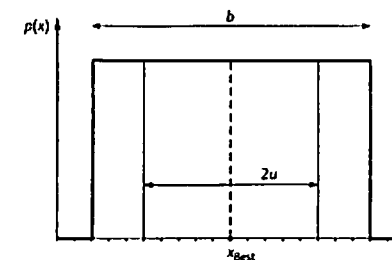
Vor allem offene Aufgabenstellungen können hier eine hilfreiche Unterstützung bieten, um die Notwendigkeit der Betrachtung von Messunsicherheiten im Physikunterricht herauszustellen, da sie nicht auf den Vergleich mit einem autoritativen Zielwert (z.B. einem Literaturwert) hinauslaufen und die Evaluation der aufgenommenen Daten auf die Abweichung von einem vorgegeben Soll reduzieren. Steht der gesuchte Wert einer Messgröße nämlich nicht bereits im Vorfeld fest, gewinnt bei der Bewertung der Ergebnisse eine Diskussion über die Verlässlichkeit der aufgenommenen Messwerte an Bedeutung. Eine gute Diskussionsgrundlage können zudem Untersuchungen bieten, die sowohl voneinander abhängige, als auch voneinander unabhängige Parameter beinhalten [12]. So können beispielsweise die Länge und die Masse eines Pendels in Bezug auf seine Periodendauer betrachtet werden, um anhand der Messunsicherheiten abzuwägen, ob eine Abhängigkeit vorliegt oder nicht.

Eine explizite Thematisierung von Messunsicherheiten im Physikunterricht könnte somit einen entscheidenden Beitrag zur Vermittlung einer adäquaten Sichtweise über die naturwissenschaftliche Arbeitsweise und Interpretation experimenteller Ergebnisse leisten.

Die Forschung zum Verständnis von der begrenzten Genauigkeit experimenteller Messungen unter Lernenden ist im deutschen Bildungsraum noch sehr jung. Aktuelle Projekte befassen sich zur Zeit mit dem Vergleich zwischen theoretischen Überlegungen zu Messunsicherheiten und dem praktischen Vorgehen von Lernenden bei der Durchführung eines Experimentes, der Entwicklung eines Kompetenzstufenmodells zum Verständnis der Thematik über die verschiedenen Klassenstufen hinweg und der Entwicklung und Erprobung von Materialien für den Unterricht, die einer expliziten Diskussion der Thematik Raum bieten.

Literatur

- [1] Acler, D. H.: Imprecision and explanation, *Philosophy of Science* 34, 1967, S. 1–9.
[2] Jamnison, J., Dyson, P., Carratt, J.: Student misconceptions, *U Chem Ed.* 5, 2001, S. 1–8.



Mögliche Anwendung Modelliert die angenommene Häufigkeitsverteilung der Messwerte beim einmaligen Ablesen von einer digitalen Anzeige (Ermittlungsmethode B)

Standardunsicherheit $u = \frac{1}{2} \frac{b}{\sqrt{3}}$



Vertrauensniveau Standardunsicherheit: 58%
Erweiterte Unsicherheit: hier ist eine Erweiterung der Unsicherheit nicht sinnvoll, da bereits für $k = 2$ die Unsicherheit größer als b wird.

Infokasten 5: Rechteckfunktion

- [3] Hon, G.: Towards a typology of experimental errors: An epistemological view, *Studies in History and Philosophy of Science* 20, 1989, S. 469–504.
[4] Henricke, S., Riess, F.: Was heißt denn hier Fehler? – Über die Terminologie der Fehlerrechnung und die Vorstellungen Lernender, eine Erweiterung des „point and set“ Modells, CD zur DPG-Frühjahrstagung 2009. Hrsg. V. Nordmeier.
[5] Guide to the Expression of Uncertainty in Measurement (GUM), International Organization of Standardization, Genf 2008
[6] Leitfaden zur Angabe der Unsicherheit beim Messen, Beuth Verlag, Berlin, Köln 1995
[7] KMK Kultusministerkonferenz (Hrsg.): Bildungsstandards im Fach Physik für den Mittleren Schulabschluss, Wolters-Kluwer, München-Neuwied 2004
[8] Blöck, P.: The purposes of science education, in: R. Hull (Hrsg.): ASE Secondary Science Teachers' Handbook, Simon & Schuster, London 1993, S. 6–22
[9] Mosnick, A. M., & Klahr, D.: Error matters: An initial exploration of elementary school children's understanding of experimental error, *Journal of Cognition and Development* 4, 2003, S. 67–98.
[10] Mosnick, A. M., & Morris, B. J.: Investigating the Development of Data Evaluation: The Role of Data Characteristics, *Child Development*, 79, 4, 2008, S. 1032–1048.
[11] Lubben, F., & Müller, R.: Children's ideas about the reliability of experimental data, *International Journal of Science Education* 18, 1996, S. 955–968.
[12] Konrat, Z., & Müller, R.: Reasoning from Data: How Students Collect and Interpret Data in Science Investigations, *Journal of Research in Science Teaching* 41, 2003, S. 748–769.

- [13] Varelas, M.: Third and fourth graders' conceptions of repeated trials and best representatives in science experiments, *Journal of Research in Science Teaching* 34, 1997, S. 853–872.
[14] Sév, M.-G., & Larcher, C.: Learning the statistical analysis of measurement errors, *International Journal of Science Education* 15, 1993, S. 427–438
[15] Albe, S., Buffler, A., Kouzda, L., Campbell, B., Lubben, F.: First-year physics students' perceptions of the quality of experimental measurements, *International Journal of Science Education* 20, 1998, S. 447–459
[16] Buffler, A., Albe, S., Lubben, F., Campbell, B.: The development of first year physics students' ideas about measurement in terms of point and set paradigms, *International Journal of Science Education* 23, 2001, S. 1137–1156
[17] Wolyn, T., Buffler, A., Albe, S., & Lubben, F.: Impact of a conventional introductory laboratory course on the understanding of measurement, *Physical Review Special Topics Physics Education Research*, 2008.
[18] Henricke, S., Riess, F.: How to cope with Gauss's Errors? – Learners' obstacles and their potential roots in the history of Data Treatment, Tagungsband zur ESE-RA Conference 2009.
[19] Cornelien: *Oberstufe Physik*, 2008, S. 527.

Anschriften der Verfasser
Susanne Henricke, Falk Riess, Carl von Ossietzky Universität Oldenburg; E-Mail: susanne.henricke@uni-oldenburg.de; falk.riess@uni-oldenburg.de
Julia Glomsick, Burkhard Priemer, Ruhr-Universität Bochum, Fakultät für Physik und Astronomie, Lehrstuhl für Didaktik der Physik, 44801 Bochum; E-Mail: priemer@physik.rub.de; glomsick@physik.rub.de