LEUPHANA
UNIVERSITÄT LÜNEBURG

Das Schätzen von Längen, Flächeninhalten, Fassungsvermögen und Rauminhalten

Entwicklung und Auswertung eines Schätztests für Kinder der vierten, fünften und sechsten Klasse

Von der Fakultät Bildung

der Leuphana Universität Lüneburg zur Erlangung des Grades

Doktorin der Philosophie

- Dr. phil. -

genehmigte Dissertation von

DANA FARINA WEIHER

geboren am 25.03.1992 in Soltau

Eingereicht am: 09.08.2022

Mündliche Verteidigung (Disputation) am: 28.03.2023

Erstbetreuerin und -gutachterin: Prof. Dr. Silke Ruwisch

Zweitgutachter: Prof. Dr. Matthias Brand

Drittgutachter: Prof. Dr. Aiso Heinze

Die einzelnen Beiträge des kumulativen Dissertationsvorhabens sind oder werden ggf. inkl. des Rahmenpapiers wie folgt veröffentlicht:

Weiher, D. F. & Ruwisch, S. (2018): Kognitives Schätzen aus Sicht der Mathematikdidaktik. *mathematica didactica 41*(1), S. 77-103. http://fox.leuphana.de/portal/files/15589131/md_2018_Weiher_Ruwisch.pdf

Weiher, D. F. (2019). Framework for the Parallelized Development of Estimation Tasks for Length, Area, Capacity, and Volume in Primary School – A Pilot Study. *Journal of Research in Science, Mathematics and Technology Education, 2*(1), S. 9-28. https://doi.org/10.31756/jrsmte.212

Weiher, D. F. & Ruwisch, S. (under review): The Assessment of Measurement Estimation Results – a Discussion of Different Scorings Regarding to Test Performance and Test Quality.

Weiher, D. F. (under review): Estimation of Length, Area, Capacity, and Volume: Results of a Written Estimation Test

Veröffentlichungsjahr: 2023

Rahmenpapier

Ich danke von Herzen allen, die mich bei meiner Promotion unterstützt haben.

Inhaltsverzeichnis

Tabellenverzeichnis

1 Inhaltliche Verortung und Ziele der Dissertation

„Estimation is a useful, practical skill, and everyone should be able to make estimates."
(O'Daffer, 1979, S. 46)

"[…] remember that estimation is essential, not extra. By all means, don't underestimate the value of estimation!"
(O'Daffer, 1979, S. 51)

Mit diesen Zitaten benennt O'Daffer in einfachen Worten die Relevanz des Schätzens für den Alltag. Weitere Forschende schließen sich dieser Auffassung an (z.B. Brand et al., 2002; Forrester et al., 1990; Joram et al., 1998; Newcombe, 2014; Sowder, 1992). Darüber hinaus besitzt das Schätzen von Größen[1] auch eine mathematikdidaktische Relevanz, nicht nur aufgrund der Nennung in den Curricula (Sowder, 1992; für Deutschland: Kultusministerkonferenz, 2005; für Niedersachsen: z.B. Niedersächsisches Kultusministerium, 2017), sondern auch durch die enge Verknüpfung mit der Entwicklung tragfähiger Größenvorstellungen und dem Messverständnis (Bright, 1976, 1979; Grassmann, 1999; Jones et al., 2012; Joram et al., 2005; O'Daffer, 1979) sowie dem Problemlösen (O'Daffer, 1979).

Trotz der interdisziplinären Einigkeit über die Relevanz des Schätzens weisen sowohl die kognitionspsychologische als auch die mathematikdidaktische Forschung Lücken bezüglich der Schätztestentwicklung auf. Dies bezieht sich insbesondere auf die Auswahl der Größen, auf die Charakteristika der Items sowie auf die Bewertung der Schätzgenauigkeit (Hogan & Brezinski, 2003; Weiher, 2019a; Weiher & Ruwisch, 2022; Weiher, 2022a, 2022b) und führt zu einem unbefriedigenden „Flickenteppich" verschiedener, teils widersprüchlicher Ergebnisse über das Schätzen von Größen (Weiher, 2022a). Insbesondere der mathematische Zusammenhang der vier visuell erfassbaren Größen[2] (die Größen Flä-

[1] In der Mathematikdidaktik wird das Schätzen von Größen, Anzahlen und Rechenergebnissen unterschieden (O'Daffer, 1979; Sowder, 1992). Diese Arbeit bezieht sich, wie in den Publikationen (Anhang A.2 bis A.5) ausführlicher dargestellt und definiert, auf das Schätzen von Größen: „Estimating is the process of arriving at a measurement or measure without the aid of measuring tools. It is a mental process", Bright, 1976, S.89).

[2] Zur besseren Lesbarkeit meint der Begriff „Größen" im Folgenden immer die visuell erfassbaren Größen Längen, Flächeninhalte, Fassungsvermögen und Rauminhalte. Wenn andere Größen gemeint sind, wird dies explizit gekennzeichnet.

cheninhalte, Fassungsvermögen und Rauminhalte können aus der Größe Länge hergeleitet werden) kann so in bestehender Literatur nur ansatzweise herausgearbeitet und für die mathematikdidaktische Forschung zum Schätzen von Größen genutzt werden.

Die übergeordneten Ziele dieser Dissertation sind daher die folgenden:

1) Entwicklung eines Rahmens, in dem die Entwicklung von Schätzaufgaben zu visuell erfassbaren Größen begründet erfolgen kann (Publikation 1 und 2)
2) Erstellung eines schriftlichen Schätztests, der auf diesem Rahmen basiert und für Kinder in vierten, fünften und sechsten Klassen geeignet ist (Publikation 2 und 3)
3) Untersuchung der Schätzgenauigkeit von Kindern in vierten, fünften und sechsten Klassen (Publikation 4)
4) Untersuchung des Zusammenhangs der Schätzgenauigkeit der visuell erfassbaren Größen (Publikation 4)

Die vorliegende Dissertation soll folglich auf zwei Ebenen zur mathematikdidaktischen Forschung beitragen: Zum einen soll die Schätztestentwicklung (Ziele 1 und 2) zu einer verbesserten Erhebung von Schätzfähigkeit im Sinne von mehr Transparenz und Begründbarkeit beitragen. Die ersten beiden Ziele sind somit eher methodisch angelegt. Zum anderen soll durch diese transparente und umfassende theoretische Fundierung des Schätztests die Theorie des Schätzens von Größen inhaltlich erweitert werden, indem für alle vier visuell erfassbaren Größen Aussagen zur Schätzgenauigkeit von Kindern sowie zu den Zusammenhängen zwischen den Größen bezogen auf Schätzgenauigkeit getroffen werden (Ziele 3 und 4). Diese Ziele sind somit eher inhaltlich angelegt. Weitere, diesen Zielen untergeordnete Ziele und die entsprechenden Forschungsfragen (FF) werden in den Publikationen der Dissertation hergeleitet und in den entsprechenden Kapiteln der Zusammenfassungen genannt.

Dieses Rahmenpapier stellt den Zusammenhang der Publikationen dieser Dissertation[3] dar. Dazu werden, eingebettet in die Beschreibung des Forschungsprozesses, wesentliche Aspekte jeder Publikation zusammengefasst. Auf dieser Basis erfolgt eine Reflexion der Publikationen und der zugrundeliegenden wissenschaftlichen Entscheidungen (Kapitel 2 bis 5). Anschließend wird der Forschungsbeitrag des ganzen Dissertations-

[3] Eine Übersucht der Publikationen der Dissertation befindet sich in Anhang A.1. Die Publikationen befinden sich in Anhang A.2 bis A.5. Eine Übersicht über weitere Publikationen, die nicht Teil der Dissertation sind, aber die Argumentation stützen, befindet sich in Anhang B.

(Kapitel 2 bis 5). Anschließend wird der Forschungsbeitrag des ganzen Dissertationsprojekts vor dem Hintergrund der gesetzten übergeordneten Ziele in einer übergreifenden Reflexion diskutiert (Kapitel 6). Im abschließenden Resümee (Kapitel 7) werden Implikationen aus den Ergebnissen und den Reflexionen dieser Dissertation abgeleitet, Limitationen aufgezeigt und weitere Forschungsmöglichkeiten benannt, bevor ein kurzes Fazit gezogen wird.

2 Theoretische Grundlegung für die Entwicklung eines Schätztests

Die erste Publikation (Anhang A.2: Weiher & Ruwisch, 2018) und der Theorieabschnitt der zweiten Publikation (Anhang A.3, S. 9-15: Weiher, 2019a) dieser Dissertation befassen sich mit der theoretischen Vorarbeit zur Entwicklung eines Schätztests. In diesem Rahmen wurden zwei Modelle entwickelt: das Modell zu am Schätzprozess beteiligten Fähigkeiten und das Modell zu parallelisierten Aufgabentypen für die vier Größen. Beide Modelle wurden mit dem Ziel entwickelt, zu einer begründbaren, transparenten und inhaltsvaliden Schätztestentwicklung beizutragen (Ziel 1).

2.1 Am Schätzprozess beteiligte Fähigkeiten (Publikation 1)

2.1.1 Zusammenfassung

Das Schätzen ist eine komplexe kognitive Tätigkeit. So beschreibt Winter (2003, S. 19) das Schätzen als ein „kompliziertes Zusammenspiel von Wahrnehmen, Erinnern, Inbeziehungsetzen, Runden und Rechnen“. Mit diesem Zitat werden die Beteiligung sowohl von eher mathematischen Fähigkeiten (Runden, Rechnen) sowie eher psychologischen Fähigkeiten (Wahrnehmen, Erinnern, Inbeziehungsetzen) am Schätzen deutlich. Für die Entwicklung eines Modells zur Darstellung der am Schätzen von visuell erfassbaren Größen beteiligten Fähigkeiten wird daher ein interdisziplinärer Ansatz gewählt, der Literatur aus beiden Bereichen einschließt. Basierend auf Definitionen beider Disziplinen, Theorien zum Ablauf eines Schätzprozesses aus der Kognitionspsychologie, der Beschreibung von Schätzstrategien aus der Mathematikdidaktik und Studienergebnissen zur Untersuchung von am Schätzprozess beteiligter Fähigkeiten wird in der ersten Publikation dieser Dissertation ein Modell präsentiert, das die Fähigkeiten, die zum Anwenden von Schätzstrategien aus theoretischer Sicht erforderlich sind, strukturiert und möglichst umfassend für die visuell erfassbaren Größen darstellt. Zunächst werden die drei Schritte *Verstehen der Aufgabe*, *Anwenden von Strategien* und *Äußern des Schätzergebnisses* unterschieden, bevor der Fokus auf das *Schätzen im engeren Sinne*, die Strategieanwendung, gelegt wird. Dabei werden drei größere Gruppen von Fähigkeiten, teilweise mit Untergruppen, benannt:

1) Mentales Operieren mit Stützpunkten
    a. Stützpunktwissen/Stützpunktvorstellungen
    b. Wissen über den Messprozess

c. Vergleichen
d. Räumliche Vorstellung

2) Exekutive Funktionen
3) Allgemeines (mathematisches Wissen) und grundlegende Fähigkeiten
    a. Geometrische Kenntnisse
    b. Arithmetische Kenntnisse
    c. Visuelle Wahrnehmung

Alle Untergruppen enthalten eine genauere Aufschlüsselung konkreter Kenntnisse oder Fähigkeiten. Die grafische Darstellung des Modells geht über eine bloße Auflistung hinaus und stellt die Fähigkeiten systematisch dar. Eine Stärke des Modells ist die grundsätzliche Eignung für alle vier visuell erfassbaren Größen. Zudem ist das Modell als Darstellung der Teilfähigkeiten des Schätzprozesses eine Ergänzung zu den häufig detailliert beschriebenen Teilfähigkeiten des Messprozesses (z.B. Stephan & Clements, 2003).
Es wird dabei nicht angenommen, dass für jede Schätzanforderung *immer alle* Fähigkeiten abgerufen werden müssen. Vielmehr handelt es sich um Fähigkeiten, die beim Lösen einer Schätzaufgabe abgerufen werden *könnten*.

2.1.2 Reflexion

Die erste Publikation trägt dazu bei, die Begriffsbestimmung des Schätzens zu vertiefen. Der interdisziplinäre Ansatz der Modellentwicklung ist insofern als Gewinn zu betrachten, als dass die Kognitionspsychologie auf Basis von Schätztests bereits den Zusammenhäng einiger (eher nicht mathematischer) Fähigkeiten mit dem Schätzen empirisch zeigen konnte. Da die Ziele dieser Studien naturgemäß nicht mathematikdidaktischer Natur sind und die Schlussfolgerungen daher ebenfalls nicht, ist eine Interpretation der Ergebnisse aus mathematikdidaktischer Sicht erforderlich. Dies gelingt in der vorliegenden Publikation durch die Verknüpfung der Definitionen aus beiden Disziplinen, aber vor allem durch die Einbeziehung der in mathematikdidaktischer Literatur viel beschriebenen Schätzstrategien. Die Vertiefung des Schätzbegriffs wird erreicht, indem von der Ebene der Strategiebeschreibung – im Sinne der Tätigkeit, wie die schätzende Person zu einem Schätzergebnis gelangt – auf die Ebene der der Tätigkeit zugrundeliegenden kognitiven Prozesse gewechselt wird. Die so teils aus der Literatur entnommenen und teils aus Strategiebeschreibungen abgeleiteten Fähigkeiten sind in dem vorliegenden Modell zusammengetragen und strukturiert.

Bei der Recherche und Auswahl der Literaturbasis, insbesondere der empirischen Studien, ist nach der Schneeballmethode vorgegangen worden. Zwar wurde eine große Anzahl an Literaturquellen, auch aus verschiedenen Jahrzehnten, zusammengetragen und verarbeitet, jedoch können diese – wie an der verwendeten Literatur späterer Publikationen der Dissertation zu sehen – durch insbesondere mathematikdidaktische Studien ergänzt werden, in denen der Zusammenhang zwischen Schätzen und weiteren Fähigkeiten untersucht wird (z.B. Corle, 1960; Gooya et al., 2011; Hildreth, 1983; Hogan & Brezinski, 2003; Jones et al., 2009; Paull, 1971).

Eine empirische Prüfung des Modells steht noch aus (und war nicht Fokus diese Dissertation[4]). Diese Limitation wird zwar in der Publikation angerissen, aber eher als Möglichkeit für weitere Forschung formuliert. Der Publikation kann an entsprechenden Stellen zwar entnommen werden, welche Fähigkeiten aus empirischen Studien stammen und welche Fähigkeiten theoretisch hergeleitet wurden, im Gesamtmodell wird diese Unterscheidung aber nicht mehr vorgenommen. Darüber hinaus wird im Resümee kritisch angemerkt, dass die Schätztests der Kognitionspsychologie aufgrund anderer Ziele aus mathematikdidaktischer Sicht Schwächen aufweisen, insbesondere die Mischung von Aufgaben zum Schätzen von Größen und Aufgaben zum Schätzen von Anzahlen. Auch vor diesem Hintergrund sollte die fehlende empirische Prüfung des Modells als Limitation aufgefasst werden.

Insgesamt überwiegen auch aus heutiger Sicht die Stärken der Publikation und insbesondere des präsentierten Modells, weshalb es weiterhin als gut geeignet für weitere Forschungsschritte angesehen wird.

2.2 Parallelisierte Aufgabenmerkmale (Publikation 2)

2.2.1 Zusammenfassung

Der Theorieabschnitt der zweiten Publikation dieser Dissertation (Anhang A.3, S. 9-15) hat die Darstellung eines Rahmens zur parallelisierten Entwicklung von Schätzaufgaben

[4] In einem Nebenprojekt zu dieser Dissertation wurde die empirische Prüfung des Zusammenhangs von Schätzgenauigkeit und Stützpunkwissen/Stützpunktvorstellungen untersucht (Weiher, 2020a). Der Spearman-Korrelationskoeffizient zwischen der Schätzgenauigkeit und der Punktzahl im Stützpunkte-Test betrug $r = -.316$, $p < .01$ und gibt somit einen Hinweis darauf, dass Kinder mit hoher Punktzahl im Stützpunktetest eine geringere Abweichung im Schätztest zeigen. Aussagen über gerichtete Zusammenhänge wurden nicht getroffen, ebenso bleibt offen, inwieweit der Stützpunktetest die Stützpunkte, die im Schätztest genutzt werden, überhaupt erfasst.

für die vier Größen Längen, Flächeninhalte, Fassungsvermögen und Rauminhalte zum Ziel.

Im theoretischen Hintergrund der Publikation wird zunächst der Schätzbegriff näher erläutert. Auf Basis einer Definition und der in der Literatur beschriebenen Schätzstrategien wird das Schätzen als mentaler Vergleichsprozess mit Stützpunkten beschrieben (z.B. Bright, 1976; Hildreth, 1983; Joram, 2003; Joram et al., 1998; Siegel et al., 1982; Winter, 2003). Anschließend erfolgt eine Verknüpfung der vier Größen sowohl auf mathematischer als auch mathematikdidaktischer Grundlage.
Der dritte Abschnitt des theoretischen Hintergrunds befasst sich mit den Merkmalen von Schätzaufgaben, die in der Literatur zu finden sind. Dabei wird das Modell von Bright (1976) als Ausgangspunkt und das Modell von Heinze et al. (2018) als dessen Weiterentwicklung für das Schätzen von Längen angeführt.

Auf dieser theoretischen Basis wird die Entwicklung eines Modells präsentiert, das sich für alle vier Größen eignet. Der Fokus wird auf die als für visuell erfassbare Größen relevant eingestuften Merkmale *Sichtbarkeit* und *Berührbarkeit* gelegt. Ergänzt wurde dies durch die Nutzung von *standardisierten* oder *nicht standardisierten* Einheiten.
Drei Objekte können Teil einer Schätzaufgabe sein, wobei das zu schätzende Objekt obligatorisch ist. Zusätzlich kann ein Objekt als Repräsentant der gewünschten Einheit (mit Maßzahl 1) und ein Objekt als Repräsentant eines Stützpunktes (mit beliebiger Maßzahl außer 1) in der Schätzaufgabe gegeben sein. Drei Modellannahmen wurden von Beginn an getroffen, um der Anforderung der Parallelisierung für die vier Größen gerecht zu werden:

1) Es werden keine Aufgaben zum Zeichnen im Modell aufgenommen, da dies für Fassungsvermögen und Rauminhalte vermutlich andere oder weitere (schätzfremde) Fähigkeiten als bei Längen oder Flächeninhalten erfordert.
2) Es werden keine Aufgaben mit Bildern im Modell aufgenommen, da durch die Verzerrung bei Fassungsvermögen und Rauminhalte auch hier die Parallelisierung zu Längen und Flächeninhalten nicht gewährleistet werden kann.
3) Das Schätzobjekt und die Einheit können Merkmale zur Sichtbarkeit und Berührbarkeit aufweisen. Für den Stützpunkt hingegen liegt die Modellannahme vor, dass er auf jeden Fall sichtbar sein und mit Größenangabe gegeben muss, um sicher zwischen Aufgaben mit und ohne gegebenem Stützpunkt zu unterscheiden.

Die verbleibenden Merkmale werden in einem Baumdiagramm dargestellt und führen laut dieser Publikation zu 84 theoretisch unterscheidbaren Aufgabentypen. Diese können im weiteren Verlauf der Schätztestentwicklung weiter reduziert werden (dies wird in Kap. 3.1 im Zusammenhang mit der Entwicklung der Pilotversion des Schätztests beschrieben).

2.2.2 Reflexion

Der Theorieteil der zweiten Publikation trägt dazu bei, eine Grundlage für eine begründete, größenübergreifende Schätztestentwicklung zu bieten. Die Relevanz dieses (theoretischen) Forschungsanliegens ist somit auch aus heutiger Sicht hoch.

Die gebotenen Informationen im Theorieabschnitt sind interessant und zeigen erneut einen interdisziplinären Ansatz. Da nur wenige Informationen für die Entwicklung des Aufgabenmodells benötigt werden, könnte der Theorieabschnitt prägnanter dargestellt werden, indem er mit dem Fokus auf die Begründung der gemeinsamen Behandlung der vier visuell erfassbaren Größen, welche gut und nachvollziehbar gelingt, strukturiert wird. Im Folgenden könnte dann die eigentlich gewünschte Fokussierung auf Aufgabenmerkmale (wie der Titel und das präsentierte Modell zeigen) vorgenommen werden. Zudem könnte die Literaturbasis breiter aufgestellt werden: Für wesentliche Aussagen (z.B. zu den Schätzstrategien und für die grundlegende Idee des Messens), sollten Literaturverweise eingefügt werden, um die Aussagen zu legitimieren.

Die gegenüber der ersten Publikation auf Fassungsvermögen erweitere Abbildung zu den Schätzstrategien trägt dazu bei, den Stützpunkt als weiteres Objekt für eine Schätzaufgabe später begründet einzuführen. Der für die spätere Modellentwicklung eigentlich relevante Theorieabschnitt der Publikation (State of the Art: Different Types of Estimation Tasks) fällt hingegen knapp aus. Es liegt nahe und ist auch aus heutiger Sicht aufgrund der Struktur des später präsentierten Modells passend, das Modell von Bright (1976) als Ausgangspunkt und ein weiteres (Heinze et al., 2018) als dessen Weiterentwicklung zu nennen, es fehlen jedoch Ausführungen zu weiteren Studien, die sich auch mit Merkmalen von Schätzaufgaben befassen (z.B. Desli & Giakoumi, 2017; Forrester & Shire, 1994; Heid, 2018; Jones et al., 2012). Die Aufnahme solcher Merkmale wurde zum Zeitpunkt der Entwicklung des Modells als zu große Spezialisierung gesehen, die das entwickelte Modell unübersichtlich und kaum untersuchbar machen würden. Aus heutiger Sicht könnten

diese Merkmale zunächst in das Modell integriert werden, bevor Aufgabentypen begründet ausgewählt werden. Auch wenn die Merkmale nicht vollständig in das Modell aufgenommen werden würden, hätte die Nennung weiterer Merkmale der Auswahl der Merkmale im später präsentierten Modell auf *Sichtbarkeit* und *Berührbarkeit* nicht im Weg gestanden, sondern zu einer erhöhten Transparenz und Begründbarkeit, dass dies eine Auswahl aufgrund der Merkmale der Größen darstellt, beigetragen.

Auch das dargestellte Modell weist Schwächen auf. Zunächst enthält es einen einfachen Fehler: Durch die genannten Merkmale sind nicht (wie in vorigen Versionen des Modells: Weiher, 2018a, 2018b, 2018c) 84, sondern nur 42 Aufgabenkategorien zu unterscheiden. Die Reduktion kommt durch die in der Zwischenzeit getroffene dritte Annahme (Kap. 2.2.1) zustande, nach der ein Repräsentant nur als sichtbares, mit einer Größe versehenes Objekt als Stützpunkt dienen kann.

Aus heutiger Sicht wird diese Annahme – wie schon zuvor – wieder als eher kritisch angesehen. Aus Gründen der Vollständigkeit des Modells und des Ziels, ein umfassendes, auch für weitere Forschende nutzbares Modell zu erstellen, sollten für das Objekt, welches den Stützpunkt repräsentiert, ebenfalls die Merkmale *sichtbar – nicht sichtbar* und *berührbar – nicht berührbar* im Modell aufgenommen werden. Dies würde ermöglichen, in weiteren Studien zu untersuchen, ob die hier auf theoretischer Basis getroffene Annahme Bestand hat. Dazu eigenen sich zum Beispiel Interviews.

Trotz der in der Publikation nicht ausreichend dargestellten Theorie zu Merkmalen von Schätzaufgaben kann das Modell selbst und dessen Darstellung auch aus heutiger Sicht als grundsätzlich geeignet für die Auswahl von Schätzaufgaben der vier Größen angesehen werden. Auch die Diskussion des etablierten Modells von Bright (1976) und dessen Verknüpfung mit einem aktuellen Modell (Heinze et al., 2018) sowie dem der vorliegenden Publikation ist als positiv hervorzuheben, da es einen Ausgangspunkt für weitere theoretische Entwicklungen bieten kann. Der Versuch einer eindeutigen Definition des Stützpunkts in Schätzaufgaben, insbesondere in Abgrenzung zur Einheit, macht deutlich, dass theoretische Vorarbeit unbedingt notwendig ist, um einen mathematikdidaktischen Schätztest zu entwickeln. Auch aus heutiger Sicht kann daher das Modell der vorliegenden Publikation, wenngleich es Potential für Verbesserungen aufweist, als grundlegend geeignet und dem damaligen eigenen Forschungsstand und -prozess entsprechend nachvollziehbar eingestuft werden.

3 Pilotstudie: Erste Version des Schätztests (Publikation 2)

Der empirische Teil der zweiten Publikation (Anhang A.3, S. 15-27) widmet sich der Erstellung eines Pilotschätztests auf Basis des Aufgabenmodells. Die Entwicklung und Prüfung der grundsätzlichen Eignung der Fragen und Form des Tests entspricht dem Ziel 2.

3.1 Zusammenfassung

Ziel der zweiten Publikation ist (neben der Modellentwicklung für Schätzaufgaben, Kap. 2.2) die Entwicklung und Erprobung einer Pilotversion des Schätztests. In diesem Zusammenhang sollen zwei Forschungsfragen beantwortet werden:

> FF 1: Sind die Aufgabentypen und Größen geeignet für Kinder der dritten und vierten Klasse?
>
> FF 2: Welche empirischen Unterschiede können zwischen den Merkmalen der Aufgaben bestimmt werden?

Auf Basis des zuvor präsentierten Gesamtmodells wird weiterführend die Auswahl der Aufgabentypen für die Pilotversion des Schätztests beschrieben. Zu den Reduktionsgründen gehören schätztheoretische, strukturelle oder test-praktische Aspekte:

- Durch ein berührbares Schätzobjekt in Kombination mit einem berührbaren weiteren Objekt sind Messprozesse möglich.
- Durch Erfüllung eines Merkmals wird ein weiteres erfüllt, das nicht erfüllt sein soll.
- Durch Erfüllung eines Merkmals wird ein anderes nicht mehr erforderlich.
- Die Nennung der Einheit ist zur Unterscheidung von Fassungsvermögen und Volumen immer nötig.
- Berührbarkeit erfordert alle entsprechenden Objekte für Fassungsvermögen und Volumen im Klassensatz. Aufgrund des hohen Materialaufwands werden Aufgaben mit berührbaren Objekten nicht mit in die Studie aufgenommen.
- Aufgrund der schwierigen Definition des Konstrukts *Stützpunkt* und der damit verbundenen unklaren Repräsentation im Schätztest werden Aufgabentypen mit Stützpunkten nicht in den Test aufgenommen.

Nach der Reduktion verbleiben acht Aufgabentypen im Schätztest (Tabelle 1).

Tabelle 1

Aufgabentypen in der Pilotversion des Schätztests

		Schätzobjekt sichtbar, nicht berührbar	Schätzobjekt nicht sichtbar
Einheit standardi-siert	Einheit sichtbar, nicht berührbar	Typ 1	Typ 2
	Einheit nicht sichtbar	Typ 3	Typ 4
Einheit nicht standardi-siert	Einheit sichtbar, nicht berührbar	Typ 5	Typ 6
	Einheit nicht sichtbar	Typ 7	Typ 8

Die Stichprobe für die Pilotierung besteht aus drei dritten und drei vierten Klassen mit insgesamt 137 Kindern. Jede Klasse bearbeitete die Pilotversion des Tests (Anhang C.1) zu zwei Größen. Die Bewertung der Schätzgenauigkeit erfolgte durch Berechnung der prozentualen Abweichung.

Der größenspezifische Vergleich der Perzentile für die Merkmale *Schätzobjekt sichtbar – nicht sichtbar*, *Einheit sichtbar – nicht sichtbar* und *Einheit standardisiert – nicht standardisiert* zeigt, dass es für die meisten Merkmale Unterschiede bei der Tendenz zur Über- bzw. Unterschätzung und bei der Schätzgenauigkeit gibt.

In der Diskussion wird neben Erklärungsversuchen der einzelnen Ergebnisse auch auf die Begrenzung der Skala der prozentualen Abweichung bei Unterschätzungen hingewiesen. Das hat zur Folge, dass die Schätzgenauigkeit nicht unabhängig von der Tendenz zur Über- oder Unterschätzung betrachtet werden kann. Zudem wird vorgeschlagen, *Schätzen von Größen* nicht als eindimensionales Konstrukt zu betrachten.

Darüber hinaus wird festgestellt, dass die Pilotversion des Schätztests nicht umfänglich für Kinder der dritten Klasse geeignet ist. Insbesondere die standardisierten Einheiten für Rauminhalte bereiteten den meisten Kindern der dritten (und auch vielen der vierten Klasse) erhebliche Schwierigkeiten.

3.2 Reflexion

Die Gründe für die Nennung der Reduzierung während der Entwicklung der Pilotversion des Schätztests werden – immer vor dem Hintergrund der angestrebten Parallelisierung – inhaltlich transparent dargestellt, lediglich das Nebengütekriterium der Ökonomie

(Moosbrugger & Kelava, 2012) hätte als Begründung für den Ausschluss aufgrund des hohen Materialaufwand ergänzt werden können. Zum leichteren Verständnis, wie aus der Gesamtanzahl der Pfade letztlich die acht verbleibenden Aufgabentypen übrig bleiben, hätte eine Nennung der reduzierten Pfade beigetragen[5].

Die Unterscheidung in schätz-theoretische, strukturelle und test-praktische Gründe impliziert eine Struktur, die deutlicher in der Gliederung der Publikation hätte berücksichtigt werden können: Die schätz-theoretischen und strukturellen Gründe gehören zur Entwicklung des Theorierahmens, auf dessen Basis Aufgaben für einen Test ausgewählt werden. Diese Auswahl erfolgt unter Berücksichtigung test-praktischer Gründe und gehört zur Entwicklung bzw. Präsentation der Pilotversion des Schätztests. Dieser Prozess bzw. sein Ergebnis sollte im Kapitel zu Testinstrumenten, welches in der Publikation ja existiert, ausgeführt werden. Beispielaufgaben und die Nennung der Realwerte (Anhang C.2) hätten zudem dazu beigetragen, dass das Testinstrument transparent und verständlich dargestellt wird (und nicht nur dessen Herleitung). Zudem hätte ergänzt werden können, dass ein Manual (Anhang C.3) und ein Beobachtungsbogen (Anhang C.4) zur Erhöhung der Durchführungsobjektivität vorliegen.

Die Entscheidung, keine Repräsentanten für Stützpunkte in den Test mit aufzunehmen, ist aus heutiger Sicht bedauerlich. So stellt die Ebene der Stützpunkte im Modell eine Neuerung dar, die Potential für eine Untersuchung geboten hätte und für diese Publikation bzw. die Pilotversion des Schätztests interessant gewesen wäre. Schwerwiegender ist aber die Begründung, die in der Publikation für den Ausschluss geboten wird: so ist eine „schwierige Definition“ aus heutiger Sicht keine ausreichende Begründung. Zudem scheint die Publikation an dieser Stelle (S. 15, letzter Absatz) die vorher geführte Argumentation für die Merkmale der Stützpunkte nicht stringent fortzusetzen, da auch die Möglichkeit eingeräumt wird, dass nur ein Aspekt (Sichtbarkeit oder Größenangabe) hilfreich sein könnte. Nachvollziehbarer wäre hier die Begründung gewesen, dass nicht sichergestellt werden kann, dass die schätzenden Personen die Stützpunkte auch tatsächlich für ihren Schätzprozess nutzen (wie in der Diskussion auch für die Objekte, die die Einheit repräsentieren, erläutert wird).

[5] Möglicherweise wäre bei diesem Vorgehen auch der Fehler, dass es sich nicht um 84, sondern nur um 42 Gesamtpfade handelt (Kap. 2.2.2), deutlich geworden.

Die Darstellung der Auswertungsmethoden ist prägnant und verständlich. Die Wahl der prozentualen Abweichung als klassische Methode der Fehlerberechnung hätte durch Literatur gestützt werden können. Die Präsentation der Ergebnisse erfolgt aspektorientiert bezogen auf den Anteil an Über- und Unterschätzungen sowie Schätzgenauigkeit und wird auch aus heutiger Sicht als nachvollziehbar strukturiert verständlich angesehen.
Inhaltlich weist die Auswertung der Pilotstudie einige Schwächen auf. Die Nutzung der Perzentile und deren Darstellung in Tabellen trägt zwar zu hoher Transparenz bei und erlaubt detaillierte Einblicke in die Verteilung der Daten, die fast ausschließlich deskriptive Auswertung lässt allerdings keine Schlüsse auf Signifikanz oder Effektstärke zu[6]. Zudem hätte für viele der Erkenntnisse, die aus den Perzentilen abgeleitet werden, auch das arithmetische Mittel oder, aufgrund der Schiefe der Daten, der Median herangezogen werden können. Dies hätte die Ergebnisdarstellung deutlich gestrafft und Raum für statistische Tests gegeben.

Der Vergleich der vier Größen, insbesondere bzgl. der Schätzgenauigkeit, ist aufgrund der Zusammensetzung der Stichprobe problematisch. Zwar scheint die ungleiche Verteilung der Jahrgänge auf die Größen (mehr dritte Klassen bearbeiteten Schätzaufgaben zu Längen und mehr vierte Klassen bearbeiteten Schätzaufgaben zu Rauminhalten) wenig Einfluss auf die erwartete Reihung der Größen zu haben (z.B. Joram et al., 1998; Gooya et al., 2011; Grassmann, 1999; Heid, 2018; Pike & Forrester, 1997; Ruwisch et al., 2015), dieser Aspekt hätte aber zwingend in der Diskussion oder als mögliche Limitation aufgegriffen werden müssen. Alternativ müsste der Vergleich der Größen gar nicht in die Publikation mit aufgenommen werden, da dies nicht Teil der Forschungsfragen ist und demnach die Stichprobenverteilung nicht darauf ausgelegt wurde.

Die Schiefe der Daten wird benannt und nachvollziehbar mit der für Unterschätzungen begrenzten Skala in Verbindung gebracht. Beides wird später als Grund angeführt, weshalb die Schätzgenauigkeit nicht unabhängig von der Tendenz zu Über- oder Unterschätzungen betrachtet werden sollte, was aus heutiger Sicht eine bedeutsame Implikation der Publikation darstellt. Leider wird in der Diskussion hauptsächlich auf inhaltliche und

[6] Im Laufe des weiteren Forschungsprozesses konnte dem mit einer weiterführenden Auswertung ansatzweise begegnet werden (Weiher, 2019b). Der Wilcoxon-Vorzeichen-Rang-Test intendiert die folgenden Tendenzen: Aufgaben zu Längen mit sichtbarem Schätzobjekt werden genauer geschätzt (eher kein Unterschied bei Fassungsvermögen, Flächen- und Rauminhalt). Aufgaben zu Flächen- und Rauminhalten mit sichtbaren Einheiten werden genauer geschätzt (kein Unterschied bei Längen und Fassungsvermögen). Aufgaben zu Fassungsvermögen, Flächen- und Rauminhalten mit standardisierter Einheit werden ungenauer geschätzt (kein Unterschied bei Längen).

nicht auf methodische Aspekte eingegangen, sodass die Relevanz dieses Aspekts nicht deutlich genug wird.

In dieser Phase des Forschungsprozesses wurde, wie durch die Vermeidung eines Scorings und der Inkaufnahme von Schwierigkeiten bei Nutzung der prozentualen Abweichung implizit erkennbar ist, eine Problematik hinsichtlich der willkürlichen Auswahl von Scorings bereits erkannt. Dies hätte im Methodenkapitel erläutert und in der Diskussion stärker aufgegriffen werden können. In diesem Zusammenhäng hätte auch die Berechnung des arithmetischen Mittels (statt der Bildung einer Summe, wenn ein Scoring genutzt worden wäre) pro Kind und Größe diskutiert oder als Limitation benannt werden müssen: Das arithmetische Mittel ist anfällig für Ausreißer, wie sie, wie auch in der Studie angedeutet wird, bei der Nutzung der prozentualen Abweichung durchaus vorkommen. Aus heutiger Sicht soll weniger die Nutzung der prozentualen Abweichung, sondern vielmehr das nicht erkannte Diskussionspotential kritisiert werden.

Vor dem Ziel der Publikation in Verbindung mit dem entwickelten Modell der Aufgabentypen sollte außerdem eine Begründung in die Publikation aufgenommen werden, weshalb keine konfirmatorische Faktoranalyse, wie sie eigentlich als strukturprüfendes Verfahren an dieser Stelle des Forschungsprozesses für ein solches Modell üblich ist (Backhaus et al., 2015) vorgenommen wurde. Der alleinige Hinweis auf die zu kleine Stichprobe wird aus heutiger Sicht als nicht zufriedenstellend angesehen. Es hätte ebenfalls angeführt werden können, dass die Anzahl der Items pro angenommenem Faktor (Aufgabentyp) zu gering ist (sie sollte mindestens vier betragen, Backhaus et al., 2015). Vor allem aber kann nicht sichergestellt werden, dass es sich beim vorliegenden Messmodell um ein reflektives Messmodell handelt, welches erforderlich für die konfirmatorische Faktoranalyse ist (Backhaus et al., 2015). Für ein reflektives Messmodell müsste davon ausgegangen werden, dass Genauigkeit bei den entsprechenden Items z.B. von der latenten *Schätzfähigkeit von nicht sichtbaren Objekten* abhängt, während für ein formatives Messmodell diese Beziehungsrichtung umgekehrt wird (Backhaus et al., 2016) und sich die latente Variable *Schätzfähigkeit von nicht sichtbaren Objekten* erst aus der Formulierung der Items ergibt (und selbst evt. Teil der übergeordneten latenten Variable *Schätzfähigkeit* ist). Für diese Annahme spricht auch die geringe Korrelation der Items untereinander (Backhaus et al., 2016). Dieser Umstand hätte auch in die Diskussion mit aufgenommen werden können, sodass diese nicht erklärend verbleibt, sondern auch the-

oriebildende Fragen aufwirft. An dieser Stelle soll nicht die Entscheidung gegen die konfirmatorische Faktoranalyse, sondern die fehlende Erläuterung bzw. Diskussion in der Publikation kritisiert werden.

Um tatsächlich einen Vergleich zwischen Aufgaben mit standardisierten und Aufgaben mit nicht standardisierten Einheiten für Flächeninhalte und Rauminhalte anstellen zu können, hätten Schülerinnen und Schüler der Sekundarstufe befragt werden müssen, wie aus den Bildungsstandards (Kultusministerkonferenz, 2005) und dem Niedersächsischen Kerncurriculum (Niedersächsisches Kultusministerium, 2017) hervorgeht. Zudem hätte zwingend die Thematisierung von Abhängigkeiten in die Diskussion aufgenommen werden müssen: Um dem ohnehin sehr hohen Materialaufwand entgegen zu wirken, wurden viele Objekte als Träger mehrerer Größen behandelt. Zudem kommen sie auch innerhalb einer Größe mehrmals vor, weshalb viele Aufgaben nicht unabhängig voneinander sind. Trotz des umfassenden Potentials zur Verbesserung trägt die Auswertung der Pilotstudie dazu bei, einen Test auf strukturierter und erprobter Basis für eine inhaltliche Fragestellung zu entwickeln. Während in vielen mathematikdidaktischen Studien Tests für inhaltliche Fragestellungen nicht erprobt werden (oder zumindest dies nicht berichtet wird) und eventuell beobachtete Schwierigkeiten mit dem Testinstrument höchstens in den Limitationen berichtet werden, können die Ergebnisse des Tests der zweiten Publikation als Anregung für die weitere Testentwicklung genutzt werden. Die Ergebnisse haben nicht nur eine Relevanz für den eigenen Forschungsprozess, sondern sind auch für andere Forschende potentiell nutzbar.

4 Hauptstudie: Zweite Version des Schätztests

Mit den Erkenntnissen aus der Pilotstudie wurde die Hauptstudie konzipiert. Ein besonderer Fokus lag auf der Entwicklung eines Bewertungsrahmens für einen schriftlichen Schätztest. Aus diesem ging die dritte Publikation (Anhang A.4: Weiher & Ruwisch, 2022) hervor. Da sowohl die dritte als auch die vierte Publikation (Anhang A.5: Weiher, 2022a; Kap. 5) auf derselben Erhebung basieren, wird im Folgenden zunächst die Aufgaben- und Stichprobenauswahl für die Hauptstudie beschrieben (Kap. 4.1), bevor ein Teil der Entwicklung des Bewertungsrahmens für den Schätztest durch die Vorstellung und Reflexion der dritten Publikation erläutert wird (Kap. 4.2).

4.1 Entwicklung der Schätzaufgaben und Auswahl der Stichprobe

4.1.1 Schätzaufgaben

Um mehr Schätzaufgaben pro Aufgabentyp zu ermöglichen und außerdem sicherzustellen, dass für die Untersuchung aller vier Größen nur eine Schulstunde benötigt wird, wurden aus den acht Aufgabentypen aus der Pilotstudie drei Aufgabentypen für die Hauptstudie ausgewählt:

1) Schätzobjekt nicht sichtbar, Einheit nicht sichtbar
2) Schätzobjekt sichtbar, Einheit nicht sichtbar
3) Schätzobjekt nicht sichtbar, Einheit sichtbar

Es wurde angenommen, dass bei Schätzprozessen dieser Aufgabentypen am meisten der Fähigkeiten aus dem Modell der ersten Publikation abgerufen werden müssen: Bei Schätzprozessen mit nicht standardisierten Einheiten kommt vermutlich dem Vergleichsprozess an sich eine erhöhte Bedeutung zu, und daher weniger den anderen Fähigkeiten des Modells. Dasselbe gilt für Aufgaben, bei denen die Repräsentanten für Schätzobjekt und Einheit gleichzeitig sichtbar sind. Durch Vermeidung dieser Aufgaben soll die Inhaltsvalidität des Tests erhöht und eine versehentliche isolierte Testung von Teilfähigkeiten des Schätzprozesses vermieden werden.

Die Auswahl der Repräsentanten für Schätzobjekte und Einheiten erfolgte unter Berücksichtigung der folgenden Kriterien:

- Angenommene Bekanntheit bei Kindern
- Gute Möglichkeit zur eindeutigen Beschreibung von nicht sichtbaren Objekten
- Keine Verwechslung mit anderen Objekten des Tests
- Für nicht sichtbare Objekte: Größenausprägung nur in einer Ausführung bzw. genaue Beschreibung möglich (z.B. die lange Seite eines 5€-Scheins)
- Für Flächeninhalte und Rauminhalte: nur rechteckige bzw. quaderförmige Objekte
- Möglichst ganzzahlige Realwerte

Um den Einfluss des Schätzobjekts an sich (z.B. durch seine Bekanntheit) bei der Untersuchung verschiedener Aufgabentypen zu verringern, wurde der Test in zwei Versionen erstellt (Version A: Anhang C.5; Version B: Anhang C.6). Die Schätzobjekte des Aufgabentyps 1 in Testversion A sind die Schätzobjekte des Aufgabentyps 3 in Testversion B (und andersherum). Die Schätzobjekte für Aufgabentyp 2 sind in beiden Testversionen gleich.

Die Einheiten sind für alle Größen in einem Antwortsatz vorgegeben und gleich über die Aufgabentypen verteilt (Länge: cm, m; Flächeninhalt: cm^2, m^2; Fassungsvermögen: ml, l; Rauminhalt: cm^3, m^3). Alle Schätzobjekte mit ihren Realwerten sind in Anhang C.7 dargestellt.

Zur Erhöhung der Durchführungsreliabilität[7] liegt ein Manual vor (Anhang C.8). Abweichungen davon und weitere für die Auswertung möglicherweise relevante Daten sind auf einem Beobachtungsbogen zu notieren (Anhang C.9).

4.1.2 Stichprobe

Ausgehend von den Ergebnissen der Pilotstudie und nach Sichtung der Bildungsstandards (Kultusministerkonferenz, 2005) und des Niedersächsischen Kerncurriculums (Niedersächsisches Kultusministerium, 2017) wurde festgelegt, dass an der Hauptstudie Kinder der vierten, fünften und sechsten Klassen teilnehmen sollen.

Die Stichprobengenerierung erfolge, indem im Wohnortumfeld der Autorin alle Schulen angeschrieben und nach einer Projektteilnahme gefragt wurden. Alle Schulen, die sich

[7] Wenige der Datenerhebungen wurden von Studierenden im Rahmen ihrer Abschlussarbeiten vorgenommen, sodass hier eine Klärung der Durchführungsschritte von besonderer Relevanz ist. Dennoch sollte durch das Manual auch sichergestellt werden, dass die Durchführung von einer Person immer gleich ist.

zurückgemeldet haben, haben auch an der Studie teilgenommen, es erfolgte keine weitere Selektion.

Es nehmen zwar verschiedene Arten der weiterführenden Schulen teil, somit wurde die Vielfalt der weiterführenden Schulen näherungsweise berücksichtigt, die Stichprobe ist aber dennoch aufgrund des Auswahlverfahrens nicht als repräsentativ für die Grundgesamtheit anzusehen.

4.2 Entwicklung der Auswertungsmethode (Publikation 3)

4.2.1 Problemstellung

Während es in der kognitionspsychologischen Forschung üblich ist, die Genauigkeit einer Schätzung im Sinne einer „normalen" Schätzgenauigkeit durch den Vergleich mit einer Normstichprobe zu beurteilen (Axelrod & Millis, 1994; Brand et al., 2003; Bullard et al. 2004; Della Sala et al., 2003; Mendez et al., 1998; Shallice & Evans, 1978), ist das klassische Vorgehen in der mathematikdidaktischen Forschung die Berechnung einer prozentualen Abweichung in Verbindung mit einem normativ gesetzten Scoring (Corle, 1963; Desli & Giakoumi, 2017; Heid, 2018; Hogan & Brezinski, 2003; Hoth et al., im Druck; Huang, 2014; Siegel et al., 1982; Swan & Jones, 1980). Da in der Literatur selten Gründe für die Auswahl eines Scorings genannt werden, sollte in der Dissertationsstudie zunächst der Schätzfehler nur mit der prozentualen Abweichung *beschrieben* werden – statt mit einem Scoring *bewertet*. Die Begrenzung der Skala der prozentualen Abweichung für Unterschätzungen führt aber zu mehreren Problemen:

- Gemeinsam mit der Tendenz zu Unterschätzungen (Weiher, 2019a, 2022b) führt die Begrenzung zu einer hohen Schiefe (bei der Betrachtung einzelner Items), welche die statistische Auswertung erschwert.
- Die Gütekriterien interne Konsistenz und Trennschärfe, welche bei der Testentwicklung bedeutsam sind (Moosbrugger & Kelava, 2012) weisen bei Anwendung der prozentualen Abweichung nicht zufriedenstellende Werte auf (Weiher, 2022b).
- Bei der Nutzung des Betrags der prozentualen Abweichung als Basis für das Scoring werden Über- und Unterschätzungen gleich bewertet. Es kann diskutiert werden: Ist die Schätzung 1,2 mm für ein Schätzobjekt mit der Länge 12 cm wirklich genauso gut wie die Schätzung 23,88 cm? Ist die prozentuale Abweichung als Basis für ein Scoring geeignet (Weiher, 2022b)?

Da die theoretische Basis für die Wahl einer Auswertungsart für den Schätztest als nicht zufriedenstellend eingestuft wurde, wurde die umfassende Stichprobe für die Untersuchung verschiedener Arten der Berechnung und Bewertung von Schätzfehlern genutzt. Ein Teil der Ergebnisse (die Untersuchung der Unterschiede zwischen den Scorings bezogen auf Testleistung und Testqualität) wurden in der dritten Publikation verarbeitet[8], welche im Folgenden beschrieben wird.

4.2.2 Zusammenfassung

Fokus der dritten Publikation ist die Untersuchung der Auswirkungen von verschiedenen Scorings auf die Testleistung und Testqualität mit dem Ziel, zur Transparenz bei der Scoringauswahl für einen schriftlichen Schätztest beizutragen und für die Notwendigkeit der Begründung zu sensibilisieren. In diesem Zusammenhang sollen folgende Forschungsfragen beantwortet werden:

> FF 1: Inwiefern unterscheiden sich die Scorings bezüglich der Testleistung, d.h. a) der Gesamtpunktzahl und b) der Rangreihenfolge der Schülerinnen und Schüler?
>
> FF 2: Inwieweit unterscheiden sich die Scorings bezüglich der Testqualität, d.h. a) der internen Konsistenz und b) der Trennschärfe?

Der Theorieabschnitt stellt zunächst die Berechnung der prozentualen Abweichung vor. Im Anschluss werden verschiedene, auf der prozentualen Abweichung beruhende Scorings aus der Literatur präsentiert und deren Eigenschaften und Unterschiede analysiert. Die theoretische Analyse legt nahe, dass die Wahl des Scorings nicht trivial ist. So unterscheiden sich die Scorings in der Anzahl der Intervalle und in der Wahl der Grenzen der Intervalle. Zudem werden Begriffe wie Genauigkeit oder Angemessenheit (im Original „accuracy“ (z.B. Huang, 2014; Siegel et al., 1982), „acceptability“ (Huang, 2014) oder „reasonableness“ (Desli & Giakoumi, 2017) nicht einheitlich verwendet.

Im Anschluss an die Analyse erfolgt die Präsentation der sechs ausgewählten Scorings, die in der Studie untersucht werden, und die Nennung der Forschungsfragen (s.o.).

[8] Die dritte Publikation liegt nach mehrfacher grundlegender Überarbeitung auf Basis verschiedener Reviews erneut den Gutachtern vor. In diesem Prozess wurden zugunsten einer klareren Argumentationslinie und auf Anregung der Reviews die Ausführungen zu alternativen Methoden der Schätzfehlerberechnung („logarithmischer Fehler“, „Teilen durch den größeren Wert“) aus dem Manuskript entfernt. Diese sind für das Dissertationsprojekt aber von entscheidender Bedeutung, weshalb sie in verkürzter Form auf einer Tagung präsentiert werden (Weiher, 2022b).

Die Stichprobe für diese Studie umfasst 615 Schülerinnen und Schüler aus 13 fünften und 13 sechsten Klassen. Der Schätztest wird vorgestellt, indem die Merkmale der Schätzaufgaben beschrieben und mit Beispielitems illustriert werden. Zudem wird der Umgang mit fehlenden Werten erklärt. Die Auswahl der vorgenommenen statistischen Analysen wird mit Bezug auf die Forschungsfragen begründet.

Bevor die Ergebnisse bezogen auf die Forschungsfragen präsentiert werden, wird ein Überblick über die Datenlage gegeben. Insgesamt lassen sich folgende zentrale Schlüsse aus den Daten ableiten:

- Schülerinnen und Schüler neigen zu Unterschätzungen und zur Verwendung ganzzahliger Schätzergebnisse (insbesondere mit Maßzahlen 5, 10 oder Vielfachen davon)
- Alle Paare von Scorings unterscheiden sich hinsichtlich der Testleistung signifikant voneinander.
- Eine Spearman-Korrelationsanalyse zeigt hohe, signifikante Korrelationen zwischen allen Paaren von Scorings. Der Korrelationskoeffizient Kendalls-tau-b ist etwas niedriger, die Reihung der Scorings bezüglich der Höhe des Koeffizienten ist aber im Wesentlichen gleich.
- Die Scorings unterscheiden sich bezüglich der internen Konsistenz und Trennschärfe. Die interne Konsistenz, aber insbesondere die Trennschärfe ist als gering bis höchstens moderat einzuschätzen.

Die Ergebnisse werden vor dem Hintergrund mathematikdidaktischer und methodischer Implikationen diskutiert. Schwerpunkt der inhaltlichen Diskussion ist die Auffassung von *Schätzgenauigkeit*, die durch die Wahl des Scorings erst gebildet wird, und die Frage nach der Wahl verschiedener Scorings für verschiedene Größen, Aufgabentypen oder Altersgruppen. Darüber hinaus wird auf die Angemessenheit der in der Studie verwendeten Methoden (z.B. Spearman-Korrelationsanalyse aufgrund von Bindungen, Cronbachs Alpha aufgrund von Dimensionalität), aber auch mögliche Auswirkungen auf zukünftige Forschung (z.B. Veränderung der Wirksamkeit bei Interventionsstudien durch unterschiedliche Rangordnung, Einfluss unterschiedlicher Reliabilität auf die Korrelation) eingegangen.
In der weiteren Diskussion wird der Einfluss des Schätzobjekts bzw. der entsprechenden Größenangabe auf die Schätzgenauigkeit und damit die systematische Über- oder Unterschreitung von Scoringintervallen in den Blick genommen.

Nach Präzisierung der Limitationen wird im Fazit herausgestellt, dass die Wahl des Scorings unbedingt vom Ziel der Studie, vielleicht auch von der Stichprobe, der verwendeten Größen und Aufgabentypen abhängig sein und sehr sorgfältig begründet werden sollte.

4.2.3 Reflexion

Die methodische Arbeit, die mit dieser Publikation präsentiert wird, wird als bedeutsam angesehen, da nahezu alle inhaltlichen Aussagen über Schätzgenauigkeit (auch in Verbindung mit Strategieeffizienz, Unterschiede zwischen Aufgabentypen, Korrelation mit weiteren Fähigkeiten) auf der Verwendung der prozentualen Abweichung in Verbindung mit einem Scoring beruhen. Die Publikation stellt somit eine Möglichkeit dar, eventuell unterschiedliche inhaltliche Ergebnisse bestehender Studien zu erklären, und ist ein Teil der Legitimierung der Wahl eines alternativen Berechnungs- und Bewertungsverfahrens der Schätzgenauigkeit in der eigenen Hauptstudie (Kap. 5).

Die Argumentation der Publikation orientiert sich durchgehend an den beiden Forschungsfragen und ist damit gut verständlich. In der Einleitung wird das Ziel transparent herausgestellt und dabei auch verdeutlicht, in welchem größeren Zusammenhang die Studie steht und weshalb daher der Fokus auf die untersuchten Aspekte gelegt wird (Testleistung und Testqualität). Die Beschreibung der theoretischen Grundlagen bezieht sich ausschließlich auf dieses Ziel und stellt daher keine übergeordneten Informationen zum Schätzen dar (wie z.B. Strategien oder Eigenschaften der Größen), sondern präsentiert lediglich in übersichtlicher Tabellenform in der Literatur genutzte Scorings. Die dann folgende Analyse der Unterschiede zwischen den Scorings erfolgt tiefgehend, aber prägnant, und bezieht sich auf nachvollziehbare Merkmale der Scorings (z.B. Intervallgrenzen und Intervallanzahl, Größen, Stichproben). Die Forschungsfragen werden schließlich auf Basis der begründet ausgewählten sechs Scorings gestellt, der rote Faden der Publikation ist auch hier erkennbar. Das weitere Vorgehen und die verwendeten Materialien werden transparent im Methodenabschnitt beschrieben. Im Abschnitt zu den statistischen Analysen könnte methodische Literatur ergänzt werden. Da an entsprechenden Stellen bei Entscheidungen, die grundsätzlich diskutabel sind (z.B. bei der Robustheit der rmANOVA gegenüber der Verletzung der Normalverteilungsvoraussetzung) Literaturangaben die getroffenen Entscheidungen stützen, wird dies als weniger gravierend angesehen.

Die verwendeten Methoden werden als passend eingeschätzt, die damit generierten Ergebnisse führen zum Ziel der Publikation. Ebenfalls positiv hervorgehoben werden kann

das Berichten der Prüfungsergebnisse für die Voraussetzungen der angewendeten Methoden. Sollte die Eignung der Methoden fraglich sein (wie z.B. bei der Berechnung der Korrelation mit Spearmans Rho), wurde eine alternative Methode durchgeführt, um die Ergebnisse abzusichern (z.B. die Berechnung der Korrelation mit Kendall-tau-b).

Die Präsentation der Ergebnisse erfolgt strukturiert und gestützt durch Abbildungen und Tabellen. Der erste Teilabschnitt, der einen Überblick über die Daten zeigt, gehört zwar nicht direkt zu einer Forschungsfrage, ist aber für die Interpretation der Ergebnisse relevant und wird durch Rückbezüge in der Diskussion inhaltlich eingebunden.
Die Diskussion erfolgt strukturiert in Bezug auf die Forschungsfragen, bezieht mathematikdidaktische und methodische Aspekte ein und stellt so die Relevanz der Untersuchungen dieser Publikation insbesondere für weitere Forschung zur Schätzgenauigkeit noch einmal deutlich hervor.

Zwei Erweiterungen bezüglich der verwendeten Methoden bzw. der berichteten Ergebnisse sollen hier vorgeschlagen werden: Bei der Beurteilung der internen Konsistenz könnte neben des Alpha Koeffizienten auch die Interitemkorrelation in den Blick genommen werden. Dies könnte bedeutsam sein, da der Test eine recht hohe Anzahl an Items aufweist und so die interne Konsistenz positiv beeinflusst wird (Bortz & Döring, 2006; Moosbrugger, 2012; Rammstedt, 2010), während die Interitemkorrelation unter Umständen weniger hoch ausfällt (Cortina, 1993; Green et al., 1977; Streiner, 2003). Bei der Beurteilung der Trennschärfe hätte zudem die Verteilung der Ergebnisse der einzelnen Items untersucht werden können, da eine geringe Varianz (zum Beispiel bei besonders schwierigen Items) eine geringere Trennschärfe zur Folge hat (Bortz & Döring, 2006; Kelava & Moosbrugger, 2012, Kemper et al., 2015). Beispielhaft könnten Items zur Darstellung ausgesucht werden, deren Trennschärfe bei Verwendung unterschiedlicher Scorings stark variiert (z.B. Item 19 oder 23 in Abbildung 5 der Publikation). In der Diskussion könnte darauf zurückgegriffen werden.

Darüber hinaus könnte die Entscheidung für Gütekriterien der klassischen Testtheorie (in Abgrenzung zur probabilistischen Testtheorie) umfassender begründet werden. Da dies aber eher der Testentwicklung als der Untersuchung verschiedener Scorings zugehörig ist, und die Testentwicklung als solche nicht in dieser Publikation besprochen wird, soll dies nur als Möglichkeit und nicht als Mangel aufgezeigt werden. Aus demselben Grund

wird auch die Analyse der Dimensionalität, z.B. durch eine Faktoranalyse, welche eigentlich vor Prüfung der von Gütekriterien wie Reliabilität und Trennschärfe durchgeführt wird (Kemper et al., 2015), in dieser Publikation nicht thematisiert. Diese Entscheidung ist auch aus heutiger Sicht nachvollziehbar.

5 Hauptstudie: Auswertung des Schätztests (Publikation 4)

Die vierte Publikation (Anhang A.5: Weiher, 2022a) fokussiert die inhaltliche Auswertung des Schätztests. Dabei werden zwei inhaltliche Aspekte berücksichtigt: die Auswertung der Schätzgenauigkeit und der Vergleich verschiedener Gruppen sowie die Untersuchung der vier Größen und deren Relationen. Dies entspricht den Zielen 3 und 4.

5.1 Zusammenfassung

Mit der vierten Publikation sollen die folgenden Forschungsfragen beantwortet werden:

FF 1: Wie genau schätzen Kinder der vierten, fünften und sechsten Klasse visuell erfassbare Größen?

FF 2: Unterscheidet sich die Schätzgenauigkeit von Viert-, Fünft- und Sechstklässlern?

FF 3: a) Korreliert die Schätzgenauigkeit von Viert-, Fünft- und Sechstklässlern mit dem Alter und b) unterscheidet sie sich zwischen den Geschlechtern?

FF 4: Unterscheidet sich die Schätzgenauigkeit zwischen den Größen?

FF 5: Korreliert die Schätzgenauigkeit der vier Größen untereinander?

FF 6: Kann die Schätzgenauigkeit von Längen als Prädiktor für die Schätzgenauigkeit der anderen Größen dienen?

In der theoretischen Grundlegung werden Studienergebnisse zur Genauigkeit beim Schätzen von Größen präsentiert. Dabei wird auf die Schätzgenauigkeit im Allgemeinen (z.B. Corle, 1960; Jones et al., 2012; Joram et al., 2005; Sowder, 1992, Swan & Jones, 1980), die Schätzgenauigkeit der Größen im Vergleich (z.B. Gooya et al., 2011; Heid, 2018; Joram et al., 1998), die Schätzgenauigkeit verschiedener Altersgruppen (z.B. Corle, 1960, 1963; Forrester & Shire, 1994; Hildreth, 1980; Harel, 2007; Ruwisch et al., 2015; Pike & Forrester, 1997) sowie die Schätzgenauigkeit der Geschlechter (z.B. Corle, 1960; Desli & Giakoumi, 2017; Harel, 2007; Paull, 1971) eingegangen.

Die Stichprobe für die Auswertung umfasst nach der Reduzierung aufgrund fehlender Werte, welche in der Publikation ausführlich beschrieben wird, 900 Kinder aus 22 vierten, 13 fünften und 13 sechsten Klassen.

Der Schätztest wird detailliert beschrieben und durch Beispielitems für jede Größe und jeden Aufgabentyp gut verständlich präsentiert. Es wird erklärt, wie die Schätzgenauigkeit mit der logarithmischen Abweichung berechnet wird. Die Bewertung erfolgt über ein ebenfalls logarithmisches Scoring. Die Methoden werden vor ihrer Anwendung kurz mit den entsprechenden Forschungsfragen verknüpft.
Die Ergebnisse werden entsprechend den Forschungsfragen präsentiert. Nach Darstellung der deskriptiven Statistik zur Beantwortung der ersten Forschungsfrage erfolgt eine detaillierte Beschreibung der weiteren Ergebnisse. Als besonders zentral sollen die folgenden hier hervorgehoben werden:

- Die Schätzgenauigkeit unterscheidet sich statistisch signifikant zwischen Viert- und Sechst- sowie Fünft- und Sechstklässlern, nicht jedoch zwischen Viert- und Fünftklässlern.
- Es kann keine statistisch bedeutsame Korrelation zwischen der Schätzgenauigkeit und dem Alter festgestellt werden, wenn für den Jahrgang kontrolliert wird.
- Für Flächeninhalt und Rauminhalte gibt es keinen statistisch signifikanten Unterschied zwischen Jungen und Mädchen. Für Längen und Fassungsvermögen ist der Unterschied statistisch signifikant, weist aber eine geringe Effektstärke auf.
- Die Unterschiede bezüglich der Schätzgenauigkeit zwischen den Größen sind für alle Jahrgänge statistisch signifikant. Etwa 70 % der Varianz kann durch die Größe erklärt werden.
- Die Schätzgenauigkeiten aller Größen korrelieren statistisch signifikant, aber moderat untereinander.
- Die Schätzgenauigkeit von Längen kann als Prädiktor für die anderen Größen genutzt werden.

In der Diskussion werden unter Rückgriff auf Literatur zum Schätzen, Messen oder dem Verständnis von Größen im Allgemeinen die Ergebnisse erklärt, bewertet und verknüpft. Auch Aspekte des äußeren Rahmens der Studie wie der Zeitpunkt der Erhebung und der Schulwechsel in Deutschland werden mit einbezogen.
Nach Nennung von zwei Limitationen, die sich auf das Testinstrument beziehen, werden abschließend Implikationen für den Mathematikunterricht sowie die weitere mathematikdidaktische Forschung abgeleitet.

5.2 Reflexion

Die Einleitung zeigt prägnant die Bedeutung des Schätzens, die Herausforderungen bei der Erfassung und Bewertung von Schätzungen und damit verbunden das Forschungsdesiderat und das Ziel der Studie. Durch die explizite Nennung der Forschungsfragen und die Fokussierung der Theoriebasis auf Aspekte zu diesen Fragen ist der rote Faden der Publikation auch aus heutiger Sicht gut verständlich. Die zweigeteilte Auswertung – ein Fokus auf Schätzgenauigkeit der Kinder, ein Fokus auf Theorien zum Schätzen der vier Größen – ist durch die Gruppierung der Fragen sowie durch die Abschnitte der Ergebnisse, die sich darauf beziehen, ebenfalls gut verständlich dargestellt. Die Forschungsfragen 2 bis 5 könnten durch die Formulierung „Inwiefern…“ verbessert werden.

Die Literaturbasis wird als umfangreich eingeschätzt und eignet sich dazu, Hypothesen zu den Forschungsfragen abzuleiten; lediglich die Studienergebnisse zur Schätzgenauigkeit in verschiedenen Altersgruppen hätte detaillierter dargestellt werden können.
Im Abschnitt zu den Methoden werden der verwendete Schätztest und die Stichprobe transparent und gut nachvollziehbar beschrieben. Dabei enthält die Publikation aber zwei Fehler: Die Erhebung der Studie fand nicht im Jahr 2020, sondern im Jahr 2019 statt, und der kleinste Realwert des Schätztests ist in jeder Größe 1 (und die korrespondierende Einheit). Die Fehler können in der nächsten Reviewrunde korrigiert werden.

Der Umgang mit fehlenden Werten wird transparent beschrieben. Die Entscheidung, die Schätzgenauigkeit von Rauminhalte der Fünftklässler aufgrund zu vieler fehlender Werte nicht mit in die Untersuchung aufzunehmen, ist vor dem Hintergrund der vierten bis sechsten Forschungsfrage nachvollziehbar: Es können keine verlässlichen Aussagen über das Schätzen verschiedener Größen getroffen werden, wenn vermutet werden muss, dass für eine dieser Größen vermehrt geraten worden ist. Für die erste bis dritte Forschungsfrage kann die hohe Anzahl der fehlenden Werte jedoch durchaus von Bedeutung sein, sodass darauf in der Diskussion noch einmal hätte verwiesen werden können.

Die verwendete Berechnung für die Schätzgenauigkeit wird präsentiert und deren Vorteile gegenüber der klassischen Methode (prozentuale Abweichung in Verbindung mit einem Scoring) erläutert. Dies ist in der Ausführlichkeit als erforderlich einzuschätzen, da es sich um ein selbst entwickeltes Verfahren handelt (mit Bezug auf Weiher, 2022b; Weiher & Ruwisch, 2022). Das verwendete Scoring wurde in zwei Schritten sorgfältig ausgewählt: Zunächst wurde die Testqualität (Interne Konsistenz und Trennschärfe) für

verschiedene Methoden der Fehlerberechnung verglichen. Da beobachtet wurde, dass eine Verbesserung der Parameter für interne Konsistenz und Trennschärfe durch ein Scoring herbeigeführt werden kann, wurden verschiedene, selbst gewählte (und damit ebenfalls teils willkürliche) Scorings auf Basis der logarithmischen Fehlerberechnung angewendet und schließlich das geeignetste in Bezug auf interne Konsistenz und Trennschärfe unter 18 geprüften Scorings ausgewählt. Dieser Prozess oder zumindest die Ergebnisse für das verwendete Scoring könnten in der Publikation noch ergänzt werden, um hier den Eindruck von Beliebigkeit zu vermeiden. Alternativ könnte in der Diskussion – oder aufgrund des Fokus der Publikation in den Limitationen – die Wahl des Scorings aufgegriffen und in Bezug zu anderen Studien gesetzt werden, da durch die Entwicklung eines eigenen Scorings die Vergleichbarkeit zu anderen Studien erschwert wird.

Die Auswahl der Methoden könnte durch entsprechende Literatur legitimiert werden.
Die Darstellung der Ergebnisse ist umfassend und bezieht sich immer auf die konkrete Forschungsfrage. So ist, trotz der vielen statistischen Verfahren, die Präsentation übersichtlich und die Argumentation gut erkennbar.

In der Analyse wurden sowohl parametrische (z.B. mixed rmANOVA, lineare und multivariate Regression) als auch nicht parametrische (z.B. Mann-U-Test, Spearman-Korrelationsanalyse) Verfahren verwendet. Diese wurden anhand der Erfüllung ihrer Voraussetzungen oder der entsprechenden Robustheit ausgewählt. Nicht immer wurde die Prüfung der Voraussetzungen in der Publikation berichtet, so fehlt z.B. für den Mann-U-Test für Forschungsfrage 1 oder den Kruskal-Wallis-Test für Forschungsfrage 2 der entsprechende Verweis. In den Fällen, in denen die Prüfung der Voraussetzungen und der Umgang mit den Ergebnissen berichtet wird, trägt dies zu einer höheren Transparenz bei[9] und zeigt, dass die Ergebnisse sorgfältig und verantwortungsvoll ausgewertet wurden.

Insgesamt werden die Ergebnisse aus heutiger Sicht als bedeutsam und empirisch belastbar eingestuft. Sie lassen sich gut vor dem Hintergrund der Literatur interpretieren und scheinen die Theorie zum Schätzen von Größen in wesentlichen Punkten zu ergänzen.
Eine noch bessere Verständlichkeit der deskriptiven Statistik, die damit verbundene aussagekräftigere Antwort auf die erste Forschungsfrage sowie eine bessere Nutzbarkeit der

[9] Durch die Veröffentlichung des Datensatzes könnte die Transparenz – und der Beitrag zur Forschung – noch erhöht werden. Aus datenschutzrechtlichen Gründen ist das leider nicht möglich: den teilnehmenden Personen wurde versichert, dass die Daten nur zum Zwecke dieser Dissertation verwendet werden.

Ergebnisse für den Mathematikunterricht hätte durch die Verwendung des logarithmischen Schätzfehlers ohne ein zusätzliches Scoring erreicht werden können. Aufgrund der Gütekriterien interne Konsistenz und Trennschärfe, die sich durch Nutzung des Scorings verbessern, war die Entscheidung zwar überlegt und auch aus heutiger Sicht nachvollziehbar, ist aber aufgrund der fehlenden Angaben in der Publikation für die lesende Person nicht transparent.

Die vorliegende Auswertung hätte um die folgenden beiden Aspekte ergänzt werden können, um die Qualität der Ergebnisse zu verbessern:

- Stichprobe: Durch die Schulklassen liegt eine Gruppierung vor, die als Klumpenstichprobe[10] interpretiert werden könnte (Bortz & Schuster, 2010). Schülerinnen und Schüler einer Klasse könnten ähnlichere Schätzgenauigkeit zueinander haben als zu Schülerinnen und Schüler anderer Klassen. Inwieweit die Klasse Einfluss auf die Schätzgenauigkeit hat, wurde nicht untersucht.
- Prädiktoren: Mit einer einfachen linearen Regression hätte zusätzlich untersucht werden können, inwieweit die Genauigkeit beim Längenschätzen als Prädiktor für alle weiteren Größen insgesamt (Summenscore aus allen Größen) oder für 3D-Objekte (Summenscore aus Fassungsvermögen und Rauminhalte) dienen kann.

Die Schlüsse, die aus den Ergebnissen gezogen werden, werden unter Bezugnahme auf eine breite Literaturbasis umfassend diskutiert. Dabei gelingt es, die Beobachtungen zu erklären und daraus Implikationen für den Mathematikunterricht sowie die weitere Forschung zum Schätzen abzuleiten. Diese sind aus heutiger Sicht als nachvollziehbar und gut begründet einzustufen. Insbesondere hervorzuheben ist hier die Verbindung der Erkenntnisse zum Messen bzw. zu Größen aus der Literatur und der Erkenntnisse zum Schätzen aus der vorliegenden Studie. An dieser Stelle hätte in der Diskussion (oder in den Limitationen) deutlicher gemacht werden können, dass die vermuteten Zusammenhänge meist auf Anwendung von Strategien oder dem Verständnis der jeweiligen Größe beruhen, welche in dieser Studie nicht untersucht worden sind. Es besteht daher die Möglichkeit der Formulierung eines weiterführenden Forschungsinteresses, das zusätzlich zu den beiden Desideraten, die im Fazit genannt werden, erwähnenswert wäre.

[10] Sollten die Schulen (und nicht die Klassen) als Klumpen interpretiert werden, würde es sich eher um eine Ad-hoc-Stichprobe handeln, da die Schulen nicht vollständig zufällig ausgewählt bzw. angeschrieben wurden (Bortz & Schuster, 2010). Die folgenden Aussagen sind dennoch gültig.

6 Übergeordnete Reflexion des Forschungsprozesses

In diesem Kapitel werden publikations- und forschungsfragenübergreifende Aspekte und Entscheidungen während des Forschungsprozesses unter Bezugnahme auf die gesetzten Ziele (Kap. 1) reflektiert.

Der **interdisziplinäre Ansatz** dieses Dissertationsprojekts ist gewinnbringend in mehreren Schritten des Forschungsprozesses. Die ausführliche Studienlage der Kognitionspsychologie konnte dazu genutzt werden, das mathematikdidaktische Konstrukt *Schätzen von Größen* detailliert darzustellen: Die beiden theoretischen Modelle zu Fähigkeiten des Schätzprozesses und Aufgabenmerkmalen basieren auf einer umfassenden Literaturlage und weisen eine durchdachte Systematik auf. Dadurch entstand ein transparenter Ausgangspunkt für die Schätztestentwicklung (Ziel 1). Die kognitionspsychologische Literatur öffnete zudem den Blick für alternative Bewertungen der Schätzgenauigkeit. Für die Beschreibung und Bewertung der Schätzgenauigkeit wurde in dieser Studie ein alternatives Verfahren entwickelt (Kap. 4.2 und 5). Diese Entscheidung ist auch aus heutiger Sicht sinnvoll: es liegt in der Literatur kein zufriedenstellender Test mit einer Normstichprobe vor, zudem entspricht es mathematikdidkatischen Konventionen, die Schätzgenauigkeit über die Abweichung zum Realwert zu berechnen. Nach Vorbild der Kognitionspsychologie könnte aber eine Normierung eines Schätztests angestrebt werden (Kap. 7). Daher trägt der interdisziplinäre Ansatz zu einer begründeten Berechnung und Bewertung der Schätzgenauigkeit bei, was Teil der Schätztestentwicklung ist (Ziel 2).

Während der Arbeit an der theoretischen Basis des Dissertationsprojekts wurde großer Wert auf **klare Definitionen wichtiger Begriffe** und deren Abgrenzung untereinander, wie z.B. *Schätzen – Messen, Stützpunkt – Einheit,* und *sichtbar – nicht sichtbar – berührbar – nicht berührbar* gelegt. Die theoretische Analyse der Literatur trägt somit schon zu Beginn des Forschungsprozesses zur mathematikdidaktischen Forschung bei (Kap. 2 und 7.1). Zudem gelang es, die Relevanz der Entwicklung eines theoretischen Rahmens für die Schätztestentwicklung und damit das theoretische Forschungsdefizit klar zu benennen. Diesem wurde durch die Entwicklung der Modelle (Fähigkeiten beim Schätzen und Aufgabentypen) entgegengewirkt (Ziel 1; Kap. 7.1). Auch im späteren Verlauf des Forschungsprozesses wird der Anspruch der Begriffsklärung deutlich: durch die Analyse verschiedener Benennungen von Schätzgenauigkeit, z.B. *accurate – accepptable –*

reasonable, wird die in der Literatur uneinheitliche Verwendung dieser Begriffe in Verbindung mit verschiedenen Scorings kritisiert und als Ausgangspunkt für die Entwicklung einer eigenen Bewertungsart im Rahmen der Testentwicklung (Ziel 2; Kap. 4.2) genutzt.

In den Publikationen wurde zudem großer Wert auf **Transparenz** gelegt. Dies bezieht sich auf getroffene Entscheidungen zu den Testinstrumenten, die erhobenen Daten und die Voraussetzungen für die Anwendung von statistischen Methoden, insbesondere bei der Bearbeitung der Ziele 2, 3 und 4. Das Anliegen war, keine „blackbox"-Publikationen zu präsentieren, in denen inhaltliche Ergebnisse generiert werden, die für andere Forschende weder prüfbar sind noch weiterverwendet werden können. Dazu gehört auch, statistische Methoden sorgfältig auszuwählen und bei größeren Zweifeln nicht anzuwenden (z.B. die konfirmatorische Faktoranalyse, Kap. 3.2). Es ist im Laufe der Zeit eine Steigerung diesbezüglich in den Publikationen sichtbar, die sich sowohl auf die eigenen Fähigkeiten der Methodenauswahl und deren Begründung, aber auch auf deren transparente Darstellung bezieht.

Bei der Entwicklung des Schätztests (Ziel 2) wurde **vom typischen Ablauf der Testentwicklung** (z.B. nach Eid & Schmidt, 2014; Kemper et al., 2015) **abgewichen**. Die Wahl eines Bewertungsrahmens für einen schriftlichen Test gehört nach Eid & Schmidt (2014) zu den vorbereitenden Tätigkeiten. Allerdings wurde erst in einem späteren Stadium des Forschungsprozesses – bei der Auswertung der Hauptstudie – endgültig deutlich, dass das übliche und zunächst auch hier angestrebte Verfahren zur Fehlerbeschreibung, nämlich die Berechnung der prozentualen Abweichung, nicht geeignet erscheint (Weiher & Ruwisch, 2022; Weiher, 2022b). Daher wurde die inhaltliche Auswertung der Ergebnisse der Hauptstudie pausiert und die Untersuchung verschiedener Arten der Berechnung und Bewertung von Schätzfehlern untersucht. Dieses Vorgehen wird im Nachhinein als unbedingt notwendig eingeschätzt, wie an der Reflexion der Auswertung der Pilotstudie (Kap. 3.2) und den Ergebnissen der dritten Publikation (Kap. 4.2) deutlich wird. Aus heutiger Sicht sind diese Probleme schon während der Pilotstudie sichtbar und werden auch in der entsprechenden Publikation angerissen. Da sie zu diesem Zeitpunkt nicht in diesem Maße problematisch eingeschätzt wurden, ist das weitere Vorgehen zunächst verständlich. Aus forschungstheoretischer Sicht wäre es jedoch sinnvoller gewesen, die dann erhobenen Daten der Hauptstudie als eine zweite Pilotstudie zur Evaluierung der Bewertungsmethoden anzusehen (als Teil von Ziel 2) und für die weitere inhaltliche Auswertung (Ziele 3 und 4) neue Daten zu erheben (Jonkisz et al., 2012). Aus zeitökonomischen Gründen

wurde nach der Erhebung für die Hauptstudie auf eine erneute Erhebung in einem solch großen Ausmaß (gewünscht waren etwa 1000 Kinder) verzichtet.
Dadurch, dass eine substanzielle Frage der Testentwicklung (Entwicklung der Bewertungsart) am gleichen Datensatz erfolgt, der für die inhaltliche Auswertung genutzt wurde, kann der Vorwurf entstehen, dass die Bewertungsmethode auf Basis gewünschter inhaltliche Ergebnisse (z.B. hohe Korrelationen zwischen den Schätzgenauigkeiten der Größen) ausgewählt wurde. Dieser Vorwurf kann dadurch entkräftet werden, dass die Untersuchung und Wahl des Scorings zeitlich vor der Bearbeitung der Ziele 3 und 4 und damit vor der inhaltlichen Auswertung liegt, und dass die erhobene Stichprobe keine Normstichprobe für den Test darstellt.

Im Laufe des Forschungsprozesses wurde wiederholt die **Frage nach der Dimensionalität des Konstrukts Schätzen** aufgegriffen. Bereits in der zweiten Publikation wird darauf hingewiesen, dass die gewünschte Parallelität zwischen den Größen nicht durch Tendenzen bezüglich Unter- und Überschätzen sowie Schätzgenauigkeit empirisch ohne weiteres sichtbar ist. Die Forderung, dass die Größen einzeln untersucht werden sollen, wird daraus abgeleitet. Bei der Untersuchung der Scorings in der dritten Publikation wird hiervon wieder Abstand genommen, d.h. alle Größen werden zu einem Test zusammengefasst und gemeinsam ausgewertet. Diese Entscheidung kann vor dem Hintergrund des gewählten Gütekriteriums, der internen Konsistenz, als kritisch angesehen werden, wie auch in den Limitationen der dritten Publikation benannt wird. Da jedoch die Untersuchung der Scorings am selben Testinstrument erfolgt, und alle Scorings somit denselben Einschränkungen bezüglich der Dimensionalität unterliegen, ist die Entscheidung dennoch nachvollziehbar (Kap. 4.2.3). In der vierten Publikation schließlich werden die Größen wieder einzeln betrachtet, was vor dem Hintergrund der Ziele 3 und 4 mit einer mathematikdidaktischen Legitimierung, aber ohne eine testtheoretische Sicherstellung durch eine Skalenanalyse geschieht. Auch diese Entscheidung ist aufgrund des Zweifels am Messmodell (Kap. 3.2) verständlich, jedoch wurde dadurch ein Ansatz, der sich eigentlich zwingend aus der systematischen Testentwicklung ergibt und damit ein Gewinn für das Dissertationsprojekt und auch für die Forschung über das Schätzen darstellen könnte, nicht verfolgt. Dennoch wird angenommen, dass die fehlende Skalenanalyse die Ergebnisse zu den Zielen 3 und 4 nicht einschränkt, da die getrennte Betrachtung von Größen in der Forschung und auch in der Schule nicht ungewöhnlich ist.

7 Resümee

In diesem Kapitel werden aus den Publikationen und deren Reflexionen (Kap. 2 bis 5) sowie aus der übergreifenden Reflexion (Kap. 6) Implikationen (Kap. 7.1) und Limitationen (Kap. 7.2) abgeleitet. Ausgehend davon werden dann weitere Forschungsmöglichkeiten benannt (Kap. 7.3). Das Kapitel schließt mit einem kurzen Fazit (Kap. 7.4).

7.1 Implikationen

Aus dem vorliegenden Dissertationsprojekt können Implikationen auf mathematikdidaktischer, methodischer und psychologischer Ebene abgeleitet werden.

Aus dem Fähigkeiten-Modell (Publikation 1) kann abgeleitet werden, dass der Schätzprozess ein komplexer kognitiver Prozess ist. Gemeinsam mit dem Aufgaben-Modell (Publikation 2), das trotz seiner Limitation auf wenige Merkmale (Kap. 3.2) umfangreich ist, wird daher deutlich, dass die Operationalisierung des Konstrukts *Schätzen* nicht trivial ist. Beide Modelle stellen somit für die mathematikdidaktische Forschung eine Möglichkeit, aber auch eine Aufforderung dar: Schätzaufgaben sollten begründet und transparent für eine Studie ausgewählt werden. Auch für die schulische Mathematikdidaktik können aus den Modellen Implikationen abgeleitet werden: Es wird sensibilisiert für Teilfähigkeiten, auf die im Mathematikunterricht geachtet werden sollte, um die Entwicklung der Schätzfähigkeit zu unterstützen. Das Aufgaben-Modell kann dazu dienen, Aufgaben zur vielfältigen Anregung für Schätzprozesse im Mathematikunterricht zu entwickeln (Weiher, 2020b).

Die Ergebnisse des Scoring-Vergleichs (Publikation 3; Kap. 4.2) oder des Vergleichs der alternativen Berechnungsmethoden der Schätzgenauigkeit (Weiher, 2022b) regen an, auch in mathematikdidaktischer Forschung außerhalb von der Verwendung oder Entwicklung standardisierter Leistungstests die verwendeten Testinstrumente und die Bewertungsmethode der Leistung begründet und transparent auszuwählen und auch in Publikationen darzustellen. Dies trifft insbesondere auf die Schätzgenauigkeit zu, da hier der Genauigkeitsanspruch nicht festgelegt ist (wie etwa beim Lösen einer Multiplikationsaufgabe). Die Entwicklung eines konkreten, alternativen Scorings auf Basis der logarithmischen (statt prozentualen) Abweichung (Weiher, 2022a, 2022b) u.a. mit der Empfehlung, die Skala für die Unterschätzung zu öffnen, kann daher als methodischer Beitrag zur Untersuchung der Schätzgenauigkeit angesehen werden.

Die Diskussion über die Bewertung der Schätzgenauigkeit hat auch unterrichtspraktische Relevanz. Wenn die Bewertung von Schätzgenauigkeit eine größere Bedeutung für Einzelpersonen hat, z.B. bei einer benoteten Mathematikleistung, ist sehr sorgfältig abzuwägen, welche Art der Berechnung und Bewertung genutzt wird und welche Höhe der Genauigkeit eigentlich von der schätzenden Person erwartet werden kann.

Durch die Untersuchung der Schätzgenauigkeit der vier Größen mit parallelisierten Aufgaben kann ein systematischer Beitrag, der über Einzelergebnisse zu verschiedenen Größen hinausgeht, zur Theorie über das Schätzen von Größen geleistet werden. Insbesondere ist dabei der Zusammenhang der vier Größen und die Möglichkeit, die Schätzgenauigkeit von Längen als Prädiktor der Schätzgenauigkeit von Flächeninhalten, Fassungsvermögen und Rauminhalten zu sehen, hervorzuheben. Für die mathematikdidaktische Forschung kann hieraus abgeleitet werden, dass das Konstrukt *Schätzen* vermutlich strukturell ähnlich zu dem Konstrukt *Messen* ist – welches im Gegensatz zum Schätzen schon vielfach untersucht und in seiner Bedeutung anerkannt wird (z.B. Clements & Bright, 2003). Das begründete Aufwerfen theoriebildender Fragen, z.B. nach der Dimensionalität des Konstrukts *Schätzen* oder nach der Bedeutung der Art der Fehlerberechnung für die Operationalisierung des Konstrukts *Schätzgenauigkeit*, wird ebenfalls als Beitrag zur Forschung angesehen.

Für den Mathematikunterricht ergibt sich die Anregung, das Schätzen umfassend in den Mathematikunterricht zu integrieren und Schätzen und Messen gemeinsam zu betrachten, um die Vorteile der Wechselwirkung zu nutzen (Grassmann, 1999). Der Zusammenhang zwischen Schätzen und Messen sowie zwischen den Größen sollte expliziert und damit für das Anwenden der Strategie nutzbar gemacht werden. Dabei ist zu entscheiden, ob das Schätzen von Größen als Mittel für die unterrichtliche Behandlung des mathematischen Zusammenhangs zwischen den Größen genutzt wird, oder ob der Zusammenhang der Größen beim Schätzen als unterrichtlicher Inhalt im Fokus steht.

Die Unterschiede der Schätzgenauigkeit zwischen den Jahrgängen implizieren, dass das Schätzen von Größen gelernt werden kann, was die Anregung, diese Lernprozesse gezielt und systematisch (eventuell auch durch Nutzen der theoretischen Modelle, s.o.) zu unterstützen, legitimiert. Zudem existiert bezüglich der Schätzgenauigkeit eine Leistungsheterogenität (sowohl zwischen einzelnen Schülerinnen und Schülern als auch für zwei Grö-

ßen zwischen Jungen und Mädchen), der durch Differenzierung begegnet bzw. die abgebaut werden sollte. Es kann daraus die Forderung abgeleitet werden, das Schätzen von Größen nicht nur kurz nebenbei, sondern als gut geplante und an die Fähigkeiten der Lerngruppe angepasste Unterrichtssequenzen in den Mathematikunterricht zu integrieren.

7.2 Limitationen

Detaillierte Limitationen bezogen auf die einzelnen Forschungsschritte und -entscheidungen dieses Dissertationsprojekts sind bereits in den Publikationen (Anhang A.2 bis A.5) sowie deren Zusammenfassungen und Reflexionen (Kap. 2 bis 5) benannt worden. An dieser Stelle sollen daher drei übergreifende Limitationen der Studie benannt werden:

1) Alle Versionen der Schätztests (Pilotstudie (Anhang C.1) sowie Version A und B der Hauptstudie (Anhang C.4 und C.5)) berücksichtigen keine Reihenfolgeneffekte, weder beim Vergleich der Aufgabentypen (Publikation 2), beim Vergleich der Größen (Publikation 2, 4) oder bei der Analyse der Testqualität (Publikation 3). Es ist nicht auszuschließen, dass von der Größe eines Objekts auf die Größe eines Objekts aus einer anderen Aufgabe geschlossen wird oder dass Objekte weiter hinten im Test aufgrund nachlassender Konzentration weniger genau geschätzt werden. Solche Effekte sind laut Jonkisz et al. (2012) zu vermeiden.
2) Die beiden zugrundeliegenden Modelle, insbesondere aber das Aufgaben-Modell, können nicht als abschließend vollständig oder als empirisch überprüft betrachtet werden. Zudem kann nicht sichergestellt werden, dass durch die gewählten Aufgaben das Konstrukt *Schätzen* tatsächlich wie gewünscht mit der Vielfalt an Fähigkeiten, wie sie im Modell dargestellt sind, abgebildet wird.
3) Es möglich, dass das Konstrukt *Schätzen*, welches mit dem vorliegenden Test untersucht wurde, erst durch den Test gebildet wurde (formatives Messmodell, Kap. 3.2). Es ist daher auch möglich, dass eine Untersuchung mit einem anderen Test, insbesondere mit anderen Aufgabentypen, andere inhaltliche Ergebnisse zur Folge hat. Die Ergebnisse zu den Zielen 3 und 4, aber auch aus der Scoringanalyse (Publikation 3), sind daher nicht ohne weiteres auf andere Schätztests übertragbar.

7.3 Ausblick

Aus den Ergebnissen und Limitationen des Dissertationsprojekts können Möglichkeiten für weiterführende methodische und inhaltliche Forschung zum Schätzen von Größen abgeleitet werden.

Um der ersten Limitation entgegenzuwirken, kann das aufgestellte Fähigkeiten-Modell empirisch überprüft werden. Dazu wird ein Schätztest mit Tests zu den jeweiligen Fähigkeiten kombiniert, um Relationen zwischen der Schätzgenauigkeit und der Ausprägung der entsprechenden Fähigkeit zu untersuchen. Dabei sollte auch Korrelationen zwischen den Fähigkeiten berücksichtigt werden. Ein Anfang stellt die Untersuchung des Zusammenhangs zwischen der Schätzgenauigkeit und des Stützpunktwissen/ der Stützpunktvorstellungen (Weiher, 2020a) sowie die Untersuchung von Prädiktoren beim Längenschätzen (Hoth et al., 2021) dar.

Um der zweiten und dritten Limitation entgegenzuwirken, kann das Aufgabentypen-Modell zunächst ergänzt werden (Kap. 2.2.2), jedoch bietet auch das vorliegende Modell Möglichkeiten für weitere Forschung. Nach Klärung des Messmodells (Kap. 3.2) bieten sich Strukturgleichungsanalysen an, um zu zeigen, ob die Aufgabentypen auch in der Empirie durch unterschiedliche Schätzgenauigkeit sichtbar sind. Alternativ kann durch Interviews geprüft werden, ob bestimmte Schätzstrategien bei bestimmten Aufgabentypen gewählt werden.

Der vorliegende Schätztest (Publikation 4) kann für weitere Studien überarbeitet und anschließend normiert werden. Eine Normierung setzt Forschung an der Klärung des Konstrukts *Schätzen* (s.o.) voraus, kann aber später dazu dienen, die Schätzgenauigkeit von Kindern nach Vorbild der Psychologie zu bewerten. So könnte der Schätztest sowohl für die Psychologie (interdisziplinärer Ansatz) als auch für die Schule nutzbar sein. Denkbar ist auch die Entwicklung eines Diagnosetests, der zeigt, welche Fähigkeiten beim Schätzen wie stark oder schwach ausgeprägt sind. Voraussetzung dafür ist zunächst, Aufgabentypen zu unterscheiden und zu untersuchen, ob bestimmte Strategien (und damit eventuell Fähigkeiten) besonders zu beobachten sind.

Anschließend an die Ergebnisse der vierten Publikation (Anhang A.5, Kap. 5) können zwei weiterführende inhaltliche Forschungsmöglichkeiten formuliert werden: Um zu untersuchen, inwieweit eine Verbesserung der Schätzgenauigkeit von Längen zur Verbesserung der Schätzgenauigkeit der anderen Größen beiträgt, kann eine Längsschnittstudie

durchgeführt werden. Zudem könnten mit Interventionsstudien Unterrichtsprozesse und -materialien hinsichtlich ihrer Wirksamkeit bezogen auf die Verbesserung der Schätzgenauigkeit untersucht werden.

7.4 Fazit

Ausgehend von den Reflexionen der Publikationen (Kap. 2.1.2, 2.2.2, 3.2 und 4.3.2), der übergeordneten Reflexion (Kap. 6) sowie der Betrachtung der Implikationen und Limitationen (Kap. 7.1 und 7.2) wird der Schluss gezogen, dass das vorliegende Dissertationsprojekt auf zwei Ebenen zur mathematikdidaktischen Forschung beitragen konnte.

Der methodische Beitrag zeichnet sich durch die Entwicklung zweier theoretische Modelle (Publikation 1 und 2) sowie eines darauf basierenden, für vier Größen parallelisierten Schätztest (Publikation 3 und 4) aus. Insbesondere ist die Bewertung des Schätzfehlers durch die logarithmische Abweichung in Verbindung mit einem logarithmischen Scoring hervorzuheben (Publikation 3 und 4, Weiher, 2022b).

Der inhaltliche Beitrag liegt in der systematischen Ergänzung der Theorie über das Schätzen von Größen. Dies geschieht zunächst ebenfalls durch die beiden Modelle (Publikation 1 und 2), durch die das Konstrukt *Schätzen* genauer definiert wird. Die Ergebnisse des Schätztests ermöglichen Aussagen über die Genauigkeit beim Schätzen von Größen bei Kindern und Aussagen über den Zusammenhang zwischen den Größen (Publikation 4). Insgesamt kann daher das Dissertationsprojekt als gewinnbringend für die mathematikdidaktische Forschung betrachtet werden.

Literaturverzeichnis

Axelrod, B. N., & Millis, S. R. (1994). Preliminary Standardization of the Cognitive Estimation Test. *Assessment, 1*(3), 269-274. https://doi.org/10.1177/107319119400100307

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2016). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung (*14. Auflage). Springer. https://doi.org/10.1007/978-3-662-46076-4

Backhaus, K., Erichson, B. & Weiber, R. (2015). *Fortgeschrittene Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (3. Auflage). Springer. https://doi.org/10.1007/978-3-662-46087-0

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4., überarbeitete Auflage). Springer. https://doi.org/10.1007/978-3-540-33306-7

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7., vollständig überarbeitete und erweiterte Auflage). Springer. https://doi.org/10.1007/978-3-642-12770-0

Brand, M., Kalbe, E., & Kessler, J. (2002). *Test zum Kognitiven Schätzen. Manual.* Beltz Test GmbH.

Brand, M., Fujiwara, E., Kalbe, E., Steingass, H-P., Kessler, J., & Markowitsch, H. J. (2003). Cognitive Estimation and Affective Judgments in Alcoholic Korsakoff Patients. *Journal of Clinical and Experimental Neuropsychology, 25*(3), 324–334. https://doi.org/10.1076/jcen.25.3.324.13802

Bright, G. W. (1976). Estimation as Part of Learning to Measure. In D. Nelson (Hrsg.), *Measurement in School* (S. 87–104). NCTM.

Bright, G. W. (1979). Estimating Physical Measurements. *School Science and Mathematics, 79*(8), 581–586.

Bullard, S. E., Fein, D., Gleeson, M. K., Tischer, N., Mapou, R. L., & Kaplan, E. (2004). The Biber Cognitive Estimation Test. *Archives of Clinical Neuropsychology, 19*(6), 835–846. https://doi.org/10.1016/j.acn.2003.12.002

Clements, D. H. & Bright, G. (Hrsg.) (2003). *Learning and Teaching Measurement. 2003 Yearbook*. NCTM.

Corle, C. G. (1960). A Study of the Quantitative Values of Fifth and Sixth Grade Pupils. *The Arithmetic Teacher, 7*(7), 333–340. https://www.jstor.org/stable/41184340

Corle, C. G. (1963). Estimates of Quantity by Elementary Teachers and College Juniors. *The Arithmetic Teacher, 10*(6), 347–353. https://www.jstor.org/stable/41184817

Cortina, J. M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. Journal of Applied Psychology, 78(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Della Sala, S., MacPherson, S. E., Phillips, L. H., Sacco, L., & Spinnler, H. (2003). How many camels are there in Italy? Cognitive estimates standardized on the Italian population. *Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 24*(1), 10–15. https://doi.org/10.3389/fpsyg.2015.00608

Desli, D., & Giakoumi, M. (2017). Children's Length Estimation Performance and Strategies in Standard and Nonstandard Units of Measurement. *International Journal of Research in Mathematics Education, 7*(3), 61–84.

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Hogrefe Verlag GmbH & Co. KG.

Forrester, M. A., Latham, J., & Shire, B. (1990). Exploring Estimation in Young Primary School Children. *Educational Psychology, 10*(4), 283–300. https://doi.org/10.1080/0144341900100401

Forrester, M. A., & Shire, B. (1994). The Influence of Object Size, Dimension and Prior Context on Children's Estimation Abilities. *Educational Psychology, 14*(4), 451–465. https://doi.org/10.1080/0144341940140407

Gooya, Z., Khosroshahi, L. G., & Teppo, A. R. (2011). Iranian Students' Measurement Estimation Performance Involving Linear and Area Attributes of Real-World Objects. *ZDM Mathematics Education,* 43: 709–722. https://doi.org/10.1007/s11858-011-0338-1

Grassmann, M. (1999). Zur Entwicklung von Zahl- und Größenvorstellungen als wichtigem Anliegen des Sachrechnens. *Grundschulunterricht, 46*(4), 31–34.

Green, S. B., Lissitz, R. W., Mulaik, S. A. (1977). Limitations of Coefficient Alpha as an Index of Test Unidimensionality. *Educational and Psychological Measurement, 37*(4), 827-838. https://doi.org/10.1177/001316447703700403

Harel, B. T., Cillessen, A. H., Fein, D. A., Bullard, S. E., & Aviv, A. (2007). It Takes Nine Days to Iron a Shirt: The Development of Cognitive Estimation Skills in School Age Children. *Child Neuropsychology, 13*(4), 309–318. https://doi.org/10.1080/09297040600837354

Heid, L.-M. (2018). *Das Schätzen von Längen und Fassungsvermögen: Eine Interviewstudie zu Strategien mit Kindern im 4. Schuljahr*. Springer Spektrum. https://doi.org/10.1007/978-3-658-18874-0

Heinze, A., Weiher, D. F., Huang, H.-M. E., Ruwisch, S. (2018): Which Estimation Situations Are Relevant for a Valid Assessment of Measurement Estimation Skills? In E. Bergqvist, M. Österholm, C. Granberg & L. Sumpter (Hrsg.): *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education Vol. 3* (S. 67-74). https://pure.ipn.uni-kiel.de/portal/files/1419366/PME2018_Heinze_Which_estimation_situations_are_relevant.pdf

Hildreth, D. J. (1980). *Estimation Strategy Uses in Length and Area Measurement Tasks by Fifth and Seventh Grade Students*. Dissertation at Ohio State University. University Microfilm International.

Hildreth, D. J. (1983). The Use of Strategies in Estimating Measurements. *Arithmetic Teacher, 30*(5), 50–54. https://www.jstor.org/stable/41192173

Hogan, T. P., & Brezinski, K. L. (2003). Quantitative Estimation: One, Two, or Three Abilities? *Mathematical Thinking and Learning, 5*(4), 259–280. https://doi.org/10.1207/S15327833MTL0504_02

Hoth, J., Heinze, A., Weiher, D. F., Ruwisch, S., & Huang, H-M. E. (2021). Das Schätzen von Längen in der Grundschule: Welche mathematischen Fähigkeiten sind prädiktiv? In K. Hein, C. Heil, S. Ruwisch, & S. Prediger (Hrsg.)*, Beiträge zum Mathematikunterricht 2021: Vom GDM-Monat 2021 der Gesellschaft für Didaktik*

der Mathematik (GDM) (S. 305-308). WTM. https://doi.org/10.17877/DE290R-22289

Hoth, J., Heinze, A., Huang, H.-M. E., Weiher, D. F. & Ruwisch, S. (im Druck). Elementary school students' length estimation skills – analyzing a multidimensional construct in a cross-country study. *International Journal for Science and Mathematics Education*.

Huang, H.-M. E. (2014). Investigating children's ability to solve measurement estimation problems. In S. Oesterle, P. Liljedahl, C. Nicol, & D. Allan (Hrsg.), *Proceedings of the Joint Meeting of PME 38 and PME-NA 36 Vol. 3* (S. 353–360). https://files.eric.ed.gov/fulltext/ED599830.pdf

Jones, G., Taylor, A., & Broadwell, B. (2009). Estimating linear size and scale: Body rulers. *International Journal of Science Education, 31*(11), 1495–1509. https://doi.org/10.1080/09500690802101976

Jones, G., Forrester, J. H., Gardner, G. E., Andre, T., & Taylor, A. R. (2012). Students' Accuracy of Measurement Estimation: Context, Units, and Logical Thinking. *School Science and Mathematics, 112*(3), 171–178. https://doi.org/10.1111/j.1949-8594.2011.00130.x

Jonkisz, E.; Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In Moosbrugger, & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Springer. https://doi.org/10.1007/978-3-642-20072-4_2

Joram, E. (2003). Benchmarks as Tools for Developing Measurement Sense. In D. H. Clements & G. Bright (Hrsg.), *Learning and Teaching Measurement* (S. 57–67). NCTM.

Joram, E., Gabriele, A. J., Bertheau, M., Gelman, R., & Subrahmanyam, K. (2005). Children's Use of the Reference Point Strategy for Measurement Estimation. *Journal for Research in Mathematics Education, 36*(1), 4–23. https://doi.org/10.2307/30034918

Joram, E., Subrahmanyam, K., & Gelman, R. (1998). Measurement Estimation: Learning to Map the Route from Number to Quantity and Back. *Review of Educational Research, 68*(4), 413–449. https://www.jstor.org/stable/1170734

Kelava, A. & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In Moosbrugger, & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 75-102). Springer. https://doi.org/10.1007/978-3-642-20072-4_2

Kemper, C. J., Ziegler, M., Krumm, S., Heene, M., & Bühner, M. (2015). Testkonstruktion. In G. Stemmler, & J. Margraf-Stiksrud (Hrsg.), *Lehrbuch Psychologische Diagnostik* (S. 157–221). Verlag Hans Huber, Hogrefe AG.

Kultusministerkonferenz (2005). *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss vom 15.10.2004*. Wolters Kluwer Deutschland GmbH. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Mathe-Primar.pdf

Mendez, M. F., Doss, R. C., & Cherrier, M. M. (1998). Use of the Cognitive Estimations Test To Discriminate Frontotemporal Dementia from Alzheimer's Disease. *Journal of Geriatric Psychiatry and Neurology, 11*(1), 2–6. https://doi.org/10.1177/089198879801100102

Moosbrugger, H. (2012). Klassische Testtheorie (KTT). In Moosbrugger, & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 103-117). Springer. https://doi.org/10.1007/978-3-642-20072-4_2

Moosbrugger, H., & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger, & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Springer. https://doi.org/10.1007/978-3-642-20072-4_2

Newcombe, N. (2014). The Origins and Development of Magnitude Estimation. *Ecological Psychology, 26*(1–2), 147–157. https://doi.org/10.1080/10407413.2014.875333

Niedersächsisches Kultusministerium (2017). *Kerncurriculum für die Grundschule. Schuljahrgänge 1-4. Mathematik*. unidruck. https://www.cuvo.nibis.de/index.php?p=detail_view&docid=1052&f0=kerncurriculum%20mathematik%20grundschule

O‘Daffer, P. (1979). A Case and Techniques for Estimation: Estimation Experiences in Elementary School Mathematics – Essential, Not Extra! *Arithmetic Teacher, 26*(6), 46–51. https://doi.org/10.5951/AT.26.6.0046

Paull, D. R. (1971). *The Ability to Estimate in Mathematics.* Dissertation, Columbia University. University Microfilms, A XEROX Company.

Pike, C. D., & Forrester, M. A. (1997). The Influence of Number-sense on Children’s Ability to Estimate Measures. *Educational Psychology, 17*(4), 483–500. https://doi.org/10.1080/0144341970170408

Rammstedt, B. (2010). Reliabilität, Validität, Objektivität. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 239-258). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92038-2

Ruwisch, S., Heid, L.-M., & Weiher, D. F. (2015). Measurement Estimation in Primary School: Which Answer is Adequate? In K. Beswick, T. Muir, & J. Wells, (Hrsg.). *Proceedings of 39th Psychology of Mathematics Education conference Vol. 4* (S. 113–120). http://fox.leuphana.de/portal/files/7691612/Ruwisch_Heid_Weiher_2015_PME.pdf

Shallice, T., & Evans, M. E. (1978). The Involvement of the Frontal Lobes in Cognitive Estimation. *Cortex, 14*(2), 294–303. https://doi.org/10.1016/S0010-9452(78)80055-0

Siegel, A. W., Goldsmith, L. T., & Madson, C. R. (1982). Skill in Estimation Problems of Extent and Numerosity. *Journal for Research in Mathematics Education, 13*(3), 211–232. https://doi.org/10.2307/748557

Sowder, J. (1992). Estimation and Number Sense. In D. A. Grouws (Hrsg.), *Handbook of research on mathematics teaching and learning. A project of the National Council of Teachers of Mathematics* (S. 371–389). Macmillan.

Swan, M., & Jones, O. (1980). Comparison of Students’ Percepts of Distance, Weight, Height, Area, and Temperature. *Science Education, 64*(3), 297–307. https://doi.org/10.1002/sce.3730640305

Stephan, M. & Clements, D. H. (2003). Linear and Area Measurement in Prekindergarten to Grade 2. In D. H. Clements & G. Bright (Hrsg.), *Learning and Teaching Measurement* (S. 3-16). NCTM.

Streiner, D. L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment, 80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Weiher, D. F. (2018a): Operationalisierung des Konstrukts „Schätzen von Längen, Flächeninhalten und Volumina“ für Grundschulkinder. In Fachgruppe Didaktik der Mathematik der Universität Paderborn (Hrsg.), *Beiträge zum Mathematikunterricht 2018* (S. 1939–1942). WTM. http://dx.doi.org/10.17877/DE290R-19764

Weiher, D. F. (2018b): Development of a measurement estimation test for length, area, and volume. In E. Bergqvist, M. Österholm, C. Granberg & L. Sumpter (Hrsg.), *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education Vol. 5* (S. 307).

Weiher, D. F. (2018c): Schätzen von Längen, Flächeninhalten und Volumina mit verschiedenen Aufgabentypen. In A. S. Steinweg (Hrsg.), *Inhalte im Fokus – mathematische Strategien entwickeln. Tagungsband des AK Grundschule in der GDM 2018.* (S. 105–108). University of Bamberg Press. https://fis.uni-bamberg.de/bitstream/uniba/44712/1/MDG8SteinwegGDM2018opusse_A3a.pdf

Weiher, D. F. (2019a). Framework for the Parallelized Development of Estimation Tasks for Length, Area, Capacity, and Volume in Primary School – A Pilot Study. *Journal of Research in Science, Mathematics and Technology Education, 2*(1), 9–28. https://doi.org/10.31756/jrsmte.212

Weiher, D. F. (2019b): Merkmale von Schätzaufgaben zu Längen, Flächeninhalten, Fassungsvermögen und Rauminhalten. In A. Frank, S. Krauss & K. Binder (Hrsg.), *Beiträge zum Mathematikunterricht 2019* (S. 1419). WTM. http://dx.doi.org/10.17877/DE290R-20738

Weiher, D. F. (2020a). Der Zusammenhang zwischen Schätzgenauigkeit und Stützpunktausprägung bei Längen. In H.-S. Siller, W. Weigel & J. F. Wörler (Hrsg.), *Beiträge zum Mathematikunterricht 2020*. (S. 1277–1280). WTM. http://dx.doi.org/10.17877/DE290R-21623

Weiher, D. F. (2020b). Die Vielfalt von Schätzaufgaben. Eine Darstellung verschiedener Merkmale von Schätzaufgaben zu visuell erfassbaren Größen. In *mathematik differenziert 3*:16-21.

Weiher, D. F. (2022a). Estimation of Length, Area, Capacity, and Volume: Results of a Written Estimation Test. [Eingereicht zur Veröffentlichung].

Weiher, D. F. (2022b). Measurement Estimation Accuracy: A Comparison of Different Approaches. In: IDMI-Primar Goethe Universität Frankfurt (Hrsg.), *Beiträge zum Mathematikunterricht 2022*. http://dx.doi.org/10.17877/DE290R-23806

Weiher, D. F. & Ruwisch, S. (2018). Kognitives Schätzen aus Sicht der Mathematikdidaktik. *mathematica didactica 41*(1), 77–103. http://www.mathematica-didactica.com/Pub/md_2018/md_2018_Weiher_Ruwisch.pdf

Weiher, D. F., & Ruwisch, S. (2022). The Assessment of Measurement Estimation Results – a Discussion of Different Scorings Regarding Test Performance and Test Quality [Eingereicht zur Veröffentlichung].

Winter, H. (2003). *Sachrechnen in der Grundschule. Problematik des Sachrechnens. Funktionen des Sachrechnens. Unterrichtsprojekte* (6. Auflage). Cornelsen Scriptor.

Anhang

Ich versichere, dass alle in diesem Anhang gemachten Angaben jeweils einzeln und insgesamt vollständig der Wahrheit entsprechen.

Lüneburg, 09.08.2022 _____

A Publikationen der kumulativen Dissertation

A.1 Übersicht der Publikationen und Beitrag der Autorinnen

Tabelle 2

Übersicht der Publikationen der Dissertation

Nr.	Titel	Autorinnen	Autorenstatus	Gewichtung	Publikationsstatus (Stand 09.08.2022)
1	Kognitives Schätzen aus Sicht der Mathematikdidaktik	Dana Farina Weiher, Silke Ruwisch	Ko-Autorenschaft	1	Publiziert in: *mathematika didactica, 41*(1), S. 77-103. http://www.mathematica-didactica.com/Pub/md_2018/md_2018_Weiher_Ruwisch.pdf
2	Framework for the Parallelized Development of Estimation Tasks for Length, Area, Capacity, and Volume in Primary School – A Pilot Study.	Dana Farina Weiher	Allein-Autorenschaft	1	Publiziert in: *Journal of Research in Science, Mathematics and Technology Education, 2*(1), 9-28.
3	The Assessment of Measurement Estimation Results – a Discussion of Different Scorings Regarding to Test Performance and Test Quality	Dana Farina Weiher, Silke Ruwisch	Ko-Autorenschaft	1	"under review" in: Journal of Research in Science, Mathematica and Technology Education (JRSMTE) https://jrsmte.com
4	Estimation of Length, Area, Capacity, and Volume: Results of a Written Estimation Test.	Dana Farina Weiher	Allein-Autorenschaft	1	"under review" in: Journal für Mathematikdidaktik (JMD) https://www.springer.com/journal/13138

Summe der Gewichte: 4

A.2 Publikation 1

Weiher, D. F. & Ruwisch, S. (2018). Kognitives Schätzen aus Sicht der Mathematikdidaktik. *mathematica didactica 41*(1), 77–103. http://www.mathematica-didactica.com/Pub/md_2018/md_2018_Weiher_Ruwisch.pdf

Kognitives Schätzen aus Sicht der Mathematikdidaktik: Schätzen von visuell erfassbaren Größen und dazu erforderliche Fähigkeiten

DANA FARINA WEIHER & SILKE RUWISCH, LÜNEBURG

***Zusammenfassung**: Sowohl die Psychologie als auch die Mathematikdidaktik befassen sich mit dem Schätzen. In diesem Beitrag wird ein theoretisches Modell zur Diskussion gestellt, in welchem die Fähigkeiten, die zum Schätzen von Größen erforderlich sind, aus mathematikdidaktischer Sicht dargestellt werden. Diese sind aus Untersuchungen der Psychologie zum kognitiven Schätzen und mathematikdidaktischen Erkenntnissen zu Schätzstrategien abgeleitet. Sie umfassen exekutive Funktionen, Wissen über den Messprozess, Stützpunktwissen und Stützpunktvorstellungen, die Fähigkeit zu vergleichen und zur räumlichen Vorstellung sowie verschiedene Gruppen allgemeinen (mathematischen) Wissens und grundlegender Fähigkeiten.*

***Abstract:** Mathematics education as well as psychology is concerned with estimation. This paper presents a theoretical model which contains those abilities, which are (from a didactical perspective) required for measurement estimations. These abilities are deducted from psychological research about cognitive estimation and from research in mathematics education about estimation-strategies. The abilities involve executive functions, knowledge about measuring, benchmarks, and, last but not least, different groups of general (mathematical) knowledge and basic abilities.*

1. Einleitung

> Wie oft am Tag stellen wir Schätzungen an und merken es gar nicht? Tatsächlich stellen wir den ganzen Tag, das ganze Leben lang andauernd Schätzungen an. […] Wir schätzen und schätzen und schätzen, dann fällen wir eine Entscheidung. (Paenza, 2012, S. 336 f.)

Dieses Zitat verdeutlicht zum einen, dass Schätzen eine Alltagskompetenz ist, die als wesentlich für die Bewältigung vieler Situationen angesehen werden kann. Zum anderen ist es ein Prozess, der unermüdlich durchgeführt wird – aber dabei nicht immer bewusst verläuft und teilweise sogar dem Bewusstsein nicht zugänglich zu sein scheint. Es ist allerdings noch offen, *wie* eine Entscheidung für ein Schätzergebnis gefällt wird.

Aus Sicht der Neuro- und Kognitionspsychologie ist kognitives Schätzen ein

> Prozess der Antwortgenerierung bei Nichtverfügbarkeit der exakten Lösung mit Hilfe des semantischen Wissens und der Anwendung von (Vergleichs-) Strategien. (Brand, Fujiwara, Kalbe, Kessler & Markowitsch, 2002, S. 282)

Aus Sicht der Mathematikdidaktik ist Schätzen ein „kompliziertes Zusammenspiel von Wahrnehmen, Erinnern, Inbeziehungsetzen, Runden und Rechnen“ (Winter, 2003, S. 19).

Die Neuro- und Kognitionspsychologie konzentriert sich in ihrer Forschung insbesondere auf die am Schätzprozess beteiligten kognitiven Prozesse und bezeichnet das Schätzen daher abgrenzend als *kognitives Schätzen*. Die Erforschung der beteiligten kognitiven Prozesse dient der Nutzung des kognitiven Schätzens als Mittel zur Feststellung von Frontalhirndefiziten (Peretti Wagner, MacPherson, Parente & Trentini, 2011, S. 203). In der Mathematikdidaktik wird das Schätzen vor allem inhaltlich ausdifferenziert. So wird zwischen dem Schätzen der Anzahl der Elemente einer Menge, dem Schätzen des Ergebnisses einer Rechenaufgabe (Überschlagen) und dem Schätzen einer Größe eines Objekts oder Vorgangs unterschieden (O'Daffer, 1979; Sowder, 1992, S. 371). Es wird davon ausgegangen, dass bei allen drei Schätzvorgängen andere Fähigkeiten und Strategien zugrunde liegen, auch wenn eine zumindest teilweise Überschneidung nicht ausgeschlossen werden kann (Sowder, 1992, S. 371). Insbesondere das Schätzen von Anzahlen und das Schätzen von Größen erfordern ähnliche Fähigkeiten und werden daher in Studien häufig zusammengefasst (Hogan & Brezinski, 2003, S. 260; Siegel, Goldsmith & Madsen, 1982, S. 214; Sowder, 1992, S. 371). In diesem Beitrag sprechen wir immer dann vom *kognitiven Schätzen*, wenn explizit auf die Neuro- und Kognitionspsychologie Bezug genommen wird. Dagegen sprechen wir vom *Schätzen*, wenn wir das Schätzen von Größen aus mathematikdidaktischer Sicht beleuchten.

Aus deren Sicht können beim Schätzen von Größen noch nicht einmal alle Größenarten[1] gemeinsam betrachtet werden, da zwischen ihnen teilweise deutliche Unterschiede bestehen. So sind Längen, Flächeninhalte und Volumina visuell erfassbar, Gewichte jedoch nicht. Geschwindigkeiten können nur begrenzt visuell wahrgenommen werden, nämlich durch das Verhältnis von Strecke und Zeit. Geld ist keine physikalische Größe, sondern eine Zählgröße (Franke & Ruwisch, 2010, S. 182). Zeitspannen können nur durch Handlungen oder Vorgänge sichtbar gemacht werden und sind außerdem flüchtig (Franke & Ruwisch, 2010, S. 217).

Dennoch haben insbesondere die Größenarten Längen, Flächeninhalte und Volumina Gemeinsamkeiten. Alle drei beruhen auf der Basisgröße Länge. Grund für die Auswahl dieser Größenarten für eine gemeinsame Betrachtung in diesem Beitrag ist darüber hinaus die Tatsache, dass

> für die Qualität der Größenvorstellungen zunächst die Art, in der die Schüler die Repräsentanten von Größen wahrnehmen, von besonderer Bedeutung ist. Bezüglich der Länge, des Flächeninhalts und des Volumens ist eine visuelle Wahrnehmung möglich. (Frenzel & Grund, 1991b, S. 18)

Hildreth (1983, S. 50) nimmt an, dass die Fähigkeiten zum Schätzen für die Größenarten Längen und Flächeninhalte ähnlich sind. Griesel (1996, S. 17) setzt für alle drei Größenarten ähnliche Grundvorstellungen voraus. Nührenbörger (2004, S. 95) bezeichnet das (Mess-)Verständnis von Längen als Voraussetzung für das (Mess-)Verständnis von Flächeninhalten und Volumina. Daher scheint eine gemeinsame Betrachtung der Fähigkeiten, die für das Schätzen dieser drei Größen erforderlich sind, sinnvoll zu sein.

Die Definitionen zum Schätzen aus beiden Wissenschaftsbereichen zeigen bereits die Komplexität des Schätzprozesses und die Vielfältigkeit an Anforderungen, die an die schätzende Person gestellt werden. Die Beschreibung des Schätzprozesses und der dahinterstehenden Fähigkeiten weist jedoch trotz der alltäglichen Bedeutung und der Präsenz in verschiedenen Disziplinen Lücken auf:

> Further research is needed in order to explore exactly which cognitive abilities are requested when someone is trying to make an estimation. (Peretti Wagner et al., 2011, S. 209)

Diesem Forschungsdefizit soll die Entwicklung eines Modells zum Schätzen von Längen, Flächeninhalten und Volumina in diesem Beitrag entgegenwirken. Dazu werden die Erkenntnisse zum kognitiven Schätzen aus der Psychologie aus mathematikdidaktischer Sicht für das Schätzen von Längen, Flächeninhalten und Volumina interpretiert. Es werden aus der Literatur beider Disziplinen Fähigkeiten, die dem Schätzen dieser Größen zugrunde liegen, abgeleitet und strukturiert bezüglich ihrer Abhängigkeiten dargestellt.

Kapitel 2 stellt den Diskussions- und Forschungsstand zum Schätzen dieser Größen aus der Mathematikdidaktik dar. Zunächst findet eine Begriffsklärung, auch durch Abgrenzung von anderen Begriffen, statt. Anschließend wird herausgearbeitet, warum im Weiteren von Stützpunkt- statt Größenvorstellungen gesprochen wird, bevor schließlich im Hauptabschnitt Schätzstrategien als Grundlage für die Erläuterung der am Schätzprozess beteiligten Fähigkeiten genauer thematisiert und systematisiert werden.

Kapitel 3 stellt zentrale Komponenten des kognitiven Schätzprozesses aus neuro- und kognitionspsychologischer Sicht vor.

In Kapitel 4 und 5 werden beide Sichtweisen zu einem theoretischen Modell des Schätzens zusammengeführt. Leitend in der Modellentwicklung und deren Präsentation in diesem Beitrag ist dabei die mathematikdidaktische Sicht, so dass eine diesbezügliche Interpretation der neuro- und kognitionspsychologischen Erkenntnisse stattfindet. Kapitel 4 liefert dementsprechend die Gesamtrahmung – das Schätzen im weiteren Sinne –, bevor sich in Kapitel 5 differenzierte Ausführungen zum Schätzen im engeren Sinne finden.

2. Schätzen von Längen, Flächeninhalten und Volumina

Im Folgenden wird der Begriff des Schätzens von Größen expliziert. Diese Klärung erfolgt auch durch Abgrenzung zum Messen auf der einen und zum Raten auf der anderen Seite.

Das Schätzen von Größen kann als

> das Ermitteln einer ungefähren Größenangabe durch gedankliches Vergleichen mit eingeprägten Repräsentanten als Stützpunkten (Franke & Ruwisch, 2010, S. 248)

verstanden werden. Wesentlich für das Schätzen ist der gedankliche Vergleich mit Stützpunkten, da beim Schätzen keine konkreten Hilfsmittel genutzt werden:

> Estimation of measurements is the use of units of measure in a strictly mental way, without the aid of measurement tools. (Bright, 1976, S. 88)

Beim Schätzen können qualitative und quantitative Schätzanforderungen unterschieden werden. Während beim quantitativen Schätzen eine Zahl oder Größenangabe als Antwort gegeben wird, wird beim qualitativen Schätzen der Vergleich zweier Objekte hinsichtlich ihrer Größenausprägung erwartet (Blankenagel, 1983, S. 319; Goldstein et al., 1996, S. 37; Ruwisch, 2014, S. 4 f.).

Darüber hinaus können zwei Arten quantitativen Schätzens unterschieden werden: Grundständige Schätzaufgaben erfordern die Nennung einer Anzahl oder einer Größenangabe, welche meist durch eine Strategie wie das gedankliche Vergleichen ermittelt werden kann (zu den Strategien, die beim Schätzen angewendet werden, vgl. Abschn. 2.2). Eingebettete Schätzaufgaben, wie etwa Fermi-Aufgaben, erfordern die Koordination verschiedener Teilschritte, die weitere Überlegungen wie die Erfahrungen zur

Sachsituation oder die Beziehung zwischen Daten mit einschließen (Franke & Ruwisch, 2010, S. 251 ff.).

Frenzel und Grund (1991b, S. 24) unterscheiden beim quantitativen Schätzen das echte Schätzen und das mittelbare Schätzen. Während beim echten Schätzen die „Größenangabe ausschließlich aufgrund von bestimmten Vorstellungen und Erfahrungen des Subjekts“ (Frenzel & Grund, 1991b, S. 23) erfolgt, werden beim mittelbaren Schätzen andere Größen geschätzt, durch die rechnerisch die gesuchte Größenangabe ermittelt werden kann.

Dem mentalen Charakter des Schätzens steht der konkrete Charakter des Messens gegenüber. Peter-Koop und Nührenbörger (2011, S. 92) fassen das Messen wie folgt:

> Es muss eine Einheit gefunden werden. Diese muss wiederholt benutzt und dabei gezählt werden, wenn das zu Messende größer ist als die Maßeinheit. Die Einheit muss systematisch untergliedert werden, wenn keine Maßeinheit aus den natürlichen Zahlen das zu Messende vollständig erfasst.

Aufgrund dieser Beschreibung bleibt lediglich unklar, *wie* mit der Einheit operiert wird. Hierzu stellt Griesel (1996, S. 17) beispielhaft für Längen fest, dass die

> Vorstellung von hintereinandergelegten Strecken, mit denen die Strecke S_1 ausgelegt wird […], charakteristisch für das Meßverfahren

ist. Das Aneinanderlegen der Einheiten muss geschehen, ohne dass es Lücken und Überschneidungen gibt (vgl. Abschn. 5.1.4). Dasselbe beschreibt Battista (2003, S. 122) für das Messen von Flächeninhalten und Volumina:

> To measure area and volume in standard measurement systems, we determine the number of unit squares or cubes in the region we are measuring.

Die Anordnung der Einheiten muss auch ohne Lücken oder Überschneidungen vorgenommen werden.

Durch die Betrachtung der Messidee als Auslegen mit Einheiten wird deutlich, dass es ein zu messendes Objekt gibt und eines, mit dem gemessen wird (Frenzel & Grund, 1991a, S. 12). Zum Messen können verschiedene Hilfsmittel genutzt werden, die auf Einheiten beruhen. Hier kann zwischen Hilfsmitteln mit standardisierten und solchen mit nichtstandardisierten Einheiten unterschieden werden. Je nach Messgerät (bzw. nach Einheit) variiert der Grad der Genauigkeit der Messung, wobei zu beachten ist, dass auch mit standardisierten Messgeräten keine exakte Bestimmung der Größe möglich ist (Bright, 1976, S. 89; Frenzel & Grund, 1991a, S. 12; Ruwisch, 2014, S. 4). Demnach macht beim Ermitteln einer Größe allein das konkrete Nutzen von Hilfsmitteln und deren Zweck als Einheit (unabhängig davon, ob es sich um standardisierte oder nichtstandardisierte Einheiten handelt) dieses zu einem Messprozess.

Diese unumstrittene Abgrenzung von Schätzen und Messen ist in Bezug auf die verwendeten Hilfsmittel nicht immer eindeutig. So sprechen manche Autorinnen und Autoren auch dann vom Schätzen, wenn mit Körpermaßen die Länge oder der Flächeninhalt bestimmt wird, da zum einen die verwendete Einheit geschätzt werden müsse, zum anderen die Größe nur „ungefähr“ bestimmt werde (Huang, 2015, S. 75). Nührenbörger (2004, S. 96) hingegen fasst Körpermaße als für Kinder intuitiv nutzbare sowie historische Messgeräte auf. Verbunden mit der Tatsache, dass auch das Messen keine exakte Größenangabe ermöglicht, zeigt dies, dass das Verwenden von Körpermaßen nicht per se als Schätzprozess aufgefasst werden kann. An dieser Stelle sei erneut auf Bright (1976, S. 88) verwiesen, der, wie oben beschrieben, das strikte mentale Vorgehen als wesentlichen Bestandteil des Schätzens beschreibt. Unabhängig von den verwendeten Hilfsmitteln handelt es sich demzufolge bei konkreten Handlungen immer um einen Mess- und keinen Schätzprozess.

Schätzen kann durch Heraushebung der mental genutzten Repräsentanten ebenfalls vom Raten abgegrenzt werden. Es bedarf mentaler Vergleichsprozesse, ansonsten kann nicht von einem Schätzprozess gesprochen werden:

> Estimating is guessing, but the guessing must be educated. Wild guessing is not true estimating. (Bright, 1976, S. 89)

Das englische Wort *guess* kann sowohl mit *raten* als auch mit *schätzen* übersetzt werden. Dies könnte die Möglichkeit der synonymen Verwendung suggerieren. Dass dies nicht der Fall ist, hat Bright mit der Beschreibung *wild guessing* für *raten* ausgedrückt. Auch Winter (2003, S. 18) verweist auf diesen Unterschied zwischen Schätzen und Raten:

> Um die Aufgabe (eine Schätzaufgabe, D.F.W. & S.R.) zu lösen, muß der Schüler auf Vorerfahrungen zurückgreifen, also sein Langzeitgedächtnis bemühen. Wenn dort nichts Passendes ist, kann nur noch geschwiegen oder geraten werden. Schätzen ist jedenfalls kein blindes Raten.

Die vorangegangenen Darlegungen führen zu folgendem Verständnis des Schätzens, welches dem in diesem Beitrag entwickelten Modell zugrunde liegt:

Schätzen ist ein mentaler Prozess, in dem die quantitative Ausprägung einer Größe eines Objekts unter Anwendung mentaler Vergleichsprozesse mit anderen Objekten ermittelt wird, ohne dass konkrete Hilfsmittel verwendet werden.

2.1 Vorstellen von Größen und Stützpunkten

Eine Größe ist zunächst eine Eigenschaft ihres Trägers (Griesel, 1996, S. 15), die es beim Schätzen (oder Messen) quantitativ zu bestimmen gilt. Die quantitative Beschreibung einer Größe erfolgt durch die Größenangabe oder den Größenwert. Dabei werden eine Maßzahl und eine Maßeinheit verwendet, welche multiplikativ zusammenhängen (Frenzel & Grund 1991a, S. 11). Die Maßzahl und die Maßeinheit bedingen einander, so ergibt sich bei Verkleinerung der Maßeinheit eine größere Maßzahl und umgekehrt (Franke & Ruwisch, 2010, S. 196).

Dem Träger kommt die Größe in genau einer Ausprägung zu, d. h. er ist Repräsentant für genau diese Größenausprägung, die mit genau dieser Größenangabe beschrieben werden kann. Im Umkehrschluss sind Größen und ihre Ausprägungen nicht ohne Träger vorstellbar:

> Bezüglich der Größen gilt grundsätzlich, dass man sich nur Träger und Konfigurationen von Trägern vorstellen kann, nicht jedoch die Größenwerte. (Grode, 2001, S. 16)

Dennoch ist der Begriff *Größenvorstellungen* in der Literatur ein gängiger Begriff. Unter Größenvorstellungen verstehen Frenzel und Grund (1991b, S. 16) neben dem Erkennen und Unterscheiden verschiedener Größenarten die Kenntnis von Repräsentanten unterschiedlicher (spezieller) Größen, das Beherrschen der Umrechnung von Größenangaben sowie Fähigkeiten im Messen, Schätzen und Überschlagen. Begleitend werden Zahlvorstellungen aufgeführt (vgl. Abb. 1).

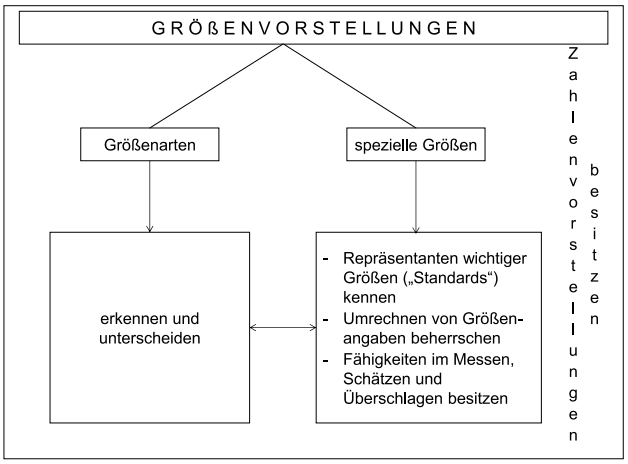


Abb. 1: Größenvorstellungen (Frenzel & Grund 1991b, S. 16)

Auch andere Autorinnen und Autoren weisen auf die Verbindung zwischen dem Schätzen von Größen und Zahlvorstellungen hin (Hope, 1989, S. 15; Kuwahara Lang, 2001, S. 463; Pike & Forrester, 1997, S. 493; Sowder, 1992, S. 371).

Franke und Ruwisch (2010, S. 235) nennen drei Aspekte, die Größenvorstellungen untergeordnet sind:

- zu Größenangaben passende Repräsentanten kennen,
- zu alltäglichen Repräsentanten die passende Größenangabe kennen,
- Stützpunktwissen beim Schätzen, Problemlösen und im Alltag flexibel nutzen.

Alle bezüglich Größenvorstellungen genannten Aspekte haben gemeinsam, dass sie sich auf das Wissen über die Größe von Objekten sowie das Nutzen dieses Wissens beziehen. Im Unterschied zum Konzept Größenvorstellungen von Frenzel und Grund (1991b) wird hier nur das Nutzen von Stützpunktwissen den Größenvorstellungen untergeordnet, während Frenzel und Grund den Mess-, Schätz- und Überschlagsprozess als Ganzes als Teil von Größenvorstellungen verstehen. Frenzel und Grund (1991b, S. 17 f.) erklären jedoch darüber hinaus:

> Vorstellungen von Größen zu erzeugen bedeutet, im Bewußtsein der Schüler adäquate Abbilder von Größen entstehen zu lassen, dafür zu sorgen, daß diese aufbewahrt werden und je nach Bedarf immer wieder reproduziert und gedanklich weiterverarbeitet werden.

Diese Abbilder können als Träger einer Größe, deren Ausprägung bekannt ist, verstanden werden. Solche Objekte sind als *Stützpunkte* zu bezeichnen und als „die Bausteine von Größenvorstellungen“ (Franke & Ruwisch, 2010, S. 235) zu verstehen.

Es ist demzufolge nicht ausreichend, lediglich Wissen über die Größe von Objekten zu besitzen. Vielmehr muss zu diesem Wissen auch ein mentales Bild vorhanden sein, um das Wissen über die Größe des Objekts und damit das Objekt beim Schätzen als Stützpunkt nutzen zu können. Joram (2003, S. 65 f.) beschreibt drei Stufen des Umgehens mit Stützpunkten. Zunächst sind Stützpunkte nur als konkrete Einheiten nutzbar, erst in einer zweiten Stufe sind sie als mentale Bilder repräsentiert:

> At the second level, […] benchmarks are represented as visual images: they can then serve as reference points when estimating.

Die dritte Stufe beschreibt die Fähigkeit, verschiedene Größenangaben an einem Stützpunkt sichtbar zu machen. So kann ein Stützpunkt, der 50 cm lang ist, auch als Stützpunkt für die Größenangabe 0,5 m genutzt werden (Joram, 2003, S. 66).

Der von Franke und Ruwisch (2010, S. 237) genutzte Begriff *Stützpunktwissen* bezieht sich auf die Kenntnis der quantitativen Größenausprägung eines Objekts. Zu unterscheiden ist in diesem Zusammenhang zwischen den Begriffen *Stützpunktwissen* und *Stützpunktvorstellungen*. Für die Größenart Gewicht nimmt Reuter (2011, S. 287) diese Unterscheidung wie folgt vor:

> Viele Kinder haben einzelnes Stützpunktwissen, bei dem es sich jedoch um rein theoretisches Wissen handelt, das nicht mit einer Vorstellung über das wahrnehmbare Gewicht verknüpft ist. […] Beim Konzept Stützpunktvorstellungen handelt es sich um ein Konzept, das sich offenbar erst mit wachsendem Umgang mit Gewichten dauerhaft innerhalb des Gewichtskonzepts aufbaut, weil dafür Stützpunktwissen mit erfahrenem wahrgenommenen Gewicht vernetzt werden muss, um sich zu Vorstellungen entwickeln zu können, auf die man zurückgreifen kann.

Bei Stützpunktwissen handelt es sich demnach um das Wissen über die Größenangabe eines Objekts, während es sich bei Stützpunktvorstellungen im Falle von visuell wahrnehmbaren Größen um ein mentales Bild der quantitativen Ausprägung handelt (zu mentalen Bildern vgl. Abschn. 3.1).

Grund unterscheidet zudem zwischen *mittelbaren* und *unmittelbaren Größenvorstellungen*. Der Unterschied liegt dabei in der Fassbarkeit und Darstellbarkeit der Größenvorstellungen: Während unmittelbare Größenvorstellungen durch „materielle Handlungen wie zeigen, umfahren, zeichnen, … wiedergegeben werden können“ (Grund, 1992, S. 42), ist das bei mittelbaren Größenvorstellungen nicht möglich. Hier muss auf sprachliche Beschreibungen mit Bezug zu anderen Größen oder auf einen Vergleich mit anderen Repräsentanten aus der Umwelt zurückgegriffen werden (Grund, 1992, S. 42 f.).

Im Folgenden wird statt des Wortes *Größenvorstellungen* das Wort *Stützpunktvorstellungen* verwendet. Dies geschieht aus zwei Gründen: Zum einen soll in besonderer Weise deutlich bleiben, dass Größen nur als Eigenschaft ihres Trägers vorstellbar sind und mentale Abbilder von Größen demnach immer mentale Bilder von Objekten sind. Zum anderen umfasst das Konzept von Größenvorstellungen, wie es in der Literatur beschrieben worden ist, nicht nur das Vorstellen von Repräsentanten von Größen, sondern darüber hinausgehende Kenntnisse wie das Wissen über die Größenangabe dieser Stützpunkte (im Folgenden als Stützpunktwissen bezeichnet) sowie die Fähigkeit, mit Größenangaben zu rechnen.

2.2 Der Einsatz von Stützpunkten: Strategien zum Schätzen von Längen, Flächeninhalten und Volumina

Um ohne konkrete Handlungen mit Körpermaßen oder anderen nichtstandardisierten Hilfsmitteln, ohne standardisierte Messgeräte und ohne zu raten zu einem vernünftigen und angemessenen Schätzwert für eine Größe eines Objekts zu kommen, ist die Anwendung von Schätzstrategien erforderlich. In der Literatur werden die Schätzstrategien hauptsächlich für das Schätzen von Längen beschrieben (Heid, 2017, S. 121 ff.; siehe auch Friebel, 1967, S. 479; Hildreth, 1983, S. 50; Joram, Subrahmanyam & Gelman, 1998, S. 415; Siegel et al., 1982, S. 226 f.), einige Autorinnen und Autoren nennen ebenfalls Strategien für das Schätzen von Volumina (Heid, 2017, S. 121 ff.) und Flächeninhalten (Hildreth, 1983, S. 51). Viele Strategien lassen sich auf die jeweils anderen Größenarten übertragen, auch wenn hierzu eine empirische Überprüfung fehlt (vgl. Abb. 2).

Aufgrund der Literaturlage lassen sich vier Kategorien von Schätzstrategien unterscheiden, die in den folgenden Abschnitten detailliert erklärt werden. Abb. 2 bietet eine Übersicht über die Kategorien (unterschiedliche Grautöne), die jeweiligen Strategien und die Größenarten, in denen sie angewendet werden können.

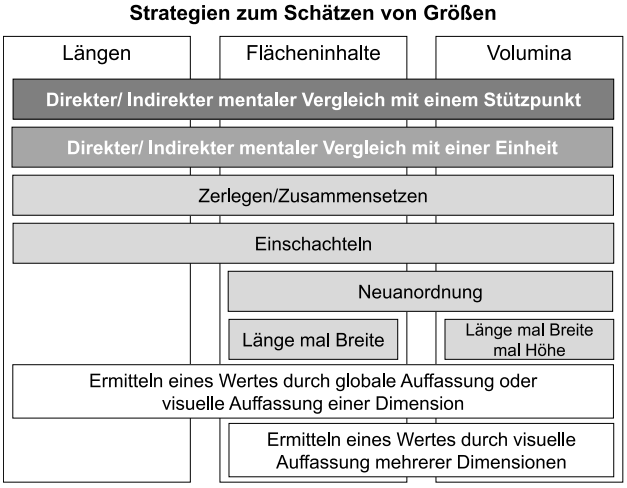


Abb. 2: Strategien zum Schätzen von Längen, Flächeninhalten und Volumina

Bereits aus der Definition des Schätzens von Größen ergibt sich die Kategorie *Direkter oder indirekter mentaler Vergleich mit einem Stützpunkt* (vgl. Abb. 2, dunkelgrau). Diese Kategorie beinhaltet solche Strategien, die allein auf dem mentalen Vergleich mit eingeprägten Repräsentanten beruhen und durch dieses Vorgehen die Ermittlung einer Größenangabe möglich machen. Die zweite Kategorie, *Direkter oder indirekter mentaler Vergleich mit einer Einheit* (vgl. Abb. 2, mittelgrau), geht auf eine Studie von Siegel et al. (1982, S. 228) zurück, in der Kinder auch durch den mentalen Vergleich mit standardisierten Einheiten zu einem Schätzergebnis kamen. Die Kategorie *Mentales Umstrukturieren der Schätzanforderung* (vgl. Abb. 2, hellgrau) beinhaltet mehrere Strategien. Diese führen in ihrer alleinigen Anwendung nicht zur Ermittlung eines Schätzwertes, Stattdessen haben sie zum Ziel, die Schätzanforderung zu vereinfachen, um die Anwendung einer anderen Strategie (in der Regel eine Vergleichsstrategie) besser zu ermöglichen. Die vierte Kategorie, *Ermitteln eines Schätzwertes durch visu-*

elle Auffassung (vgl. Abb. 2, weiß), berücksichtigt die von Heid (2017, S. 129) erkannten Reaktionen von Kindern auf eine Schätzanforderung, aufgrund des äußeren Erscheinungsbildes des zu schätzenden Objekts auf dessen Größe zu schließen.

2.2.1 Direkter oder indirekter mentaler Vergleich mit einem Stützpunkt

Mit Stützpunkten kann unterschiedlich mental operiert werden, um zu einem Schätzergebnis zu gelangen. Unterschieden werden hier das direkte und das indirekte mentale Vergleichen.

Die Strategie *direkter mentaler Vergleich mit einem Stützpunkt* wird in der englischsprachigen Literatur als *Benchmark Comparison* bezeichnet (Bright, 1976, S. 100; Heid, 2017, S. 124; Siegel et al., 1982, S. 226). Auch wenn das Vergleichsobjekt physisch anwesend und damit nicht nur in der Vorstellung repräsentiert ist, wird das Vorgehen dieser Strategie zugeordnet (Hildreth, 1983, S. 50).

Der Stützpunkt kann auch mental vervielfacht oder zerlegt werden, bevor ein Vergleich mit dem zu schätzenden Objekt vorgenommen wird. Dieses Vorgehen wird angewendet, wenn der Stützpunkt zu groß oder zu klein ist, um sinnvoll bei einem direkten mentalen Vergleich mit dem zu schätzenden Objekt zu einer Größenangabe zu führen. Heid (2017, S, 126) bezeichnet das Zerlegen oder Vervielfältigen eines Stützpunktes als *indirekten mentalen Vergleich*. Wenn der Stützpunkt zerlegt wird, wird dies in englischsprachiger Literatur mit *Fractional Benchmark* bezeichnet. Wird er vervielfältigt, heißt die Strategie *Multiple Benchmark* (Siegel et al., 1982, S. 213).

Beide Vorgehensweisen werden in der Literatur für die Größenart Längen beschrieben. Denkbar ist aber auch eine Übertragung auf die Größenarten Flächeninhalte und Volumina: So kann auch z. B. die Fläche eines Fußballfeldes oder das Fassungsvermögens eines Putzeimers als Stützpunkt dienen. Dennoch ist zu diskutieren, inwieweit das Vergleichen von Längen mit dem Vergleichen von Flächeninhalten und Volumina identisch ist. Beim Vergleich der Form von Flächenstücken und Körpern kann auf einen Vergleich der Seiten- bzw. Kantenlängen zurückgegriffen werden, was nicht einem *echten* Vergleich der Größenarten Flächeninhalte und Volumina entspricht.

2.2.2 Direkter oder indirekter mentaler Vergleich mit einer Einheit

Auch wenn Größen nicht ohne ihre Träger vorstellbar sind (vgl. Kapitel 2.1), finden sich in der Literatur Strategien, die auf dem Vergleich mit Einheiten beruhen. So führen Siegel et al. (1982, S. 228) eine als *Benchmark* bezeichnete Strategie auf. Sie weisen explizit darauf hin, dass die Kinder in der entsprechenden Studie standardisierte Einheiten zum gedanklichen Ausmessen nutzen.

Hildreth (1983, S. 50) nennt ebenfalls eine Strategie, die auf dem gedanklichen Ausmessen mit Einheiten beruht und von ihm als *Unit Iteration* (für Längen) bzw. *Repeated Addition* (für Flächeninhalte) bezeichnet wird. Denkbar ist demnach auch, dass das Volumen durch den Vergleich mit Einheiten ermittelt werden kann. Es wird jedoch nicht beschrieben, ob bei beiden Strategien lediglich das Bild einer Einheit ohne deren Träger verstanden werden soll oder ob als Einheit auch nichtstandardisierte Einheiten und insbesondere Stützpunkte aufgefasst werden können. Aufgrund der von ihm genannten Strategie *Comparison* (Hildreth, 1983, S. 50), die sich auf den Vergleich mit Repräsentanten bezieht, ist anzunehmen, dass zu *Unit Iteration* nur standardisierte Einheiten gezählt werden.

Beim Schätzen einer Größe verläuft die Grenze zwischen den Strategien *Mentaler Vergleich mit einem Stützpunkt* und *Mentaler Vergleich mit einer Einheit* fließend. Dies wird insbesondere durch das folgende Zitat deutlich:

> In order to form an estimation then, one must have a mental reference unit, that is, a mental ‚picture' or ‚feel' for the size of the unit. (Sowder, 1992, S. 371)

Die Einheit, die hier als Stützpunkt genutzt wird, kann entweder als mentales Bild (vgl. Abschn. 3.1) oder als „Gefühl" [2] für die Größe der Einheit vorliegen. Denkbar wäre auch, dass in den untersuchten Fällen, in denen diese Strategie auftritt, die schätzenden Personen die eventuell verwendeten Stützpunkte so verinnerlicht haben, dass sie ihnen als Objekt nicht bewusst waren, oder dass reflektierende oder sprachliche Fähigkeiten nicht ausreichend ausgeprägt waren, um die Verwendung von Stützpunkten zu beschreiben.

2.2.3 Mentales Umstrukturieren der Schätzanforderung

Dieser Abschnitt beinhaltet die Beschreibung von Strategien, die die ursprüngliche Schätzanforderung verändern, um anschließend eine weitere Strategie einfacher durchführen zu können. Diese Strategien sind somit als vorgeschaltete Strategien zu verstehen, die durch alleinige Verwendung nicht zu einem Schätzwert führen.

Zerlegen/Zusammensetzen. Das Zerlegen und anschließende Zusammensetzen des zu schätzenden Objekts benennen Siegel et al. (1982, S. 213) als *Decomposition/Recomposition*. Hildreth (1983, S. 50) bezeichnet diese Strategie als *Chunking*. Dabei wird das zu schätzende Objekt zuerst in Segmente

zerlegt, deren Größen einzeln geschätzt werden. Anschließend erfolgt eine mentale Zusammenfügung der ermittelten Größenangaben der einzelnen Segmente, um das Schätzergebnis zu erhalten. Diese Strategie ist dann geeignet, wenn für das Objekt als Ganzes kein entsprechender Stützpunkt vorliegt, der zu einem mentalen Vergleich herangezogen werden kann.

Der Prozess *Zerlegen* kann auf verschiedene Arten geschehen. Dies hängt i. d. R. mit dem zu schätzenden Objekt zusammen. *Regular Decomposition* liegt vor, wenn die einzelnen Segmente die gleiche Größe besitzen und regelmäßig angeordnet sind. Wenn eine dieser Voraussetzungen nicht gegeben ist, wird von *Irregular Decomposition* gesprochen (Siegel et al., 1982, S. 213). Die Unterteilung des Objekts kann bereits durch das Objekt selbst gegeben sein. Dies bezeichnet Hildreth (1983, S. 50) als *Use of subdivision clues*. Diese Unterteilungsmöglichkeit kann ebenfalls entweder regelmäßig oder unregelmäßig erfolgen, je nachdem, welche Struktur das zu schätzende Objekt aufweist (Siegel et al., 1982, S. 227).

Hildreth (1983, S. 50) führt darüber hinaus eine Strategie an, die sich auf das Wissen über das zu schätzende Objekt bezieht. *Prior knowledge* bedeutet, über einen Teil des Objekts bereits Informationen zu besitzen, so dass für einen Schätzprozess (unter anderem mittels der Strategie Zerlegen/ Zusammensetzen) kein weiterer Stützpunkt mehr erforderlich ist.

Der Prozess *Zusammensetzen* beinhaltet das Zusammenfügen der einzelnen Segmente, deren Größe geschätzt wurde. Siegel et al. (1982, S. 228) beschreiben dies als Multiplikation: Die Anzahl der Segmente (die gezählt oder ebenfalls geschätzt wird) wird mit der geschätzten Größenangabe multipliziert. Dieses Vorgehen ist jedoch nur zielführend, wenn die Segmente gleich groß sind (Regular Decomposition). Ansonsten müsste eine Addition der verschiedenen Schätzwerte vorgenommen werden. Es wird nicht eindeutig formuliert, ob sowohl das zu schätzende Objekt mental wieder zusammengefügt wird oder ob sich das Zusammensetzen nur auf das Rechnen mit den Größenangaben der einzelnen Segmente bezieht. Es ist zwangsläufig notwendig, die einzelnen Größenangaben zu einer Größenangabe für das gesamte Objekt zusammenzufügen, es ist aber nicht unbedingt notwendig, auch das mental zerlegte Bild wieder zu einem Bild zusammenzufügen. Um diese Strategie sinnvoll anzuwenden, ist jedoch Verständnis darüber erforderlich, dass das mental zerlegte Bild des Objekts wieder zusammengesetzt werden könnte und die Größenangaben der einzelnen Objektteile daher summiert die Größenangabe des zu schätzenden Objekts ergeben.

Die Strategie *Zerlegen/Zusammensetzen* ist keine Strategie, die allein zu einem Schätzergebnis führt, da es nur durch das Zerlegen und Zusammensetzen des zu schätzenden Objekts nicht zu einem Schätzergebnis kommen kann. Es ist eine weitere Strategie notwendig, um die Größe einzelner Segmente des zu schätzenden Objekts zu schätzen. Dies verdeutlichen Siegel et al. (1982, S. 228) in ihrem Modell durch den Verweis auf eine der anderen Strategien nach der Handlungsaufforderung, die Größe eines Segments zu schätzen. Auch Heid (2017, S. 137) betrachtet die Strategie *Zerlegen/Zusammensetzen* als vorgeschaltete Strategie.

Die Strategie *Zerlegen/Zusammensetzen* wird vor allem für das Schätzen von Längen beschrieben, kann aber auch für das Schätzen von Volumina angewendet werden (Heid, 2017, S. 122). Auch für das Schätzen von Flächeninhalten ist diese Strategie geeignet, da auch Flächen in kleinere Flächen zerlegt und anschließend wieder zusammengesetzt werden können.

Einschachteln. Um das Ermitteln eines Schätzwertes zu vereinfachen, kann ein Intervall gebildet werden, zwischen dessen Grenzen der gesuchte Schätzwert liegen soll. Siegel et al. (1982, S. 226) bezeichnen dies als *Range*, Hildreth (1983, S. 50) nennt das Vorgehen *Squeezing*, von Heid (2017, S. 123) wird es als *Einschachteln* bezeichnet. Auch diese Strategie ist nicht ohne den mentalen Vergleich mit Stützpunkten möglich: So muss sowohl für den Schätzwert der unteren Grenze als auch für den Schätzwert der oberen Grenze ein Schätzprozess stattfinden. Diese Strategie ist daher dann sinnvoll, wenn kein geeigneter Stützpunkt für einen Vergleich mit dem Objekt, dessen Größe geschätzt werden soll, vorliegt oder wenn die Strategie *Zerlegen/Zusammensetzen* nicht angewendet werden soll oder kann.

Frenzel und Grund (1991b, S. 32) beschreiben die „Angabe einer unteren und einer oberen Schranke für die gesuchte Größe“ als *Abschätzen*, welches als „Operieren (Vereinfachen, Vernachlässigen...) mit gegebenen Größen bei Beachtung bestehender Gesetzmäßigkeiten bzw. Zusammenhänge“ definiert und damit vom echten Schätzen abgegrenzt wird. Dieser Unterschied zur sonstigen Literatur ist möglicherweise damit zu erklären, dass die Strategie nicht ohne eine weitere Strategie funktioniert und eigentlich zwei Schätzungen durchgeführt werden: eine für die obere und eine für die untere Grenze. Da die bloße Notwendigkeit einer weiteren Strategie nicht ausschließt, dass ein Vorgehen ebenfalls zu den Schätzstrategien gehört (siehe Zerlegen/Zu-

sammensetzen und Neuanordnung), wird hier das Ermitteln eines Schätzwertes durch Festlegen einer oberen und einer unteren Grenze ebenfalls als den Schätzstrategien zugehörig verstanden.

Neuanordnung. Die Strategie *Rearrangement* wird laut Hildreth (1983, S. 51) für das Schätzen von Flächeninhalten angewendet. Dabei wird die zu schätzende Fläche zerlegt und so wieder neu zusammengesetzt, dass ihr Flächeninhalt leichter geschätzt werden kann. Im Unterschied zum Zerlegen/Zusammensetzen wird bei dieser Strategie nicht die Größe eines Segments geschätzt, sondern die gesamte neu entstandene Fläche. Diese Strategie ist auch für das Schätzen von Volumina geeignet. So kann ein Körper zerlegt und so wieder zusammengesetzt werden, dass dessen Volumen leichter geschätzt werden kann. Für das Schätzen von Längen ist die Strategie *Neuanordnung* nicht in diesem Sinne möglich. Jedoch kann auch der Repräsentant einer Länge mental umstrukturiert werden, damit die Anwendung weiterer Strategien leichter fällt. So kann z. B. eine gekrümmte Linie (z. B. eine Straße oder der Umfang eines Objektes) in der Vorstellung zu einer geraden Linie neu angeordnet werden.

Anwendung findet diese Strategie dann, wenn der Repräsentant einer Größe eine andere äußere Struktur (z. B. Krümmung) aufweist als der Stützpunkt und daher eine Angleichung erfolgen soll. Auch um weitere Strategien anwenden zu können, kann eine Umstrukturierung im Sinne der Neuanordnung hilfreich sein.[3]

Länge mal Breite (mal Höhe). Eine Strategie zum Schätzen des Flächeninhalts eines Rechtecks ist das Schätzen und anschließende Multiplizieren der Längen der Seiten. Dieses Vorgehen wird in englischsprachiger Literatur als *Length Times Width* bezeichnet (Hildreth, 1983, S. 51). Auch für das Schätzen der Volumina von Quadern ist diese Strategie denkbar, indem die dritte Dimension ergänzt wird (die Strategie heißt dann entsprechend *Length Times Width Times Height*).

Die Strategie könnte mit dem entsprechenden Formelwissen auch auf andere Flächen und Körper übertragen werden.

Zum Schätzen der Längen der Seiten bzw. Kanten muss sowohl bei Flächeninhalten als auch bei Volumina eine weitere Strategie angewendet werden. Die Umstrukturierung der Schätzanforderung bezieht sich hier nicht auf die Form des zu schätzenden Objekts, sondern auf die Veränderung der erforderlichen Vergleichsprozesse. So beziehen sich die eigentlichen Schätzstrategien des Vergleichens nicht auf die Größenarten Flächeninhalte oder Volumina, sondern auf die Größenart Längen. Daraus ergibt sich, dass für das Schätzen von Flächeninhalten oder Volumina nicht zwangsläufig Stützpunkte aus diesen Größenarten erforderlich sind, sondern diese durch Stützpunkte zur Größenart Längen (die sich auch auf das zu schätzende Objekt beziehen können) ersetzt werden können.

2.2.4 Ermitteln eines Schätzwertes durch visuelle Auffassung

Neben den Strategien, die auf einem mentalen Vergleich mit Stützpunkten beruhen, gibt es Strategien, die nur auf der Wahrnehmung des zu schätzenden Objekts beruhen. Hierzu zählt die Strategie *Eyeball*, bei der das Schätzergebnis aufgrund der subjektiven Wahrnehmung der Größen und deren Verknüpfung mit einer als groß oder klein empfundenen Größenangabe ermittelt wird (Siegel et al., 1982, S. 226). Heid (2017, S. 112) bezeichnet dies als *visuelles Vorgehen* und unterscheidet drei Arten: *das Ermitteln eines Schätzwertes durch globale Auffassung*, *durch die visuelle Auffassung einer Dimension* sowie *durch die visuelle Auffassung zweier Dimensionen*. Die globale Auffassung des zu schätzenden Objekts ist vergleichbar mit der Strategie Eyeball, da von den Versuchspersonen als Begründung für den Schätzwert lediglich Wörter wie groß/klein und viel/wenig genutzt werden. Bei der Berücksichtigung von einer bzw. zwei Dimensionen wird auf die Gestalt des zu schätzenden Objekts eingegangen. Dies wird bei einer Dimension durch Wörter wie z. B. lang/kurz, bei zwei und mehr Dimensionen durch Wörter wie z. B. dick/dünn, breit/schmal ausgedrückt, wie Heid (2017, S. 129) in einer Untersuchung zu Längen und Fassungsvermögen feststellte. Zum Schätzen eines Flächeninhalts bzw. eines Volumens könnte diese Strategie ebenso angewendet werden, wenn es sich um eine (näherungsweise) rechteckige Fläche bzw. um einen (näherungsweise) quaderförmigen Körper handelt.

Die visuelle Auffassung geschieht in der Regel ohne den mentalen Vergleich mit Stützpunkten und ist daher im eigentlichen Sinne keine Schätzstrategie.[4]

3. Zentrale kognitive Komponenten des Schätzprozesses

In der Kognitions- und Neuropsychologie wird das kognitive Schätzen als eine Tätigkeit genutzt, durch deren Einschränkung Defizite im Frontalhirn festgestellt werden können (D'Aniello, Castelnuovo & Scarpina, 2015, S. 1). Die Vielzahl an Schätztests, die zur Untersuchung und Unterscheidung verschiedener Defizite im Frontalhirn sowie zur Feststellung von Korrelationen mit anderen Fähigkeiten entwickelt wurden (Axelrod & Millis, 1994; Brand et al., 2002; Bullard et al., 2004; Goldstein et al., 1996; Liss, Fein, Bullard & Robins, 2000; MacPherson et

al., 2014; Mendez, Doss & Cherrier, 1998), gehen auf den Cognitive Estimation Test (CET) von Shallice und Evans (1978) zurück, die als erste einen Zusammenhang der kognitiven Schätzfähigkeit mit exekutiven Funktionen feststellten.

Das erste Modell, welches den Prozess des kognitiven Schätzens aus Sicht der Kognitionspsychologie abbildet, zeigt die Beteiligung des Arbeitsgedächtnisses, des deklarativen Langzeitgedächtnisses und einer zentralen Kontrollinstanz am Schätzprozess (Brand et al., 2002, vgl. Abb. 3). Dieses Modell wurde von D'Aniello et al. (2015) weiterentwickelt (vgl. Abb. 4). Hierzu zählt insbesondere die Spezifizierung der Vorgänge im Arbeitsgedächtnis, welche sich in Reasoning und Monitoring unterscheiden lassen. Zusätzlich werden das semantische Wissen und die zentrale Exekutive des Arbeitsgedächtnisses der kristallinen Intelligenz (vgl. Abschn. 3.2) zugeordnet.

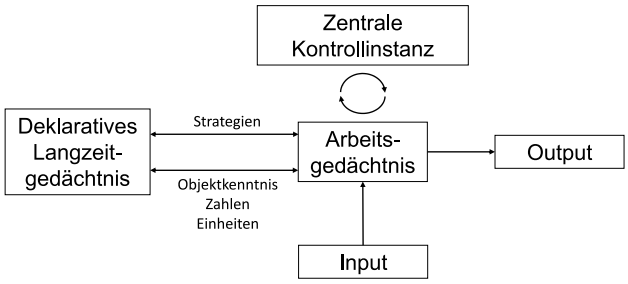


Abb. 3: Modell zum kognitiven Schätzen (Brand et al., 2002, S. 283)

Die Schätzanforderung trifft zunächst auf das *Arbeitsgedächtnis* (Brand et al., 2002, S. 283) bzw. auf die *zentrale Exekutive des Arbeitsgedächtnisses* (D'Aniello et al., 2015, S. 2). Die Verarbeitung der aufgenommenen Informationen im *Arbeitsgedächtnis* geschieht unter Rückgriff auf das *deklarative Langzeitgedächtnis* (Brand et al., 2002, S. 283), aus dem *semantisches Wissen* (D'Aniello et al., 2015, S. 2) in Form von Wissen über *Strategien, Objekte, Zahlen* und *Einheiten* (Brand et al., 2002, S. 283) abgerufen wird. Die Prozesse des Arbeitsgedächtnisses können in *Reasoning* und *Monitoring* differenziert werden. Das *Reasoning* beinhaltet das Formulieren einer Hypothese, welche im zweiten Schritt durch das *Monitoring* auf Plausibilität geprüft wird. Im Falle einer bizarren Antwort wird der Prozess erneut durchlaufen. Sofern die Hypothese als nachvollziehbar anerkannt wird, wird die Antwort als *Output* ausgegeben. Zusätzliche Faktoren, die auf das Reasoning einwirken, sind *räumliche Fähigkeiten,* die Fähigkeit zum *Planen und Problemlösen, numerische Fertigkeiten* sowie *Erfahrungen* (D'Aniello et al., 2015, S. 2).

Als die drei wesentlichen Schritte des Schätzprozesses können aus beiden Modellen das Verständnis der Frage im Sinne einer Repräsentation der Schätzanforderung im Arbeitsgedächtnis, das Reasoning als Verarbeitung der wahrgenommenen und aus dem Langzeitgedächtnis abgerufenen Informationen im Arbeitsgedächtnis und das Monitoring als Überprüfung des aus dem Reasoning hervorgehenden Ergebnisses unter Nutzung einer zentralen Exekutive abgeleitet werden.

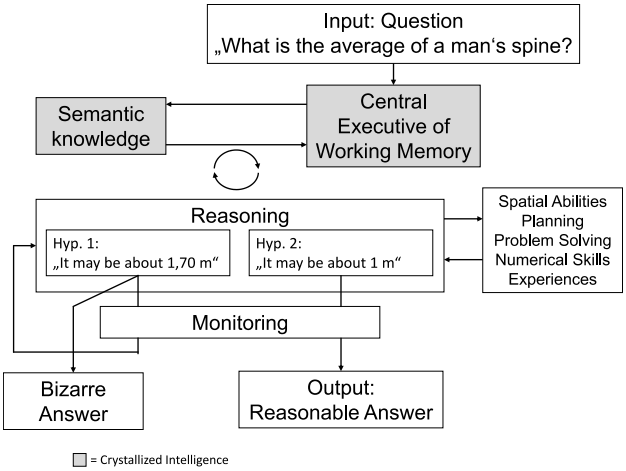


Abb. 4: Erweitertes Modell zum kognitiven Schätzen (D'Aniello et al., 2015, S. 2)

Die drei am Schätzprozess beteiligten kognitiven Komponenten sind daher das *Arbeitsgedächtnis*, das *Langzeitgedächtnis* und eine *zentrale Kontrollinstanz*, deren Tätigkeiten während des Schätzprozesses im Folgenden genauer erläutert werden.

3.1 Repräsentation der Schätzanforderung im Arbeitsgedächtnis

Kennzeichnend für das Arbeitsgedächtnis ist die Repräsentation von Informationen für eine kurze Zeit, die zur Weiterverarbeitung dieser Informationen benötigt wird. Diese Informationen werden entweder neu aufgenommen oder aus dem Langzeitgedächtnis abgerufen (Goswami, 2001, S. 248). Das bloße Abrufen von Informationen aus dem Langzeitgedächtnis ist jedoch nicht ausreichend, um ein Schätzergebnis zu erhalten. Vielmehr müssen diese im Arbeitsgedächtnis kombiniert werden (Brand et al., 2002, S. 283). Das Anwenden von Schätzstrategien kann als eine solche Kombination von Informationen verstanden werden (Brand et al., 2002; D'Aniello et al., 2015).

In der Forschung über das Arbeitsgedächtnis fehlt eine die Forschungsansätze verbindende Definition (Berti, 2010, S. 3; Ullsperger & von Cramon, 2006, S. 482). Beispielhaft seien hier einerseits Baddeley und Hitch sowie andererseits Cowan genannt. Baddeley und Hitch verstehen das Arbeitsgedächtnis als Mehrkomponentenmodell (Baddeley, 1992, S. 556). Cowan vertritt eine eher weite Definition und rechnet alle Hilfsmittel, die den Denkprozess unterstüt-

zen, dem Arbeitsgedächtnis zu. Dazu gehört auch das Verwenden von Stift und Papier (zitiert nach Berti, 2010, S. 3).

Aufgrund dieser Spannbreite ist ein expliziter Bezug auf eines der Verständnisse des Arbeitsgedächtnisses für die Untersuchung der am Schätzprozess beteiligten Komponenten erforderlich. Brand et al. (2002, S. 282) berufen sich auf das Arbeitsgedächtnis im Sinne von Baddeley und Hitch. Demnach besteht das Arbeitsgedächtnis aus mehreren Komponenten. Einer *zentralen Exekutive* sind zwei bzw. drei Subkomponenten untergeordnet. Für die Verarbeitung von verbalen Informationen ist die *phonologische Schleife* zuständig, während Bilder und räumliche Informationen vom *visuell-räumlichen Notizblock* verarbeitet werden (Baddeley, 1992, S. 556). Als dritte Subkomponente fügte Baddeley (2000, S. 421) später den *episodischen Speicher* hinzu. Dieser kann kurzzeitig Informationen aus beiden Bereichen bereithalten und dient als Speicherkomponente für exekutive Prozesse. Diese Erweiterung ist aber umstritten (Hagendorf, 2006, S. 343).

Das Modell von D'Aniello et al. (2015, S. 1) bezieht sich ebenfalls auf die Arbeitsgedächtnisdefinition nach Baddeley und Hitch. Durch die Darstellung (vgl. Abb. 4) wird suggeriert, dass die Prozesse *Reasoning* und *Monitoring* laufend während des Austausches zwischen zentraler Exekutive des Arbeitsgedächtnisses und semantischem Wissen geschehen. Das Monitoring wird also der zentralen Exekutive des Arbeitsgedächtnisses untergeordnet: Dies kann so interpretiert werden, dass die zentrale Exekutive „Ort" dieser Plausibilitätsprüfung ist. Die beiden Subsysteme visuell-räumlicher Notizblock und phonologische Schleife werden nicht explizit genannt.

D'Aniello et al. (2015, S. 1) weisen auf die Relevanz mentaler Bilder in Form von *Mental Imagery* hin, auf welches beim Reasoning zurückgegriffen wird. Mental Imagery bedeutet, mit Hilfe mentaler Bilder eine Lösung für ein Problem zu finden, ohne dass die verwendeten Informationen tatsächlich der Wahrnehmung zur Verfügung stehen (Kosslyn, Thompson & Ganis, 2006, S. 3). Ein mentales Bild kann wie folgt definiert werden:

> A mental image occurs when a representation of the type created during the initial phases of perception is present but the stimulus is not actually being perceived; such representations preserve the perceptible properties of the stimulus and ultimately give rise to the subjective experience of perception. (Kosslyn et al., 2006, S. 4).

Mental Imagery ist dabei nicht begrenzt auf visuelle Aspekte, sondern kann ebenso auditive und taktile Elemente enthalten (Kosslyn et al., 2006, S. 4). In Bezug auf die visuell erfassbaren Größen ist aber vor allem das visuelle mentale Bild von Relevanz. Daher ist für das Schätzen dieser Größen auch der visuell-räumliche Notizblock so bedeutend, obwohl dieser im Modell von D'Aniello et al. (2015, S. 1) nicht explizit erwähnt wird. Der visuell-räumliche Notizblock hält Informationen bereit, die auf der Wahrnehmung sichtbarer Eindrücke beruhen und als ikonische Codes bezeichnet werden (Posner & Koppitz, 1976, S. 38).

Die Subsysteme des Arbeitsgedächtnisses werden von der zentralen Exekutive gesteuert, das heißt, dass Informationen von einem System ins andere übertragen, in diese eingebracht oder aus diesen abgerufen werden (Anderson, 2013, S. 121). Darüber hinaus bestimmt die zentrale Exekutive, wie die Untersysteme genutzt werden, sie lenkt die Aufmerksamkeit und stellt die Verbindung zum Langzeitgedächtnis her. Eine weitere Aufgabe der zentralen Exekutive des Arbeitsgedächtnisses ist es, Informationen entsprechend ihrer Kodierung den Subsystemen zuzuordnen, damit sie weiterverarbeitet werden können (Smith & Jonides, 1999, S. 1659).

Die Aktivität der zentralen Exekutive wird im frontalen Kortex des Gehirns vermutet (Goswami, 2001, S. 257). Dies verdeutlicht, weshalb sie in beiden Modellen zum kognitiven Schätzen gesondert aufgeführt ist. Die Bedeutung der zentralen Exekutive für das Schätzen wird in Abschn. 3.3 erläutert.

Die zentrale Exekutive des Arbeitsgedächtnisses muss als kognitive Funktion verstanden und daher vom Monitoring, einer Tätigkeit, die unter Verwendung verschiedener kognitiver Funktionen durchgeführt wird, abgegrenzt werden (vgl. Abschn. 3.3).

3.2 Abruf von Informationen aus dem Langzeitgedächtnis

Das Anwenden von Strategien zum Lösen der Schätzaufgabe erfordert das Abrufen von Informationen aus dem Langzeitgedächtnis (Brand et al., 2002, S. 283; D'Aniello et al., 2015, S. 2). Beide beschriebenen Modelle des kognitiven Schätzens legen durch ihre Darstellung nahe, dass das Arbeitsgedächtnis bzw. die zentrale Exekutive des Arbeitsgedächtnisses mit dem Langzeitgedächtnis in Verbindung steht, um die erforderlichen Informationen zur Verwendung abzurufen. Diese Prozesse sind wesentlich für das Formulieren einer ersten Hypothese über den Schätzwert. Brand et al. beziehen sich dabei auf das *deklarative Langzeitgedächtnis*, D'Aniello et al. weisen dem *semantischen Wissen* eine große Relevanz zu.

Das Langzeitgedächtnis enthält Informationen, die dauerhaft gespeichert sind. Im Unterschied zum Arbeitsgedächtnis weist das Langzeitgedächtnis vermutlich keine Kapazitätsbeschränkung auf (van der Meer, 2006, S. 346). Das Langzeitgedächtnis wird in ein *deklaratives Langzeitgedächtnis* und ein *nichtdeklaratives Langzeitgedächtnis* unterteilt (vgl. Abb. 5 als Darstellung der Strukturzusammenhänge). Das deklarative Langzeitgedächtnis

> enthält Wissen über Tatsachen. Dieses Wissen kann bewusst erinnert und genutzt werden [...]. Deklaratives Wissen ist flexibel, d. h. auch in neuartigen Situationen einsetzbar. (van der Meer, 2006, S. 347)

Im Unterschied dazu sind die Inhalte des nichtdeklarativen Langzeitgedächtnisses nicht unbedingt dem Bewusstsein zugänglich, sondern bestimmen das Verhalten und die Individualität (van der Meer, 2006, S. 348 f.). Wissen, das sich auf die „Ausführung von (automatisierten) Fertigkeiten“ (Schermer, 2006, S. 128) bezieht, ist im *prozeduralen Gedächtnis* gespeichert. Auch diese sind weitgehend nicht bewusst zugänglich, selbst wenn sie das Verhalten beeinflussen. *Prozedurales Gedächtnis* kann als eine ältere Bezeichnung für das nichtdeklarative Langzeitgedächtnis gelten (Schermer, 2006, S. 128).

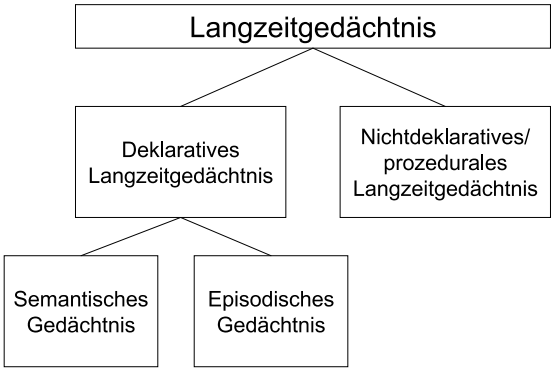


Abb. 5: Gliederung des Langzeitgedächtnisses

Für das Reasoning wird auf semantisches Wissen, räumliche Fähigkeiten und numerisches Wissen zurückgegriffen (D’Aniello et al., 2015, S. 2). Semantisches Wissen wird als „Wissen über Wortbedeutungen (Konzepte), allgemeines Faktenwissen über die Welt [...] und über Regelsysteme“ (van der Meer, 2006, 348) definiert. Demnach kann numerisches Wissen im Sinne von erlerntem mathematischem Wissen ebenfalls als semantisches Wissen aufgefasst werden. Was genau unter numerischem Wissen verstanden wird, geht aus dem Modell jedoch nicht hervor. Räumliche Fähigkeiten sprechen „verschiedenartige gedankliche Leistungen an, die vielfach zu einer Differenzierung in Teilbereiche führen“ (Franke & Reinhold, 2016, S. 61). Ein genauer Bezug, welche Differenzierung und damit welche gedanklichen Leistungen unter räumliche Fähigkeiten gefasst werden, geht aus dem Modell ebenfalls nicht direkt hervor. Der Verweis auf Horazek et al. (2010, S. 189 ff.) legt jedoch nahe, dass die Fähigkeit, sich bewegte und unbewegte Objekte vorzustellen, für das kognitive Schätzen von Bedeutung ist. Genauer spezifiziert wird insbesondere die mentale Rotation (Horazek et al., 2010, S. 191).

Das *semantische Gedächtnis* wird als Subsystem des deklarativen Langzeitgedächtnisses aufgefasst (van der Meer, 2006, S. 347), sodass D’Aniello et al. hier eine Spezifizierung des Modells von Brand et al. vornehmen.

Mit dem Rückgriff auf Erfahrungen (D’Aniello et al., 2015, S. 2) wird auch das zweite Subsystem des deklarativen Langzeitgedächtnisses, das *episodische Langzeitgedächtnis*, als am Schätzprozess beteiligt angesehen. Das episodische Langzeitgedächtnis speichert autobiographische Ereignisse mit den raum-zeitlichen Charakteristika, d. h. „wann und wo die Information erworben wurde“ (van der Meer, 2006, S. 348).

Es ist fraglich, inwieweit nur das deklarative Langzeitgedächtnis am Schätzprozess beteiligt ist. Da automatisierte Fähigkeiten, wie das Ausführen einer einfachen Rechnung, dem nichtdeklarativen Langzeitgedächtnis zugeordnet werden, ist davon auszugehen, dass dieses auch am Schätzprozess beteiligt ist.

Im Modell von D’Aniello et al. (2015, S. 2) sind das semantische Wissen und die zentrale Exekutive des Arbeitsgedächtnisses als *kristalline Intelligenz* gekennzeichnet. Dieser Begriff geht auf Cattell (1963, S. 2) zurück:

> Crystallized ability loads more highly those cognitive performances in which skilled judgment habits have become crystallized […] as the result of earlier learning application of some prior, more fundamental general ability to these fields. […] Fluid general ability, on the other hand, shows more in tests requiring adaption to new situations, where crystallized skills are of no particular advantage.

Auch Brand et al. (2002, S. 284) vermuten, dass für Schätzaufgaben die kristalline Intelligenz wichtiger ist als die *fluide Intelligenz*, auch wenn das in ihrem Modell nicht extra gekennzeichnet ist. Insofern scheinen erlernte Prozesse und erworbenes Wissen maßgeblich für die Schätzfähigkeit zu sein. Aus Sicht der Mathematikdidaktik ist dies von Bedeutung, weil das Entwickeln von Stützpunktwissen und Stützpunktvorstellungen sowie damit verbundener Schätzstrategien als Prozess gesehen wird, der Zeit in Anspruch nimmt. Objekte können nicht einfach so als Stützpunkt dienen, sondern der Einsatz von Schätzstrategien muss gelernt und durch die

Lehrperson unterstützt und initiiert werden (Joram, 2003, 65 f.).

3.3 Überprüfung der Antwort durch eine zentrale Kontrollinstanz

Schon Shallice und Evans (1978, S. 298) verweisen auf die Bedeutung der Kontrolle des ersten ermittelten Schätzwertes:

> A patient unable to obtain an appropriate strategy […] or who has inappropriate error-checking is more likely to produce a very incorrect response.

Der dritte wesentliche Schritt im Schätzprozess ist daher die Überprüfung der Hypothese durch eine zentrale Kontrollinstanz, deren Aufgabe es ist,

> mögliche Fehler zu entdecken sowie die weitere Suche nach passenden Antworten mit Hilfe anderer Strategien und/oder anderen Informationen aus dem Langzeitgedächtnis einzuleiten. (Brand et al., 2002, S. 283)

Dieses Vorgehen wird von D'Aniello et al. (2015, S. 1 f.) als *Monitoring* bezeichnet:

> The hypothesis […] is then checked by the second main level of executive processing, that is monitoring: at this stage, the formulated hypothesis is compared with all the information gleaned during the process in order to verify its consistency […]. In case the hypothesis is considered as an inconsistent or inadequate one, it is sent back to the reasoning process.

Auch Siegel et al. (1982, S. 226 f.) verweisen auf die Relevanz der Überprüfung des durch Schätzstrategien ermittelten Schätzwertes. In ihrem Modell zur Anwendung der verschiedenen Strategien wird dies durch die vor dem Output stehende Frage „Does it seem to fit?" oder „Reasonable?" verdeutlicht. Wird diese Frage verneint, verweisen die Pfeile des Modells auf vorhergehende Schritte der Strategieanwendung (vgl. Abb. 6). Wie genau diese Überprüfung und damit die Beurteilung der Plausibilität stattfindet, geht aus dem Modell jedoch nicht hervor. Denkbar wäre der Vergleich der ermittelten Größenangabe mit einem weiteren Stützpunkt oder das wiederholte Anwenden der eingeschlagenen Strategie.

Das Monitoring wird als zweites Level exekutiver Funktionen bezeichnet (D'Aniello et al., 2015, S. 2), womit verdeutlicht werden soll, dass nicht nur das Reasoning (als erstes Level exekutiver Funktionen), sondern auch das Monitoring auf die exekutiven Funktionen zurückgreift und diese daher an mehreren Stellen relevant für den Schätzprozess sind (vgl. Abschn. 5.2). Auch Shallice und Evans (1978, S. 301) benennen die Relevanz der exekutiven Funktionen und weisen ebenfalls auf verschiedene Ebenen hin:

> The selection and regulation of cognitive planning is one of the main functions of the human frontal lobes. Such planning functions would presumably by mediated through high-level programs which control the operation of lower-level cognitive programs.

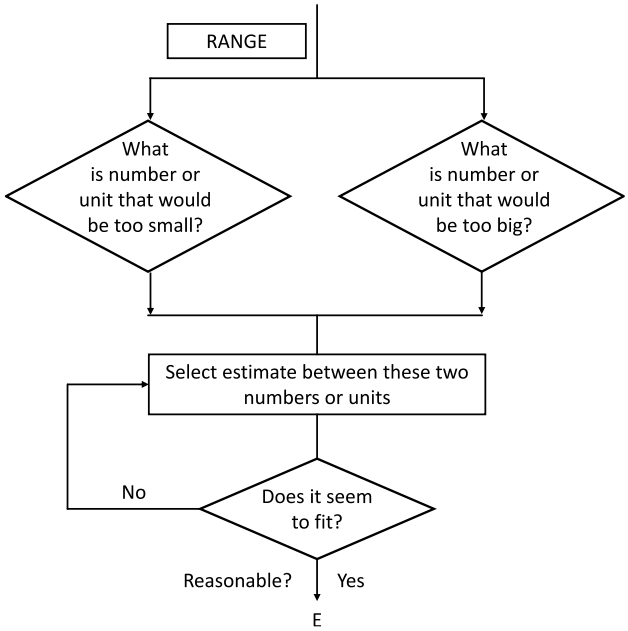


Abb. 6: Überprüfung der Plausibilität des Schätzergebnisses am Beispiel der Strategie „Range" (Siegel et al., 1982, S. 226)

Es wird aus dem Modell von D'Aniello et al. (2015) nicht deutlich, ob die Unterscheidung in Levels derjenigen Unterscheidung entspricht, die auch Shallice und Evans vornehmen. Denkbar wäre, dass beim Reasoning eher Tätigkeiten ausgeführt werden, die den Tätigkeiten auf dem unteren Level nach Shallice und Evans entsprechen. Dies könnte z. B. arithmetisches Rechnen als Routinetätigkeit umfassen (Shallice & Evans, 1978, S. 301). Die Plausibilitätsprüfung durch die zentrale Kontrollinstanz entspräche demnach eher Tätigkeiten auf dem höheren Level nach Shallice und Evans.

Im Modell von D'Aniello et al. (2015) ist die Kontrolle zweifach dargestellt: einmal als kognitive Funktion (zentrale Exekutive des Arbeitsgedächtnisses) und einmal als Tätigkeit (Monitoring). Beide Kontrollinstanzen bedienen sich exekutiver Funktionen, was erneut zeigt, dass diese an verschiedenen Schritten des Schätzprozesses von Bedeutung sind (vgl. Abschn. 5.2).

Für das kognitive Schätzen ist die zentrale Exekutive des Arbeitsgedächtnisses von besonderer Relevanz, da sie die Ausführung des Reasonings und Monitorings steuert und für die Auswahl des geeigneten Plans verantwortlich ist (D'Aniello et al., 2015, S. 1 f.)

4. Kognitives Schätzen aus mathematikdidaktischer Sicht: Kontext und Rahmung eines theoretischen Modells

Die Kenntnis über Fähigkeiten, die für einen Schätzprozess erforderlich sind, ist unter anderem für das schulische Lernen von Bedeutung. Grassmann (1999, S. 34) vermutet eine Wechselwirkung beim Erwerb von Stützpunktvorstellungen und Stützpunktwissen und der Schätztätigkeit. Zudem kann das Schätzen von Größen als Möglichkeit dienen, den physikalischen Messprozess zu verstehen (Joram, Gabriele, Bertheau, Gelman & Subrahmanyam, 2005, S. 4). Andererseits ist das Verständnis des Messprozesses auch Voraussetzung für den Schätzprozess (Joram et al., 1998, S. 415, vgl. auch Abschn. 5.1.4).

Weil Schätzen in der Mathematikdidaktik wie auch in der Kognitionspsychologie mit verschiedenen Schwerpunktsetzungen beforscht wird, ist für eine umfassende Betrachtung des Schätzprozesses und der darin involvierten Fähigkeiten die Kombination der Erkenntnisse aus beiden Wissenschaften gewinnbringend. Dabei ist zwischen einem *Schätzen im engeren Sinne* und einem *Schätzen im weiteren Sinne* zu unterscheiden (vgl. Abb. 7).

Wesentlich für das Durchführen eines Schätzprozesses ist aus mathematikdidaktischer Sicht das *Anwenden von Schätzstrategien*. Das Ausführen dieser Strategien kann aus kognitionspsychologischer Perspektive dem Reasoning zugeordnet werden. Damit verschränkt ist das Monitoring, welches die Plausibilität des durch die Strategien ermittelten Ergebnisses beurteilt. In kognitionspsychologischen Modellen werden zusätzlich der Input, das heißt die Schätzanforderung (meist in Form einer Schätzfrage), und der Output, das heißt der geäußerte Schätzwert, berücksichtigt. In einem Modell aus mathematikdidaktischer Perspektive steht das Schätzen im engeren Sinne, unter welchem die Strategieanwendung einschließlich deren Überprüfung verstanden wird, im Fokus.

Dem Schätzprozess im engeren Sinne geht das *Verstehen der Fragestellung bzw. der Schätzanforderung* voraus (vgl. Abb. 7, linke Spalte). Dies erfordert beim quantitativen Schätzen neben *sprachlichen Fähigkeiten* für das Verständnis einer Schätzfrage ebenso das Wissen darüber, dass eine Größenangabe als Antwort erwartet wird. Außerdem muss ein Konzept der zu schätzenden Größe vorhanden sein, um den sprachlichen Informationen zu entnehmen, welche Größe geschätzt werden soll (Hildreth, 1983, S. 50). Dies kann als *allgemeines (mathematisches) Wissen* (vgl. Abschn. 5.3.1) verstanden werden. Letztendlich muss diese Größe an einem konkreten oder mentalen Bild wiedererkannt werden können.

Nach der Anwendung von Strategien und der Überprüfung der daraus resultierenden Ergebnisse erfolgt *das Äußern des Schätzergebnisses* als Output (vgl. Abb. 7, rechte Spalte). Hierfür sind wiederum *sprachliche Fähigkeiten* erforderlich, die mit *allgemeinem (mathematischem) Wissen* (vgl. Abschn. 5.3.1) kombiniert werden müssen, um eine Größenangabe zu formulieren. Der Schätzprozess als Ganzes, so, wie er in der Kognitionspsychologie abgebildet wird, kann demnach aus mathematikdidaktischer Sicht als Schätzen im weiteren Sinne bezeichnet werden (vgl. Abb. 7 gesamt).

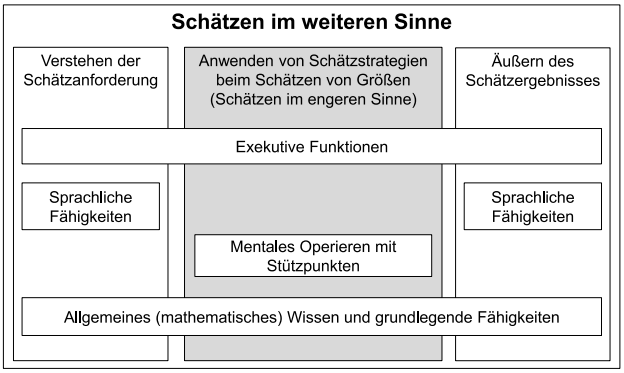


Abb. 7: Schätzen im weiteren Sinne

Das Schätzen im engeren Sinne (vgl. Abb. 7, mittlere Spalte) beruht auf dem *mentalen Operieren mit Stützpunkten* unter Anwendung *exekutiver Funktionen* sowie *allgemeinen mathematischen Wissens und grundlegender Fähigkeiten*. Diese drei Aspekte sind wesentlich für das Anwenden von Schätzstrategien beim Schätzen von Längen, Flächeninhalten und Volumina und werden daher in dem im nächsten Kapitel vorgestellten Modell weiter spezifiziert. Die Verbindung des Schätzens aus mathematikdidaktischer Sicht mit dem kognitiven Schätzen aus psychologischer Sicht wird hier durch die Verschränkung von Fähigkeiten aus beiden Disziplinen sowie durch die mathematikdidaktische Interpretation psychologischer Fähigkeiten ausgedrückt. Das hier vorgestellte Modell bezieht sich auf die Fähigkeiten, die für das Schätzen von Längen, Flächeninhalten und Volumina erforderlich sind. Eine Übertragung auf das Schätzen von anderen Größen, Anzahlen oder Rechenergebnissen ist zumindest teilweise nicht ausgeschlossen, aber aufgrund der Besonderheiten der visuell erfassbaren Größen nicht intendiert. Ebenso müsste für das qualitative Schätzen von Größen eine Anpassung einzelner Fähigkeiten vorgenommen werden, insbesondere auf der sprachlichen Ebene. [5]

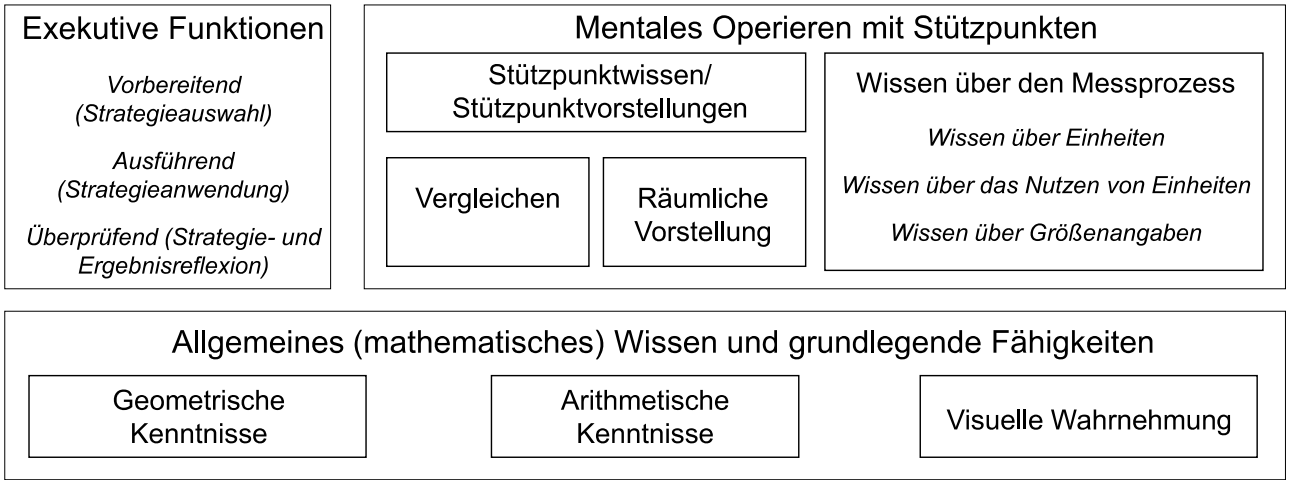


Abb. 8: Schätzen im engeren Sinne und dazu erforderliche Fähigkeiten (auf Hauptkategorien reduziertes Modell)

5. Schätzen im engeren Sinne: Modell zum Schätzen visuell wahrnehmbarer Größen

Aus mathematikdidaktischer Sicht ist insbesondere die detaillierte Betrachtung der Strategieanwendung beim Schätzen von Größen interessant. In der Forschung über Schätzstrategien herrscht zwar Konsens über die Existenz der meisten Schätzstrategien (vgl. Kapitel 2.2), welche Fähigkeiten zur Ausführung erforderlich sind, wurde jedoch bisher in mathematikdidaktischer Literatur noch nicht detailliert erfasst. Es lassen sich drei Gruppen von Fähigkeiten, die zur Anwendung von Schätzstrategien erforderlich sind, differenzieren: das *mentale Operieren mit Stützpunkten*, *die exekutiven Funktionen* sowie *allgemeines (mathematisches) Wissen und grundlegende Fähigkeiten*.

Abbildung 8 zeigt eine erste Ausdifferenzierung dieser drei Gruppen, die im Folgenden näher betrachtet und dadurch weiter spezifiziert werden.

Eine Gesamtdarstellung des vollständig ausdifferenzierten Modells befindet sich im Anhang.

5.1 Mentales Operieren mit Stützpunkten

Schätzstrategien im mathematikdidaktischen Sinne beruhen hauptsächlich auf dem Operieren mit Stützpunkten (vgl. Abschn. 2.2). Dies umfasst neben dem direkten mentalen Vergleich auch das Zerlegen und Zusammensetzen sowie Teilen und Vervielfachen von Stützpunkten. Für diese Tätigkeiten sind vier Aspekte von Relevanz: das *Wissen über Stützpunkte*, die Fähigkeit zu *vergleichen*, die Fähigkeit zur *räumlichen Vorstellung* sowie das *Wissen über den Messprozess* (vgl. Abb. 9).

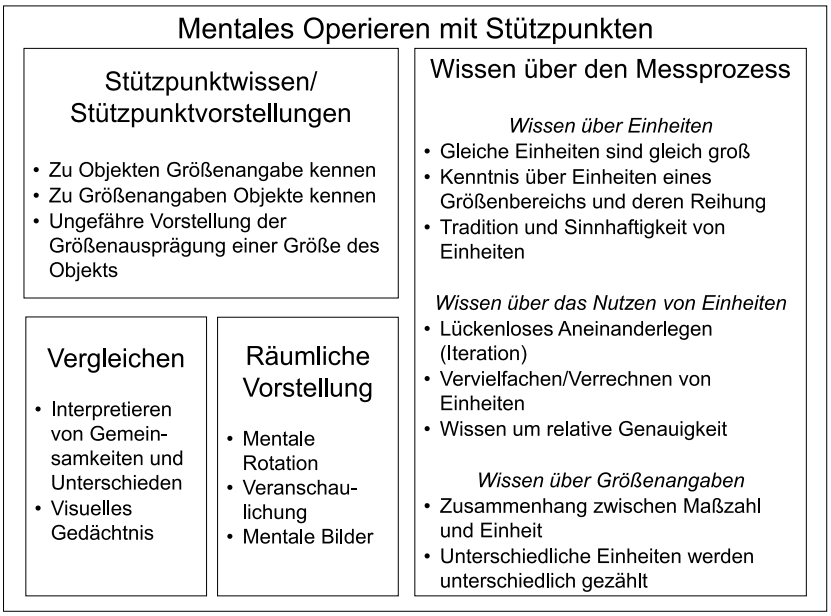


Abb. 9: Mentales Operieren mit Stützpunkten im Schätzprozess

5.1.1 Stützpunktwissen und Stützpunktvorstellungen

Stützpunkte sind die Objekte, mit denen das zu schätzende Objekt verglichen werden soll (vgl. Abschn. 2.1). Damit ein Objekt als Stützpunkt verwendet werden kann, ist es erforderlich, die *Größenangabe der entsprechenden Größe zu kennen* und auch mit der entsprechenden Größe am Objekt zu verbinden (vgl. Abb. 10). Beides setzt voraus, dass die schätzende Person weiß, dass Objekte Träger mehrerer Größen (und anderer Eigenschaften) sind und zusätzlich in der Lage ist, die erforderliche Größe (als Eigenschaft) von den anderen Eigenschaften zu abstrahieren. So muss der Fokus auf der zu schätzenden Größe liegen und alle anderen, für die Schätzung unwichtigen Eigenschaften müssen ausgeblendet werden (Nunes, 1992, S. 558). Stützpunktwissen sollte in zwei Richtungen abrufbar sein: Zum einen sollten zu *Repräsentanten die entsprechenden Größenangaben* genannt werden können, zum anderen sollten *zu Größenangaben passende Repräsentanten* genannt werden können.

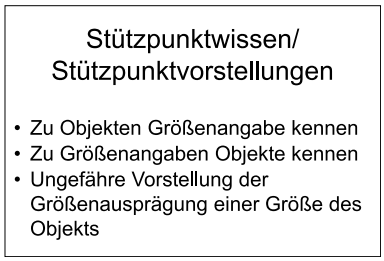


Abb. 10: Stützpunktwissen und Stützpunktvorstellungen

Weil Stützpunkte in der Regel nicht physisch anwesend sind und weil das Operieren mit Stützpunkten nur auf mentaler Ebene erfolgt, sind Stützpunktvorstellungen (neben Stützpunktwissen) von zentraler Relevanz für den Schätzprozess. Diese Vorstellungen werden als

> mentale Vorstellungen […] in einem weiteren Sinne als die Verarbeitung von wahrnehmungsähnlichen Informationen in Abwesenheit externer Quellen perzeptueller Information (Anderson, 2013, S. 75)

definiert. Als Besonderheit gelten die mentalen Bilder, die „als interne Repräsentation von visuellen und räumlichen Informationen" (Anderson, 2013, S. 75) verstanden werden können.

Die schätzende Person sollte demnach mentale Bilder der Stützpunkte aus dem Gedächtnis abrufen können. Dies ist nicht nur für die mentalen Operationen wie mentale Rotation, Zerlegung oder Vervielfältigung relevant, sondern auch für die Auswahl eines passenden Stützpunktes für den Vergleich. Bezogen auf die Größe bedeutet dies, dass nicht nur die quantitative Ausprägung in Form einer Größenangabe bekannt, sondern auch die *ungefähre qualitative Größenausprägung des Stützpunktes* in Form eines mentalen Bildes repräsentiert sein sollte (genauere Ausführungen zum (mentalen) Vergleichen befinden sich in Abschn. 5.1.2).

Auch für das Monitoring sind Stützpunkte hilfreich. So könnte das Ergebnis eines Schätzprozesses auch mit anderen Stützpunkten verglichen werden, um die Plausibilität zu beurteilen. Entweder kann dies ein anderer Stützpunkt sein, der für die ermittelte Größenangabe naheliegend ist (auch deshalb ist die Fähigkeit, Stützpunkte in zwei Richtungen verwenden zu können, von Bedeutung) und mit dem zu schätzenden Objekt nach Anwendung der Strategie Zerlegen bzw. Vervielfachen eines Stützpunktes mittels eines einfachen Vergleichs in Beziehung gesetzt wird, oder es handelt sich um das qualitative Vergleichen der Größen der Objekte.

Auch Hope (1989, S. 15) verweist auf die Bedeutung von Stützpunkten sowohl für das Ermitteln eines Schätzwertes als auch für das Überprüfen einer Antwort:

> A knowledge of a wide variety of everyday measurement referents […] is the foundation of good measurement sense […]. If children cannot refer to some meaningful equivalence, they will find it difficult to produce estimates as well as to judge the reasonableness of a quantitative statement.

In einigen Fällen wird auch der Vergleich mit einer Einheit als Schätzstrategie beschrieben (vgl. Abschn. 2.2.2). In diesem Fall kann die entsprechende Einheit, die zum Vergleich herangezogen wird, als Stützpunkt verstanden werden. Dies ist jedoch eher mit einem „Gefühl" für Größenausprägungen zu erklären (Hope 1989, S. 15; Joram 2003, S. 57; Kuwahara Lang, 2001, S. 462), da, wie in Abschn. 2.1 dargestellt wird, Größen ohne Träger nicht vorstellbar sind und somit auch in Gedanken nicht mit Einheiten allein operiert werden kann.

5.1.2 Vergleichen

Wesentlich für das Anwenden der Schätzstrategien, die auf dem In-Beziehung-Setzen von Stützpunkten und zu schätzendem Objekt beruhen, ist das Vergleichen (vgl. Abb. 11). Auch der Messprozess, der dem Zerlegen oder Vervielfachen eines Stützpunktes zugrunde liegt, ist ein Vergleich eines Stützpunktes (als Einheit) mit dem Objekt, dessen Größe ermittelt werden soll. Der Vergleichsprozess erfordert die *Wahrnehmung von Gemeinsamkeiten und Unterschieden und deren Interpretation.*

In der Regel ist bei kognitiven Schätzaufgaben nur eines oder sogar keines der Objekte physisch anwesend. Für den direkten Vergleich zweier Objekte, von denen eines nur in der Vorstellung repräsentiert

ist, ist das *visuelle Gedächtnis* von besonderer Relevanz.

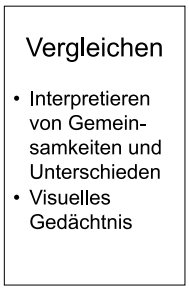


Abb. 11: Vergleichen

Diese Fähigkeit ermöglicht es, Merkmale eines nicht mehr präsenten Objekts auf ein präsentes Objekt zu beziehen (Franke & Reinhold, 2016, S. 59). Dies gilt auch, wenn beide Objekte (Stützpunkt und zu schätzendes Objekt) nur in der Vorstellung repräsentiert sind, weil

> der mentale Prozeß, der der Imagination zugrunde liegt, […] dem Prozeß der Wahrnehmung von Objekten oder Bildern ähnlich (Brander, Kompa & Peltzer, 1989, S. 78)

ist.

Neben der Wahrnehmung von Gemeinsamkeiten oder Unterschieden ist es von Relevanz, die wahrgenommenen Gemeinsamkeiten und Unterschiede des Objekts und des Stützpunktes bezogen auf die entsprechende Größe zu interpretieren. Dies wiederum setzt das Wissen über die Stützpunkte und allgemeine geometrische Kenntnisse (vgl. Abschn. 5.3.1) voraus.

Der Vergleich von Größenausprägungen, die sich quantitativ wenig unterscheiden, fällt in der Regel schwerer als der Vergleich von Größenausprägungen, die sich quantitativ deutlicher unterscheiden. Dennoch werden sich Objekte, die eine ähnliche Größenausprägung aufweisen, häufiger bildlich vorgestellt als Objekte, die einen hohen Größenunterschied aufweisen (Anderson, 2013, S. 80).

Es wird deutlich, dass der visuell-räumliche Notizblock des Arbeitsgedächtnisses an der Lösung von Schätzaufgaben beteiligt ist, da visuelles Informationsmaterial zum Abruf bereitgestellt wird (Cassels, 1995, S. 171). Um diese Informationen abrufen zu können, sind die Komponenten der visuellen Wahrnehmung (vgl. Abschn. 5.3.3) grundlegend.

5.1.3 Räumliche Vorstellung

Für das mentale Operieren mit Stützpunktvorstellungen ebenfalls von Relevanz sind Fähigkeiten der Raumvorstellung (vgl. Abb. 12). Diese umfassen die Fähigkeit, sich bewegte und unbewegte Objekte vorzustellen sowie die Existenz mentaler Bilder. Das gedankliche Bewegen von Objekten bezieht sich insbesondere auf die Fähigkeit, Objekte mental zu rotieren (Horazek et al., 2010, S. 189 ff., vgl. Abschn. 3.2). Maier (1999, S. 14) differenziert aus Sicht der Mathematikdidaktik insgesamt sechs Aspekte der räumlichen Vorstellung. Für das Schätzen von Größen bedeutsam ist, analog zu den Forschungsergebnissen aus der Neuro- und Kognitionspsychologie, die Fähigkeit zur *mentalen Rotation*. Sie „charakterisiert die Fähigkeit, sich Rotationen von zwei- oder dreidimensionalen Objekten vorzustellen“ (Maier, 1999, S. 12).

Abb. 12: Räumliche Vorstellung

Auch eine weitere Komponente der Raumvorstellung beinhaltet das gedankliche Operieren mit Objekten und wird ähnlich eingeordnet wie die mentale Rotation (Maier, 1999, S. 14). Dies ist die Fähigkeit zur *Veranschaulichung*, welche „die gedankliche Vorstellung von Aktivitäten wie Verschiebungen, Faltungen oder Schnitten von räumlichen Objekten oder Objektteilen“ (Maier, 1999, S. 11) umfasst.

Diese Fähigkeiten stehen in einem engen Zusammenhang mit der Fähigkeit zu vergleichen. So kann eine mentale Rotation nützlich sein, um den Vergleich zu erleichtern, etwa wenn die Ausrichtung des Merkmals der Stützpunktvorstellung nicht der Ausrichtung des Merkmals des zu schätzenden Objekts entspricht. Genauso können andere gedankliche Transformationen notwendig sein, etwa wenn zwei Gefäße, deren Fassungsvermögen miteinander verglichen werden soll, eine unterschiedliche Form besitzen (Franke & Ruwisch, 2010, S. 187). Auch für die Schätzstrategie Zerlegen/Zusammensetzen sind räumliche Fähigkeiten von besonderer Relevanz: So muss das zu schätzende Objekt gedanklich zerlegt und wieder zusammengesetzt werden, um die Ermittlung einer Größenangabe durch weitere Strategien erst zu ermöglichen.

Räumliche Fähigkeiten sind für alle drei Größenarten von Relevanz. Dies zeigt sich insbesondere im Umgang mit Stützpunkten: Auch Stützpunkte für Längen und Flächeninhalte haben drei Dimensionen, die in der Vorstellung interpretiert und verarbeitet werden müssen. Darüber hinaus sind die Teilkomponenten der Raumvorstellung nach Maier sowohl für zwei- als auch für dreidimensionale Objekte formuliert. Dies geht zum einen direkt aus der

Definition jedes Teilbereiches hervor, zum anderen wird mit Aufgaben zur Raumvorstellung häufig das Erkennen eines Zusammenhangs zwischen zwei- und dreidimensionalen Objekten getestet (Maier, 1999, S. 11 f.).

Für das Schätzen mittels des mentalen Vergleichs mit Stützpunkten ist es erforderlich, überhaupt ein *mentales Bild* der entsprechenden Objekte abrufen zu können, weshalb das Wissen über Stützpunkte und das Operieren mit ihnen in einem engen Zusammenhang stehen.

5.1.4 Wissen über den Messprozess

Ohne Kenntnis über den Messprozess ist es nicht möglich, mit Stützpunkten als nichtstandardisierte Einheiten einen mentalen Messprozess durchzuführen (Sowder, 1992, S. 383; Joram, Subrahmanyam & Gelman, 1998, S. 433). Insbesondere die Strategie Zerlegen/Vervielfachen eines Stützpunktes greift auf einen mentalen Messprozess zurück (vgl. Abschn. 5.3.1). An dieser Stelle überschneidet sich das Wissen über den Messprozess mit der Fähigkeit zu vergleichen, da sowohl im Mess- als auch im Schätzprozess die Grundidee des Vergleichens mit Einheiten verankert ist (Bright, 1976, S. 88; Frenzel & Grund, 1991a, S. 12).[6]

Allgemein liegen dem Messprozess drei Tätigkeiten zugrunde: Festlegen einer Einheit, die unabhängig von Zeit und Raum ist, das wiederholte Benutzen dieser Einheit, wenn das zu Messende größer als die Einheit ist, sowie das systematische Untergliedern der Einheit, wenn das zu Messende kleiner ist als die Einheit (Peter-Koop, 2008, S. 21). Darüber hinaus muss für das Angeben des Messergebnisses der Zusammenhang zwischen Maßzahl und Einheit bekannt sein. Es lassen sich daher drei Aspekte ausmachen, die für das Verständnis des Messprozesses erforderlich sind: das *Wissen über Einheiten*, das *Wissen über das Nutzen von Einheiten* sowie das *Wissen über Größenangaben* (vgl. Abb. 13).

Um Einheiten nutzen zu können, ist zunächst ein Konzept von Einheiten als solche erforderlich. Dies bedeutet zunächst zu wissen, dass *gleiche Einheiten gleich groß* sind. Für den Umgang mit Einheiten, die größer sind als das zu schätzende Objekt, ist es notwendig zu wissen, dass Einheiten ebenfalls Träger einer Größe sind, die wiederum aus Einheiten aufgebaut sind. Dies wird bei der Strategie Zerlegen eines Stützpunktes deutlich: Der Stützpunkt, verstanden als Einheit, muss zerlegt werden (dies setzt die Kenntnis darüber voraus, dass auch Einheiten immer weiter untergliedert werden können) und diese, nun kleinere Einheit, wird zum Vergleich mit dem zu schätzenden Objekt herangezogen.

Zum Wissen über Einheiten gehören darüber hinaus die *Kenntnis der standardisierten Einheiten der entsprechenden Größenart* einerseits und *deren Reihung* andererseits. Dies ist Voraussetzung dafür, die Ausprägung der entsprechenden Größe in der richtigen, zur Größenart passenden Einheit angeben und, falls erforderlich, damit rechnen zu können (siehe Wissen über das Nutzen von Einheiten).

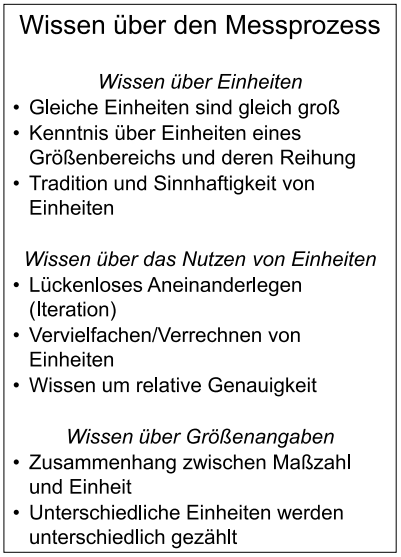


Abb. 13: Wissen über den Messprozess

Wenn Kinder das Wissen über die Einheiten der Größenarten Flächeninhalte und Volumina im Sinne von abgeleiteten Einheiten besitzen, erleichtert es ihnen das Anwenden von Strategien wie Länge mal Breite bzw. Länge mal Breite mal Höhe. So kann die Einheit für den Flächeninhalt aus Einheiten von Längen gebildet werden, indem diese quadriert werden. Bei Einheiten für die Größenart Volumina ist zwischen der abgeleiteten Einheit Kubikmeter und den Einheiten für Fassungsvermögen Liter und Milliliter zu unterscheiden und eventuell deren Zusammenhang zu nutzen.

Um die Entscheidung treffen zu können, in welcher Einheit eine Größenangabe erfolgen soll, ist ebenfalls die *Kenntnis der üblichen Konventionen und Sinnhaftigkeit von Einheiten* notwendig.

Das Wissen über Einheiten ist die Voraussetzung dafür, dass ein Stützpunkt als Einheit genutzt werden und im Sinne des Messprozesses wiederholt mental am zu schätzenden Objekt abgetragen werden kann. Dieser Prozess, das Nutzen von Einheiten, erfolgt beim Schätzen mental, berücksichtigt aber dieselben Grundideen wie ein konkret durchgeführter Messprozess. So muss dieses Abtragen lückenlos und ohne Überlappung geschehen. Dies wird für Längen als *Iterating* bezeichnet (Stephan & Clements, 2003, S. 4). Für Flächeninhalte gilt das Prinzip *structuring an array*, welches besagt, dass beim

Auslegen einer Fläche mit kleineren Flächen ein bestimmtes Gitter berücksichtigt werden muss (Stephan & Clements, 2003, S. 4 f.). Gleiches gilt für das Messen von Volumina (Battista, 2003, S. 123, Battista & Clements, 1996, S. 263; vgl. Abschn. 2). Generell lassen sich die Grundideen des Messens und die Entwicklung des Messverständnisses der Größenart Längen auch auf die Größenarten Flächeninhalte und Volumina übertragen (Clarke, Cheeseman, McDonough & Clarke, 2003, S. 75; Lehrer, Jaslow & Curtis, 2003, S. 109 und S. 118).

Beim mentalen Vergleich mit Einheiten oder beim Schluss von Stützpunkten auf standardisierte Einheiten muss eine zur Größe passende Einheit ausgewählt werden. Dazu kann es erforderlich sein, zwischen zwei Einheiten einer Größenart zu wechseln: Wenn der Stützpunkt 30 cm lang ist und achtmal gezählt wurde, ist es wahrscheinlich, dass das Schätzergebnis als 2,40 m oder 2 m und 40 cm angegeben wird (und nicht als 240 cm), weil es im kulturellen Kontext üblicher ist. Neben arithmetischen Grundkenntnissen ist hier die Fähigkeit erforderlich, *mit Einheiten zu rechnen*, was die Kenntnis der Reihenfolge und die entsprechenden zahlenmäßigen Beziehungen zwischen den Einheiten eines Größenbereichs (Umrechnungszahlen) voraussetzt.

Das Rechnen mit Einheiten wird bei der Strategie Länge mal Breite oder Länge mal Breite mal Höhe direkt gefordert. Indirekt ist die relationale Betrachtung der Einheiten zueinander auch bei Strategien der visuellen Auffassung mit mehreren Dimensionen von Bedeutung: So muss bei der Beurteilung des Fassungsvermögens eines Zylinders der Durchmesser des Gefäßes mit der Höhe ins Verhältnis gesetzt werden, um zu einem begründeten Schätzergebnis zu gelangen. Bezogen auf den Messprozess für Flächeninhalte bedeutet dies, dass Flächeninhalte von rechteckigen Figuren nicht nur durch iteratives Auslegen, sondern auch durch Messen der Seitenlängen und das Rechnen mit den entsprechenden Größenangaben ermittelt werden können. Für das Messen von Volumina wird diese Rechnung (bei quaderförmigen Objekten) durch eine dritte Dimension ergänzt. Beim Messen von Fassungsvermögen wird der indirekte Charakter des Formelwissens deutlich, wenn die Größenangabe durch Höhe des Füllstandes an einer Skala abgelesen werden kann. Der mögliche Füllstand variiert durch das Verändern des Durchmessers des Gefäßes.

Sowohl beim Schätzen als auch beim Messen ist es erforderlich, über die *relative Genauigkeit* des jeweiligen Prozesses informiert zu sein und dies auf das eigene Nutzen der Einheiten zu beziehen, um zu entscheiden, wann das Nutzen der Einheit im Sinne eines Messprozesses zu einem zufriedenstellenden Ergebnis geführt hat. Damit das Nutzen der Einheiten am Ende zu einer Größenangabe führt, müssen die genutzten Einheiten gezählt werden, um die Größe als Vielfaches der genutzten Einheit angeben zu können. Dies erfordert ein Verständnis des *Zusammenhangs zwischen Maßzahl und Maßeinheit*. Darüber hinaus muss bei Verwendung unterschiedlicher Einheiten auch das Verständnis vorhanden sein, dass bei der Verwendung mehrerer, verschieden großer Einheiten der Zählprozess nicht einfach fortgeführt werden darf (Stephan & Clements, 2003, S. 4 f.). Dies baut auf das Verständnis des Konzepts von Einheiten auf und besagt, dass *unterschiedliche Einheiten unterschiedlich gezählt* werden.

5.2 Exekutive Funktionen

Während des Schätzprozesses sind die *exekutiven Funktionen* von besonderer Relevanz (vgl. Abb. 14). Defizite in den exekutiven Funktionen bedeuten meist auch eine defizitäre Schätzleistung, wie eine Vielzahl an Studien aus der Kognitionspsychologie belegt (vgl. Abschn. 3).

Unter exekutiven Funktionen versteht man

> mentale Prozesse höherer Ordnung, die immer dann ins Spiel kommen, wenn wir Handlungen planen oder Absichten/Ziele über mehrere Schritte hinweg verfolgen. (Konrad, 2007, S. 300)

Welche kognitiven Tätigkeiten als exekutive Funktionen verstanden werden, ist hingegen nicht eindeutig festgelegt. Erwähnt werden Aktivitäten wie Planen, Problemlösen, Inhibition und Flexibilität (Kaufmann, Nuerk, Konrad & Willmes, 2007, S. 404). Andere Autorinnen und Autoren nennen Aufmerksamkeit und Inhibition, Aufgaben-Management, Handlungsplanung, Handlungsüberwachung (Ullsperger & von Cramon, 2006, S. 480), Inhibition, kognitive Flexibilität (Kubesch & Walk, 2009, S. 309) sowie Ablauforganisation, Aufmerksamkeit und Inhibition, Planen, Überwachung und Kodierung (Smith & Jonides, 1999, S. 1657). Gemeinsam haben alle Aktivitäten, dass Handlungs- bzw. Denkschritte geplant und kontrolliert sowie Prozesse, die dem Ziel nicht dienlich sind, unterbunden werden.

In Kapitel 3.3 wurde bereits herausgearbeitet, dass der Rückgriff auf exekutive Funktionen an vielen Stellen des Schätzprozesses geschieht. Je nach Zeitpunkt der Anwendung im Schätzprozess lassen sich daher drei Gruppen von exekutiven Funktionen unterscheiden: solche, die *vorbereitend* und zur Strukturierung der Schätzanforderung angewendet werden, solche, die den Einsatz von Strategien *begleiten*, und solche, die für die *Prüfung des Ergebnisses* erforderlich sind (vgl. Abb. 14).

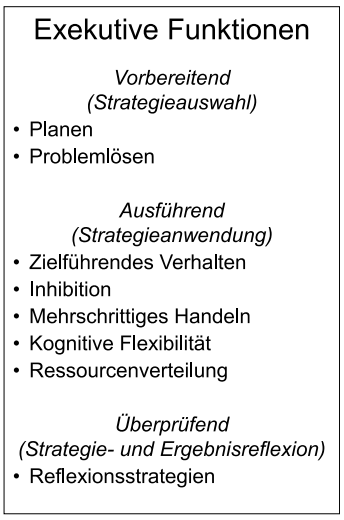


Abb. 14: Exekutive Funktionen im Schätzprozess

Wird die schätzende Person mit der Schätzanforderung konfrontiert, so muss sie die Fragestellung verstehen und eine geeignete Vorgehensweise wählen, mit der sie der gewünschten Lösung näherkommt.

Der Übergang zwischen dem Wählen und dem Ausführen einer Strategie, also vom Verständnis der Frage zum Reasoning im Sinne des kognitiven Schätzens aus psychologischer Sicht, ist nicht eindeutig festgelegt. So wird sowohl während der Strategieauswahl als auch während der Strategieanwendung auf die exekutiven Funktionen *Planen* und *Problemlösen* zurückgegriffen, welche durch die zentrale Exekutive des Arbeitsgedächtnisses ausgeführt werden (vgl. Abschn. 3.1).

Die Fähigkeit zum *mehrschrittigen Handeln* geht mit der Fähigkeit zum Planen einher. Sie zeigt sich besonders bei der Verwendung von Strategien, die das Anwenden einer weiteren Strategie erfordern (vgl. Abschn. 2.2). Hier müssen mehrere Schritte hintereinander und vorausschauend ausgeführt werden. Strategien wie Einschachteln oder Zerlegen und Zusammensetzen werden weniger angewendet als andere Strategien (Heid, 2017, S. 139). Eine mögliche Erklärung wäre die zusätzliche Anforderung, die durch die Notwendigkeit des Anwendens einer weiteren Strategie entsteht. Diese kann durch das Anwenden von Strategien, die die exekutiven Funktionen kognitive Flexibilität und mehrschrittiges Handeln nicht in solch besonderer Weise herausfordern, vermieden werden.

Insbesondere während des Ausführens der Strategien sind *zielführendes Verhalten* und *Inhibition* von Relevanz, um den Schätzprozess weiter- und zu Ende zu führen. Personen mit einem hohen Ablenkungspotential bzw. geringer Inhibitionskontrolle zeigen schlechtere Schätzleistungen (Brand et al., 2002, S. 288). D'Aniello, Scarpina, Albani, Castelnuovo und Mauro (2015, S. 1428) stellten hingegen keine signifikante Korrelation zwischen der Interferenzanfälligkeit und der Schätzleistung fest.

Kognitive Flexibilität ist dann gefragt, wenn die vorliegende Aufgabe nicht bekannt ist und/oder mit Hilfe bekannter Techniken nicht gelöst werden kann. Dazu muss in einer Situation erlerntes Wissen auf eine andere Situation übertragen werden können (Mandl, Kopp & Dvorak, 2004, S. 21). Darüber hinaus ist kognitive Flexibilität dann erforderlich, wenn Aufmerksamkeit zwischen verschiedenen Anforderungen aufgeteilt werden muss (Bellebaum, Thoma & Daum, 2012, S. 71 f.). Beide Aspekte sind für das Bearbeiten einer Schätzaufgabe relevant. Die Strategie zur Ermittlung der Größe ist nicht auf den ersten Blick ersichtlich und die Größenangabe muss aus bereits erlerntem Wissen abgeleitet werden. Im Kontext des Schätzens kann das Stützpunktwissen als bereits erlerntes Wissen, das auf die unbekannte Größe übertragen werden muss, verstanden werden. Darüber hinaus kann kognitive Flexibilität bei Strategien, die eine Umstrukturierung der Schätzanforderung beinhalten, relevant sein.

Zu den exekutiven Funktionen gehört an dieser Stelle auch die *Verteilung von Ressourcen*, die in der Regel ebenfalls durch die zentrale Exekutive des Arbeitsgedächtnisses vorgenommen wird (Goswami, 2001, S. 257, vgl. Abschn. 3.1).

Exekutive Funktionen sind nicht nur während der Ausführung der Schätzung, sondern ebenfalls bei der Überprüfung der ermittelten Größenangabe von Relevanz (vgl. Abschn. 3.3). So sind *Reflexionsstrategien* erforderlich, mit denen entschieden werden kann, ob der Schätzprozess erneut ausgeführt werden muss oder ob das Schätzergebnis als endgültig geäußert werden kann. Diese Strategien sind jedoch bisher noch nicht für das Schätzen beschrieben worden. Denkbar wäre, wie in Kapitel 3.3 beschrieben, ein Abgleich mit anderen Stützpunkten oder das Anwenden einer anderen Strategie.

Darüber hinaus ist auch für das Monitoring die Fähigkeit des Planens und Problemlösens relevant, da unter Umständen der Schätzprozess nicht ganz von Anfang, sondern nur von einer bestimmten Stelle des Plans oder der Strategie neu begonnen werden muss.

5.3 Allgemeines (mathematisches) Wissen und grundlegende Fähigkeiten

Neben den Fähigkeiten, die direkt für den Schätzprozess erforderlich sind, gibt es Fähigkeiten, die das Verwenden der genannten Fähigkeiten erst ermöglichen und so grundlegend sind, dass sie an fast allen Stellen des Schätzprozesses Anwendung finden. Diese sind in Abb. 15 dargestellt und lassen

sich in drei Gruppen differenzieren: *Geometrische Kenntnisse, Arithmetische Kenntnisse* und die Fähigkeit zur *visuellen Wahrnehmung*.

Die geometrischen und arithmetischen Kenntnisse können einerseits als allgemeines Wissen bezeichnet werden, andererseits gibt es aus Sicht der Mathematikdidaktik klare geometrische bzw. arithmetische Bezüge. Daher wird es hier als *allgemeines (mathematisches) Wissen* bezeichnet. Die visuelle Wahrnehmung wird als grundlegende Fähigkeit bezeichnet, da der Schätzprozess von visuell erfassbaren Größen auf der Wahrnehmungsfähigkeit beruht.

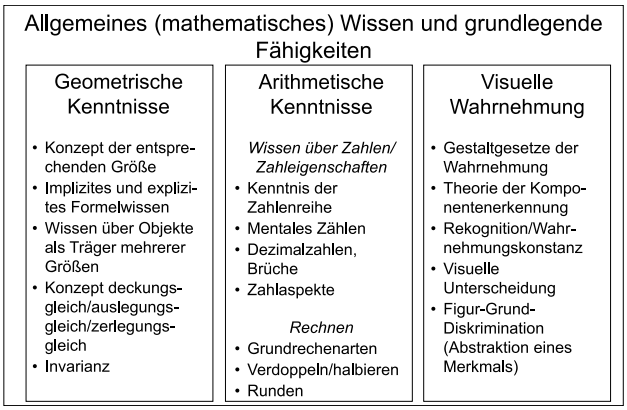


Abb. 15: Allgemeines (mathematisches) Wissen und grundlegende Fähigkeiten im Schätzprozess

5.3.1 Geometrische Kenntnisse

Als grundlegend für den gesamten Schätzprozess von Längen, Flächeninhalten und Volumina sowie als allgemeines (geometrisches) Wissen verstanden werden die in Abb. 16 dargestellten Fähigkeiten.

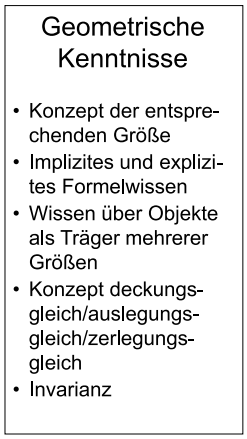


Abb. 16: Geometrische Kenntnisse

Für das Schätzen von Größen ist ein grundlegendes Verständnis darüber, was Größen sind bzw. wie sie beschrieben werden können, erforderlich. Insbesondere für das Verständnis der Schätzanforderung und die Auswahl einer passenden Strategie ist ein *Konzept der entsprechenden Größe* notwendig (Hildreth, 1983, S. 50).[7]

Für die Vorstellung von Repräsentanten als Träger der entsprechenden Größe spielt somit *implizites Formelwissen* im Sinne der Bewusstheit von Ein-, Zwei- oder Dreidimensionalität eine Rolle, während für das Anwenden der Schätzstrategie Länge mal Breite (mal Höhe) auch *explizites Formelwissen* und das Inbeziehungsetzen verschiedener Größen relevant ist. Gleichzeitig ist hier auch das Wissen über *Objekte als Träger mehrerer Größen* in beiden Schritten des Schätzprozesses relevant: Zum einen muss die Größe, deren Ausprägung geschätzt werden soll, von den anderen Größen des Objekts abstrahiert werden. Dies erfordert zusätzlich die Fähigkeit, diese verschiedenen Größen auch unterscheiden zu können und mit den visuell sichtbaren Eigenschaften (Seite, Kante, Fläche) zu verknüpfen. Zum anderen können verschiedene Größen eines Objekts zueinander in Beziehung gesetzt werden, wenn die schätzende Person weiß, dass Objekte Träger mehrerer Größen sind und an welcher Stelle des Objekts die Ausprägung der Größe sichtbar ist.

Für die Auswahl eines passenden Stützpunktes ist es sowohl erforderlich, die entsprechende Größe am Stützpunkt zu abstrahieren, als auch zu wissen, dass der Stützpunkt ebenfalls Träger mehrerer Größen ist. So ist für die Frage nach der Größe einer Länge ein Stützpunkt, dessen Volumen bekannt ist, nicht unbedingt hilfreich (vorausgesetzt, die Kantenlängen können nicht aus dem Volumen errechnet werden). Andersherum kann ein Stützpunkt, dessen Länge bekannt ist, durch explizites Formelwissen und der Anwendung der Strategie Länge mal Breite (mal Höhe) für die Ermittlung eines Flächeninhalts oder eines Volumens hilfreich sein, wenn auch die Länge des zu schätzenden Objekts abstrahiert werden kann und das Objekt die Anwendung der entsprechenden Formel ermöglicht.

Die Grundidee des gedanklichen Ausmessens mit Stützpunkten, für die das Wissen über den Messprozess vorausgesetzt wird, erfordert neben der Auswahl eines geeigneten Stützpunktes durch Abstrahieren der entsprechenden Größe ein *Konzept über Deckungs- bzw. Auslegungsgleichheit.* Dies ist erforderlich, um die Äquivalenz zweier Größen festzustellen und zu beurteilen, ob zwei Größen demnach die gleiche Ausprägung haben oder ob eine kleiner oder größer ist als die andere. Für das Schätzen und den Vergleich bzw. das gedankliche Ausmessen mit Stützpunkten ist dennoch zu erwähnen, dass fehlende Kongruenz nicht bedeuten muss, dass zwei Größen eine unterschiedliche Ausprägung haben. So können zwei unterschiedlich gekrümmte Linien trotzdem gleich lang sein, auch wenn sie sich in ihrer Form unterscheiden (gleiches gilt für unterschiedliche Repräsentanten von Flächeninhalten und Volumina). Insbesondere für das Schätzen von Flä-

cheninhalten und Volumina ist daher das Verständnis von *Zerlegungsgleichheit* nützlich.

Grundlegend für das gedankliche Nutzen von Stützpunkten und für die mentale Rotation von Objekten ist das Verständnis darüber, dass sich die Größe eines Objekts bei Bewegung oder Transformation nicht ändert. Dies wird von Stephan und Clements (2003, S. 5) als *Conservation* bezeichnet, Piaget und Szeminska (1975, S. 22) sprechen von *Invarianz*. Beim Messen ist das Verständnis von Invarianz deshalb von Bedeutung, weil im physischen Messprozess der Träger der verwendeten Einheit eventuell nicht ausreichend oft vorhanden ist, um ihn an der gesamten Größe abzutragen. So muss verstanden werden, dass auch ein Träger der Einheit wiederholt angelegt werden kann und sich dessen Größe dadurch nicht ändert (auf Längen bezogen vgl. Castle & Needham, 2007, S. 219). Dies ist auch für die Strategie Vervielfachen eines Stützpunktes von Bedeutung, da das Anlegen nicht physisch erfolgen kann und die angelegten Objekte nicht tatsächlich gezählt werden können. Für den Messprozess bei Fassungsvermögen ist darüber hinaus von Bedeutung, dass das Zählen der Einheiten gleichzeitig mit dem Umschütten erfolgen muss, da sie im Nachhinein in der Regel nicht mehr ablesbar sind (Franke & Ruwisch, 2010, S. 227). Auch bei Strategien, die das Umstrukturieren der Schätzanforderung erfordern, ist ein Verständnis der Invarianz erforderlich. Dazu gehört das Wissen darüber, dass sich die Größe eines Objekts nicht ändert, wenn es zerlegt und wieder zusammengesetzt wird oder wenn die Ausrichtung verändert wird. Dies gilt auch für das Objekt, welches ausgemessen wird (z. B. beim Anwenden der Strategie Neuanordnung, vgl. Abschn. 2.2.3).

5.3.2 Arithmetische Kenntnisse

Die für den Schätzprozess relevanten arithmetischen Kenntnisse umfassen zum einen das *Wissen über Zahlen und Zahleigenschaften* und zum anderen die *Fähigkeit zum Rechnen* (vgl. Abb. 17).

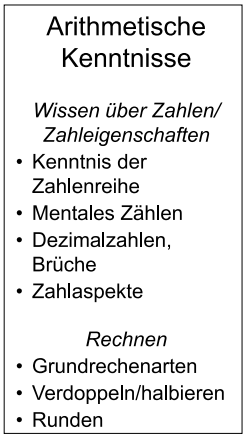


Abb. 17: Arithmetische Kenntnisse

Das Wissen über Zahlen beinhaltet die *Kenntnis der Zahlenreihe* und die Fähigkeit, diese *mental beim Zählen von Objekten zu durchlaufen*. Das Zählen von Objekten ist für das Verständnis des Messprozesses von Relevanz, wenn die Einheiten gezählt werden, um die Maßzahl für die Größenangabe zu ermitteln. Aber auch die Strategie Einschachteln erfordert die Kenntnis der Zahlenreihe: So muss die Größenangabe des zu schätzenden Objektes zwischen den Größenangaben zweier Stützpunkte eingeordnet werden. Hier spielt insbesondere auch die Verknüpfung der Maßzahl mit der entsprechenden Einheit eine Rolle: 30 cm ist eine kleinere Größenausprägung als 1 m, obwohl 30 die größere Zahl ist. Soll die Einheit zwischen diesen beiden Größenangaben Meter sein, so ist zusätzlich *Wissen über Dezimalzahlen* oder *Brüche* erforderlich.

Das Wissen über Zahlen beinhaltet auch das Wissen darüber, dass dieselbe Zahl in unterschiedlichen Zusammenhängen unterschiedliche Bedeutungen haben kann. Das Wissen über *Zahlaspekte* beinhaltet das Wissen darüber, dass die Zahl vor einer Einheit als Maßzahl zu bezeichnen ist und die Anzahl der entsprechenden Einheiten angibt (Maßzahlaspekt). Beim Rechnen mit Einheiten ist darüber hinaus zwischen der Maßzahl und der Rechenzahl zu unterscheiden, wenngleich dies nicht explizit benannt werden muss.

Das Rechnen im Schätzprozess umfasst zunächst die *Grundrechenarten*. Alle Strategien, die den Vergleich mit einem Stützpunkt beinhalten, die aber eine standardisierte Einheit hervorbringen sollen, erfordern nach dem Vergleich das Umrechnen der Größenangabe des Stützpunktes in die gewünschte Größenangabe mit standardisierter Einheit. Dies geschieht nicht unbedingt streng nacheinander: So wird bei der Strategie Vervielfachen eines Stützpunktes vermutlich nicht die Größenangabe mit dem Stützpunkt als Einheit geäußert, sondern die Umrechnung geschieht parallel mit dem Zählprozess der Stützpunkte. Hierbei kann auch das Zählen in Schritten (in der Größenangabe des Stützpunktes) hilfreich sein. Dennoch muss ein Verständnis von Multiplikation und Addition bei der Ermittlung der Größe durch die Strategie Vervielfachen eines Stützpunktes vorhanden sein, wie auch Siegel et al. (1982, S. 227) beschreiben. Für die Strategie Zerlegen eines Stützpunktes ist zusätzlich das Verständnis der Division erforderlich. So ist es unwahrscheinlich, dass beim Halbieren des Stützpunktes mit 0,5 multipliziert wird, sondern es ist eher praktikabel, durch zwei zu teilen. Die Subtraktion ist beim Zerlegen eines Stützpunktes nicht so intuitiv vorzunehmen wie die Addition beim Vervielfachen eines Stützpunktes: So muss erst die Größe des Teils vom Stützpunkt, der nicht verwendet wird, geschätzt

werden, um sie dann von der Größe des gesamten Stützpunktes zu subtrahieren.

Auch beim Rechnen mit Größenangaben sind die Grundrechenarten relevant. So kann es erforderlich sein, innerhalb eines Größenbereiches in eine andere Einheit umzurechnen oder zwei Größenangaben aus unterschiedlichen Größenbereichen miteinander zu verrechnen (hierzu auch Abschn. 5.1.4).

Das *Verdoppeln und Halbieren* ist neben den geometrischen auch für die arithmetischen Fähigkeiten von Relevanz. Sollte ein Stützpunkt gedanklich verdoppelt oder halbiert werden, um seine Größe an die des zu schätzenden Objekts anzupassen, muss dies auch mit der Größenangabe geschehen. Das Verdoppeln und Halbieren ist nicht unbedingt rechnerisch erforderlich, so kann bekannt sein, dass 100 das Doppelte von 50 ist (mit entsprechend passender Größeneinheit), ohne das bewusst eine Multiplikation mit zwei erforderlich ist. Dies ist jedoch nur ein Spezialfall im Umgang mit Zahlen, der aufgrund der in der Grundschule unterrichteten Rechenstrategien eine Relevanz haben könnte und nicht direkt mit dem Schätzen in Verbindung stehen muss.

Auch das *Runden* spielt beim Schätzen eine Rolle: So ist es nicht zweckdienlich, eine geschätzte Größe z. B. mit 4,36 m anzugeben. Hier ist abzuwägen, wie exakt die Schätzung sein muss, um die Maßzahl entsprechend anzupassen (Sowder, 1992, S. 373).[8]

5.3.3 Visuelle Wahrnehmung

Die Besonderheit und Gemeinsamkeit der Größenarten Längen, Flächeninhalte und Volumina liegt in deren visueller Wahrnehmbarkeit. Aus diesem Grund ist die *Fähigkeit zur visuellen Wahrnehmung* fundamental bedeutend für das Schätzen dieser drei Größenarten und findet Anwendung nicht nur beim Verständnis der Schätzanforderung, sondern auch bei der Strategieanwendung (vgl. Abb. 18).

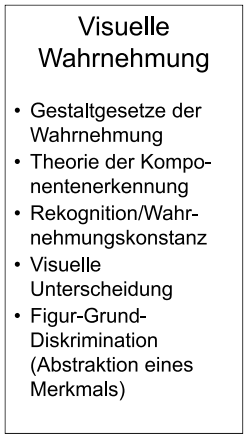


Abb. 18: Visuelle Wahrnehmung

Die Wahrnehmung mentaler Bilder erfolgt in ähnlicher Weise wie die Wahrnehmung realer Objekte (Anderson, 2013, S. 82). Aus diesem Grund ist visuelle Wahrnehmung zunächst die Grundlage dafür, an den (vorgestellten oder physisch anwesenden) zu schätzenden Objekten die Größe auszumachen, deren quantitative Ausprägung ermittelt werden soll, und trägt somit zum Verständnis der Schätzanforderung bei. Bei Anwendung der Schätzstrategien ist visuelle Wahrnehmung erforderlich, um Eigenschaften an Stützpunkten zu erkennen. Für das Überprüfen des Schätzwertes kann visuelle Wahrnehmung herangezogen werden, wenn Eigenschaften des zu schätzenden Objekts mit der ermittelten Größenangabe gedanklich in Beziehung gesetzt werden.

Visuelle Wahrnehmung beruht auf der Verarbeitung visuell dargebotener Informationen. Um Objekte wahrzunehmen, reicht es nicht aus, Linien im Raum wahrzunehmen, sondern sie müssen darüber hinaus gruppiert werden (Anderson, 2013, S. 30 ff.). Bei dieser Ordnung sind die *Gestaltgesetze der Wahrnehmungsorganisation* relevant. Diese umfassen das Prinzip der Nähe, nach dem nahe beieinanderliegende Elemente als Einheit wahrgenommen werden, das Prinzip der Ähnlichkeit, nach dem ähnliche Objekte gruppiert werden, das Prinzip des glatten Verlaufs, nach dem das zusammengehört, was augenscheinlich zusammenpasst, und das Prinzip der Geschlossenheit, nach dem Objekte, die nicht vollständig sichtbar sind, passend zu den sichtbaren Strukturen mental ergänzt werden (Anderson, 2013, S. 33). Um nun ein Objekt nicht nur wahrzunehmen, sondern als Objekt zu erkennen, ist die *Theorie der Komponentenerkennung* von Bedeutung. Hiernach findet das Erkennen von Objekten in drei Stufen statt: auf der ersten Stufe wird das Objekt in Teilobjekte untergliedert. Auf der zweiten Stufe erfolgt die Klassifizierung dieser Teilobjekte (für eine ausführlichere Betrachtung siehe Anderson, 2013, S. 37 ff.). Auf der dritten Stufe werden die erkannten Teilobjekte wieder zu einem Objekt zusammengefügt, welches dann erkennbar ist (Anderson, 2013, S. 37).

Sollte sich bei der gedanklichen Zerlegung eines Objekts, zum Beispiel bei der Strategie Zerlegen/Zusammensetzen oder beim Zerlegen eines Stützpunktes, an der Struktur orientiert worden sein, so kann durch das Prinzip der Nähe oder das Prinzip der Ähnlichkeit gedanklich eine Strukturierung der Teilobjekte des Objekts stattfinden oder wahrgenommen werden. Um die Zerlegung des zu schätzenden Objekts am Vorwissen zu Teilen des Objekts festzumachen, sind genau diese Fähigkeiten ebenfalls relevant. Darüber hinaus ist aber ein Wiedererkennen der bekannten Segmente des Objekts Voraussetzung, da diese in der Regel nicht einzeln vorliegen, sondern als Teilobjekte des zu schätzenden Objekts. Das Wiedererkennen von Informationen, auch als *Rekognition* bezeichnet, muss nicht

bewusst erfolgen (Cassels, 1995, S. 163). So kann eine regelmäßige oder eine unregelmäßige Unterteilung des zu schätzenden Objekts unter Zuhilfenahme der Strukturierung des Objekts oder des Wissens über einzelne Teile des Objekts geschehen. Das Wiedererkennen von Objekten – auch in unterschiedlicher Lage oder aus unterschiedlichen Perspektiven – wird in Anlehnung an Frostig als *Wahrnehmungskonstanz* bezeichnet (Franke & Reinhold, 2016, S. 57). Weil aber nicht nur das Wahrnehmen von Gemeinsamkeiten, sondern auch von Unterschieden eine Rolle spielt, ist in Anlehnung an Hoffer auch die *visuelle Unterscheidung* von Relevanz (Franke & Reinhold, 2016, S. 59).

Um überhaupt einen Vergleich zweier Objekte vornehmen zu können, ist auch die Fähigkeit zur *Figur-Grund-Diskrimination* notwendig, um ein Objekt in der Umwelt wahrzunehmen oder um einzelne Segmente einer Figur oder eines Körpers zu erkennen (Franke & Reinhold, 2016, S. 55). Eine bestimmte Größe unabhängig von anderen Eigenschaften zu betrachten, kann darüber hinaus als *Fähigkeit zur Abstraktion* verstanden werden.

6. Resümee

Aus schulischer Sicht ist die Fähigkeit zu schätzen aus mindestens zwei Gründen von Bedeutung: Zum einen kann das Schätzen als Grundlage für ein tiefgehendes Verständnis des Messprozesses gesehen werden (Bright, 1976, S. 87; Joram et al., 1998, S. 414; Joram et al., 2005, S. 4). Zum anderen bedarf das Schätzen aufgrund seiner alltäglichen Bedeutung (Joram et al., 2005, S. 4) nicht unbedingt einer weiteren Legitimation, um Unterrichtsthema zu sein.

Im Gegensatz zum Schätzen ist das Messen vielfach untersucht und durch Einzelfähigkeiten dargestellt worden (z. B. Stephan & Clements, 2003). Aufgrund der hohen Bedeutung des Schätzens für das Verständnis des Messprozesses, für das Entwickeln von Stützpunktwissen und Stützpunktvorstellungen sowie für den Alltag ist es daher unerlässlich, auch für das Schätzen die Einzelfähigkeiten zu untersuchen.

Die Modelle zum kognitiven Schätzen und die darauf aufgebauten Studien sind für das mathematikdidaktische Verständnis des Schätzprozesses nicht zufriedenstellend. So liegt hier der Fokus auf den beteiligten kognitiven Komponenten und weniger auf Einzelkompetenzen und deren Verknüpfung. Die entwickelten Tests enthalten eine wenig fundierte Auswahl an Schätzfragen, die neben der gleichzeitigen Abfrage von quantitativen und qualitativen Schätzanforderungen sowohl das Schätzen von Größen (aller Art) und Anzahlen enthalten. Die Konstrukte des Schätzens, die hier zugrunde liegen, können aus diesem Grunde einer mathematikdidaktischen Betrachtung nicht genügen.

Die aus mathematikdidaktischer Sicht für das Schätzen relevanten Aspekte beziehen sich insbesondere auf die Strategieanwendung. Dies wird als Schätzen im engeren Sinne bezeichnet, um es vom Schätzen aus kognitionspsychologischer Sicht abzugrenzen.[9] Die dafür erforderlichen Fähigkeiten sind hauptsächlich, aber nicht nur, mathematisches Wissen und mathematische Fähigkeiten. Das entwickelte Modell zeigt, welche grundlegenden Fähigkeiten vorhanden sein müssen, um überhaupt die Strategien zum Schätzen von Längen, Flächeninhalten oder Volumina ausführen zu können.[10]

Insbesondere die mathematischen Fähigkeiten können gut an bestimmten Punkten des Schätzens im engeren Sinne (der Strategieanwendung) festgemacht werden, während die Fähigkeiten, die aus psychologischer Sicht beschrieben werden, häufig den gesamten Schätzprozess betreffen und hauptsächlich für die Koordination der Strategieanwendung erforderlich sind. Da es sich bei den Größenarten Längen, Flächeninhalte und Volumina um visuell erfassbare Größen handelt, ist die Fähigkeit zur visuellen Wahrnehmung grundlegend für den Schätzprozess und kann als verbindendes Element zwischen mathematischen und psychologischen Fähigkeiten angesehen werden.

Das Modell geht über eine bloße Auflistung der für das Schätzen von Größen erforderlichen Fähigkeiten hinaus, indem Abhängigkeiten und Hierarchien durch die Anordnung in Kästchen und deren Verschachtelung dargestellt werden. Weitere Verknüpfungen werden nicht grafisch aufgezeigt, da das Modell sonst an Übersichtlichkeit verlöre und die wesentliche Aussage, nämlich die der interdisziplinären Betrachtung von Fähigkeiten, verloren ginge.

Die Erkenntnis aus der Kognitionspsychologie, dass bei der Anwendung von Schätzstrategien das Wissen und die Fähigkeiten einzeln nicht zielführend sind, sondern erst in ihrer Kombination ein begründbares Schätzergebnis ermöglichen (vgl. Abschn. 3.1), führt zur Betrachtung von Prozessmodellen. Mit der Interpretation aus mathematikdidaktischer Sicht und, damit verbunden, mit dem Fokus auf der Strategieanwendung und den dafür erforderlichen Fähigkeiten liegt der Schwerpunkt des hier vorgestellten Modells aspektorientiert auf den Inhalten eines Schätzprozesses und weniger auf der zeitlichen Abfolge der Schritte des Schätzens von Größen. Denkbar wäre daher in weiteren Modellen eine Ergänzung der Fähigkeiten, die für das Verständnis der Schätzanforderung und der Formulierung eines Outputs erforderlich sind (Schätzen im weiteren

Sinne), bevor ein vollständiges Prozessmodell formuliert werden könnte.

Der nächste Schritt für die empirische Untersuchung des Modells ist dessen Operationalisierung. Hierfür kann die Unterscheidung verschiedener Schätzaufgaben nach Bright (1976, S. 90) herangezogen werden, um eine möglichst breite Aufgabenvielfalt zu gestalten. Es ist aber auch zu berücksichtigen, dass alle genannten Fähigkeiten und deren Klassifizierung Anwendung in den Schätzaufgaben finden.

Anmerkungen

[1] Das Wort „Größenart" wird verwendet, wenn die physikalischen Eigenschaften bei der Nennung im Vordergrund stehen (z. B. bei der Betonung, ob es sich um Längen, Flächeninhalte oder Volumina handelt). „Größenbereich" wird bei Nennung aus mathematischer Fundierung genutzt (z. B. bei Betonung der mathematischen Strukturen).

[2] Es ist nicht abschließend geklärt, wie bei Menschen, die keinen Stützpunkt äußern, die Größe im Gedächtnis repräsentiert ist. In Anlehnung an *Number Sense* wurde der Begriff *Measurement Sense* eingeführt. Measurement Sense bedeutet, ein Gefühl für die Größe von Einheiten oder Objekten zu besitzen (Hope 1989, S. 15; Joram, 2003, S. 57; Kuwahara Lang, 2001, S. 462), so dass der Begriff „Gefühl" für hier gewählt wurde.

[3] Denkbar wäre auch eine Veränderung der äußeren Form des Stützpunktes, obwohl dann nicht mehr die Schätzanforderung, sondern die Mittel zur Bearbeitung gedanklich umstrukturiert werden würden.

[4] Die Nennung eines Vorgehens, das auf rein visueller Wahrnehmung beruht, könnte dem *Measurement Sense* zugeordnet werden (vgl. Endnote 2).

[5] Hier ist vermutlich semantisches Wissen in Form von Begriffswissen und semantischen Netzen von Bedeutung. Darüber hinaus kommt dem Objektwissen eine erhöhte Bedeutung zu, da nicht nur Wissen über ein zu schätzendes Objekt vorhanden sein muss, sondern ebenfalls über alle Objekte, die der Anforderung nicht entsprechen (z. B. bei der Frage: Welches ist der größte Gegenstand in einem Haus?).

[6] Trotzdem sollte der Messprozess nicht nur als ein Vergleich aufgefasst werden. Dies ist damit zu begründen, dass zwischen der zu messenden Größe und der Größe, mit der gemessen wird, unterschieden werden muss (Frenzel & Grund, 1991a, S. 12).

[7] Hier ist insbesondere das Konzept von Volumina von Relevanz, da verschiedene Vorstellungen verbunden mit verschiedenen Begriffen zugrunde liegen können. So kann das Volumen als Fassungsvermögen aufgefasst werden, zu dem in der Grundschule die Einheiten Liter und Milliliter gehören. Die zweite Vorstellung umfasst die „Vorstellung von Körpern, welche mit Einheitswürfelchen ausgemessen oder aber konstruiert werden" (Ruwisch, 2012, S. 40). Hier ist insbesondere die Formel Länge mal Breite mal Höhe von Relevanz. Die dritte Vorstellung ist die der Verdrängung. Hier muss erkannt werden, dass ein Körper, der in einem mit Flüssigkeit gefüllten Gefäß mehr Flüssigkeit verdrängt als ein anderer Körper, ein größeres Volumen hat als dieser Körper (Ruwisch, 2012, S. 43).

[8] Forrester, Latham und Shire (1990, S. 284) unterscheiden drei Ebenen, in denen Schätzungen vorgenommen werden können. Diese beziehen sich auf den Kontext der Schätzaufgabe. Die soziale bzw. diskursive Ebene beruht auf der Annahme, dass das kindliche Verständnis des Diskurses im Unterricht durch das Wiedererkennen bestimmter Themen gekennzeichnet ist. Das bedeutet, dass es notwendig ist, eine Schätzsituation zu erkennen und mit der Art und Weise, wie eine Schätzung durchgeführt wird, zu verbinden. Eine Schätzaufgabe kann etwa als Trick-Aufgabe gestellt werden oder es kann gelernt worden sein, dass Schätzen bedeutet, ein exaktes Ergebnis so zu verändern, dass es ungenau ist. Die zweite Ebene ist die Risiko- bzw. Wahrscheinlichkeitsebene. In diesem Kontext kommt es darauf an, die Folgen der Schätzung absehen zu können und danach Risiken abzuwägen. Die dritte Ebene kann als idealisiertes oder neutrales Schätzen bezeichnet werden. Dabei geht es ausschließlich um das Ermitteln eines möglichst genauen Schätzwertes. Sollte auf dieser Ebene eine Risikoabwägung stattfinden, so ist sie hypothetisch.

[9] Für das *kognitive Schätzen* müsste mindestens der Beginn genauer dargestellt werden, der das Verständnis der Frage und die Aufbereitung der Anforderung im Arbeitsgedächtnis enthält. Hier ist der visuell-räumliche Notizblock von besonderer Relevanz. Ebenfalls interessant wäre die Einbindung von weiteren prozessualen Aspekten, etwa wie die Auswahl des ersten Stützpunktes geschieht und wie die Entscheidung für einen endgültigen Output fällt.

[10] Offen bleibt, inwieweit das Modell auf das Schätzen anderer Größen oder das Schätzen von Anzahlen übertragbar ist.

Literatur

Anderson, J. R. (2013). *Kognitive Psychologie*. Berlin: Springer.

Axelrod, B. N. & Millis, S. R. (1994). Preliminary Standardization of the Cognitive Estimation Test. *Psychological Assessment, 1*(3), 269-274.

Baddeley, A. (1992). Working Memory. *Science, 255(5044)*, 556-559.

Baddeley, A. (2000). The episode buffer: a new component of working memory? *Trends in cognitive sciences, 4*(11), 417-423.

Battista, M. T. (2003). Understanding Students' Thinking about Area and Volume Measurement. In D. H. Clements & G. Bright (Hrsg.), *Learning and Teaching Measurement* (S. 122-142). Reston, VA.: National Council of Teachers of Mathematics.

Battista, M. T. & Clements, D. H. (1996). Students' understanding of three-dimensional rectangular arrays of cubes. *Journal for Research in Mathematics Education, 27*(3), 258-292.

Bellebaum, C.; Thoma, P. & Daum, I. (2012). *Neuropsychologie*. Wiesbaden: Springer.

Berti, S. (2010). Arbeitsgedächtnis: Vergangenheit, Gegenwart und Zukunft eines theoretischen Konstrukts. *Psychologische Rundschau, 61*(1), 3-9.

Blankenagel, J. (1983). Schätzen, Überschlagen, Runden. Bestandsaufnahme, Reflexion und Möglichkeiten (2). *Sachunterricht und Mathematik in der Primarstufe, 11*(9), 315-322.

Brand, M.; Fujiwara, E.; Kalbe, E.; Kessler, J. & Markowitsch, H. J. (2002). Kognitives Schätzen bei Alzheimer- und Korsakow-Patienten. *Praxis Klinische Verhaltensmedizin und Rehabilitation,15*(60), 282-291.

Brander, S.; Kompa, A. & Peltzer, U. (1989). *Denken und Problemlösen. Einführung in die kognitive Psychologie.* 2., durchgesehene Auflage. Opladen: Westdeutscher Verlag.

Bright, G. W. (1976). Estimation as Part of Learning to Measure. In National Council of Teachers of Mathematics (Hrsg.), *Yearbook* 38 (S. 87-104). Reston, VA: National Council of Teachers of Mathematics.

Bullard, S. E.; Fein, D.; Gleeson, M. K.; Tischer, N.; Mapou, R. L. & Kaplan, E. (2004). The Biber Cognitive Estimation Test. *Archives of Clinical Neuropsychology, 19*(6), 835-846.

Cassels, A. (1995). *Erinnern und Vergessen*. In P. Banyard; J. Gerstenmaier & P. Holler (Hrsg.), Einführung in die Kognitionspsychologie (S.153-193), München: Reinhardt.

Castle, K. & Needham, J. (2007). First Graders' Understanding of Measurement. *Early Childhood Education Journal, 35*(3), 215-221.

Cattell, R. B. (1963). Theory of Fluid and Crystallized Intelligence: A critical Experiment. *Journal of Educational Psychology, 54*(1), 1-22.

Clarke, D.; Cheeseman, J.; McDonough, A. & Clarke, B. (2003). Assessing and Developing Measurement with Young Children. In D. H. Clements & G. Bright (Hrsg.), *Learning and Teaching Measurement* (S. 68-80). Reston, VA.: National Council of Teachers of Mathematics.

D'Aniello, G. E.; Castelnuovo, G. & Scarpina, F. (2015). Could cognitive estimation ability be a measure of cognitive reserve? *Frontiers in Psychology, 6*, 1-4.

D'Aniello, G. E.; Scarpina, F.; Albani, G.; Castelnuovo, G. & Mauro, A. (2015). Disentangling the relationship between cognitive estimation abilities and executive functions: a study on patients with Parkinson's disease. *Neurological Sciences, 36*(8), 1425-1429.

Forrester, M.A.; Latham, J. & Shire, B. (1990). Exploring Estimation in Young School Children. *Educational Psychology, 10*(4), 283-300.

Franke, M. & Reinhold, S. (2016). *Didaktik der Geometrie in der Grundschule*. 3. Auflage. Heidelberg: Springer Spektrum.

Franke, M. & Ruwisch, S. (2010). *Didaktik des Sachrechnens in der Grundschule*. 2. Auflage. Heidelberg: Springer Spektrum.

Frenzel, L. & Grund, K.-H. (1991a). Umgang mit Größen: Fachlich exakt – trotzdem schülergemäß. *mathematik lehren, 45*, 10-14.

Frenzel, L. & Grund, K.-H. (1991b). Wie „groß" sind Größen? *mathematik lehren, 45*, 15-24 u. 31-34.

Friebel, A. C. (1967). Measurement understandings in modern school mathematics. *The Arithmetic Teacher, 14*(6), 476-480.

Goldstein, F. C.; Green, J.; Presley, R. M.; O'Jile, J.; Freeman, A.; Watts, R. & Green, R. C. (1996). Cognitive Estimation in Patients with Alzheimer's Disease. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 9*(1), 35-42.

Goswami, U. (2001). *So denken Kinder. Einführung in die Psychologie der kognitiven Entwicklung*. Bern: Huber.

Grassmann, M. (1999). Zur Entwicklung von Zahl- und Größenvorstellungen als wichtigem Anliegen des Sachrechnens. *Grundschulunterricht, 46*(4), 31-34.

Griesel, H. (1996). Grundvorstellungen zu Größen. *mathematik lehren, 78*, 15-19.

Grode, H.-P. (2001). Mathematik, Physik. In P. Kiehl & N. Breutmann (Hrsg.), *Einführung in die DIN-Normen* (S. 935-965). Stuttgart: Teubner.

Grund, K.-H. (1992). Größenvorstellungen – eine wesentliche Voraussetzung beim Anwenden von Mathematik. *Grundschule, 24*(12), 42-44.

Hagendorf, H. (2006). Arbeitsgedächtnis. In J. Funke & J. Bengel (Hrsg.), *Handbuch der Allgemeinen Psychologie – Kognition* (S. 340-345). Göttingen: Hogrefe.

Heid, M. (2017). *Das Schätzen von Längen und Fassungsvermögen: Eine Interviewstudie zu Strategien mit Kindern im 4. Schuljahr*. Wiesbaden: Springer Spektrum.

Hildreth, D. J. (1983). The Use of Strategies in Estimating Measurements. *The Arithmetic Teacher, 30*(5), 50-54.

Hogan, T. P.; Brezinski, K. L. (2003). Quantitative Estimation: One, Two, or Three Abilities? *Mathematical Thinking and Learning, 5*(4), 259-280.

Hope, J. (1989). Promoting Number Sense in School. *The Arithmetic Teacher, 36*(6), 12-16.

Horazek, J.; Preiss, M.; Tintera, J.; Laing, H.; Kopecek, M.; Spaniel, F.; Brunovsky, M.; Höschl, C. (2010). A Functional Magnetic Resonance Imaging Study of the Cognitive Estimation. *Activitas Nervosa Superior Redivia*, *52*(3), 187-192.

Huang, H.-M. E. (2015). Children's Performance in Estimating the Measurement of Daily Objects. In K. Beswick, T. Muir & J. Wells (Hrsg.). *Proceedings of 39th Psychology of Mathematics Education conference,* (Bd. 3, S. 73-80). Hobart, Australia: PME.

Joram, E. (2003). Benchmarks as Tools for Developing Measurement Sense. In D. H. Clements & G. Bright (Hrsg.), *Learning and Teaching Measurement* (S. 57-67). Reston, VA.: National Council of Teachers of Mathematics.

Joram, E.; Gabriele, A. J.; Bertheau, M.; Gelman, R. & Subrahmanyam, K. (2005). Children's Use of the Reference Point Strategy for Measurement Estimation. *Journal for Research in Mathematics Education, 36*(1), 4-23.

Joram, E.; Subrahmanyam, K. & Gelman, R. (1998). Measurement Estimation: Learning to Map the Route from Number to Quantity and Back. *Review of Educational Research, 68*(4), 413-449.

Kaufmann, L.; Nuerk, H.-C.; Konrad, K. & Willmes, K. (2007). *Kognitive Entwicklungspsychologie.* Göttingen: Hogrefe.

Konrad, K. (2007). Entwicklung von Exekutivfunktionen und Arbeitsgedächtnisleistungen. In L. Kaufmann; H.-C. Nuerk; K. Konrad & K. Willmes (Hrsg.), *Kognitive Entwicklungspsychologie* (S. 300-320). Göttingen: Hogrefe.

Kosslyn, S. M.; Thompson, W. L. & Ganis, G. (2006). *The Case for Mental Imagery*. New York: Oxford UP.

Kubesch, S. & Walk, L. (2009). Körperliches und kognitives Training exekutiver Funktionen in Kindergarten und Schule. *Sportwissenschaft 39*(4), 309-317.

Kuwahara Lang, F. (2001). What is a "Good Guess," Anyway? Estimation in Early Childhood. *Teaching children mathematics* 7(8), 462-466.

Lehrer, R.; Jaslow, L. & Curtis, C. (2003). Developing an Understanding of Measurement in the Elementary

Grades. In D. H. Clements & G. Bright (Hrsg.), *Learning and Teaching Measurement* (S. 100-121). Reston, VA.: National Council of Teachers of Mathematics.

Liss, M.; Fein, D.; Bullard, S. & Robins, D. (2000). Brief Report: Cognitive Estimation in Individuals with Pervasive Developmental Disorders. *Journal of Autism and Developmental Disorders, 30*(6), 613-618.

MacPherson, S. E.; Wagner, G. P.; Murphy, P.; Bozzali, M.; Cipolotti, L. & Shallice, T. (2014). Bringing the Cognitive Estimation Task into the 21st Century: Normative Data on Two New Parallel Forms. Abgerufen von http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0092554&type=printable (Zugriff 11.05.2017).

Maier, P. H. (1999). Raumgeometrie mit Raumvorstellung – Thesen zur Neustrukturierung des Geometrieunterrichts. *Der Mathematikunterricht, 45*(3), 4-18.

Mandl, H.; Kopp, B. & Dvorak, S. (2004). Aktuelle theoretische Ansätze und empirische Befunde im Bereich der Lehr-Lern-Forschung – Schwerpunkt Erwachsenenbildung. Abgerufen von https://www.die-bonn.de/esprid/dokumente/doc-2004/mandl04_01.pdf (Zugriff 11.05.2017)

Mendez, M. F.; Doss, R. C. & Cherrier, M. M. (1998). Use of the Cognitive Estimation Test to Discriminate Frontotemporal Dementia from Alzheimer's Disease. *Journal of Geriatric Psychiatry and Neurology, 11*(1), 2-6.

Nührenbörger, M. (2004). Children's measurement thinking in the context of length. In Törner, G.; Bruder, R.; Peter-Koop, A.; Neill, N.; Weigand, H.-G. & Wollring, B. (Hrsg.). *Developments in Mathematics Education in German-Speaking Countries. Selected Papers of the Annual Conference on Didactics of Mathematics, Ludwigsburg, March 5-9, 2001* (S. 95-106). Hildesheim: Franzbecker.

Nunes, T. (1992). Ethnomathematics and everyday cognition. In D. A. Grouws (Hrsg.), *Handbook of research on mathematics and learning. A project of the National Council of Teachers of Mathematics* (S. 557-574). New York: Macmillan.

O'Daffer, P. (1979). A Case and Techniques for Estimation: Estimation Experiences in Elementary School Mathematics – Essential, Not Extra! *The Arithmetic Teacher, 26*(6), 46-51.

Paenza, A. (2012). *Mathematik durch die Hintertür. Faszinierende Reisen in die Wunderwelt der Zahlen*. Köln: Anaconda-Verlag.

Peretti Wagner, G.; MacPherson, S. E.; Parente, M. A. M. P.; Trentini, C. M. (2011). Cognitive estimation abilities in healthy and clinical populations: the use of the Cognitive Estimation Test. *Neurological Sciences, 32*(2), 203-210.

Peter-Koop, A. (2008). Eine unbekannte Größe? Entwicklung von Kompetenzen im Bereich Größen und Messen. *Grundschule, 40*(4), 20-22.

Peter-Koop, A. & Nührenbörger, M. (2011). Größen und Messen. In G. Walther; M. van den Heuvel-Panhuizen; D. Granzer & O. Köller (Hrsg.), *Bildungsstandards für die Grundschule: Mathematik konkret* (S. 89-117). 5. Auflage. Berlin: Cornelsen.

Piaget, J. & Szeminska, A. (1975). *Die Entwicklung des Zahlbegriffs beim Kinde*. Stuttgart: Klett.

Pike, C. D. & Forrester, M. A. (1997). The Influence of Number-sense on Children's Ability to Estimate Measures. *Educational Psychology* 17(4), 483-500.

Posner, M. I. & Koppitz, W. J. *(1976). Kognitive Psychologie. Grundfragen der Psychologie.* München: Juventa.

Reuter, D. (2011). *Kindliche Konzepte zur Größe Gewicht und ihre Entwicklung. Theoretische Modellierung und zwei Einzelfallstudien mit Drittklässlern*. Abgerufen von http://edoc.hu-berlin.de/dissertationen/reuter-dinah-2011-05-23/PDF/reuter.pdf (Zugriff 23.05.2017).

Ruwisch, S. (2012). Hohlmaß, Fassungsvermögen, Rauminhalt oder Volumen? *Grundschule Mathematik, 34*, 40-43.

Ruwisch, S. (2014). Das Ungefähre zu schätzen wissen. Die Bedeutung des Ungefähren und wie man sich ihm nähert. *Grundschule Mathematik, 42,* 4-5.

Schermer, F. J. (2006). *Lernen und Gedächtnis*. Stuttgart: Kohlhammer.

Siegel, A. W.; Goldsmith, L. T. & Madson, C. R. (1982). Skill in Estimation Problems of Extent and Numerosity. *Journal for Research in Mathematics Education, 13*(3), 211-232.

Shallice, T. & Evans, M. E. (1978). The Involvement of the Frontal Lobes in Cognitive Estimation. *Cortex: a Journal Devoted to the Study of the Nervous System and Behaviour, 14*(2), 294-303.

Smith, E. E. & Jonides, J. (1999). Storage and Executive Processes in the Frontal Lobes. *Science, 283*(1654), 1657-1661.

Sowder, J. (1992). Estimation and Number Sense. In D. A. Grouws (Hrsg.*), Handbook of research on mathematics teaching and learning. A project of the National Council of Teachers of Mathematics* (S. 371-389). New York: Macmillan.

Stephan, M. & Clements, D. H. (2003). Linear and Area Measurement in Prekindergarten to Grade 2. In D. H. Clements & G. Bright (Hrsg.), *Learning and Teaching Measurement* (S. 3-16). Reston, VA.: National Council of Teachers of Mathematics.

Ullsperger, M. & von Cramon, D. Y. (2006). Funktionen frontaler Strukturen. In H.-O. Karnath & P. Thier (Hrsg.), *Neuropsychologie* (S. 479-488). 2. Auflage. Heidelberg: Springer.

Van der Meer, E. (2006). Langzeitgedächtnis. In J. Funke & J. Bengel (Hrsg.). *Handbuch der Allgemeinen Psychologie – Kognition* (S. 346-355). Göttingen: Hogrefe.

Winter, H. (2003). *Sachrechnen in der Grundschule. Problematik des Sachrechnens. Funktionen des Sachrechnens. Unterrichtsprojekte.* 6. Auflage. Frankfurt am Main: Cornelsen Scriptor.

Anschrift der Verfasser

Dana Farina Weiher
Leuphana Universität Lüneburg
Institut für Mathematik und ihre Didaktik
Universitätsallee 1
21335 Lüneburg
weiher@leuphana.de

Silke Ruwisch
Leuphana Universität Lüneburg
Institut für Mathematik und ihre Didaktik
Universitätsallee 1
21335 Lüneburg
ruwisch@uni.leuphana.de

Anhang

Anwenden von Schätzstrategien beim Schätzen von Größen (Schätzen im engeren Sinne)

A.3 Publikation 2

Weiher, D. F. (2019a). Framework for the Parallelized Development of Estimation Tasks for Length, Area, Capacity, and Volume in Primary School – A Pilot Study. *Journal of Research in Science, Mathematics and Technology Education, 2*(1), 9–28. https://doi.org/10.31756/jrsmte.212

JOURNAL OF RESEARCH IN SCIENCE, MATHEMATICS AND TECHNOLOGY EDUCATION
DOI: 10.31756/jrsmte.212

http://www.estej.com/

Framework for the Parallelized Development of Estimation Tasks for Length, Area, Capacity, and Volume in Primary School – A Pilot Study

Dana Farina Weiher

Leuphana University Lueneburg, Germany

Abstract: The purpose of this study is to present a framework for the development of parallelized estimation tasks for the visible measures length, area, capacity, and volume. To investigate if there are differences between the estimation types of task, a written estimation test for 3rd- and 4th-graders was developed. It includes eight different types of task for each measure. The percentage deviation of the estimated value from the real value (the measured size) of 137 students indicates that there are differences between the four measures as well as within the types of task that affect over- and underestimation and the estimation accuracy. Further research could address relations between the estimation of visible measures and the investigation of more characteristics in an estimation task, using a written estimation test that is based on this valid framework.

Keywords: *Estimation test; Estimation tasks; Measurement estimation; Parallelized items; Visible measures*

Introduction

Estimation in General

Estimation in general from a psychological point of view means answering a question whose exact answer is unknown. Therefore, it is necessary to use cognitive skills like, among others, developing an appropriate estimation strategy, reasoning, general knowledge, and executive functions (Brand et al., 2003; D'Aniello, Castelnuovo, & Scarpina, 2015; MacPherson et al., 2014). According to Winter (2003), who defines estimation from a mathematics education point of view, estimation can be described as a complex interaction between perceiving, remembering, correlating, rounding and calculating.

Psychological researchers focus on the parts of the brain involved during the estimation process. This focus results in a process model (Figure 1) that contains the working memory, the semantic long-term memory and the executive functions as parts for estimation (Brand et al. 2003; D'Aniello et al., 2015).

Mathematics education researchers mainly investigate the use of estimation strategies. They differentiate three kinds of estimation: number estimation, measurement estimation and computational estimation (Hogan & Brezinski, 2003; O'Daffer, 1979). Within measurement estimation, another distinction can be made. The estimating person can perceive the object's attributes (measures) such as length, area, capacity and volume with the eye. Therefore, these measures are characterized as visible measures. Other attributes of objects such as speed, time, weight and temperature are limited visible or not visible. Either visibility is expressed via other quantities, such as distance covered in a certain time (speed), or they must be made visible by performing an action (time), or they can only be perceived with other senses (weight, temperature).

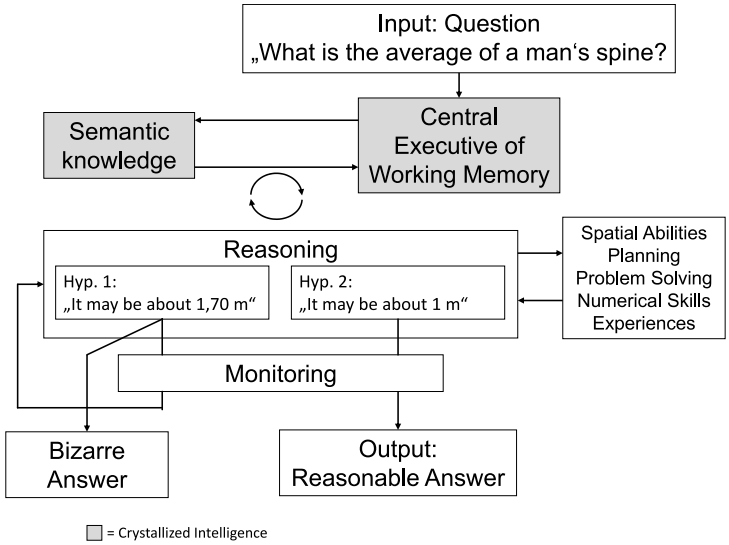


Figure 1. Model of Cognitive Estimation (D'Aniello et al., 2015).

Both psychological and mathematics education researchers use estimation tests for their investigations. Nevertheless, a valid theoretical base of how the tests were developed is often missing (Heinze, Weiher, Huang, & Ruwisch, 2018). In general, psychological tests mix up number and measurement estimation and, in addition, the different measures, without giving any further explanation (e.g. D'Aniello et al., 2015; MacPherson et al., 2014). Mathematics educations tests consider different kinds of estimation, but do not provide any references about the relation of different measures (e.g. Siegel, Goldsmith, & Madson, 1982).

Research Goal

This article focuses on the estimation of visible measures. Visible measures are length, area, capacity (within the meaning of liter and milliliter), and volume (within the meaning of cubic volume).

In current mathematical didactic research, there are no estimation tests based on a theoretically-based selection of different types of tasks, taking into account the possible differences of estimation of different sizes. For describing the estimation competence and the investigation of involved other competences like, for example, measuring competence or executive functions, it is necessary to use a valid estimation test. In addition, an estimation test for primary school aged children is missing.

The first aim of this article is to present a broad framework for the parallelized development of estimation tasks for length, area, capacity, and volume.

In addition, empirical results of the first test use are presented and discussed in order to obtain information on the suitability of the use in 3rd and 4th grade. More specifically, I sought answers for questions below:

- Are the types of task and measures suitable for 3rd- and 4th-graders?
- Which empirical differences between the characteristics of the tasks can be determined?

Theoretical Background

Measurement Estimation

Bright (1976) defines measurement estimation as "the process of arriving at a measurement without the aid of measuring tools. It's a mental process" (p.

89). Crucial in his explanation is the word *mental*. Already by doing a concrete measurement activity, the process is no longer seen as an estimation, but a measurement process.

This mental process is characterized by the comparison of the to-be-estimated-object (TBEO) with another object whose size and measure are known. These objects for comparison are named *benchmarks* (Joram, 2003). Most estimation strategies described for length, area, capacity, and volume are based on the process of comparison with benchmarks (Joram, Subrahmanyam, & Gelman, 1998; Siegel et al., 1982): Either the benchmark could be nearly the same size as the TBEO, or it has to be divided or multiplied. If this is not possible because the TBEO is too big or has a different shape, the estimator can simplify the estimation situation. Therefore, he could divide the TBEO to get an appropriate benchmark. Then, the person estimates the parts of the TBEO. Finally, the estimator merges the parts and their results to get the complete estimation result. This strategy is named decomposition/recomposition (Siegel et al., 1882). To bring the TBEO into a similar shape as the benchmark, the estimator can rearrange the TBEO mentally to simplify the comparison with the benchmark (Hildreth, 1983).

Additionally, a strategy for the estimation of area and volume exists: Length-Times-Width for area and Length-Times-Width-Times-Height for volume (Hildreth, 1983). These strategies are based on the formulas for rectangles and cubes. Therefore, the lengths of the sides of the rectangle respectively the edges of a cube are estimated and merged to get the result. Again, to estimate the length of the sides or edges, benchmark knowledge is required.

Even psychological studies indirectly refer to the use of benchmarks: They describe general knowledge to be relevant for accurate estimation results (e.g. Brand et al., 2003; D'Aniello et al., 2015). General knowledge could be seen as one part of the essential knowing about benchmarks (which means knowing the size of objects and be familiar with them).

Length, Area, Capacity, and Volume as Visible Measures

For the description of the relation between length, area, capacity, and volume two approaches can be used: first, as they are mathematics measures, the relation could be mathematically-derived. Second, because the estimation process of these measures is based, among others, on the understanding of the measurement process (Nührenbörger, 2004), an approach from the view of mathematics education can be useful.

The measures length, area, and volume (volume in the sense of a cube) are part of the same mathematical size system, with length as the base size. That means that all other measures can be derived from length. This becomes apparent by looking at the formulas for the computation of the size of an area or of a volume. For computing the size of an area, two lengths are multiplied, for computing the size of a volume, three lengths are multiplied. This fact is also reflected in the estimation strategies for area and volume. Capacity (in the sense of the content of a vessel) is not part of this size system, hence, it is more important in everyday life of young children than volume. Furthermore, it is part of the (German) primary school curriculum, so this meaning of volume should be part of a (German) visible measure estimation test.

In mathematics education research on estimation and measurement, some different reasons can be named for a joint investigation of estimation of length, area, capacity and volume.

First, all these measures are visible, and therefore some strategies that are already described for length can also be used to estimate the other measures (Figure 2). This is a result of the benchmark idea and the fact that length, area, capacity, and volume are visible measures. For all visible measures, the estimator can compare the TBEO with another object (the benchmark).

Second, the comprehension of the measurement process, which is one important aspect for the comprehension of the estimation process, includes similar aspects for the different measures. The most important aspect of measuring a size can be seen in iterating a unit of the same size. This idea is also possible for all visible measures. For length, e.g. lines as representatives of a unit can be laid end to end to one another. For area, squares as representatives for the unit can be used in the same manner, which creates a grid pattern. For volume, cubes can be used in the same way for a result without gaps and overlaps. Even if the process is not visible anymore after completion, this idea also works for capacity: A vessel can be used as a unit to measure the size of a bigger vessel. The number of pouring e.g. water from the smaller in the bigger vessel determines the size indication. After doing that, the number of pouring-activity is not visible anymore, but the idea of repeating a same sized unit is the same.

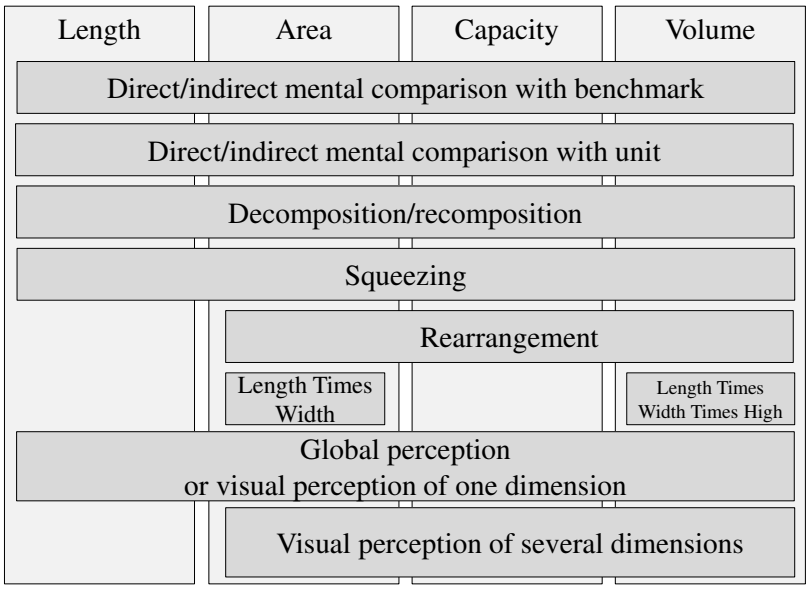


Figure 2. Estimation Strategies for the Measures Length, Area, Capacity, and Volume.

Third, Nührenbörger (2004) claims that the comprehension of the measurement process for length is not only similar to, but also fundamental for the comprehension of the measurement process from other measures, especially area and volume.

State of the Art: Different Types of Estimation Tasks

The first approach for structuring estimation tasks originates with Bright (1976). According to him, two objects are part of an estimation task: the TBEO, which measure should be estimated, and a unit. Both

of these objects can be physically present or absent. Another type of task includes a measure to which an appropriate object should be found. For this task, a list with possible objects can be given (or not). Overall, Bright formulated eight types of estimation tasks (Figure 3).

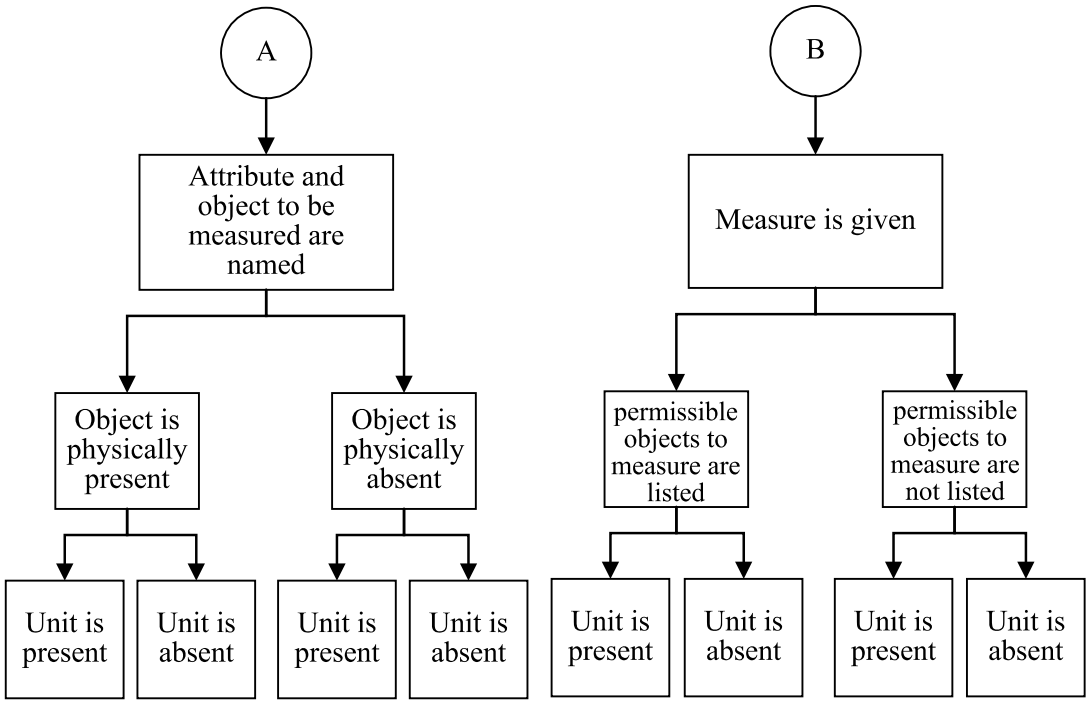


Figure 3. Eight Kinds of Estimation (Bright, 1976).

Heinze et al. (2018) used part A of Bright's model as initial position, to develop a broader framework for length estimation tasks. The aspect *physically present* distinguishes between *just visible* objects and *visible and touchable* objects. Besides that, *not physically present* automatically means *not visible*, but there is also the possibility to present the object with help of a picture (which means that it is *not physically present in real size*).

Another addition to Bright's model is a third object that can be part of an estimation task: the benchmark. This is an object of a given size that can be used as an object for a comparison. If the benchmark is given, the same characteristics as for the other objects are possible. It can be *physically present* or *not physically present*. If it is physically present, it can be *just visible* or *visible and touchable*.

The length estimation framework also includes construction tasks. These tasks require drawing a line with a length given. This entails that the TBEO is not visible at the beginning (because it is not constructed yet), but it is visible after the drawing process. Since these characteristics change during the working process, the distinction between visible and not visible is not appropriate for drawing tasks. Nevertheless, the other objects (unit and benchmark) can have the same characteristics as described above.

This framework is currently still restricted to length estimation. For creating parallelized items for the estimation of length, area, capacity, and volume, further development of the framework is needed.

A New Framework

Characteristics of Measurement Estimation Tasks

For developing an estimation test to investigate the visible measures length, area, capacity and volume, parallelized test items are desirable. This would be an improvement of existing tests (which mixed up both number and measurement estimation, and the different measures) and allows to get valid results. Furthermore, no valid estimation test exists for primary school aged students (who are the target group of this study) exists.

Parallelized items are characterized by equal task characteristics. They should also require the same or similar competences over the measures.

As described above, three objects can be part of all estimation tasks: the TBEO (which has to be named), a unit, and a benchmark. These objects can either be visible or not. For being visible, two possibilities exist: Either it is physically present (visible in real size), or it is shown on a picture (visible, but not in real size). If it is physically present, it can be touchable or not.

For developing parallelized items for all visible measures, three restrictions can be made from the very start.

First, the drawing tasks, which are included in the length estimation framework described above, do not fit the demand to address the same competences for every visible measure. It seems to be much easier to draw a line than to draw a cube, especially in a defined size. Because drawing is not a competence that is usually needed for estimation, it should not affect estimation tasks. By drawing lines, the impact could be left aside because all students should principally be able to draw a line. Therefore, drawing tasks could work well in estimating lengths, but not in estimating area, capacity, or volume. Consequently, the framework for all visible measures does not include drawing tasks.

Second, the framework does not include tasks with pictures of the TBEO, benchmark or unit. For length and area, there are no problems to perceive the real sizes and the relation between them from a picture, but for capacity and volume, there are. Due to the projection of three-dimensional objects on a two-dimensional surface, the estimator cannot perceive all real sizes or all real relations between them.

Third, if a benchmark is given, and should be usable as a benchmark, both the object and its size must be given. To use an object as a benchmark, the estimator has to have a clue about both the real size of the object and its measured size. If only one aspect is given, the benchmark may still be helpful (because it is more than nothing), but if both aspects are unknown to the estimator, the object given cannot be used as a benchmark. To ensure the difference between tasks with and without given benchmark, tasks that only include a picture of the benchmark or tasks that only name a size are not included in this framework.

Following these deliberations, the fundamental structure for possible estimation tasks for length, area, capacity, and volume, includes the characteristics shown in Figure 4.

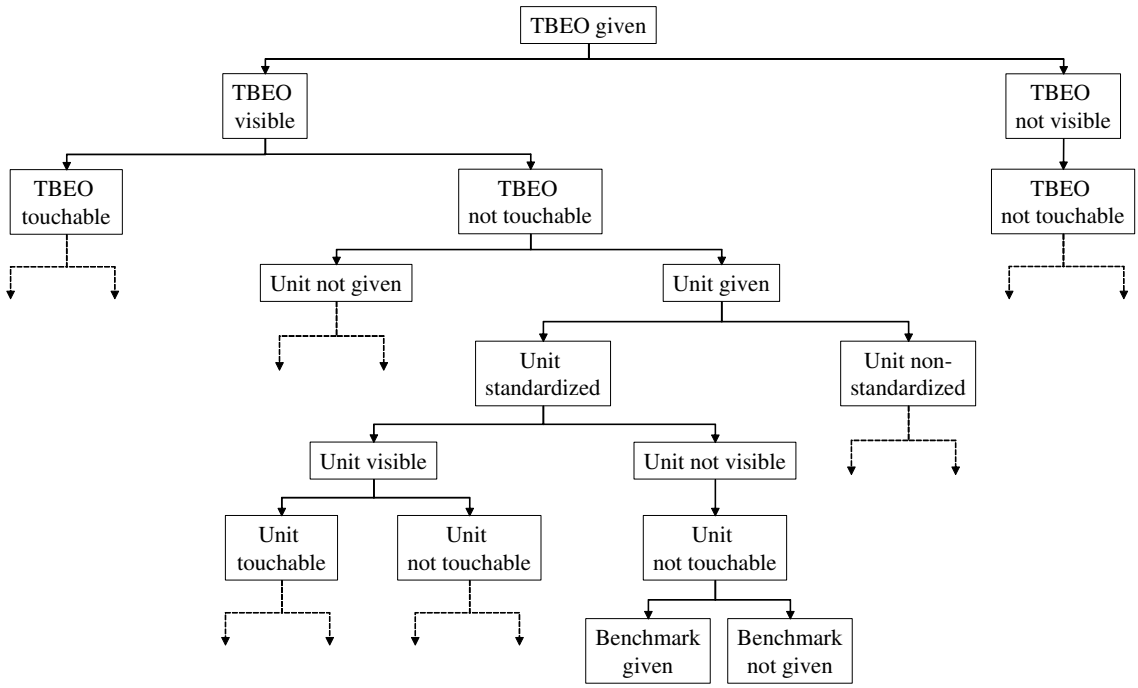


Figure 4. Possible Characteristics for Estimation Tasks.

By combination of the characteristics named above, overall 84 types of task result. Not all of them are appropriate for a written estimation test. The next chapter gives reasons for the exclusion of types of task.

Types of Task for the Parallelized Development of Estimation Tasks for Visible Measures

The most important reason to exclude a task from the estimation test is the possibility to do a measure. This applies to tasks that include two touchable objects (the TBEO and unit or TBEO and benchmark). Either the unit or the benchmark can be used to measure the size of the TBEO directly, or a third object, e.g. a finger, can be used to measure the unit or the benchmark first and the TBEO second. In both cases, this measuring process should be avoided. Consequently, all types of task with two touchable objects were excluded.

In some cases, a given characteristic entails another characteristic that should not be given. This is true if the unit should not be named, but the size of a benchmark is given, because the unit of the benchmark's size reveals the unit of the TBEO. To ensure the distinction between tasks with given a unit and tasks with no unit given, there are no benchmarks given if the unit is not named.

It can happen that more characteristics are given than needed. This is true if not only a visible unit, but also a benchmark is given. The benchmark is not necessary if the unit is visible, because for the estimation process, the unit could be used. So, if the unit is visible, a benchmark is redundant.

In tasks without a visible unit, it is theoretically possible to name an object that is intended to be a benchmark. Actually, both the size and the measure should be given (which means the object has to be visible) to make an object usable as a benchmark, but it could also be helpful to only know one aspect. This distinction is not made in the framework to avoid the situation that a student could not use the object in the intended way. Due to this unclear definition of "giving a benchmark" and,

furthermore, the need of material and the number of items is high anyway, this type of task is excluded from the test.

Since we want to distinguish clearly between capacity and volume, the unit has to be named. If tasks with no given units are used for these measures it cannot be evaluated which concept the children are referring to. To ensure the parallelism between the four measures, this type of task is excluded not only for capacity and volume tasks, but also for length and area tasks.

Last but not least, the test is intended to be reasonable and understandable regarding the material. For length and area, touchable objects can be printed in the test booklet. For capacity and volume, printing is not possible because objects are three-dimensional. Consequently, real objects must be given to each student for all tasks that include touchable TBEOs, units, and benchmarks. This would increase the need of material (and therefore the costs and grasp) in an unacceptable way. Therefore, this kind of task was excluded for all measures.

Finally, eight of the 84 possible types of tasks were chosen (see Table 1).

			TBEO visible Not touchable	TBEO not visible
Unit standardized	Unit visible	Not touchable	Type 1	Type 2
	Unit not visible		Type 3	Type 4
Unit non-standardized	Unit visible	Not touchable	Type 5	Type 6
	Unit not visible		Type 7	Type 8

Figure 5. Eight Types of Task for Parallelized Items for Length, Area, Capacity, and Volume.

For this test, the TBEO can either be visible or not, but not touchable. There are two possibilities to fulfill these conditions: the TBEO may not be physically present or may physically be present, but will be shown at the blackboard. The pupils can see these objects, but they are not allowed to go to the front and touch it. The unit is always named. It is either visible or not, too, so for objects that represent the unit, the same characteristics apply as for the TBEO. The unit can be standardized or non-standardized.

Methodology for Testing the Suitability

Instrument

For investigating the suitability of estimation tasks for 3rd- and 4th-graders, four written estimation tests, one test for each measure, were developed. Each test includes eight types of task as described above. Three items represent each type of task. Overall, each test includes 24 items. The objects in these items should be familiar to students in 3rd and 4th grade. For each test, 45 minutes (one lesson) are provided.

The test includes materials, which are presented in the classroom. Minor changes in the array of

materials are possible due to different furniture in the classrooms.

Participants

In this pilot-study, 137 children from three 3rd and three 4th grade classes were involved. The sampling was convenient. Not all children participated in all estimation tests; each class worked out two tests (which means two measures). Some students only filled in one test due to absence at one of the two dates of testing. This leads to different sample-sizes per measure (see Table 2).

Table 1.

Distribution of the Sample.

	Length	Area	Capacity	Volume
Class 3 girls	22	20	9	14
Class 3 boys	17	15	9	10
Class 4 girls	16	10	26	25
Class 4 boys	8	11	15	22

Data analysis

Missing values and outliers were counted to get information about the suitability of the measures for 3rd- and 4th-graders. Outliers were identified by using boxplots for each item (values below Q_1-3 IQR or above Q_3+3 IQR). They were deleted for further analysis to avoid distortion.

In order to investigate the suitability of the estimation tasks for 3rd- and 4th-graders, for each item the percentage deviation from the real value (P_r) of the TBEO was computed. Therefore, the real value (V_r) and the estimated value (V_e) is needed:

$$\frac{V_e - V_r}{V_r} \cdot 100\% = P_r. \qquad (1)$$

The deviation from the real value can be negative or positive, so "0%" would be the best result. A negative percentage deviation indicates an underestimation, whereas a positive percentage deviation means that there is an overestimation. To investigate the accuracy of an estimation, it is therefore necessary to use the modulus of the percentage deviation from the real size:

$$| P_r |. \qquad (2)$$

The Shapiro-Wilk test for testing normal distribution was used because of its high statistical power for small samples (Shapiro & Wilk, 1965). The results of the study are not normally distributed (Shapiro-Wilk test for length p = 0,018, area p = 0,091, capacity p = 0, volume p = 0).

Because the results are not normally distributed, percentiles of the percentage deviation were used to investigate over- and under-estimation and the accuracy of estimates for the different measures and different task characteristics.

Findings

Missing Values and Outliers

For each item, the missing values were counted to get information about the "attempting" to solve this item. Table 3 shows the number of missing values per measure.

Table 2.

Missing Values and Percentage per Measure.

Measure	Number of missing values (percentage)
Length	138 (9.12%)
Area	165 (12.27%)
Capacity	57 (4.03%)
Volume	634 (37.21%)

The high number of missing values concerning volume indicates that students have difficulties to estimate volume, or to understand what they have to do. This might be caused by not knowing the standardized units (cm^3, m^3, dm^3) that are used in the test booklet. Due to the observations during the test, and because it is not part of the curriculum of German primary school, it is possible that pupils do not really have a concept of this measure.

Regarding the lower numbers of missing values for length, area, and capacity, these measures seem to be less problematic. The difference between length and area on the one and capacity with less than half of missing values on the other side might be caused by the distribution of the sample. Two 3rd classes and one 4th class solved the length and area tasks, whereas one 3rd class and two 4th classes solved the capacity tasks (and the volume tasks).

Table 4 contains the number of outliers from the solved tasks. The number of extreme outliers is higher than the number of mild outliers for all measures. The highest amount of mild and extreme outliers are found in capacity tasks. Area estimation tasks have the lowest amount of mild outliers, while volume estimation tasks have the lowest amount of extreme outliers. This might be explained by the German curriculum again: because of the unknown standardized units or a missing concept of the measure, the competences concerning these measures might be quite similar. Consequently, a lower amount of outliers is the result.

Table 3.

Number of Outliers and Percentage per Measure.

Measure	Mild outliers (below Q_1-1,5 IQR or above Q_3+1,5 IQR)	Extreme outliers (below Q_1-3 IQR or above Q_3+3 IQR)
Length	38 (2.77%)	97 (7.06%)
Area	19 (1.62%)	62 (5.26%)
Capacity	60 (4.42%)	111 (8.17%)
Volume	36 (3.36%)	37 (3.46%)

Over- and Underestimation

To investigate the over- and underestimations, the percentage deviation from the real size was used. The arithmetic mean of the percentage deviation was computed for each student (and each measure). A negative arithmetic mean implies that this student tends towards under-estimation for this measure, whereas a positive arithmetic mean implies that this student tends towards over-estimation for this measure. For the investigation of the general over-

and under-estimation of each measure, the arithmetic mean of all tasks of this measure for all students is used.

The arithmetic means of all students for length, capacity and volume are not normally distributed (verified using the Shapiro-Wilk-Test, length: p = 0.018, capacity and volume: p = 0.000), whereas for area, they are (p = 0.091). All curves are slightly positively skewed (skewness for length = 0.839, area = 0.741, capacity = 1.093, volume = 1.580). The skewness can be unattended because of the used scale is open to the right, but closed to the left (see discussion).

Table 5 shows the descriptive statistics of the percentage deviation from the real size for all measures. The arithmetic mean and the median indicate that length are rather over-estimated (positive arithmetic mean and median), while area, capacity and volume were rather under-estimated (negative arithmetic mean and median).

Table 4.

Descriptive Statistics for the Percentage Deviation from the Real Value per Measure.

Measure	*N*	*R*	*Min.*	*Max.*	*M*	*Med.*	*SD*
Length	63	142.20	-29.75	112.19	20.19	13.55	29.64
Area	55	145.31	-88.31	57.00	-28.79	-31.25	24.89
Capacity	59	127.14	-66.50	60.64	-25.85	-36.65	32.15
Volume	70	166.07	-93.00	73.07	-45.60	-58.08	39.02

By looking at the percentiles for each measure, this interpretation can be supported (Table 6).

Table 5.

Percentiles of the Arithmetic Mean of the Percentage Deviation from the Real Value per Measure.

	Length	Area	Capacity	Volume
N	63	55	58	69
Percentile 10	-14.23	-58.21	-58.65	-84.25
Percentile 20	-3.45	-48,52	-53.39	-74.07
Percentile 30	2.34	-42.01	-46.25	-69.5
Percentile 40	9,00	-34.84	-40.96	-65.11
Percentile 50	13.54	-31.25	-36.67	-58.75
Percentile 60	22.81	-26.80	-26.70	-47.42
Percentile 70	31.02	-19.85	-16.80	-35.60
Percentile 80	41.82	-6.67	-7.27	-24.5
Percentile 90	64.76	2.55	30.67	2.54

For length, all negative values (and probably a few positive values) are within the 30th percentile. This means, at most 30% of all students tend towards underestimation for the length of an object. For the other measures, the percentiles with all negative values are within the 90th percentile. This indicates that most students (nearly all) under-estimate these measures.

Over- and underestimation can be investigated also for the different types of task. In each case only one characteristic was investigated (independent of the other characteristics), because of the small number of items for each type of task and the small sample. Therefore, the tasks of each measure were divided up into two groups (e.g. TBEO visible or not). The arithmetic mean of the percentage deviation from the real size was computed for each group per student.

For visible TBEOs that length should be estimated, no consistent trend to over- or under-estimation could be indicated (Table 7). The transition from positive to negative arithmetic means is between the 50th and 60th percentile. That means that nearly the same number of students under- and over-estimate the length of a visible TBEO. Concerning the not visible TBEOs, the results show that students tend towards over-estimate this length. Less than 20% of the students under-estimate the length of a not visible TBEO.

In all other conditions, the measures of the TBEOs were in general under-estimated. More than 80% of the students under-estimate the size of a not visible TBEO in an area estimation task. Not visible TBEOs for capacity and volume were under-estimated in more than 90% of the cases. In the visible conditions, the number of students who generally under-estimate are slightly lower, but still the trend to under-estimation is obvious (area and volume more than 80%, capacity more than 60%).

Table 6.

Percentiles of the Arithmetic Mean of the Percentage Deviation from the Real Value in Estimation Tasks with Visible and Not Visible TBEO.

	Length TBEO visible	Length TBEO not visible	Area TBEO visible	Area TBEO not visible	Capacity TBEO visible	Capacity TBEO not visible	Volume TBEO visible	Volume TBEO not visible
N	63	63	55	55	59	59	69	70
Percentile 10	-24.96	-12.42	-77.96	-53.80	-48.50	-74.60	-86.83	-85.81
Percentile 20	-17.03	4.05	-65.49	-43.88	-39.91	-71.11	-73.67	-74.44
Percentile 30	-11.90	11.12	-58.80	-37.94	-35.17	-64.83	-70.78	-71.50
Percentile 40	-7.80	25.71	-52.28	-31.77	-26.50	-60.50	-63.50	-67.04
Percentile 50	-2.64	40.55	-45.33	-23.56	-15.81	-51.55	-57.67	-62.25
Percentile 60	2.53	51.35	-37.38	-17.90	-8.16	-48.42	-53.33	-57.10
Percentile 70	8.66	68.86	-26.62	-11.80	5.00	-36.13	-40.50	-43.65
Percentile 80	11.33	81.82	-19.77	-5.01	36.90	-31.50	-21.67	-31.47
Percentile 90	17.67	124.53	36.90	15.53	71.70	-11.33	39.41	-3.79

As the TBEO, the unit can be visible or not. The results (Table 8) indicate that there is no great difference between tasks with a visible unit and tasks with a not visible unit: In tasks with a visible unit, less than 50% of the students under-estimate the length of the TBEO. In tasks with a not visible unit, less than 40% of the students under-estimate the length of the TBEO.

Unlike to length estimation, it seems to make an obvious difference in the other measures if the unit is visible or not. More than 90% of the students under-estimate the TBEO in tasks with a unit that is not visible. For visible units, the number of under-estimating students is lower: for area, less than 40% of the students under-estimate the size of the TBEO if the unit is visible. For capacity and volume, less than 70% of the students under-estimate the size of the TBEO if the unit is visible.

Table 7.

Percentiles of the Arithmetic Mean of the Percentage Deviation from the Real Value in Estimation Tasks with Visible and Not Visible Unit.

	Length Unit visible	Length Unit not visible	Area Unit visible	Area Unit not visible	Capacity Unit visible	Capacity Unit not visible	Volume Unit visible	Volume Unit not visible
N	63	63	55	55	59	59	69	70
Percentile 10	-18.80	-21.29	-26.80	-83.09	-46.33	-77.00	-75.00	-95.49
Percentile 20	-13.20	-15.27	-13.17	-80.38	-41.08	-70.58	-57.83	-92.79
Percentile 30	-3.00	-6.17	-2.56	-77.00	-33.50	-67.40	-49.83	-90.59
Percentile 40	-0.16	0.99	2.80	-72.13	-23.00	-62.17	-38.67	-89.10
Percentile 50	4.33	8.75	15.00	-67.28	-16.40	-57.58	-32.25	-86.33
Percentile 60	19.82	21.30	21.33	-62.07	-7.27	-52.00	-19.83	-83.03
Percentile 70	36.78	34.41	35.04	-56.65	5.75	-40.10	4.00	-79.76
Percentile 80	49.34	43.61	66.76	-53.38	42.90	-28.27	32.33	-75.299
Percentile 90	97.7	55.53	102.29	-46.52	96.91	-14.00	93.33	-65.20

As a second characteristic, the estimation tasks include standardized or non-standardized units. Table 9 shows the percentiles of the arithmetic mean of the percentage deviation for the measures with the distinction of standardized and non-standardized units.

The percentiles indicate that for length, area, and capacity, more under-estimations were made in estimation tasks with standardized units. For length estimation tasks with standardized units, less than 50% of the students under-estimate the TBEO, so a tendency to over-estimation could be indicated. Nevertheless, the number of students that under-estimate the TBEO in tasks with non-standardized units, is lower (between 20% and 30%). For area and capacity, the number of students who under-estimate are higher: More than 90% under-estimate the area of the TBEO in tasks that include a standardized unit, and more than 80% under-

estimate the capacity of the TBEO in tasks that include a standardized unit. The percentage of students who under-estimate the size of the TBEO when the unit is non-standardized is (slightly) lower: between 80% and 90% for area, and between 60% and 70% for capacity.

Table 8.

Percentiles of the Arithmetic Mean of the Percentage Deviation from the Real Value in Estimation Tasks with Standardized and Non-standardized Unit.

	Length Unit stand.	Length Unit non-stand.	Area Unit stand.	Area Unit non-stand.	Capacity Unit stand.	Capacity Unit non-stand.	Volume Unit stand.	Volume Unit non-stand.
N	63	61	52	55	57	59	36	70
Percentile 10	-27.64	-9.73	-90.83	-56.35	-75.38	-49.88	-96.90	-87.44
Percentile 20	-19.84	-2.26	-87.67	-48.74	-71.25	-40.58	-71.90	-78.18
Percentile 30	-12.28	6.50	-83.30	-34.20	-64.13	-35.73	-63.25	-75.61
Percentile 40	-2.34	13.41	-75.86	-28.14	-61.58	-30.25	-50.07	-70.36
Percentile 50	5.71	21.50	-68.14	-20.27	-52.92	-24.25	-36.40	-67.61
Percentile 60	11.09	37.03	-45.99	-13.57	-47.64	-6.92	7.77	-62.48
Percentile 70	17.89	47.13	-34.49	-6.27	-42.63	1.75	36.75	-57.17
Percentile 80	31.56	65.07	-19.58	-0.08	-26.50	22.18	81.41	-48.95
Percentile 90	62.43	105.50	-1.14	18.01	23.09	39.58	187.09	-28.55

In contrast, more than 90% of the students underestimate volume estimation tasks with non-standardized units, whereas less than 60% of the students underestimate volume estimation tasks with standardized units. These percentiles indicate that there is only a tendency to under-estimate the volume of an object if the unit is standardized, while there is an obvious trend to under-estimate the volume of an object if the unit is non-standardized.

Estimation Accuracy

For the investigation of the estimation accuracy, the modulus of the arithmetic mean of the percentage deviation from the real size was computed. Therefore, the modulus of the percentage deviation from the real size was computed for each task. The arithmetic mean of these values was computed for each child to evaluate the middle percentage deviation from the real value (the accuracy). Table 10 shows the descriptive statistics of the modulus of the percentage deviation from the real size per measure.

Length estimation tasks have in general the smallest percentage deviation from the real size (both the arithmetic mean and median are smaller than the others), while volume has the highest percentage deviation. The deviation for area and capacity is quite similar (both arithmetic mean and median).

Table 9.

Descriptive Statistics of the Modulus of the Percentage Deviation from the Real Value per Measure.

Measure	*N*	*R*	*Min.*	*Max.*	*M*	*Med.*	*SD*
Length	63	138.59	18.83	157.42	59.33	53.24	27.07
Area	55	68.45	48.83	117.28	70.04	68.80	13.47
Capacity	59	186.13	42	228.13	74.88	65.79	28.57
Volume	70	159.13	43.63	202.76	86.58	74.91	35.45

The standard deviation is the smallest for area (SD = 13.47) and the highest for volume (SD = 35.45), while the range is the highest for capacity (186.13). This indicates that student's estimates are nearly similar for area and quite different for volume and capacity estimation tasks. The standard deviations for length (SD = 27.07) and capacity (SD = 28.57) are quite similar. This indicates that the student's percentage deviations are similar within length and capacity.

The arithmetic means from the modulus of the percentage deviation from all students for all measures are not normally distributed (verified by using the Shapiro-Wilk Test, length: $p = 0.001$, area: $p = 0.011$, capacity and volume: $p = 0.000$). All curves are slightly positively skewed (skewness for length = 1.153, area = 0.959, capacity = 3.106, volume = 1.949). The skewness can be unattended because of the used scale is open to the right, but closed to the left (using the modulus of the percentage deviation might reinforce this effect).

For investigating the accuracy of the estimates per measure, Table 11 shows the percentiles of the modulus of the percentage deviation from the real size for each measure.

Table 10.

Percentiles of the Arithmetic Mean of the Modulus of the Percentage Deviation from the Real Value per Measure.

	Length	Area	Capacity	Volume
N	63	55	59	70
Percentile 10	31.14	52.96	54.20	59.03
Percentile 20	37.20	57.39	58.96	62.90
Percentile 30	42.27	60.36	60.05	66.85
Percentile 40	48.13	65.89	61.34	71.57
Percentile 50	53.23	68.80	65.79	74.91
Percentile 60	60.98	72.81	70.72	82.09
Percentile 70	70.06	77.63	76.25	88.53
Percentile 80	77.24	79.67	85.17	99.93
Percentile 90	101.45	87.60	114.30	131.57

These results indicate that area estimation tasks were estimated most accurate (90% of the students show a maximal deviation of 87.6% from the real value) followed by length (90% of the students have a maximal deviation of 101.45% from the real value). However, it is conspicuous that the best 10% of

students estimate area with a maximal deviation of 52.96%, while in length estimation tasks, 50% of the students reached a similar deviation (53.23%).

The lowest estimation accuracy is shown for capacity (90% of the students have a maximal deviation of 114.3%) and volume (90% of the students have a maximal deviation of 131.57%). This is supported by the fact that only 10% of the students have a deviation of 54.2% (capacity) or 59.03% (volume). For length and area, the percentage of students with similar deviation is higher.

The accuracy of the estimates can also be investigated for each type of task. The percentiles indicate that for area and volume, it makes no noticeable difference if the TBEO is visible or not. The percentage deviation from the real value is nearly the same for both type of task (Table 12).

Table 11.

Percentiles of the Arithmetic Mean of the Modulus of the Percentage Deviation from the Real Value in Estimation Tasks with Visible and Not Visible TBEO.

	Length TBEO visible	Length TBEO not visible	Area TBEO visible	Area TBEO not visible	Capacity TBEO visible	Capacity TBEO not visible	Volume TBEO visible	Volume TBEO not visible
N	63	63	54	55	59	59	69	70
Percentile 10	21.41	30.96	51.94	50.59	44.64	58.89	53.33	59.08
Percentile 20	26.63	39.05	56.17	58.11	49.33	61.75	57.67	64.00
Percentile 30	30.60	46.00	60.52	59.42	50.60	66.33	64.33	69.37
Percentile 40	34.20	54.60	62.58	63.98	56.00	70.80	68.75	71.12
Percentile 50	36.82	58.90	67.81	68.57	56.67	73.67	73.67	75.95
Percentile 60	40.93	78.71	73.33	72.29	59.67	77.00	78.50	80.73
Percentile 70	47.81	101.98	77.85	77.33	66.17	78.82	86.58	85.72
Percentile 80	56.83	117.50	80.67	81.47	95.80	85.41	93.42	91.63
Percentile 90	64.78	152.15	92.95	90.43	128.00	95.87	132.83	131.63

In contrast to area and volume, for length and capacity there seem to be a difference. Most of the students have a deviation of 64.78 % or less if the TBEO is visible, whereas they have a deviation of up to 152.15% if the TBEO is not visible. This indicates that length estimation tasks with a visible TBEO were estimated more accurate than length estimation tasks with not visible TBEOs. For capacity, it is the other way round. If the TBEO is visible, 90% of the students show a deviation up to 128%. If the TBEO is not visible, 90% of the students have a deviation of 95.87%.

The estimation accuracy varied within the different characteristics for the unit (Table 13). For length estimation tasks, the deviation is lower if the unit is visible (90% of the students have a deviation of 114.43% or less) than if the unit is not visible (90% of the students have a maximum deviation of 118.77%). For area, capacity and volume, most of the students have a smaller percentage deviation in

tasks with visible object. This trend reverse not until the 80th percentile (area, volume) and the 60th percentile (capacity).

Table 12.

Percentiles of the Arithmetic Mean of the Modulus of the Percentage Deviation from the Real Value in Estimation Tasks with Visible and Not Visible Unit.

	Length Unit visible	Length Unit not visible	Area Unit visible	Area Unit not visible	Capacity Unit visible	Capacity Unit not visible	Volume Unit visible	Volume Unit not visible
N	63	63	61	55	59	59	69	70
Percentile 10	27.37	23.28	39.60	56.73	49.13	49.82	42.60	68.82
Percentile 20	31.23	33.27	46.04	61.60	56.11	60.08	46.33	75.77
Percentile 30	33.34	40.29	52.30	67.01	59.44	64.73	49.83	84.33
Percentile 40	38.05	48.93	58.30	69.57	64.43	69.45	56.71	86.07
Percentile 50	44.75	56.00	63.00	75.45	69.25	72.42	61.00	88.44
Percentile 60	49.86	65.08	67.25	77.18	76.33	74.33	71.33	90.17
Percentile 70	61.85	75.62	73.39	79.18	83.57	75.58	81.25	91.45
Percentile 80	84.93	96.53	83.95	81.98	94.50	77.67	115.67	92.98
Percentile 90	114.43	118.77	91.25	85.74	111.50	82.10	178.36	96.84

The last characteristic to look for intensively is the difference between standardized or non-standardized unit (Table 14). The percentage deviation from the real size is higher for estimation tasks for area, capacity, and volume, if the unit is non-standardized. This is valid for all percentiles, but especially obvious for volume estimation tasks (90% of the students have a percentage deviation of 295% from the real size if the unit is standardized, whereas the deviation is only 87% if the unit is non-standardized). In capacity estimation tasks, this trend is also noticeable.

For length estimation tasks, the trend is not that noticeable. For the most accurate estimates, the unit is non-standardized, but with increasing inaccuracy, the tasks with standardized units were estimated better than tasks with non-standardized units.

Table 13.

Percentiles of the Arithmetic Mean of the Modulus of the Percentage Deviation from the Real Value in Estimation Tasks with Standardized and Non-standardized Unit.

	Length Unit stand.	Length Unit non-stand.	Area Unit stand.	Area Unit non-stand.	Capacity Unit stand.	Capacity Unit non-stand.	Volume Unit stand.	Volume Unit non-stand.
N	63	61	52	55	57	59	36	70
Percentile 10	29.74	23.10	67.25	38.99	55.64	45.58	67.63	54.72
Percentile 20	33.82	29.76	71.05	43.45	59.73	50.91	75.12	62.23
Percentile 30	38.47	38.98	74.58	49.65	64.30	53.58	87.05	64.44
Percentile 40	43.65	42.98	80.13	54.23	67.00	59.17	92.23	67.09
Percentile 50	47.25	49.71	83.75	62.44	71.00	61.18	99.17	71.03
Percentile 60	51.80	60.57	87.16	66.88	73.81	64.88	115.10	74.15
Percentile 70	57.03	71.86	88.65	70.07	79.65	68.73	173.31	76.52
Percentile 80	72.53	92.63	90.83	76.18	96.67	78.80	187.92	80.95
Percentile 90	108.53	112.80	96.19	88.60	133.84	97.42	295.48	87.44

Discussion

Estimation of length, area, capacity, and volume have different results concerning under- and overestimation and estimation accuracy. Under- or over-estimation of the measures in general might be caused by the limitations or inexperience by using a higher number range. For area, capacity and volume, the measure values were naturally higher because of more dimensions. A consequence for not feeling safe using higher numbers (for everyday-sized objects) might be choosing lower numbers as measured values which results in an under-estimation. The everyday-experience with length may compensate the insecurity with higher numbers. Students might feel safer using higher length-sizes than high numbers for measures they do not have much experience with. This conjecture is supported by the fact that the mean of the modulus of the percentage deviation is smaller for length than for the other measures (see below).

Higher accuracy for not visible TBEOs than for visible TBEOs in capacity estimation tasks might be explained with the higher number of under-estimations in these tasks. An under-estimation has a maximal deviation of 100%. When the TBEO is visible, over-estimations are more frequently. This result has an impact on the highest deviation (over-estimations are not percentage limited). The reverse argumentation can be used for length: Tasks with not visible TBEOs were in general over-estimated, therefore, the accuracy is lower. In addition, it is more difficult to estimate the size of a not visible object because more cognitive processes (like getting an appropriate image of the object and its size from the memory) are necessary.

In tasks with not visible units, the area, capacity or volume of an object was more often under-estimated than in tasks with visible units. This difference might be caused by mental benchmarks that are too small or because by the absence of any benchmark for these measures (and wild guessing causes

estimated that are too low because of the higher number range, see above). If a unit is visible, the estimation is supported by a hint of the real magnitude, so it is easier to estimate. These results might cause a better accuracy in tasks with not visible units for area, capacity and volume.

Estimation tasks for capacity and area with non-standardized have a better accuracy, although these tasks were in general underestimated. For area, this might be explained with the unknown standardized units, but the units for capacity actually should be known. The accuracy in volume estimation tasks with standardized units is noticeable lower than in volume estimation tasks with non-standardized units. This might be explained on the one hand with the unknown standardized unit, on the other hand by the general over-estimation in volume estimation tasks with standardized units.

Because of the different results concerning over- and under-estimation and estimation accuracy, the parallelism between the tasks and the measures should be discussed. On the one hand, estimation of visible measures seems to require similar competences because the estimation strategies could be used for all measures. On the other hand, even these visible measures have differences like, e.g., the dimensionality. The results indicate that measurement estimation in general is not a valid construct and should be theoretically divided in length estimation, area estimation, capacity estimation and volume estimation.

The different types of task per measure cause different evaluations of the appropriateness for 3rd- and 4th-graders. The high amount of missing values in volume estimation indicates that these tasks are not appropriate for 3rd- and 4th-graders. This seems to be limited to tasks with standardized units: 70 from 71 students attempt tasks with non-standardized units, whereas only 36 students attempt tasks with standardized units. For area, this difference is not noticeable.

Conclusion

The most important conclusion is the fact, that the one and only estimation task does not exist. Different types of task result in different (strong) tendencies for over- and under-estimation and different estimation accuracies. This pilot study shows that an estimation test must include different types of task to get a valid result.

Furthermore, this pilot study indicates that 3rd- and 4th-graders are able to solve estimation tasks for length, area, capacity, and volume. Except tasks with standardized units for volume, they are able to deal with all types of task. However, because this might be caused by the German curriculum which does not involve standardized units for this measure in primary school, in general, these tasks might be usable in higher grades.

Further research is needed to investigate the relations of all characteristics among each other. Therefore, a bigger sample and even more items per type of task are needed to allow factor analysis for the eight types of task in this estimation test.

Finally, the question of scoring is not answered yet. Different studies in psychology and mathematics education research do not only use different tasks, but also different kinds of scoring. Consequently, the comparison of these studies' results is difficult (Clayton 1996). Further research is needed to improve and develop an appropriate scoring for (parallelized) items for all visible measures.

References

Bright, G. W. (1976). Estimation as Part of Learning to Measure. In D. Nelson (Ed.), *Measurement in School. Mathematics 1976 Yearbook* (pp. 87–104). Reston, VA: National Council of Teachers of Mathematics.

Brand, M., Fujiwara, E., Kalbe, E., Steingass, H.-P., Kessler, J., & Markowitsch, H. J. (2003). Cognitive Estimation and Affective Judgments in Alcoholic Korsakoff Patients. *Journal of Clinical and Experimental Neuropsychology, 25* (3), 324-334. doi: http://dx.doi.org/10.1076/jcen.25.3.324.13802

D'Aniello, G. E., Castelnuovo, G., & Scarpina, F. (2015). Could Cognitive Estimation Ability Be a Measure of Cognitive Reserve? *Frontiers in Psychology, 6,* 1-4. doi: http://dx.doi.org/10.3389/fpsyg.2015.00608

Heinze, A., Weiher, D. F., Huang, H.-M., & Ruwisch, S. (2018). Which Estimation Situations are Relevant for a Valid Assessment of Measurement Estimation Skills? *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education (Vol. 1)*. Umeå, Sweden: PME.

Hildreth, D. J. (1983). The Use of Strategies in Estimating Measurements. *Arithmetic Teacher, 30* (5)*,* 50-54.

Hogan, T. P., & Brezinski, K. L. (2003). Quantitative Estimation: One, Two, or Three Abilities? *Mathematical Thinking and Learning, 5* (4), 259-280. doi: http://dx.doi.org/10.1207/S15327833MTL0504_02

Joram, E. (2003). Benchmarks as Tools for Developing Measurement Sense. In D. H. Clements & G. Bright (Eds.), *Learning and Teaching Measurement. 2003 Yearbook* (pp. 57-67). Reston, VA: National Council of Teachers of Mathematics.

Joram, E., Subrahmanyam, K., & Gelman, R. (1998). Measurement Estimation: Learning to Map the Route from Number to Quantity and Back. *Review of Educational Research, 68*(4), 413-449. doi: http://dx.doi.org/10.3102/00346543068004413

MacPherson, S. E., Wagner, G. P., Murphy, P., Bozzali, M., Cipolotti, L., & Shallice, T. (2014). Bringing the Cognitive Estimation Task into the 21st Century: Normative Data on Two New Parallel Forms. PLoS ONE 9(3): e92554. doi: http://dx.doi.org/10.1371/journal.pone.0092554

Nührenbörger, M. (2004). Children's Measurement Thinking in the Context of Length. In G. Törner, R. Bruder, A. Peter-Koop, N. Neill, H. G. Weigand, & B. Wollring (Eds.) *Developments in Mathematics Educations in German-speaking Countries. Selected Papers from the Annual Conference on Didactics of Mathematics. Ludwigsburg 2001.* (pp. 95-106).

O'Daffer, P. (1979). A Case and Techniques for Estimation: Estimation Experiences in Elementary School Mathematics – Essential, Not Extra!. *The Arithmetic Teacher, 26* (6), 46-51.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Sample). *Biometrika*, *52* (3/4), 591-611. doi: http://dx.doi.org/10.2307/2333709

Siegel, A. W., Goldsmith, L. T., & Madson, C. R. (1982). Skill in Estimation Problems of Extent and Numerosity. *Journal for Research in Mathematics Education, 13* (3), 211–232. doi: http://dx.doi.org/10.2307/748557

Winter, H. (2003). *Sachrechnen in der Grundschule. Problematik des Sachrechnens. Funktionen des Sachrechnens. Unterrichtsprojekte*. (6th ed.). Frankfurt am Main: Cornelsen Scriptor.

Corresponding Author Contact Information:

Author name: Dana Farina Weiher

University, Country: Leuphana University Lueneburg, Germany

Email: weiher@leuphana.de

Please Cite: Weiher, D. F. (2019). Framework for the Parallelized Development of Estimation Tasks for Length, Area, Capacity, and Volume in Primary School – A Pilot Study. *Journal of Research in Science, Mathematics and Technology Education*, *2*(1), 9-28. doi: 10.31756/jrsmte.212

Received: October 19, 2018 ▪ Accepted: January 10, 2019

A.4 Publikation 3

Weiher, D. F., & Ruwisch, S. (2022). The Assessment of Measurement Estimation Results – a Discussion of Different Scorings Regarding Test Performance and Test Quality [Eingereicht zur Veröffentlichung].

The Assessment of Measurement Estimation Results – a Discussion of Different Scorings Regarding to Test Performance and Test Quality

Die Bewertung von Ergebnissen beim Schätzen von Größen – eine Diskussion verschiedener Scorings im Hinblick auf Testleistung und Testqualität

Dana Farina Weiher
Leuphana University, Lüneburg, Germany
Universitätsallee 1
21335 Lüneburg
Germany

weiher@leuphana.de
ORCID: 0000-0001-7292-7321

Silke Ruwisch
Leuphana University, Lüneburg, Germany
Universitätsallee 1
21335 Lüneburg
Germany

ruwisch@leuphana.de
#49-4131-677-1731

Declarations

Funding

The authors did not receive support from any organization for the submitted work.

No funding was received to assist with the preparation of this manuscript.

No funding was received for conducting this study.

No funds, grants, or other support was received.

Conflicts of interest/Competing interests

The authors have no relevant financial or non-financial interests to disclose.

The authors have no conflicts of interest to declare that are relevant to the content of this article.

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

The authors have no financial or proprietary interests in any material discussed in this article.

Compliance with ethical standards

This research involves Human Participants. All participating students have been informed that participation is voluntary and will not suffer any disadvantages, regardless of whether they take part in the study or not. The teachers of the learning groups were present during the written test. The test itself represents a typical assessment situation in mathematics classrooms of the children.

Abstract

To evaluate the accuracy of estimation results when estimating measures, it is common to assign points depending on the percentage deviation from the real value (scoring). In mathematics educational research, different scorings are used, which differ with respect to the number of intervals and the interval limits. The reasons for choosing a scoring and the effects of this choice on the results of a study remain intransparent. Therefore, the aim of this article is to compare and discuss different scorings by investigating their influence on the test performance and the test quality of the same estimation test for the measures length, area, capacity and volume. For this purpose, different scorings were applied to the same data set of a written estimation test for 615 children of the 5th and 6th grade. All scorings differ from each other with respect to the amount of points achieved, rank order, internal consistency, and discriminatory power. Scorings with multiple intervals, which award points even with higher percentage deviation, seem to have slightly better values. However, general statements about a higher suitability of dichotomous or polytomous, strict or lenient scorings cannot be made. That is the reason why the choice of a scoring must always be carefully justified with respect to the aim of the study. The results also show overall rather low values for internal consistency and discriminatory power, which could partly be attributed to the use of the percentage deviation as the basis for the scorings. Due to this fact, further research can also look for alternatives to the percentage deviation as a basis for scoring.

Zusammenfassung

Zur Bewertung der Genauigkeit von Schätzergebnissen beim Schätzen von Größen ist es üblich, Punkte abhängig von der prozentualen Abweichung vom Realwert zu vergeben (Scoring). In der mathematikdidaktischen Forschung werden verschiedene Scorings, die sich hinsichtlich der Intervallanzahl- und Intervallgrenzen unterscheiden, angewendet. Die Gründe für die Auswahl eines Scorings und die Auswirkungen dieser Wahl auf die Ergebnisse einer Studie bleiben intransparent. Ziel dieses Artikels ist daher der Vergleich und die Diskussion verschiedener Scorings, indem deren Einfluss auf das Testergebnis und die Testqualität desselben Schätztests für die Größen Länge, Flächeninhalt, Fassungsvermögen und Volumen untersucht wird. Dazu werden verschiedene Scorings auf denselben Datensatz eines schriftlichen Schätztests für 615 Kinder der 5. und 6. Klasse angewendet. Alle Scorings unterscheiden sich hinsichtlich des Anteils der erreichten Punkte, der Rangordnung, der internen Konsistenz und der Trennschärfe voneinander. Dabei scheinen Scorings mit mehreren Intervallen, die Punkte auch bei höherer prozentualer Abweichung vergeben, leicht bessere Werte aufzuweisen. Allgemeine Aussagen über eine bessere Eignung von dichotomen oder polytomen, strengen oder nachsichtigen Scorings können jedoch nicht getroffen werden, weshalb die Wahl eines Scorings immer vor dem Ziel der Studie sorgfältig begründet werden muss. Die Ergebnisse zeigen zudem insgesamt eher niedrige Werte für interne Konsistenz und Trennschärfe, was sich teilweise auch auf die Verwendung der prozentualen Abweichung als Basis für die Scorings zurückgeführen lässt. Aufgrund dessen könnte sich weitere Forschung auch mit Alternativen zur prozentualen Abweichung als Basis für Scorings befassen.

Keywords

Measurement Estimation, Scoring, Estimation Test, Estimation Accuracy, Evaluation, Estimation Result

Schlüsselwörter

Größen schätzen, Bepunktung, Schätztest, Schätzgenauigkeit, Bewertung, Schätzergebnis

1 Introduction

Measurement estimation is important for the daily life (Jones et al., 2009) and necessary in different situations (Sowder, 1992). By estimating the amount of water needed for cooking noodles, one estimates capacity. For wrapping a gift properly, one estimates the area of the paper needed or the length of a ribbon. When thinking about the duration of a way, one considers the length of that way. Standing in the hardware store and having forgotten the measure of an object one can estimate its length, area, or volume. Measurement estimation processes can also help to develop a feeling for the object sizes that surround us (Joram et al., 2005). Therefore, even psychologists use this mathematical ability to investigate lesions in the frontal lobe (Axelrod & Millis, 1994; Brand et al., 2003; Bullard et al., 2004; D'Aniello et al., 2015; Della Sala et al., 2003; Mendez et al., 1998, Shallice & Evans, 1978).

Measurement estimation (as one of three kinds of estimation, beside number and computational estimation, O'Daffer, 1979) can be described as a "process of arriving at a measurement or measure without the aid of measuring tools. It is a mental process" (Bright, 1976, p. 89). Therefore, the estimation process or its result is not directly accessible to researchers, but must be described by the person doing the estimation.

In daily situations, the success of estimation processes is visible in the results: the amount of water fits to the noodles, or the gift is wrapped properly. In scientific research on estimation, where e.g. written estimation tests were used, the quality of an estimation result cannot be derived from any contextual embedding that easy. Therefore, mathematics education researchers refer to the estimation accuracy: the lower the deviation from the real value, the higher the quality of an estimation result (Corle, 1963; Desli & Giakoumi, 2017; ; Heid, 2018; Hildreth, 1980Hogan & Brezinski, 2003; Hoth et al., 2019; in press; Joram et al., 2005; Siegel et al., 1982; Swan & Jones, 1980). Even if mathematics education researchers are more interested in the estimation process itself and therefore in the use of strategies, they also look for their efficiency and whether the estimation (results and processes) can be described as good or not (Bright, 1976; Desli & Giakoumi, 2017; Forrester et al., 1990; Heid, 2018; Hildreth, 1980; 1983; Jones et al., 2012; Joram et al., 1998; Siegel et al., 1982). Furthermore, estimation accuracy was used to describe the development of estimation performance (Harel et al., 2007), or the effect of different task characteristics (Desli & Giakoumi, 2017; Forrester & Shire, 1994; Weiher, 2019).

In these studies, the *description* of the estimation error is made consistently by using the percentage deviation from the real value, while the *evaluation* of the estimation error is often made differently. This leads to a variety of different scorings with different characteristics regarding the number and limits of the intervals. Consequently, the meanings of 'accuracy' vary, as do the terms used. In the literature dealing with the adequacy of estimated values, most researchers used 'accuracy' (e.g. Huang, 2014, Siegel et al., 1982, Swan & Jones, 1980). Huang (2014) also used 'acceptability', a term that already includes the scope for decision making by the researcher. Other researchers used 'reasonable estimates' (Desli & Giakoumi, 2017) or 'reasonableness' (Siegel et al., 1982) as well, but even these terms differ in their meaning. Whereas Desli and Giakoumi (2017) used it as a more suitable term for 'accuracy', Siegel et al. (1982) called comprehensible estimations 'reasonable'. Siegel et al. (1982) mentioned that accuracy and reasonability are not the same constructs. An estimation result can be reasonable, but not accurate: "if a cracker is 2 inches long, an estimate of 3 ½ inches might be inaccurate; but it is certainly not unreasonable" (Siegel et al., 1982, p. 217).

Nonetheless, most studies do not provide an explanation for the selection of thescoring they used. They neither derive it theoretically based on the aim or characteristics of their study, nor are there empirical derivations, e.g. a

standard sample (such as in psychology, e.g., Axelrod & Millis, 1994; Brand et al., 2003; Bullard et al., 2004; Della Sala et al., 2003; Mendez et al., 1998, Shallice & Evans, 1978). Therefore, the quality of an estimation result is not that easy to evaluate, even when using accuracy in terms of the percentage deviation from the real value as the indicator. Furthermore, it remains unclear, if the results of different studies that used different scorings are comparable, both nationally and internationally.

In the context of developing a written estimation test for length, area, capacity, and volume, the purpose of this article is to compare and discuss the effects of different scorings on the test performance (amount of achieved points and rank order of students) and the test quality (internal consistency and discriminatory power) in order to contribute to a transparent selection and justification when choosing approaches for the evaluation of measurement estimation results. Therefore, the classic evaluation of estimation accuracy using the percentage deviation in combination with a scoring is presented. The description is followed by a theoretical discussion of different scorings with regard to their characteristics. Since the scorings were quite different, their impact on test performance and test quality was investigated by applying six of them to the same data set using a estimation test including length, area, capacity and volume estimation tasks[1].

2 Evaluation of estimation accuracy: Overview of different scorings in mathematics education literature

To describe and evaluate the accuracy of an estimate, its deviation from the real value has to be determined. In mathematics education, it is customary to compute the percentage deviation from the real value. Based on this, different scorings with various intervals of the percentage deviation are used to assign points to the estimate. This chapter first presents the computation of the percentage deviation, followed by the description and theoretical analysis of different scorings used in mathematics educational research.

2.1 Percentage deviation from the real value

Since the 1960s, the *percentage deviation from the real value* was used as a measure for the estimation ability (Joram et al., 1998, 2005). Equation (1) is the *percentage deviation from the real value* D_r, with the estimated value e and the real value r:

$$D_r = \frac{e - r}{r} \cdot 100\,\% \tag{1}$$

If the normalized deviation from the real value D_r is specified in percent and therefore multiplied by 100 %, it is called the percentage deviation of the real value. This can be illustrated by the following example in equation (2): A chocolate bar has a length of 11 cm. The estimation result 15 cm leads to an approximate percentage deviation D_r of +36.4 % (which is an overestimation since the estimated value is higher than the real value and therefore, the percentage deviation is positive):

$$\frac{15\ cm - 11\ cm}{11\ cm} \cdot 100\,\% \approx 36.4\,\% \tag{2}$$

[1] Although it is conceivable that there are differences in the estimation accuracy of different variables (Heid, 2018) or task types (Weiher, 2019), all measures and task types are considered together in this article. The reason for this is to find a scoring that is suitable for an estimation test that includes all measures and task types. Results that include the comparison of the four measures or different task types with respect to estimation accuracy is not the aim of this article.

Estimating 4 cm for the same chocolate bar results in an approximate percentage deviation D_r of –63.7 % as can be seen in equation (3) (which is an underestimation since the estimated value is lower than the real value and therefore, the percentage deviation is negative):

$$\frac{4\ cm - 11\ cm}{11\ cm} \cdot 100\ \% \approx -63.7\ \% \tag{3}$$

The range of possible values for the percentage deviation D_r is between $-100\ \% < D_r < \infty\ \%$. The example of the chocolate bar can be used to illustrate this: With a length of 11 cm, any estimated value greater than 11 cm is an overestimate and results in a positive percentage deviation. An overestimation is theoretically infinitely possible, so that the positive percentage deviation can assume values up to ∞ %. In contrast, a value of –100 % can only be achieved theoretically or by rounding, because the estimated value would have to be 0 cm, which is not a meaningful specification for any object.

If over- and underestimation should be treated equally, which means that the same scoring is used for over- and underestimations, the *absolute value of the percentage deviation from the real value* can be used. The range of possible values is between $0\ \% < |D_r| < \infty\ \%$. Again, the example of the chocolate bar can illustrate this: Estimated values of 12.1 cm (overestimation) and 9.9 cm (underestimation) both result in an absolute value of the percentage deviation from the real value of 10 %.

In all equations (1) to (3), 0 % deviation represents an estimated value that corresponds to the real (measured) value.

2.2 Scorings based on the percentage deviation from the real value

In mathematics education, to evaluate an estimation result, first the percentage deviation from the real value is calculated[2] and then a scoring is assigned based on this. Table 1 gives an overview of the scorings used in different studies[3]. Although all scorings were based on the absolute value of the percentage deviation from the real value, the sample, test material and measures they contained differed as well as the scorings themselves.

Table 1: Scorings in measurement estimation studies based on the absolute percentage deviation from the real value $|D_r|$ (sorted in time from old to new).

Study	Sample	Measure	Scoring
Corle 1963	Primary school teachers, college students	length, weight, time, capacity, volume, temperature	Nearly accurate, correct: $\|D_r\| \leq 10\ \%$ Not named, but investigated: $10\ \% < \|D_r\| \leq 100\ \%$
Hildreth 1980	Students age 10, 12 and 18	length, area	Number of items on which the relative error was less than 1/3
Swan & Jones 1980	4th–12th grade; college students	length, mass, temperature	1 point: $\|D_r\| < 25\ \%$
Siegel et al. 1982	2nd–8th grade	length, (number)	Accurate: $\|D_r\| \leq 50\ \%$

[2] The formulas of the calculation of the percentage deviation are only mentioned in rare cases. Often there is only a description of the calculation, which more or less directly indicates this type of calculation.
[3] The studies in the table are studies known to the authors after careful literature research. It cannot be ruled out that there may be further scorings.

			Reasonable: $\lvert D_r \rvert$ within the order of magnitude
Hogan & Brezinski 2003	College students	length, weight, volume[1], time	3 points: $\lvert D_r \rvert \leq 10$ % 2 points: 10 % < $\lvert D_r \rvert \leq 20$ % 1 point: 20 % < $\lvert D_r \rvert \leq 30$ %
Della Sala et al. 2003[2]	Adults	weight, area, length, time, speed, capacity, (number)	0 points: $\lvert D_r \rvert \leq 30$ % 1 point: 30 % < $\lvert D_r \rvert \leq 90$ % 2 points: $\lvert D_r \rvert > 90$ % (inverse allocation of points)
Huang 2014	4th–6th grade	length, area	Accurate, 2 points: $\lvert D_r \rvert \leq 10$ % Acceptable, 1 point: 10 % < $\lvert D_r \rvert \leq 25$ %
Desli & Giakoumi 2017	3rd and 5th grade	length	Reasonable, 1 point: $\lvert D_r \rvert \leq 30$ %
Heid 2018	4th grade	length capacity	1 point: $\lvert D_r \rvert \leq 25$ % 1 point: $\lvert D_r \rvert \leq 50$ %
Hoth et al. 2019; in press	4th grade	length	3 points: $\lvert D_r \rvert \leq 10$ % 2 points: 10 % < $\lvert D_r \rvert \leq 25$ % 1 point: 25 % < $\lvert D_r \rvert \leq 50$ %

[1]Volume estimation is assessed differently here than in other studies. As the example given in the paper shows ("How many pennies would it take to fill this one-quart container [show container]?" (Hogan & Brezinski, 2003, p. 271)), it is more about estimating numbers as a relation between different volumes.
[2]This psychological study focuses on executive functions when estimating. Therefore, different measures and the estimation of numbers are included in the estimation test without any differentiation between the types of estimates or measures.

2.3 Analyzing the scorings

When looking closer to the scorings in the literature, a first rough distinction can be made between *dichotomous* (Desli & Giakoumi, 2017; Heid, 2018; Hildreth, 1980; Swan & Jones, 1980) and *polytomous* scorings, using three (Corle, 1963; Huang, 2014; Siegel et al., 1982) or four intervals (Hogan & Brezinski, 2003; Hoth et al., 2019; in press).

In all studies, the *boundaries for the intervals* differed, regardless from the number of intervals included in the evaluation process. The dichotomous studies used 25 %, 30 %, 33 %, or 50 % as their limit of accuracy. The studies with a polytomous scoring used not only different boundaries – from 10 % to 100 % –, but also different *bandwidths* for their intervals. Whereas Hogan and Brezinski (2003) used 10 % increments, all others used different increments in their studies (table 1).

If the scorings should separate the extreme answers from the rest in the middle, two scorings can be used from the literature. When scoring according to Corle (1963), the most accurate estimation results (10 % deviation) are distinguished from the moderate (10 - 100 % deviation) and the least precise (> 100 % deviation) by points. Della Sala et al. (2003), on the other hand, see the first interval up to a deviation of 30 %, i.e. much less strict than Corle (1963), whereas they evaluate the middle interval only up to a deviation of 90 %.

To illustrate the impact of the different scorings, the example of the chocolate bar with the length of 11 cm is used again. An estimate of 13 cm (approximate percentage deviation from the real value of 18.2 %) results in 1 point

using the scoring of Huang (2014) and 2 points using the scoring of Hogan and Brezinski (2003) and Hoth et al. (2019; in press). Huang (2014) would describe that estimation result as *acceptable*, Siegel et al. (1982) would use the term *accurate* (in contrast to Huang (2014), who used this term for estimates within a deviation of 10 %) and Desli and Giakoumi (2017) would name that estimate result *reasonable*. If one estimated the length of the chocolate bar at 15 cm (approximate percentage deviation of 27.2 %), this leads to 0 points according to Huang (2014), whereas Desli and Giakoumi (2017) and Hogan and Brezinski (2003) would give 1 point. According to Siegel et al. (1982), this estimate is still *accurate*, whereas Huang (2014) would describe it as *no longer acceptable*.

Although some of the scorings are from studies with adults (Corle, 1963; Della Sala et al., 2003; Hogan & Brezinski, 2003), no clear trend toward the use of more strict scorings in adults can be seen here. This is already evident when comparing the two studies that include only adult subjects: whereas the scoring of Hogan and Brezinski (2003) is relatively strict, those of Della Sala et al. (2003) and Corle (1963) are not strict at all. Furthermore, some studies use the same scoring for different, rather more widely spreaded age groups (Hildreth, 1980; Siegel et al., 1982; Swan & Jones, 1980). Finally, even if only children's estimation accuracy was scored, this was done more strictly (Desli & Giakoumi, 2017; Heid, 2018 for lengths, Huang, 2014) or less strictly (Heid 2018, for volume; Hoth et al., 2019, Siegel et al., 1982).

3 Research Questions

For the analysis in this paper, six scorings were used (table 2). Scorings 1 to 5 were derived from the literature (table 1). Scoring 1 was also used by Swan and Jones (1980) and Heid (2018). Scoring 2 derived from the boundary "accurate" from Siegel et al. (1982) as well as from Heid (2018). Scoring 3 was used by Della Sala et al. (2003), whereas the points have been inverted to match the other scorings. Scoring 4 corresponds to the scoring used by Hogan and Brezinski (2003). Scoring 5 is the same used by Hoth et al. (2019; in press). The scorings with boundary 30 % (Desli & Giakoumi (2017) and "relative error less than 1/3" (Hildreth, 1980) were not selected because of their similarity to the 25 % boundary and because the 30 % boundary is already represented in scoring 3 and 4. Scoring 6 as an additional one was created independently from the literature in order to investigate a scoring with more and equidistant boundaries for comparison purpose. All scorings relied on the absolute value of percentage deviation from the real value.

Table 2: Scorings based on the absolute value of percentage deviation $|D_r|$ applied on the data of this study.

Scoring 1	Scoring 2	Scoring 3	Scoring 4	Scoring 5	Scoring 6
1 point:	1 point:	2 points:	3 points:	3 points:	10 points:
$\|D_r\| \leq 25$ %	$\|D_r\| \leq 50$ %	$\|D_r\| \leq 30$ %	$\|D_r\| \leq 10$ %	$\|D_r\| \leq 10$ %	$\|D_r\| \leq 10$%
		1 point:	2 points:	2 points:	9 points:
		30 % $< \|D_r\| \leq 90$ %	10 % $< \|D_r\| \leq 20$ %	10 % $< \|D_r\| \leq 25$ %	10% $< \|D_r\| \leq 20$%
			1 point:	1 point:	8 points:
			20 % $< \|D_r\| \leq 30$ %	25% $< \|D_r\| \leq 50$%	20% $< \|D_r\| \leq 30$%
					…
					1 point:
					90% $< \|D_r\| \leq 100$%

Scoring 1 and 2 are dichotomous, whereas the others are polytomous with three (scoring 3), four (scoring 4 and 5) and eleven (scoring 6) intervals. The polytomous scorings also differ in their bandwidth: Scoring 4 and 6 use equidistant intervals, whereas scoring 3 and 5 show different bandwidths.

Furthermore, scoring 1 and 4 can be characterized as strict, whereas especially scoring 3 seems to evaluate the results rather smoothly. Scoring 6 also gives points for higher deviations, but takes a much more differentiated approach.

Since all scorings differ in various aspects, no information about the suitability of a scoring for a written estimation test can be derived from the literature. The following research questions will be answered according to the aim of the article:

1) To what extent do the scorings differ in terms of test performance, namely a) the total test score and b) the rank order of students?
2) To what extent do the scorings differ in terms of test quality, especially a) internal consistency and b) discriminatory power?

Milder boundaries result in higher scores. Due to the combination of different number of intervals and different stringency of intervals, it is suspected that the mean test performance is lowest for scoring 1, and highest for scoring 3. Due to the complexity of the differences, it is not possible to make a differentiated statement regarding the ranking of the other scorings. It is presumed that the ranking of the children remains essentially the same, since the scores are all based on the absolute value of the percentage deviation.

It is assumed that internal consistency and discriminatory power are highest for scoring 6, since the higher number of intervals contributes to a higher differentiation of test performance. Lowest internal consistency and discriminatory power is expected for scoring 1, since presumably many children do not receive any point at all and thus no differentiation is made between the children.

4 Material and methods

In this section, information about the sample of this study, the used estimation test, the treatment of missing data and the statistical methods were presented.

4.1 Sample

The study took place at various school types in northern Germany. In total, 615 students from 13 fifth and 13 sixth classes took part in the estimation test. The mean age was 11.1 years (SD = 0.53) in fifth grade and 12.2 years (SD = 0.57) in sixth grade.

Two versions of an estimation test (section 4.2) were given out to the children in turns. The allocation of the test books thus depended on the seating arrangement in the classrooms. Altogether, 310 students worked on test book A, and 305 students worked on test book B (table 3).

Table 3: Sample sizes per grade and test book.

	Test book A	Test book B
All (f/m)	310 (158/151)[1]	305 (163/142)
Grade 5 (f/m)	155 (79/76)	154 (89/65)
Grade 6 (f/m)	155 (79/75)[1]	151 (74/77)

[1] One value is missing because one child did not specify a gender.

4.2 Estimation Test

The estimation test contained a total of 48 test items, 12 items for each visually perceptible measure (length, area, capacity, volume). An estimation task can include three objects: the to-be-estimated-object (TBEO), an object representing a standardized unit, and an object representing a benchmark (Weiher, 2019). The TBEO must always be named in the estimation task, while objects for the unit or a benchmark do not necessarily have to be named. Whether an object is considered a representative for the unit or a benchmark depends on its size specification: an object with size 1 in the corresponding unit is to be understood as a representative for the unit, while all other size specifications lead to the object being a representative for a benchmark.

By combining the three objects with different characteristics (visibility, touchability), different task types are created (Weiher, 2019), of which three task types were selected for this estimation test (table 4). The characteristic *visible* means that the object (TBEO or unit) can be viewed, but not touched. The visible TBEOs and the visible units were placed in the front of the classroom. *Non-visible* objects are not present in the classroom. Each task type was represented by four items per measure in the test.[4]

The sizes of the TBEOs ranged between 1 cm and 269 m for lengths, 1 ml and 240 l for capacity, 1 cm^2 and 500 m^2 for area and 1 cm^3 and 140 m^3 to 260 m^3 for volume[5].

Table 4: Example items and their characteristics (task types) per measure.

	Length	Area	Capacity	Volume
Task type 1: TBEO and unit not visible	Approximately how long is a balance beam? A balance beam is about __m long.	Approximately how big is an unfolded handkerchief? An unfolded handkerchief is about __cm^2 in size.	Approximately how much water fits into a yellow plastic garbage bag? About __l of water fit into a yellow plastic garbage bag.	What is the approximate size of a shipping container? A shipping container is about __m^3 in size.
Task type 2: TBEO visible, unit not visible	Approximately how long is the red stripe on the board? The red stripe on the blackboard is about ___cm long.	Approximately how big is the calendar on the board? The calendar on the blackboard is about __cm^2 in size.	Approximately how much water fits in the shower gel bottle on the board? About __ml of water fits into the shower gel bottle on the board.	Approximately how big is the game pack on the board? The game pack on the board is about __cm^3 in size.
Task type 3: TBEO not visible, unit visible	The yellow ribbon on the board is 1 cm long. What is the approximate length of the long side of a 5 € bill? The long side of a 5 € bill is about __cm long.	The brown square on the board is 1 m^2 in size. What is the approximate size of a ping-pong table? A ping-pong table is about __m^2 in size.	1 ml of water fits into the green bottle on the board. How much water fits into a small tea light? About __ml of water fits into a small tea light.	The cube on the board is 1 cm^3 in size. What is the approximate size of a packet of butter? A packet of butter is about __cm^3 in size.

[4] Two parallelized test forms were developed. They contained the same number of items per measure and the same task characteristics, but the not visible TBEOs (task type 1 and 3) were switched: The TBEOs from task type 1 in version A became the TBEOs of task type 3 in version B and the other way round. As an example: in test version A, the balance beam is TBEO in task type 1, whereas in test version B, it is TBEO in task type 3.The visible TBEOs (task type 2) were the same in both test versions.

[5] The largest real value of volume is the size of the classroom in which the test took place. Since all classes performed the test in their own room, the value varies and the range of classroom sizes is given here.

4.3 Treatment of missing data

From 29520 possible answers (615 students, 48 estimation tasks), 24757 answers were used as data base. The main reason for missing data is that students did not work on the task at all. In addition, answers were deleted that, for various reasons, suggested that no estimation process had taken place. This includes e.g. answers with an additional unit (beside the one given in the test booklet), when this unit does not correspond to the measure (e.g. 5 cm for the estimation of area), when students gave answers that contain two or more commas, or when the estimation result is obviously not the result of a serious estimation process (results like filling in zeros until the end of the page). Moreover, based on the real size of the estimation objects, it is assumed that estimates < .1 and > 10000 are not results of a valid estimation process.

4.4 Statistical Analyses

The raw data in the form of the results of the paper-and-pencil test were transferred to SPSS 26 for further processing.

The overview of the data contains a *characterization of the 24757 estimated values* (raw values) and the amount of exact, over- and underestimates is determined on the basis of the *percentage deviation*. Furthermore, based on the *absolute value of the percentage deviation*, the distribution of the data over the *descriptive statistics* and the *percentiles* is considered.

Afterwards, the six scorings (table 2) were applied to the data. Due to the fact, that the absolute numbers of points that could be achieved in the scorings differed, *the relative amount of points achieved was calculated* and used for the following steps of the analyses.

To address research question 1a, *descriptive statistics for each scoring* are presented first. The significance of the observed differences was tested using a *mixed repeated measurement analysis of variance (mixed rmANOVA) and post-hoc analysis*, with the scoring as the repeated within-subject variable and the test booklet as between subject variable. The results are validated with the *Friedmann test*. *Spearman rank correlation analysis* is used for research question 1b, in order to determine rank order differences between the scorings.

To address research question 2a, *Cronbach's alpha* is used as a measure of internal consistency. In addition, *the discriminatory power* is calculated in regard to research question 2b. Both of them, internal consistency and discriminatory power, are considered to be indicators of the quality of a test instrument in the context of classical test theory during test development (Moosbrugger & Kelava, 2012) and are therefore of importance in this study.

5 Results

First, an overview of the data is presented including the amount of exact, of under- and of overestimations and percentiles of absolute values of percentage deviation. Second, after the descriptive statistics of data using the scorings, the results of the mixed rmANOVA for testing the effects of different scorings on the total test score (research question 1a) and spearman correlation coefficient in regard to the rank order of students (research question 1b) were presented. Third, the results of the investigation of internal consistency (research question 2a) and discriminatory power (research question 2b) of the six scorings were reported.

5.1 Overview of the data

Initial observations of the data show that they have a discrete character (93.5 % are natural numbers). More interesting is the observation that students tended to give their estimates as multiples of 10 – or 5, respectively. The

chocolate bar of 11 cm length can serve as an example. Although students' estimates ranged from 1 cm to 100 cm, 26.6 % of the students answered 10 cm. The second most common answer was 5 cm (12.0 %), followed by 15 cm (8.4 %). 2.5 % of the students gave the real value 11 cm as their estimated value.

The percentage deviation from the real value shows that – according to the sign – students tended to underestimate: 68.4 % of all answers were underestimations, 23.2 % were overestimations, and 8.4 % of the estimates pinpointed the real value.

Afterwards, the absolute values of the percentage deviation were cumulated up to the different boundaries from the literature (table 5) to get an impression of how the available data is distributed within these boundaries.

Table 5: Amount of answers (in %) per criterion.

Absolute percentage deviation in % ≤	10	20	25	30	33	40	50	60	70	75	80	90	100
Amount of answers (in %)	12	18	21	22	24	29	35	41	46	50	54	63	90

Firstly, approximately 10 % of all answers showed a greater deviation than 100 %. These answers included only overestimations because the underestimation scale is limited to 100 %. On average, the most accurate estimates showed a similar order of magnitude: 12 % of all answers were within the 10 % criterion. Secondly, it becomes visible that there is hardly any difference between 30 % and 33 % in terms of the amount of responses; only 2 % of responses are added if the 33 % criterion is used instead of the 30 % criterion. The largest increase appears to be between the 90 % and 100 % criterion, in which 27 % of the responses lie.

The distribution of the answers (table 5) can be used to draw conclusions about the number of answers that would not receive any point. Using hard scorings as scoring 1 and 4, responses with a deviation greater than 30 % (or even stricter) were rated with 0 points. Based on the data of this study, only about 22 % of the answers would receive any point. Scoring 2 and 5 used the 50 % criterion as the strictest boundary, so about 35 % of the answers would get a point. Using rather smooth scorings, as scoring 3 and 6, results in evaluating 60 % to 90 % of the answers in this study with at least one point. However, in scoring 6, the number of points differs more than in other scorings, e.g. an answer with less than 10 % deviation is rated 10 times as much as an answer between 80 % and 90 % deviation. Using scoring 3, the better answer is only rated with twice as many points as the worse one.

5.2 Differences between the scorings concerning the test performance

Total test score. Table 6 shows the descriptive statistics for each scoring. Since the point totals differ between the scorings, the relative amount of achieved points is used for comparison. Performance was the highest using scoring 3 (M = 35.8 %, SD = 10.43 %), and decreased by using scoring 6 (M = 34.08 %, SD = 9.47 %), scoring 2 (M = 29.75 %, SD = 10.24 %), scoring 5 (M = 19.17 %, SD = 7.18 %), scoring 1 (M = 17.82 %, SD = 7.94 %), and scoring 4 (M = 14.62 %, SD = 6.36 %). It is also noteworthy that even in the scoring, with which the most points were achieved, only 67.71 % of the points were scored at most (scoring 3). The minimum points achieved vary between all scorings from 1.39 % to 5.21 %, which is the same order as for the mean values.

Table 6: Descriptive statistics per scoring (amount of achieved points in %).

	Minimum (in %)	Maximum (in %)	Arithmetic mean (in %)	SD (in %)
Scoring 1	2.08	45.83	17.82	7.94
Scoring 2	4.17	62.5	29.75	10.24

Scoring 3	5.21	67.71	35.8	10.43
Scoring 4	1.39	36.11	14.62	6.36
Scoring 5	3.47	41.67	19.17	7.18
Scoring 6	4.38	61.67	34.08	9.47

The boxplots (figure 1) illustrate the differences between the scorings. It can also be seen that there are only mild outliers in the data.

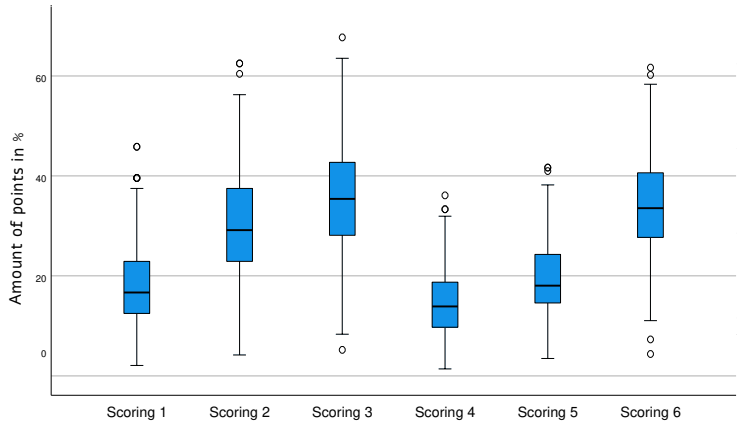


Figure 1: Differences between the scorings concerning the amount of achieved points in %.

The mixed rmANOVA[6] with a Greenhouse-Geisser correction does not result in a significant interaction effect between the scoring and the test booklet ($F(1.908, 1169.720) = .433$, $p = .639$). This indicates that differences between scorings were the same in both test books. Therefore, for the following analysis, both test booklets were considered together. The main effect scoring becomes significant ($F(1.91, 1172.06) = 5374.44$, $p < .001$), what means performance level differ statistically significant between the scorings. Bonferroni-adjusted post-hoc analysis revealed a significant difference ($p < .001$) in performance between all pairs of scorings (table 7). The Friedman-test carried out for comparison confirmed these results.

Table 7: Pairwise comparison with bonferroni-corrected post-hoc analysis.

Scoring	Scoring	Mean difference	Standard error	95% CI
1	2	-.119*	.002	-.126/-.113
1	3	-.180*	.002	-.186/-.173
1	4	.032*	.001	.029/.035
1	5	-.013*	.001	-.016/-.011
1	6	-.163*	.002	-.168/-.157
2	3	-.061*	.002	-.066/-.055
2	4	.151*	.002	.144/.158
2	5	.106*	.002	.101/.111
2	6	-.043*	.001	-.047/-.039

[6] As a prerequisite for the mixed rmANOVA, the data were tested for normal distribution. The data were only normally distributed if using scoring 3 and 6 ($p > .001$) as assessed by the Shapiro-Wilk-test. Because of the large sample size of N = 615 which is well over N = 30 and because rmANOVA reacts robustly to the violation of this requirement (Schmider et al., 2010; Blanca et al., 2017), it was nevertheless carried out. Furthermore, there was homogeneity of the error variances, as assessed by Levene's test ($p > .05$), and there was homogeneity of covariances, as assessed by Box's test ($p = .947$).

3	4	.212*	.002	.205/.219
3	5	.166*	.002	.161/.172
3	6	.017*	.001	.015/.020
4	5	-.045*	.001	-.048/-.043
4	6	-.195*	.002	-.201-.188
5	6	-.149*	.002	-.154/-.145

* $p < .001$

Rank order. There is little variation in students' ranks between the scorings, as can be inferred from the high Spearman rank correlation coefficients[7] (table 8). All correlation coefficients can be interpreted as high according to Cohen (1988) and indicate that the rank order of the students is nearly the same for all scorings.

Table 8: Correlation coefficients of the Spearman rank correlation analysis.

Scoring	1	2	3	4	5
2	.851*				
3	.857*	.900*			
4	.974*	.831*	.844*		
5	.962*	.941*	.908*	.963*	
6	.853*	.939*	.979*	.848*	.928*

*$p < .01$

The highest correlations was found between scoring 6 and 3 ($r = .979$, $p < .01$) and between scoring 1 and 4 ($r = .974$, $p < .01$). The scatterplot (figure 2) illustrate the correlation between scoring 6 and scoring 3.

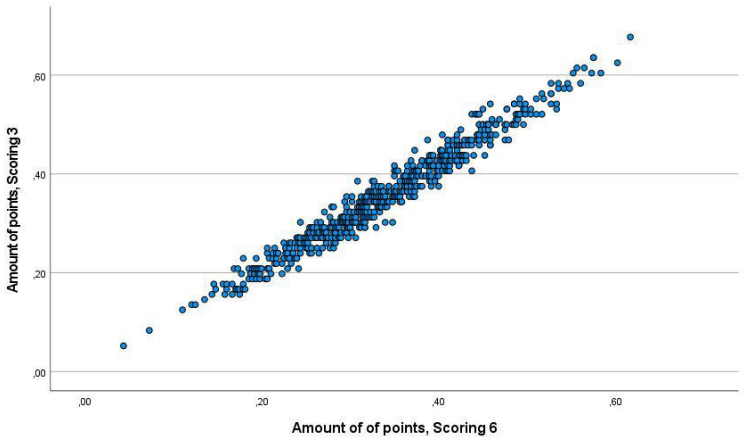


Figure 2: Scatter plot of the relation between scoring 3 (y-axis) and scoring 6 (x-axis).

Lowest correlation was found between scoring 4 and 2 ($r = .831$, $p < .01$), but can be nevertheless be classified as high according to Cohen (1988). The scatterplot (figure 3) illustrates the correlation.

[7] Since the proportion of ties is very high with all six scorings, Kendall-tau-b correlation analysis could be an alternative for calculating the correlation between the scorings. The correlation coefficients are slightly lower, but still they can be classified as high (between $r = .662$ and $r = .898$, all $p < .01$) according to Cohen (1988). The order of scoring pairings is almost identical, scoring 1 and 4 have the highest correlation ($r = .898$, $p < .01$), followed by scoring 3 and 6 ($r = .891$, $p < .01$). Scoring 2 and 4 correlates lowest ($r = .898$, $p < .01$). Since over-estimation of the correlation coefficient due to the ties can be understood here as a more cautious approach (fewer differences between the rank order), the Spearman correlation coefficient is reported in detail.

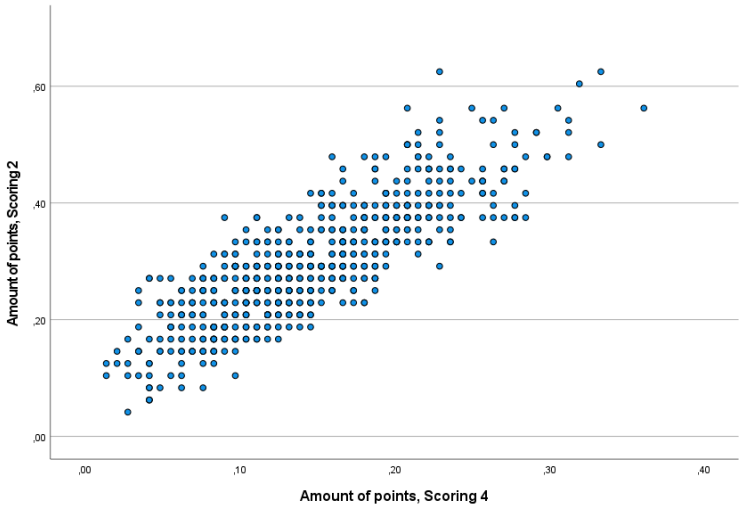


Figure 3: Scatter plot of the relation between scoring 2 (y-axis) and scoring 4 (x-axis).

However, the proportion of tied ranks (equal ranks) should be noted. Whereas in scoring 1 all ranks are tied, in scoring 6, 91.8 % of the ranks are tied. The percentage of tied ranks of the other scorings lies between these two: 98.8 % (scoring 3), 99.2 % (scoring 5), 99.4 % (scoring 4), and 99.6 % (scoring 2). However, the number of different ranks must also be considered as can be illustrated with the following example: For scoring 1, 20 ranks exist, whereas for scoring 6, there are 180 ranks. Thus, although scoring 6 has also a high percentage of tied ranks, fewer students shared the same single rank as in scoring 1. This can be seen in the scatterplot (figure 4): The data from scoring 4 (shown on the y-axis) form horizontal "lines". Vertical lines (for scoring 6, shown on the x-axis) cannot be seen because there are fewer cases per rank. The scatter plot that showed the correlation between scoring 2 and 4 (figure 3) illustrates a high number of cases per rank for both scorings, as indicated by visible "lines" in both the horizontal and vertical directions.

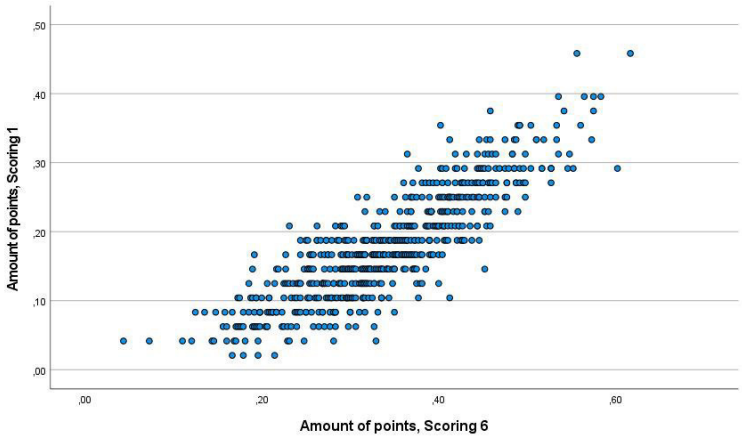


*Figure 4: Scatter plot of the relation between scoring 1 (y-axis) and scoring 6 (x-axis)***.**

5.3 Differences between the scorings concerning the test quality

Internal consistency. Because the individual items were used to calculate the internal consistency and these differed slightly between the test booklets, they are considered separately. Nevertheless, concerning the total test scores, the internal consistency is quite similar in both test books (table 9).

Table 9: Cronbach's Alpha for each scoring and both test books.

Scoring	Cronbach's Alpha	
	Test book A	Test book B
Scoring 1	.620	.638

Scoring 2	.715	.732
Scoring 3	.816	.822
Scoring 4	.587	.617
Scoring 5	.692	.707
Scoring 6	.817	.828

Highest internal consistency will be achieved by using rather smooth scorings: scoring 6, followed by scoring 3 result in an overall good internal consistency according to Cortina (1993), who describes a internal consistency of .7 as a measure of high internal consistency (whereas he not intended to create a cut-off score for internal consistency). Both scorings 2 and 5 seem to be acceptable concerning internal consistency. The rather hard scorings 1 and 4 do not seem to be that suitable, both scorings show rather low internal consistency.

Discriminatory power. In order to discuss the scorings with regard to their suitability for a written test, the discriminatory power was examined as an essential feature of the test development. Overall, the discriminatory power was rather low, independent from the scoring used. Since the discriminatory power is comparable in both test books, only the results from test book A are presented here (figure 5).

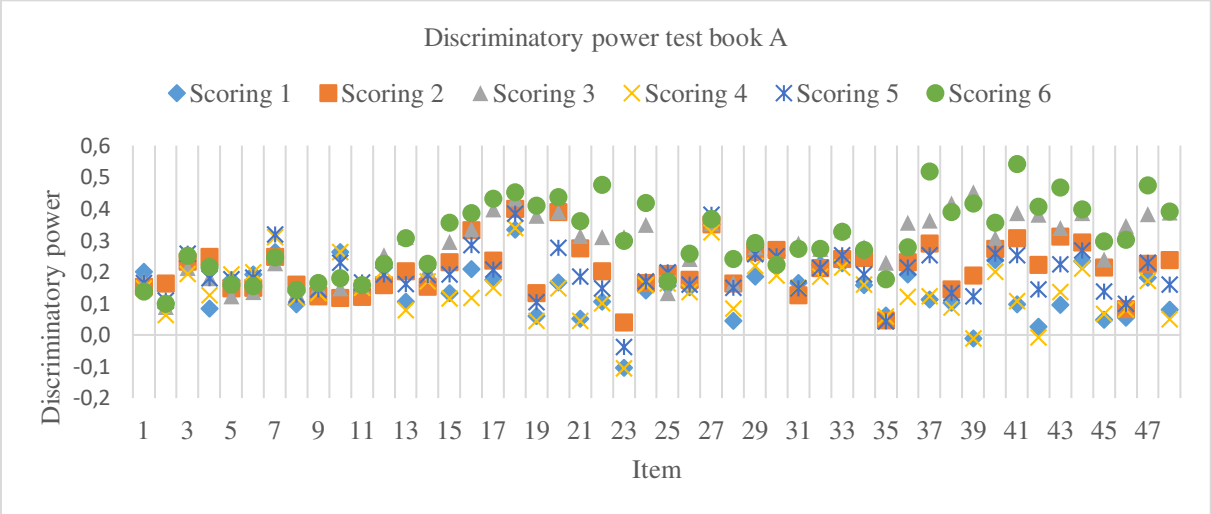


Figure 5: Discriminatory power in test book A.

Comparing the scorings with regard to their discriminatory power, the same groups and order could be observed as when considering the internal consistency. Scoring 6 and 3 seem to have the highest discriminatory power in both test books. The discriminatory power is very low for scorings 1, 4, and 5, with even negative values using scoring 4 or 5.

6 Discussion

The following discussion is based on the results of the application of different scorings and the characteristics of the underlying data. Therefore, first the suitability of the scorings for scientific studies is discussed, then the influence of over- and underestimation and the discrete character of the data is illuminated, before some limitations of the study are presented.

6.1 Differences of the scorings in terms of test performance

Total test score. All scorings that were applied to the data differed significantly from each other in terms of test performance. Overall, it can be stated: The smoother the scoring, the better the performance. However, the question

of how mild or hard a scoring is cannot be answered so simply; both the number of intervals and the boundaries of the intervals must always be considered, as the following examples show. The comparison of scoring 2 and 5, both using the deviation of 50 % for at least one point, shows that, although the total test score will be higher in using scoring 5 – more points can be achieved for better estimates – it is harder in the mean test performance: With a deviation of e.g. 50 % a student will already get the maximum number of points when scoring 2 is used. In our sample, 35 % of the results got the maximum number of points, whereas only 12 % got the maximum number of points by using scoring 5. Furthermore, it is more difficult to compare scorings if one is harder in one aspect and the other is harder in another one. Looking to scoring 3 and 6 will serve as an example: Scoring 6 uses with 100 % deviation a smoother boundary for one point, but scoring 3 only has three intervals. Although in the first case 90 % of our sample got 1 point, in contrast to 63 % in the second case, the average performance is more affected by the number of intervals and the hardness of these intervals in scoring 6.

With the variety of different scorings and due to the result that all scorings differ significantly in terms of the amount of points achieved, it becomes clear that the choice of a scoring need to be carefully thought through. The aim of the study should be the decisive factor: If the best estimators are to be separated from the others, a hard scoring is required. To identify the weakest estimators, a more lenient scoring should be used. In order to represent the estimation accuracy as broadly as possible, it seems suitable to use several intervals, whereas a differentiation of the intervals better than 10 % deviation and worse than 100 % deviation does not seem to be necessary. This view is shared by the literature: none of the scorings differentiated estimates better than 10 % deviation and worse than 100 % deviation.

Besides the general discussion of the suitability of a scoring for a specific aim, it can be discussed whether different scorings should be used for different measures (length, area, capacity, volume), task types, or target groups (children, adults). Heid (2018) used a stricter scoring for the evaluation of length estimates than for the evaluation of capacity estimates due to the assumption that estimating capacity is more difficult. This approach seems fair to the estimator's performance, but does not seem useful, for example, for comparing estimation accuracy for length and volume. In scientific contexts, the use of different scorings could lower the validity of e.g. comparisons. For example, the expected higher estimation accuracy when estimating lengths compared to estimating volumes might be less or no longer visible due to stricter scoring. Uniform scoring is also required to examine differences between task types for the same reasons. In the school context, on the other hand, using a milder scoring for supposedly more difficult estimation tasks can contribute to fairness. Here, however, it would first be necessary to clarify what differences actually exist between, for example, the measures or the task types before selecting different scorings. The differences between the scorings (e.g. the number of intervals or the boundaries) should not be determined by feeling, but should at least correspond approximately to the differences between the measures or task types. The extent to which e.g. one measure is harder to estimate than another should therefore be reflected equally in the hardness of one scoring relative to another. However, this problem is currently unsolved and cannot be answered in the context of this study. Furthermore, as described above, it is not possible to derive from the literature which scoring characteristics are used or appropriate for which age group. In addition to the different scorings, the instruments used to assess estimation accuracy also vary, and thus no systematically derived statement can be made as to what estimation accuracy can be expected at what age. Consequently, no basis exists on which to justify selecting a stricter or milder scoring for a particular age group.

In principle, it should also be said that the discussion of a scoring primarily relates to scientific purposes, since accuracy itself plays a subordinate role in school contexts. In mathematics classrooms, precision itself is normally not the most interesting point. The focus is laid on the use of strategies (Desli & Giakoumi, 2017; Friebel, 1967; Heid, 2018; Hildreth, 1980; 1983; Joram et al., 1998; Siegel et al., 1982): if the estimation result can be explained by referring to suitable benchmarks and appropriate strategies, it is seen as a reasonable estimation process. Nevertheless, the question of how *estimation ability* is assessed also arises in the school context. If estimation ability is assessed in examinations, this can be done via *estimation accuracy* (in which case the question of scoring is also relevant in school) or via other formats, such as describing estimation strategies.

Rank order. Since the correlation coefficients between all pairs of scorings are classified as high, it is assumed that the choice of a scoring has little or no effect on the ranking among the children in terms of their test performance. Since all scorings are based on the absolute value of the percentage deviation from the real value, this result was expected and indicates that for correlation and regression analyses all scorings lead to similar results.

The high number of ties or the high number of cases per rank for some scorings (for those with few intervals) leads to a high loss of information. This is well illustrated in the scatter plots (figure 2 to 4), especially in the comparisons with scoring 6, which uses a much finer gradation than the other scorings. For example, in figure 4, the horizontal "lines" make it clear that the amount of points for scoring 6 is quite different for all students that have the same amount of points using scoring 1. In this context, it must therefore be discussed to what extent the correlation coefficients can actually be classified as high enough. Even though the rankings of the students between the scores are very similar (high correlation), they are just not the same. This could have implications for the results of further research, e.g., on effectiveness in intervention studies: effectiveness could be higher with one scoring than with another scoring. This problem is amplified if the lower Kendall-tau-b coefficient is interpreted instead of Spearman-rho-coefficient (see below).

Furthermore, from a methodological point of view, the number of ties should be pointed out here. As expected, scorings with fewer intervals (and thus, lower points and a lower total score) have the most ties. However, not only do they have the highest amount of ties, they also have more cases per tie than scorings with more intervals. The problem for the correlation analysis between the scorings presented here, but also for further correlation analyses in the literature or in future research (e.g. between estimation accuracy and mathematical ability or other skills), is the lower informative value of the correlation coefficient when the proportion of ties is high. Thus, it seems advisable to use a small-step scoring with a high number of intervals for studies that want to investigate such relations. In further research, special attention should also be paid to ties concerning method choice (e.g. using Kendall-tau-b correlation analysis instead of Spearman-rank-correlation analysis).

6.2 Differences of the scorings in terms of test quality

Internal consistency. Although internal consistency differs greatly between the scorings, no tendencies concerning differences in the internal consistency were found comparing dichotomous (scoring 1 and 2) with polytomous scorings (scoring 3 to 6) or comparing scorings with equidistant bandwidths (scoring 4 and 6) with those that used non-equidistant bandwidths (scoring 3 and 5). Therefore, no general recommendation can be made as to whether dichotomous or polytomous scoring or scoring with equal or different bandwidths should be used.

From a psychometric point of view, the values of Cronbach's Alpha were on the threshold between an acceptable alpha and one that requires revision of the test instrument (Cortina, 1993). Since the number of items is rather

high, the moderate values indicate problems with unidimensionality (Cortina, 1993). Due to the variety of measures, strategies, and task characteristics, it can be assumed in principle that the construct *measurement estimation* is multidimensional. Initial research on this question shows that the size specification and the touchability of an object represent different dimensions of the construct length estimation (Hoth et al., in press). Since there are no touchable objects in the estimation test used for this article, and the study by Hoth et al. (in press) only investigated length estimation, the three-dimensional model was not useful for this article.

Besides the question of dimensionality, the estimation process itself is a complex cognitive process. For example, estimation consists of a variety of cognitive processes (Weiher & Ruwisch, 2018) and allows for the use of different strategies (Heid, 2018; Hildreth, 1983; Siegel et al., 1982; Weiher & Ruwisch, 2018). It is unclear which of these processes and strategies are elicited by which estimation objects or task types, so the construct *measurement estimation* might be rather broad. Lower interitem correlation (and therefore lower internal consistency) is typical for broader constructs (Clark & Watson, 1995; Streiner, 2003). A comparison with other estimation tests could help to find out to what extent the internal consistency measured here for estimation tasks is within a 'normal' range. Unfortunately, the internal consistency of the estimation tests is rarely reported: the authors are aware of only one estimation test (Brand et al., 2002), which has an alpha coefficient of .767, slightly higher than the values reported in this article (although with a different way of assessing estimation accuracy and by using length, weight, and time estimation, combined with number estimation).

Nevertheless, the fact that the internal consistency is so different should give pause for thought. The fact that the scoring for the same tasks produces such large differences in internal consistency indicates that it is an essential part of the operationalization of the construct *estimation accuracy*. Thus, it is shown once again that the choice of a scoring cannot be seen as "an arbitrary evaluation method" of an estimation test, but has to be motivated by content and adapted to the purpose of the study.

In addition, different internal consistency also has methodological implications. The correlation coefficient is limited by the internal consistency as shown in equation (4) with the *correlation coefficient r* and the *internal consistency R* (Danner, 2015):

$$r_{max} = \sqrt{R} \tag{4}$$

Thus, a correlation coefficient between an estimation accuracy (test book A) rated with scoring 6 and another ability can be at most $r = .90$; between an estimation accuracy (test book A) rated with scoring 1 and another ability the internal consistency can be at most $r = .78$.

Discriminatory power. As with internal consistency, it can be observed that the rather smooth scorings 3 and 6 have better discriminatory power. The hard scorings 1 and 4 seems to have the lowest discriminatory power. The students' estimation accuracy could be used as an explanation: too many children do not receive a point at all because they estimate too inaccurately. In the case of scoring 1, it can thus be concluded that the cut-off value, a maximum deviation of 25 %, is set too strictly. The cut-off value of scoring 2, a maximum deviation of 50 % for at least one point, seems to be more appropriate, although the discriminatory power still is rather low. In this context, it is interesting to note that scoring 5 tends to have lower discriminatory power than scoring 2, although here, too, the deviation for a point may not exceed 50 %. Thus, the additional intervals do not seem to add any insight but, on the contrary, make it more difficult to distinguish between accurate and inaccurate estimators. In

the case of scoring 3, the intervals also do not achieve a suitable differentiation of accurate and inaccurate estimators, presumably also because the limit for at least one point (< 30 % deviation) is chosen too strictly.

An improvement of the discriminatory power seems to be achieved by many intervals with a simultaneously high cut-off value (scoring 6). However, it remains to be discussed to what extent the number of intervals is meaningful – are there really 11 levels of estimation skill (accuracy)? This scoring differentiates so precisely that the percentage deviation could also be used without further scoring (which, however, is not common in the mathematics educational literature and also entails many disadvantages, see Weiher 2022).

Overall, the discriminatory power must be considered as rather low; the value of .3 described as acceptable in the literature (Döring & Bortz, 2016) is not reached by most items – except when scoring 6 is used. A possible explanation for this could lie in the distribution of the data together with rather strict cut-off values: If many children estimate with a deviation higher than the highest deviation at which points are still available, this results in an insufficient differentiation between strong and weak estimators. This may also be exacerbated by the high amount of underestimates in combination with the scale being closed to underestimates: children "accumulate" between 90 % and 100 % deviation even though they have different estimates.

Despite the low discriminatory power, items should not be excluded from the test prematurely, because the low discriminatory power does not automatically mean that the item does not belong to the construct (Kemper et al., 2015). Since the focus in this article is not on test development, but on presenting the effects of different scorings, all items have remained in the analysis. In the context of test development, it would have to be carefully examined, also under consideration of construct validity, whether the discriminatory power of all items is satisfactorily high after the selection of a scoring and whether items should be eliminated or exchanged.

Finally, no tendencies concerning differences in the discriminatory power were found comparing dichotomous (scoring 1 and 2) with polytomous scorings (scoring 3 to 6) or comparing scorings with equidistant bandwidths (scoring 4 and 6) with those that used non-equidistant bandwidths (scoring 3 and 5), so, as with the internal consistency, no general recommendation can be made.

6.3 Other impacts on the test performance and the test quality

Under- and overestimation. There were about three times as many underestimates as overestimates in the dataset. One explanation for this fact could lie in the age of the participants: children in grades 5 and 6 may not yet be as confident in the higher number range – especially in connection with units of measurement – and therefore tend to use smaller numbers for their estimation results. The greater amount of underestimations has an impact on the test result when using a scoring based on the percentage deviation from the real value: since the scale of possible values for underestimates is limited (-100 %) and the absolute value of the percentage deviation from the real value is used for the scores, all underestimates lie within the 100 % interval – in addition to those overestimates that also lie in that interval. The overestimates can theoretically be infinitely large. This disparity leads to the following question: Is an estimate of 1 mm for an 11 cm long chocolate bar as good as an estimate of 21.9 cm (both showing the same absolute value of percentage deviation from the real value, about 99 %)? If it is assumed that the estimation process also includes a reflection phase as well as the application of benchmarks for certain units (e.g. 1 mm), then it should be immediately obvious that the chocolate bar is much too long for 1 mm. The same can of course be said about the estimation value 21.9 cm, so that a normative setting about whether overestimation and underestimation should be equally evaluated, i.e. scored, is necessary. Therefore, the suitability of the absolute value of

the percentage deviation from the real value as a basis for scorings or at least the symmetrical application of the scorings (equal points for over- and underestimations) need to be reconsidered.

Tendency to estimate integers and multiples of ten and five. The discrete character of the estimates in our sample and the preference of multiples of tens and fives drew attention to a not planned, but possible influence of the concrete size of the TBEOs themselves on the estimation accuracy: The estimation accuracy might be higher for such TBEOs with a real value of five, ten or multiples of those numbers than for objects with other sizes. Furthermore, the boundaries of the scorings are usually set normatively and the concrete sizes of the TBEOs and the expected estimation accuracy due to the tendency of estimate integers are not taken into account. This opens up the possibility that certain boundaries are regularly exceeded or undershot.

6.4 Limitations

Two groups of limitations of the study will be addressed to here: limitations concerning the dataset and methodological limitations.

The observed tendency to give multiples of five or ten, but in any case whole numbers as estimation results, in combination with corresponding real values, can falsify a result. In this study, 47 of 48 real values were integers, of which 16 were multiples of five or ten. Only one value was a decimal number. A distortion due to many real values as multiples of five or ten will rather not have occurred here. Nevertheless, the results were limited to estimation items with mostly whole numbers. This question also arises for other studies, even if they do not refer to the effect or the corresponding real values.

Furthermore, some methodological limitations can be described: First, the high amount of ties when looking at total test scores means that results of Spearman-rho correlation analysis should be viewed with caution. Spearman-rho was reported to provide the best possible fit to other studies. However, the results of Kendall-tau-b correlation analysis as an alternative due to the high proportion of ties shows that the level of correlation is rather low and tends to amplify the problem pointed out (the differences between scores). Second, the dimensionality of the construct "measurement estimation" has not been clarified (as described above), so it is conceivable that the rather low internal consistency can also be explained by the violation of the one-dimensionality requirement for Cronbach's alpha. A scale analysis seems to be necessary. Until then, Cronbach’s alpha as a measure of quality should only be considered to a limited extent. However, from the authors' point of view, the differences in internal consistency can illustrate the differences between the scorings, since they are all based on the same estimation test. Third, these statistical analyses are based on classical test theory and are always a measure of a specific test instrument and sample. Therefore, the results cannot be directly transferred to other test instruments and samples. This applies in particular to the level of internal consistency and discriminatory power.

7 Resume

Aim of this paper was to present information about the differences between scorings used in measurement estimation research and their impact on the test performance and the test quality.

The following initial conclusions can be drawn: There are significant differences between all investigated scorings in terms of test performance. According to our data, a differentiation for estimates with a deviation less than 10 % and a deviation greater than 100 % does not appear to be necessary. Since in our data the number of estimates with a deviation less than 30 % was approximately as high as the number of answers in the interval from 90 % to 100 % deviations, a more precise breakdown of the estimated values for larger deviations might also be conceivable.

Test quality also differs between the scorings. The fact that hardly any uniform tendency of test performance and test quality for scoring characteristics (dichotomous or not, strict or lenient) could be identified shows that the assessment of estimation accuracy in mathematics educational research is not straightforward and requires careful justification.

It becomes clear that different approaches for the evaluation of estimation accuracy should not only be a question of choosing a method, but also a subject of mathematics educational research itself. The results of this article are therefore intended to raise awareness of the differences in scoring at various aspects, both in terms of mathematics education (research and classroom) and test theory. Moreover, they may provide an occasion to address the characteristics of the scoring used or, even more generally, the researchers underlying notion of estimation accuracy and to provide a transparent clarification. This includes not only the scoring, but likewise the calculation of the estimation error. Furthermore, the article can likewise serve as a starting point for further discussion on the assessment of estimation ability via estimation accuracy. Thus, besides the percentage deviation, alternatives to the calculation of the error are conceivable, even if not common (Weiher, 2022). However, the choice of a scoring or any other assessment or evaluation method for the estimation accuracy should not only be based on higher internal consistency and discriminatory power, but especially with regard to the aim of the corresponding study and possible other aspects, such as equal evaluation of over- and underestimation, age of the sample, measure to be estimated, or size and size specification of the TBEOs. First and foremost should be the own, transparent und justified conceptual clarification of *estimation accuracy*.

Literature

Axelrod, B. N., & Millis, S. R. (1994). Preliminary Standardization of the Cognitive Estimation Test. *Assessment 1*(3), 269–274. https://doi.org/10.1177/107319119400100307

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema 29*(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Brand, M., Kalbe, E., & Kessler, J. (2002). *Test zum Kognitiven Schätzen. Manual.* [Cognitive Estimation Test. Manual.] Göttingen: Beltz Test GmbH (in german).

Brand, M., Fujiwara, E., Kalbe, E., Steingass, H-P., Kessler, J., & Markowitsch, H. J. (2003). Cognitive Estimation and Affective Judgments in Alcoholic Korsakoff Patients. *Journal of Clinical and Experimental Neuropsychology 25*(3), 324–334. https://doi.org/10.1076/jcen.25.3.324.13802

Bright, G. W. (1976). Estimation as Part of Learning to Measure. In D. Nelson (Ed.), *Measurement in School.* (pp. 87–104). Reston, VA.: NCTM.

Bullard, S. E., Fein, D., Gleeson, M. K., Tischer, N., Mapou, R. L., & Kaplan, E. (2004). The Biber Cognitive Estimation Test. *Archives of Clinical Neuropsychology 19*(6), 835–846. https://doi.org/10.1016/j.acn.2003.12.002

Clark, L. A., &Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7, 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum.

Corle, C. G. (1963). Estimates of quantity by elementary teachers and college juniors. *The Arithmetic Teacher, 10*(6), 347–353.

Cortina, J. M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Danner, D. (2015). Reliabilität. Die Genauigkeit einer Messung [Reliability. The Accuracy of a Measurement]. Mannheim, GESIS Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines). https://www.gesis.org/fileadmin/upload/SDMwiki/Reliabilitaet_Danner_08102015_1.1.pdf (in german).

D'Aniello, G. E., Castelnuovo, G., & Scarpina, F. (2015). Could cognitive estimation ability be a measure of cognitive reserve? *Frontiers in Psychology 6*:608. https://doi.org/10.3389/fpsyg.2015.00608

Della Sala, S., MacPherson, S. E., Phillips, L. H., Sacco, L., & Spinnler, H. (2003). How many camels are there in Italy? Cognitive estimates standardized on the Italian population. *Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 24*(1), 10–15. https://doi.org/10.3389/fpsyg.2015.00608

Desli, D., & Giakoumi, M. (2017). Children's length estimation performance and strategies in standard and non-standard units of measurement. *International Journal of Research in Mathematics Education, 7*(3), 61–84.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* [Research Methods and Evaluation in the Social and Human Sciences]. Berlin, Heidelberg: Springer Spektrum (in german).

Forrester, M. A., Latham, J., & Shire, B. (1990). Exploring Estimation in Young Primary School Children. *Educational Psychology, 10*(4), 283–300. https://doi.org/10.1080/0144341900100401

Forrester, M. A., & Shire, B. (1994). The Influence of Object Size, Dimension and Prior Context on Children's Estimation Abilities. *Educational Psychology, 14*(4), 451–465. https://doi.org/10.1080/0144341940140407

Friebel, A. C. (1967). Measurement understandings in modern school mathematics. *The Arithmetic Teacher 14*(6), 476–480.

Harel, B. T., Cillessen, A. H., Fein, D. A., Bullard, S. E., & Aviv, A. (2007). It takes nine days to iron a shirt: the development of cognitive estimation skills in school age children. *Child Neuropsychology*, *13*(4), 309–318. https://doi.org/10.1080/09297040600837354

Heid, M. (2018). *Das Schätzen von Längen und Fassungsvermögen: Eine Interviewstudie zu Strategien mit Kindern im 4. Schuljahr* [Estimation of Length and Capacity: Strategies of 4th-graders in an Interview Study]. Wiesbaden: Springer Spektrum (in German).

Hildreth, D. J. (1980). *Estimation Strategy Uses in Length and Area Measurement Tasks by Fifth and Seventh Grade Students*. Dissertation at Ohio State University. Ann Arbor: University Microfilm International.

Hildreth, D. J. (1983). The use of strategies in estimating measurements. *Arithmetic Teachers, 30*(5), 50–54.

Hogan, T. P., & Brezinski, K. L. (2003). Quantitative estimation: one, two, or three abilities? *Mathematical Thinking and Learning*, *5*(4), 259–280. https://doi.org/10.1207/S15327833MTL0504_02

Hoth, J., Heinze, A., Weiher, D. F., Ruwisch, S., & Huang, H.-M. E. (2019). Primary school students length estimation competence: A cross-country comparison between Taiwan and Germany. In J. Novotná, & H. Moraová (Eds.), *Opportunities in Learning and Teaching Elementary Mathematics* (p. 201-211). Prague: Charles University, Faculty of Education.

Hoth, J., Heinze, A., Huang, H.-M. E., Weiher, D. F. & Ruwisch, S. (in press). Elementary school students' length estimation skills – analyzing a multidimensional construct in a cross-country study. *International Journal for Science and Mathematics Education*.Huang, H.-M. E. (2014). Investigating children's ability to solve measurement estimation problems. In S. Oesterle, P. Liljedahl, C. Nicol, & D. Allan (Eds.), *Proceedings of the Joint Meeting of PME 38 and PME-NA 36* (Vol. 3, pp. 353–360). Vancouver, Canada: PME.

Jones, G., Forrester, J. H., Gardner, G. E., Andre, T., & Taylor, A. R. (2012). Students' Accuracy of Measurement Estimation: Context, Units, and Logical Thinking. *School Science and Mathematics, 112*(3), 171–178. https://doi.org/10.1111/j.1949-8594.2011.00130.x

Jones, G., Taylor, A., & Broadwell, B. (2009). Estimating linear size and scale: Body rulers. *International Journal of Science Education, 31*(11), 1495–1509. https://doi.org/10.1080/09500690802101976

Joram, E., Gabriele, A. J., Bertheau, M., Gelman, R., & Subrahmanyam, K. (2005). Children's use of the reference point strategy for measurement estimation. *Journal for Research in Mathematics Education, 36(1)*, 4–23. https://doi.org/10.2307/30034918

Joram, E., Subrahmanyam, K., & Gelman, R. (1998). Measurement Estimation: Learning to Map the Route From Number to Quantity and Back. *Review of Educational Research, 68*(4), 413–449. https://doi.org/10.3102/00346543068004413

Kemper, C. J., Ziegler, M., Krumm, S., Heene, M., & Bühner, M. (2015). Testkonstruktion. In G. Stemmler, & J. Margraf-Stiksrud (Eds.), *Lehrbuch Psychologische Diagnostik* (pp. 157–221) [Test Construction. In G. Stemmler, & J. Margraf-Stiksrud (Eds.), Textbook Psychological Diagnostics (pp. 157–221)]. Bern: Verlag Hans Huber, Hogrefe AG. (in german).

Mendez, M. F., Doss, R. C., & Cherrier, M. M. (1998). Use of the Cognitive Estimations Test To Discriminate Frontotemporal Dementia from Alzheimer's Disease. *Journal of Geriatric Psychiatry and Neurology 11*(1), 2–6. https://doi.org/10.1177/089198879801100102

Moosbrugger, H., & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger, & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 7–26) [Quality requirements for a psychological test (test quality criteria). In H. Moosbrugger, & A. Kelava (Eds.), Test theory and questionnaire construction (pp. 7–26)] Berlin, Heidelberg, New York: Springer. https://doi.org/10.1007/978-3-642-20072-4_2

O'Daffer, P. (1979). A Case and Techniques for Estimation: Estimation Experiences in Elementary School Mathematics – Essential, Not Extra! *Arithmetic Teacher, 26*(6), 46–51. https://doi.org/10.5951/ AT.26.6.0046

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it Really Robust? Reinvestigating the Robustness of ANOVA Against Violations of the Normal Distribution Assumption. *Methodology 6*(4), 147–151. https://doi.org/10.1027/1614-2241/a000016

Shallice, T., & Evans, M. E. (1978). The Involvement of the Frontal Lobes in Cognitive Estimation. *Cortex 14*(2), 294–303. https://doi.org/10.1016/S0010-9452(78)80055-0

Siegel, A. W., Goldsmith, L. T., & Madson, C. R. (1982). Skill in Estimation Problems of Extent and Numerosity. *Journal for Research in Mathematics Education, 13*(3), 211–232. https://doi.org/10.2307/748557

Sowder, J. (1992). Estimation and Number Sense. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning: A Project of the National Council of Teachers of Mathematics* (pp. 371–389). Macmillan: NCTM.

Streiner, D. L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment, 80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Swan, M., & Jones, O. E. (1980). Comparison of Students' Percepts of Distance, Weight, Height, Area, and Temperature. *Science Education, 64*(3), 297–307. https://doi.org/10.1002/sce.3730640305

Weiher, D. F. (2018). Development of a measurement estimation test for length, area, and volume. In E. Bergqvist, M. Österholm, C. Granberg, & L. Sumpter (Eds.), *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education Vol. 5* (p. 307). Umeå Sweden: PME.

Weiher, D. F. (2019). Framework for the Parallelized Development of Estimation Tasks for Length, Area, Capacity, and Volume in Primary School – A Pilot Study. *Journal of Research in Science, Mathematics and Technology Education, 2*(1), 9–28. https://doi.org/10.31756/jrsmte.212

Weiher, D. F. (2022). Measurement Estimation Accuracy: A Comparison of Different Approaches. Submitted for Publication.

Weiher, D. F., & Ruwisch, S. (2018). Kognitives Schätzen aus Sicht der Mathematikdidaktik: Schätzen von visuell erfassbaren Größen und dazu erforderlichen Fähigkeiten [Cognitive Estimation from a mathematic educational point of view: Estimation of visually perceptible measures and skills to do so] *mathematica didactica, 41*(1), 77–103 (in german).

A.5 Publikation 4

Weiher, D. F. (2022a). Estimation of Length, Area, Capacity, and Volume: Results of a Written Estimation Test. [Eingereicht zur Veröffentlichung].

Estimation of Length, Area, Capacity, and Volume: Results of a Written Estimation Test

Dana Farina Weiher[1]

[1]Institute of Mathematics and its Didactics, Leuphana University Lueneburg

Author Note

Funding

The author did not receive support from any organization for the submitted work. No funding was received to assist with the preparation of this manuscript. No funding was received for conducting this study. No funds, grants, or other support was received.

Conflicts of interest/Competing interests

The author have no relevant financial or non-financial interests to disclose. The author have no conflicts of interest to declare that are relevant to the content of this article. The author certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The author have no financial or proprietary interests in any material discussed in this article.

Compliance with ethical standards

This research involves Human Participants. All participating students have been informed that participation is voluntary and will not suffer any disadvantages, regardless of whether they take part in the study or not. The teachers of the learning groups were present during the written test. The test itself represents a typical assessment situation in mathematics classrooms of the children.

Correspondence

Correspondence concerning this article should be addressed to Dana Farina Weiher, Institute of Mathematics and its Didactics, Leuphana University Lueneburg, Universitaetsalle 1, 21335 Lueneburg. Email: weiher@leuphana.de

Abstract

This study describes the estimation accuracy of the four visible measures lengths, area, capacity, and volume of 4th, 5th, and 6th graders as well as the relations between the estimation accuracy of the four measures. The estimation test contains a total of 48 items whose characteristics are parallelized and evenly distributed among the four measures. The analysis of estimation results from 900 students shows that the estimation accuracy differs significantly between 4th and 6th graders, 5th and 6th graders, and between the measures. Lengths are estimated most accurately, followed by capacity, area, and volume. All measures correlate moderately with each other. The estimation accuracy of lengths can serve as a predictor for the estimation accuracy of the other measures.

Keywords: Measurement estimation, estimation accuracy, estimation ability, estimation test

Introduction

Measurement estimation (besides computational and numerical estimation, O'Daffer, 1979) is defined as "the process of arriving at a measurement or measure without the aid of measurement tools. It is a mental process" (Bright, 1976, p. 89). Newcombe (2014) describes estimation as requirement to interact "effectively with the physical world", and other authors also emphasize the importance of estimation for everyday life (Brand et al., 2002; Forrester et al., 1990; Grund, 1992; Joram et al., 1998; O'Daffer, 1979; Sowder, 1992). In addition to its everyday relevance, estimation also has a mathematics educational relevance since it can contribute to the learning of other mathematical concepts. First of all, estimating measures is closely intertwined with measurement: "Understanding estimation is important for understanding measurement. Every measurement is an approximation, or if you will, an estimate. […] Estimation and measurement are mentally reinforcing" (Bright, 1979). He also states that "measurement can be taught more successfully if estimating is one kind of instructional activity" (Bright, 1976, p. 87). Other authors are of the same opinion (Jones et al., 2012; Joram et al., 2005; O'Daffer, 1979). Some other reasons for learning to estimate are the development of a concept of the estimated measure (Grassmann, 1999; O'Daffer, 1979), the development of problem solving skills and motivation to solve problems (O'Daffer, 1979), and the development of a "positive attitude towards mathematics (O'Daffer, 1979), which is why estimation is included in curricula as a topic for teaching mathematics (Sowder, 1992).

In general, estimation accuracy of children is said to be too low (e.g. Corle 1960, 1963; Desli & Giakoumi, 2017; Jones et al., 2012; Swan & Jones, 1971, 1980). Nevertheless, the existing research on estimating measures is broad and usually focuses on other topics rather than detailed descriptions of estimation accuracy. Several studies describe strategies and their efficiency, (e.g., Heid, 2018; Hildreth, 1983; Siegel et al., 1982), estimation-related or underlying skills (e.g. Forrester & Shire, 1994; Hoth et al., 2019), or task characteristics (e.g. Desli & Giakoumi, 2017; Forrester et al., 1990; Forrester & Shire,

1994; Weiher, 2019). Due to these different aims, some studies focus on a specific measure, usually lengths (Desli & Giakoumi, 2017; Hoth et al., 2019; Jones et al., 2012; Siegel et al., 1982), while others examine multiple measures individually (Heid, 2018; Swan & Jones, 1971, 1980; Weiher, 2019) or even mix them, sometimes with different numbers of tasks per measure (Corle, 1960, 1963; Forrester et al., 1990; Harel et al., 2007; Hildreth, 1980). In addition, the term "estimation accuracy" is thus at least implicitly interpreted in different ways (see a detailed description of different scorings in Weiher & Ruwisch, 2022). It is therefore not straightforward to assume that, e.g., estimating lengths, estimating temperatures, or a mixture of estimating lengths, weight, and time, results in the same estimation accuracy. This makes statements about "the" estimation accuracy of children inaccurate.

In this study, the measures length, area, capacity, and volume are focused. These are the visible measures, in contrast to time, weight/mass, or temperature. The same or very similar estimation strategies can be used (Heid, 2018; Hildreth, 1983; Weiher & Ruwisch, 2018) because the measures are also mathematically related and therefore the measurement process follows the same basic idea at least for lengths, areas and volumes (Carpenter et al., 1975). Capacity could be derived from volume (a cube with an edge length 10cm has a volume of $10cm^3$, which corresponds to 1l). Despite these commonalities, an understanding of each measure, as well as an understanding of its measurement process, must be established. This poses different challenges, for example the grid structure for area and volumes (Battista & Clements, 1996; Battista et al., 1998; Huang, 2014) or the recognition of an object as a unit (Carpenter et al., 1975; Nunes et al., 1993). These skills, among others, are the basis for estimation (Jones et al., 2012; Weiher & Ruwisch, 2018). It therefore seems appropriate to look at the estimation accuracy for the measures individually. In order to allow comparisons between the measures, an assessment with parallelized items is also necessary. Such an approach contributes to a valid assessment of the estimation accuracy of visible measures and only then enables implications for mathematical research.

The aim of this study is therefore twofold: First, the estimation accuracy of students is described for each visible measure separately. Second, the study contribute to a better understanding of the relations between the estimation accuracy of the visible measures. Therefore, a written estimation test is developed that includes the same types of tasks across measures, contains equal numbers of tasks per measure, and calculates and scores accuracy equally across measures.

Estimating visible measures

Estimation accuracy

In mathematics educational research, measurement estimation accuracy is usually reported based on the percentage deviation of the estimated value from the actual size of the to-be-estimated-object (TBEO; Desli & Giakoumi, 2017; Heid, 2018; Hildreth 1980, 1983; Hogan & Brezinski, 2003; Hoth et al., 2019; Huang, 2014; Joram et al., 1998; Joram et al., 2005; Siegel et al., 1982; Swan & Jones, 1980). For this purpose, the following formula for the "percentage deviation from the real value" D_{perc} is usually used, although it is not explicitly stated in most studies, with the estimated value *e* and the real value *r*:

$$D_{\text{perc}} = \frac{e-r}{r} \quad (1)$$

The sign of the calculated deviation indicates whether it is an underestimate (negative sign) or an overestimate (positive sign). The value range of the deviation is between -100 %, which is reached at an estimated value of 0, so is only theoretically meaningful, and +∞.

Based on the percentage deviation, different scorings are applied in the studies, which assign points to different intervals of the percentage deviation. They differ in the number of intervals, the interval width, and the interval limits and thus assess the estimation accuracy with different degrees of rigor (for a detailed consideration of different types of scoring, see Weiher & Ruwisch, 2022).

Because studies examining measurement estimation differ in terms of type of estimation task, type of assessment of estimation results, measure studied, age of subjects, and survey form (written or interviews), it is not surprising that it is difficult to derive consistent results across studies of children's

and young adults' estimation accuracy. Nevertheless, an attempt is made here to present as broadly as possible study results on accuracy in estimating measures.

Accuracy of estimating visible measures

Estimation ability, equated with estimation accuracy in studies, seems to be not satisfactorily developed in both children (Corle, 1960, 1963; Jones et al., 2012; Joram et al., 2005; Swan & Jones, 1980; Sowder, 1992) and adults (Sowder, 1992, Swan & Jones, 1971, 1980), even if adults are more accurate estimators (Corle, 1963; Heid, 2018; Ruwisch et al., 2015).

Nevertheless, length seems to be more easy to estimate than any of the other measures named here[1]. In more detail, length estimation accuracy is higher than accuracy while estimating area (Gooya et al., 2011; Grassmann, 1999; Pike & Forrester, 1997), capacity (Heid, 2018; Joram et al., 1998; Ruwisch et al., 2015), and volume (Joram et al., 1998). These findings are not always confirmed for all grades: Huang (2014) stated that 4th graders estimation accuracy is higher for estimating length than estimating area, but did not find any difference in accuracy of length and area estimation for 5th and 6th graders.

Estimation accuracy of female and male students

Several studies found no difference in estimation accuracy between male and female students (Desli & Giakoumi, 2017; Forrester & Shire, 1994; Harel, 2007; Hildreth, 1980; Paull, 1971; Siegel et al., 1982). Swan & Jones (1980) stated that male students outperform female students while estimating distance and height, but not while estimating weight and temperature. Corle (1960) found that male students outperform female students, but not differentiated between measures. The test used included items for weight, length, time, temperature, and capacity.

[1] Only temperature is estimated more accurate than length (Joram, 1998). However, it should be remembered that temperature is not a visible measure. In addition, the range of values of temperatures indoors is naturally very limited - by knowing about this, a very high estimation accuracy can already be achieved. Moreover, in Joram's (1998) study, only two questions were asked about temperature (indoor and outdoor).

Estimation accuracy of different age groups

In most studies, estimation accuracy increase with age or grade (Corle, 1960, 1963; Desli & Giakoumi, 2017; Forrester & Shire, 1994; Harel, 2007; Huang, 2014; Ruwisch et al., 2015; Siegel et al., 1982; Swan & Jones, 1980), although the results concerning grades are not consistent across studies. For example, Harel et al. (2007) observed a steep increase in achieved scores on an estimation test between 5 and 9 year olds, and a less pronounced but visible increase between 9 and 16 year olds which seems to be in line. Corle (1963) notes a difference in estimation accuracy among children in the 5th and 6th grade which fits and at least does not contradict Harel's et al. (2007) results. In contrast, Huang (2014) observed significant differences in terms of scores achieved on a test between 4th and 6th graders, but not between 4th and 5th, and 5th and 6th graders, which contradicts results from Corle (1963). Other studies found no difference between the investigated grade or age groups when investigating estimation ability of 5th, 7th, and 13th grade (Hildreth, 1980) or 6 years to 11 years olds (Pike & Forrester, 1997).

Research Goal

As described in the introduction, first aim of this study is to contribute to the systematic description of children's estimation accuracy. The following research questions should be answered:

1) How accurate do children in 4th, 5th and 6th grade estimate visible measures?
2) Does the estimation accuracy vary between 4th, 5th, and 6th graders?
3) Does the estimation accuracy of 4th, 5th, and 6th graders a) correlate with age or b) vary between gender?

It is hypothesized that, consistent with the state of the research, a wide variance of estimation accuracy is observed with, on average, rather poor accuracy. Accuracy presumably increases with age/grade level. A correlation with age is hypothesized, as presumably due to learning growth and therefore grade level.

Second, the study aims to describe the relation between the visible measures. The following research questions should be answered:

4) Does the estimation accuracy vary between the measures?

5) Does the estimation accuracy of the measures correlate with each other?

6) Can length estimation accuracy be a predictor for accuracy of the other measures?

According to literature, it is hypothesized that accuracy is highest for lengths. Capacity will be estimated more accurately than area and volume due to higher familiarity in Germany. A correlation between all measures is assumed due to their mathematical relation and the likely similar strategy application. It is also assumed that the estimation accuracy of lengths can serve as a predictor for the estimation accuracy of the other measures, in line with literature that considers the understanding of length as a prerequisite for the understanding of other measures.

Material and Methods

Sample

The study took place at various school types[2] in northern Germany in spring 2020. A total of 1040 students from 22 4th, 13 5th and 13 6th grades classes took the test (for the explanation of the test books, see 4.2). After exclusion of students with more than 50% missing values overall or per measure (see 4.4), the sample for analysis contains 900 students. The demographic data of the sample for each test book can be found in Table 1.

[2] In Germany, six secondary school types (5th grade and higher) exist. All but one secondary school type was included in this study. Since all schools participated in the study voluntarily and all schools that re-registered were included, the proportion of students per school in the study does not necessarily correspond to the population's proportion.

Table 1

Sample Size per Test Book and Grade.

		Test Book A			Test Book B		
		All	Female	Male	All	Female	Male
All Grades	n	459[1]	229	229	441	221	220
	Mean Age (SD)	11.00[2] (1.03)	10.99 (1.06)	11.00[2] (0.99)	10.98 (1.03)	10.93 (1.01)	11.04 (1.05)
Grade 4	n	213	107	106	205	100	105
	Mean Age (SD)	10.15 (0.52)	10.07 (0.48)	10.23 (0.55)	10.15 (0.57)	10.10 (0.57)	10.20 (0.57)
Grade 5	n	114	54	60	107	59	48
	Mean Age (SD)	11.20[3] (0.53)	11.25 (0.57)	11.15[3] (0.49)	11.15 (0.55)	11.10 (0.45)	11.21 (0.64)
Grade 6	n	132[1]	68	63	129	62	67
	Mean Age (SD)	12.21[3] (0.55)	12.22 (0.51)	12.21[3] (0.60)	12.17 (0.58)	12.11 (0.60)	12.23 (0.55)

[1]*Annotation.* Contains an extra value because one student did not specify a gender.

[2]*Annotation*. Two values are missing because two students did not specify their age.

[3]*Annotation.* One value is missing because one student did not specify his age.

The drawing of the sample is quasi random and depends on the place of residence of the author. Nevertheless, there were no preferences in terms of secondary school type or catchment area.

The measurement estimation test

The measurement estimation test is a written group test. For each measure, the test contains 12 items evenly distributed among three task types. Overall, the test includes 48 items. 4th graders answered only length and capacity estimation tasks. 5th and 6th graders gave estimates to all four measures, but 5th graders struggled to estimate volume (see 4.4).

Based on the theoretical framework for different task types and after conducting 137 3rd and 4th graders in a pilot study with several more task types (Weiher 2019), three task types have proven to be theoretically useful (since they presumably require an interaction of the skills needed to estimate) and practically feasible. These task types are shown – with examples for each measure – in Table 2. For each

task type, two units per measure were chosen: cm and m for lengths, ml and l for capacity, cm^2 and m^2 for area, and cm^3 and m^3 for volume. Objects were selected as TBEOs that could be described unambiguously, had an (approximate) integer dimension number, and were likely to be known to children without knowing their size specification. The real values for the size of these objects are between 11cm and 269m for lengths, 9ml and 240l for capacity, $46cm^2$ and $500m^2$ for area and $4cm^3$ and $140m^3$ to $260m^3$ for volume[3].

Table 2

Example Items and Their Characteristics (Task Types) per Measure.

	Length	Area	Capacity	Volume
Task Type 1: TBEO and unit not visible	Approximately how long is the ship "Titanic"? The ship "Titanic" is about __m long.	Approximately how big is an unfolded handkerchief? An unfolded handkerchief is about __cm^2 in size.	Approximately how much water fits into a yellow plastic garbage bag? About __l of water fit into a yellow plastic garbage bag.	What is the approximate size of a shipping container? A shipping container is about __m^3 in size.
Task Type 2: TBEO visible, unit not visible	Approximately how long is the red stripe on the board? The red stripe on the blackboard is about ___cm long.	Approximately how big is the calendar on the board? The calendar on the blackboard is about __cm^2 in size.	Approximately how much water fits in the shower gel bottle on the board? About __ml of water fits into the shower gel bottle on the board.	Approximately how big is the game pack on the board? The game pack on the board is about __cm^3 in size.
Task Type 3: TBEO not visible, unit visible	The yellow ribbon on the board is 1cm long. What is the approximate length of the long side of a 5 € bill? The long side of a 5 € bill is about __cm long.	The brown square on the board is $1m^2$ in size. What is the approximate size of a ping-pong table? A ping-pong table is about __m^2 in size.	1 ml of water fits into the green vial on the board. How much water fits into a small tea light? About __ml of water fits into a small tea light.	The cube on the board is $1cm^3$ in size. What is the approximate size of a packet of butter? A packet of butter is about __cm^3 in size.

[3] The largest real value of volume is the size of the classroom in which the test took place. Since all classes performed the test in their own room, the value varies and the range of classroom sizes is given here.

The test was presented in two forms (A and B). They include the same task types, but the TBEOs are swapped between two task types 1 and 3 (the non-visible TBEOs) to minimize distortion due to the concrete objects while comparing task types in further research.

Students were informed by the author that they are not allowed to use a ruler. They are allowed to stand up to take a closer look at the objects placed on the board or on the floor. However, they are not allowed to measure the objects with their hands, steps or similar.

Scoring of estimation accuracy

The raw data in the form of the results of the paper-and-pencil test were transferred to SPSS 26 for further processing.

To standardize the estimation error for making comparisons between different items, measures, and task types possible, the "logarithmic deviation" $D_{\log}$ was computed as described in equation (2) with the estimated value *e* and the real value *r*:

$$D_{\log} = \log_{10} \frac{e}{r} \tag{2}$$

The percentage deviation from the real value used in other studies has several disadvantages, which can be avoided by using the logarithm error score (Weiher, 2022; Weiher & Ruwisch, 2022). The main advantage is the open scale for underestimations, which prevents underestimates from tending to be more accurate than overestimations. The logarithm of ten also allows an easy interpretation of the estimation deviation: A value of -1 or 1 represents an under- or overestimation of one order of magnitude. One order of magnitude is defined as a deviation by a factor of 10 from the real value[4].

[4] Although every logarithm could be used, it is more difficult to interpret others. For example: by using the natural logarithm, e.g., -1 and 1 mean a deviation of the factor of Euler's number *e*.

In order to avoid more or less arbitrary intervals for scoring points as it is common by using discrete scorings for estimation accuracy (Weiher & Ruwisch, 2022) a continuous scoring is selected.

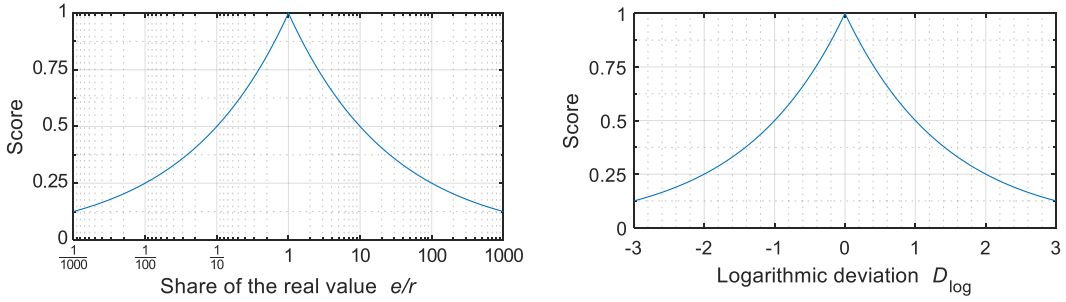
The logarithmic deviation $D_{\log}$ is scored with the following formula (3) to get the "score" S for each item:

$$S = (0.5)^{|D_{\log}|} \tag{3}$$

An exact estimation is scored with 1 point. Half of the achievable points (0.5 points) are awarded for a deviation of one order of magnitude (factor of ten). Every further order of magnitude leads to a halving of points. In between, the point values decrease exponentially (Figure 1).

Figure 1

Left: Scores for the Share of the Real Value. Right: Scores for the Logarithmic Deviation.



Treatment of missing data

There were two types of missing values in this test. First, and this was almost exclusively the case, missing values arose when students did not answer at all. Second, missing data arose when students gave answers that are invalid according to the test manual. This includes e.g. answers with an additional unit (beside the one given in the test booklet), when this unit does not correspond to the measure (e.g. 5cm for estimation of area), when students gave answers that contain two or more commas, or when the estimation result is obviously not the result of a serious estimation process (results like filling in zeros until the end of the page). Furthermore, it was assumed that no reasonable

estimation process underlies estimates < 0.1 and > 10000. Therefore, they were excluded as extreme values.

The tasks underlying the second type of missing values have actually been processed. The answers are not assessed as estimation result due to a lack of cognitive skills, and therefore scored with 0. A scoring of 0 for missing data of the first type would distort the results (especially with regard to the frequency of cases). Analysis of missing data showed no problems with data from 4th graders, whereas 5th graders struggled to estimate volume. As missing data showed, about 46% of all 5th graders did not answer to a single question (but about 43% answered all questions). For area, about 20% of students did not answer any question, but 60% only had up to two missing values. It could be assumed that area is more familiar to 5th graders than volume, so volume was excluded for 5th graders from the analysis[5]. For 5th and 6th graders, only about 75% of students completed the length estimation tasks without missing data. Nevertheless, about 98% of students had at least three missing values, so no problem with length could be observed. No anomalies were found for any of the other measures.

In order to avoid distortion, students with more than 50% missing values (each measure or overall test result) were excluded. This procedure resulted in the sample described above, which is the basis for the following analysis.

Statistical Analyses

The Man-Whitney U-test was used to test whether there were differences in central tendencies between the two test books for each measure in each grade. Descriptive statistics of estimation accuracy for each measure in each grade are presented (research question 1). To assess differences in

[5] Even though it can be assumed that some 5th graders have a concept of the measure volume, the high number of missing values and especially the high number of children who did not answer any question about volume at all shows that they probably do not have a concept of the measure volume and consequently cannot estimate it.

estimation accuracy between the grades, Kruskal-Wallis-test with Dunn-Bonferroni post-hoc tests and Man-Whitney-U-test were conducted (research question 2). Correlations of estimation accuracy with age were investigated using (Partial) Spearman-Rho Correlation analysis (research question 3a). The Man-Whitney-U-test was conducted to assess differences between genders (research question 3b). To assess if differences between measures were the same for the test books and to test if there were significant differences between measures (research question 4), a mixed ANOVA with repeated measurement (mixed rmANOVA) was conducted. The Greenhouse-Geisser adjustment was used to correct for violations of sphericity. The relation between the measures (research questions 5 and 6) was assessed using Spearman-Rho Correlation analysis, simple linear regression and multivariate regression.

Results

Estimation accuracy of 4th, 5th and 6th graders (research question 1)

The estimation accuracy of 4th, 5th, and 6th graders was described based on the total score achieved for each measure. The descriptive statistics can be found in Table 3. The Man-Whitney-U test showed no significant differences between the two versions of the test booklets for all measures in all grades[6], except for length in grade 4 ($p < .05$). Since the median test did not become significant ($p = .171$) and the effect size of the difference according to Man-Whitney-U test ($r = .13$) according to Cohen (1988) is to be assessed as weak, both test booklets were nevertheless considered together for all grades and measures.

[6] Grade 4: Man-Whitney-U-test indicated no differences between the test versions for capacity ($p = .763$). Grade 5: Man-Whitney-U-test indicated no differences between the test versions for length ($p = .240$), capacity ($p = .799$) and area ($p = .340$). Grade 6: Man-Whitney-U-test indicated no differences between the test versions for length ($p = .784$), capacity ($p = .893$), area ($p = .669$) and volume ($p = .788$).

Table 3

Descriptive Statistics per Measure in Every Grade.

Grade	Measure	Median	Mean	SD	Range
4 (n = 418)	Length	9.80	9.57	1.13	4.44–11.28
	Capacity	7.33	7.31	1.35	2.89–10.46
5 (n = 221)	Length	9.80	9.60	1.03	4.31–11.33
	Capacity	7.41	7.18	1.55	3.42–10.71
	Area	6.81	6.74	1.05	2.99–9.75
6 (n = 261)	Length	10.12	9.91	0.94	6.60–11.63
	Capacity	7.97	7.86	1.48	3.97–11.06
	Area	7.21	7.18	1.11	3.17–9.96
	Volume	6.96	6.88	1.15	3.04–10.15

The median also shows that 4th and 5th graders scored about the same on estimating lengths and capacity, but 6th graders scored slightly higher. 6th graders also scored about half a point more for estimating area than 5th graders.

The results observed here are examined further below.

Variation of estimation accuracy between the grades (research question 2)

According to the median (Table 3), the mean points achieved differ between grades for each measure, with 4th graders always scored the least and 6th graders always scored the most. The extent to which these differences are statistically significant is examined below.

A Kruskal-Wallis test showed that there was a statistically significant difference in length estimation accuracy score between the different grades, χ^2 (2) = 18.94, $p < .001$. Dunn-Bonferroni post-hoc test showed a statistically significant difference between 4th and 6th graders ($z = -4.00$, $p < .001$, $r = .15$) and between 5th and 6th graders ($z = -3.58$, $p < .05$, $r = .16$). No significant difference was found between 4th and 5th graders, χ^2 (2) = .134, $p = 1.00$. Significance values were adjusted by the Bonferroni correction for several tests. According to Cohen (1988), effect size is rather weak for both differences.

A Kruskal-Wallis test showed that there was also a statistically significant difference in capacity estimation accuracy score between the different grades, χ^2 (2) = 30.87, $p < .001$. Dunn-Bonferroni post-hoc test showed a statistically significant difference between 4th and 6th graders ($z = -4.90$, $p < .001$, $r = .19$) and between 5th and 6th graders ($z = -4.82$, $p < .001$, $r = .22$). No significant difference was found between 4th and 5th graders, χ^2 (2) = .651, $p = 1.00$. Significance values were adjusted by the Bonferroni correction for several tests. According to Cohen (1988), effect size is rather low for both differences.

A Mann-Whitney-U-Test indicated that there was also a statistically significant difference in area estimation accuracy score between 5th and 6th graders, $U = 35433.00$, $z = 4.33$, $p < .001$, $r = .20$. According to Cohen (1988), the effect size is rather low.

Relation of estimation accuracy and age (research question 3a)

Taken all grades into account, length estimation ability does not correlate with age, as assessed by Spearman-Rho ($r = .06$, $p = .062$). The same could be observed for capacity estimation ability: even if the effect is significant, it is rather low ($r = .08$, $p < .05$). If controlled by grade, the non-parametric partial correlation showed a significant negative correlation between length estimation accuracy and age ($r = -.08$, $p < .05$) and capacity estimation accuracy and age ($r = -.09$, $p < .05$). Both correlation coefficients are to be classified as low according to Cohen (1988).

For the assessment of the correlation between area estimation ability and age only 5th and 6th graders were taken into account. Area estimation ability and age correlate significantly, as assessed by Spearman-Rho ($r = .13$, $p < .001$.). Nevertheless, the effect is rather low according to Cohen (1988). Controlled by grade, non-parametric partial correlation showed a non significant effect ($r = -.08$, $p = .083$), which is negative, but also rather low.

For the assessment of the correlation between volume estimation ability and age only 6th graders were taken into account. Volume estimation ability and age do not correlate significantly, as assessed by Spearman-Rho ($r = -.05$, $p = .439$).

Variation of estimation accuracy between gender (research question 3b)

A Mann-Whitney-U-Test[7] was calculated to determine if there were differences in estimation accuracy between male and female students.

Male students (Median = 10.02) outperformed female students (Median = 9.80) significantly while estimating lengths, according to Mann-Whitney-U-Test (U = 115563.00, z = 3.74, p < .001, r = .13). The same could be observed for estimating capacity: male students (Median = 7.86) outperformed female students (Median = 7.17) significantly according to Mann-Whitney-U-Test (U = 123611.00, z = 5.80, p < .001, r = .194).

For area estimation accuracy, only 5th and 6th graders were taken into account. No significant differences in median area estimation accuracy could be found between female (Median = 6.91) and male (Median = 7.11) students according to Mann-Whitney-U-Test (U = 31422.00, z = 1.64, p = .100).

For volume estimation accuracy, only 6th graders were taken into account. There was no significant difference in volume estimation accuracy between female (Median = 6.97) and male (Median = 6.96) students according to Mann-Whitney-U-Test (U = 9039.00, z = .97, p = .331).

Variation of estimation accuracy between the measures (research question 4)

A mixed rmANOVA[8] with test booklet as between-subjects factor and measure as within-subjects factor was conducted for each grade in order to investigate if differences between measures were the same for both test books.

---

[7] The requirements for a Man-Whitney-U-test test have been checked: Kolmogorov-Smirnov-Test was used to assess differences in distribution (not median) between female and male students' estimation accuracy. Therefore, standardized values of variables were used. The distributions of length, capacity, area, and volume estimation accuracy did not differ between both groups, length: p = .894 and capacity: p = .622, area: p = .986, and volume: p = .067.

[8] The requirements for mixed rmANOVA have been checked: The Greenhouse-Geisser adjustment was used to correct for violations of sphericity (except for grade 4, here the sphericity was not taken into account because there were only two variables). There was no homogeneity of the error variances, as assessed by Levene's test (p > .05). There was homogeneity of covariances, as assessed by Box's test (grade 4: p = .067, grade 5: p = .974,

First, there was no statistically significant interaction between test booklet and measure for all groups considered (grade 4: $F(1, 416) = 1.203$, $p = .273$; grade 5: $F(1.731, 379.198) = .681$, $p = .487$; grade 6: $F(2.808, 727.240) = .028$, $p = .992$). This leads to the conclusion that the differences between the variables do not vary between the test booklets, so they are considered together for the following analyses.

The main effect test booklet also did not become significant in any grade (grade 4: $F(1, 416) = 3.018$, $p = .083$; grade 5: $F(1, 219) = .4361$, $p = .510$; grade 6: $F(1, 259) = .061$, $p = .805$).

The main effect measure becomes significant in all grades:

Grade 4: There was a significant main effect for measure, $F(1, 416) = 1010.546$, $p < .001$, partial $\eta^2 = .708$. Bonferroni-adjusted post-hoc analysis revealed significantly ($p < .001$) higher performance scores for length estimation than for capacity estimation ($M_{Diff} = 2.263$, 95%-CI[2.123, 2.403]).

Grade 5: There was a significant main effect for measure, $F(1.731, 379.198) = 494.505$, $p < .001$, partial $\eta^2 = .693$. Bonferroni-adjusted post-hoc analysis revealed significantly ($p < .001$) higher performance scores for length estimation than for capacity estimation ($M_{Diff} = 2.426$, 95%-CI[2.178, 2.674]) and area estimation ($M_{Diff} = 2.860$, 95%-CI[2.674, 3.046]). Furthermore, performance scores for capacity estimation were significantly ($p < .001$) higher than for area estimation ($M_{Diff} = .434$, 95%-CI[0.167, .702]).

Grade 6: There was a significant main effect for measure, $F(2.808, 727.240) = 518.181$, $p < .001$, partial $\eta^2 = .684$. Bonferroni-adjusted post-hoc analysis revealed significantly ($p < .001$) higher performance scores for length estimation than for capacity estimation ($M_{Diff} = 2.051$, 95%-CI[1.815, 2.287]), area estimation ($M_{Diff} = 2.725$, 95%-CI[2.533, 2.918]), and volume estimation ($M_{Diff} = 3.028$, 95%-

---

grade 6: $p = .777$). Normal distribution of the residuals is assumed due to the sample size ($N > 30$) according to the central limit theorem.

CI[2.831, 3.225]). Furthermore, performance scores for capacity estimation were significantly ($p < .001$) higher than for area estimation ($M_{Diff} = 0.674$, 95%-CI[0.435, 0.914]) and volume estimation ($M_{Diff} = 0.976$, 95%-CI[.747, 1.206]). Performance scores for area estimation were significantly ($p < .001$) higher than for volume estimation ($M_{Diff} = 0.302$, 95%-CI[0.103, 0.501]).

In summary, the scores achieved in all grades varied significantly between all measures. In line with the descriptive statistics mentioned above (Table 3), most points were achieved for estimating lengths, followed by capacity, area and volume. The partial Eta-square can be characterized as high and indicates that in 4th grade 70.8%, in 5th grade 69.3%, and in 6th grade 68.4% of the variance can be explained by the measure.

Relation of estimation accuracy between measures (research question 5 and 6)

As seen above, the estimation accuracy differs between the measures in all grades. Because the sample size can influence the correlation analysis (smaller samples would tend to result in higher correlations) and the sample is not the same size for all measures, each grade was examined individually here. This also means that correlations between measures should not be compared between grades.

Grade 4: Length estimation accuracy and capacity estimation accuracy correlate significantly, as assessed by Spearman Rho ($r = .34$, $p < .001$). According to Cohen (1988), the correlation coefficient can be described as moderate.

A simple linear regression with robust standards errors[9] was calculated to predict the capacity estimation score based on the length estimation score (Table 4). A significant regression equation was

[9] The requirements for simple linear regression have been hcecked: The error term of the variable capacity was normally distributed, as indicated by the Kolmogorov-Smirnov test ($p = .200$) and the graphical inspection of the Q-Q diagram. The error term of the variable length was not normally distributed, as shown by the Kolmogorov-Smirnov test ($p < .001$) and the graphical inspection of the Q-Q diagram. Graphical inspection of the scatterplot of standardized dependent variable and standardized residuals also suggested heteroskedasticity as well as autocorrelation. Therefore, parameter estimation was performed with robust standard errors (HC4).

found ($F(1;416) = 46.783$, $p < .001$), with an R^2 of .101. Students' predicted capacity estimation score is equal to 3.664+.381 (length estimation score) points when the length estimation score is measured in points. Students' capacity estimation score increased 0.381 for each point of their length estimation score. The effect size is $f = .335$ and can be classified as moderate according to Cohen (1988).

Table 4

Details of Simple Linear Regression (Capacity Estimation Score Based on Length Estimation Score) for 4th Grade (n = 418)

	Effect on Capacity Estimation Score		
Variables	Unstandardized	Standardized	Robust Standard Error (HC4)
Constant	3.664***		0.605***
Length Estimation Score	0.381***	0.318***	0.063***
R^2	.101		
R^2 (corrected)	.099		
F (df=1; 416)	46.783***		

***$p < .001$

Grade 5: All correlations between the measures can be seen in Table 5. According to Cohen (1988), the correlation coefficients can be described as moderate. The highest correlation could be observed between length and capacity ($r = .36$, $p < .01$), the lowest between area and capacity ($r = .31$, $p < .01$).

Table 5

Spearman Rho for Each Correlation Between Measures (5th grade, n = 221)

	Area	Capacity
Length	.34**	.36**
Area		.31**

** The correlation is significant at the .01 level (two-sided).

A multivariate linear regression with robust standard errors[10] was calculated to predict the area and capacity estimation score based on the length estimation score (Table 6).

A significant regression equation was found for capacity: ($F(1;219) = 29.790$, $p < .001$), with an R^2 of .120. Students' predicted capacity estimation score is equal to 2.171+0.521 (length estimation score) points when the length estimation score is measured in points. Students' capacity estimation score increased 0.521 for each point of their length estimation score. The effect size is $f = .369$ and can be classified as rather high according to Cohen (1988).

A significant regression equation was found for area: ($F(1;259) = 40.493$, $p < .001$), with an R^2 of .156. Students' predicted area estimation score is equal to 2.839+0.406 (length estimation score) points when the length estimation score is measured in points. Students' area estimation score increased 0.406 for each point of their length estimation score. The effect size is $f = .430$ and can be classified as high according to Cohen (1988).

Table 6

Details of Multivariate Linear Regression (Capacity and Area Estimation Score Based on Length Estimation Score) for 5th Grade (n = 221).

Variables	Unstandardized	Standardized	Robust Standard Error (HC 4)
	Effect on Capacity Estimation Score		
Constant	2.171*		0.853*
Length Estimation Score	0.521***	0.346***	0.088***
R^2	.120		
R^2 (corrected)	.116		
F (df=1; 219)	29.790***		

[10] The requirements for multivariate linear regression have been checked: The error terms of all variables were not normally distributed, as indicated by the Kolmogorov-Smirnov-test (length: $p < .001$; area and capacity: $p < .05$). This could be confirmed by the graphical inspection of the Q-Q diagrams. Graphical inspection of the scatter plot of the standardized dependent variables and standardized residuals could not rule out heteroskedasticity and autocorrelation, so robust error terms were also used here for parameter estimation (HC4).

	Effect on Area Estimation Score		
Constant	2.839***		0.649***
Length Estimation Score	0.406***	0.395***	0.066***
R^2	.156		
R^2 (corrected)	.152		
F (df=1; 219)	40.493***		

***p < .001, *p < .05

Grade 6: All correlations between the measures can be seen in Table 7. According to Cohen (1988), the correlation coefficients can be described as moderate. The highest correlation could be observed between area and volume (r = .47, p < .01) the lowest between lengths and volume (r = .29, p < .01).

Table 7

Spearman Rho for Each Correlation Between Measures (6th grade, n = 261).

	Area	Capacity	Volume
Length	.37**	.37**	.29**
Area		.46**	.47**
Volume		.44**	

** The correlation is significant at the .01 level (two-sided).

A multivariate linear regression with robust standard errors[11] was calculated to predict the area, capacity, and volume estimation score based on the length estimation score (Table 8).

A significant regression equation was found for capacity: (F(1;259) = 40.302, p < .001), with an R^2 of .135. Students' predicted capacity estimation score is equal to 2.130+0.578 (length estimation score)

[11] The requirements for multivariate linear regression have been checked: The error terms of length and volume were not normally distributed, as indicated by the Kolmogorov-Smirnov-test (length: p < .001; volume: p < .05), whereas they were normally distributed for area (p = .200) and capacity (p = .200).This could be confirmed by the graphical inspection of the Q-Q diagrams. Graphical inspection of the scatter plot of the standardized dependent variables and standardized residuals could not rule out heteroskedasticity and autocorrelation, so robust error terms were also used here for parameter estimation (HC4).

points when the length estimation score is measured in points. Students' capacity estimation score increased 0.578 for each point of their length estimation score. The effect size is f = .395 and can be classified as high according to Cohen (1988).

A significant regression equation was found for area: (F(1;259) = 39.568, p < .001), with an R^2 of .133. Students' predicted area estimation score is equal to 2.909+0.431 (length estimation score) points when the length estimation score is measured in points. Students' area estimation score increased 0.431 for each point of their length estimation score. The effect size is f = .392 and can be classified as high according to Cohen (1988).

A significant regression equation was found for volume: (F(1;259) = 37.500, p < .001), with an R^2 of .126. Students' predicted volume estimation score is equal to 2.579+0.434 (length estimation score) points when the length estimation score is measured in points. Students' volume estimation score increased 0.434 for each point of their length estimation score. The effect size is f = .380 and can be classified as rather high according to Cohen (1988).

Table 8

Details of Multivariate Linear Regression (Capacity, Area, and Volume Estimation Score Based on Length Estimation Score) for 6th Grade (n = 261).

Variables	Unstandardized	Standardized	Robust Standard Error (HC4)
Effect on Capacity Estimation Score			
Constant	2.130*		0.848*
Length Estimation Score	0.578***	0.367***	0.085***
R^2	.135		
R^2 (corrected)	.131		
F (df=1; 259)	40.302***		
Effect on Area Estimation Score			
Constant	2.909***		0.640***
Length Estimation Score	0.431***	0.364***	0.065***
R^2	.133		
R^2 (corrected)	.129		

F (df=1; 259)	39.568***		
	Effect on Volume Estimation Score		
Constant	2.579***		0.695***
Length Estimation Score	0.434***	0.356***	0.070***
R^2	.126		
R^2 (corrected)	.123		
F (df=1; 259)	37.500***		

***p < .001, *p < .05

Discussion

As expected, the estimation accuracy of lengths is the highest in all grades. With a mean of slightly more than 80% of the possible points, this can also be described as satisfactory – although, taken the high range into account, there is still some potential for improving the estimation accuracy of lengths. In contrast, the estimation accuracy of the other measures is rather poor, with around 60% of the achievable points each. The only exception here is the estimation accuracy of capacity in 6th grade, which is slightly higher, but with about 66 % of the points achieved, cannot satisfy either. These results are in line with previous research, which, as mentioned above, describes estimation ability as rather poorly developed (Corle, 1960, 1963; Jones et al., 2012; Joram et al., 2005; Swan & Jones, 1980; Sowder, 1992). However, against the background that even older students and adults show high estimation deviations (Corle, 1963; Heid, 2018; Ruwisch et al., 2015), one has to ask whether these deviations are "normal" estimation accuracy and therefore acceptable anyway.

As expected, estimation accuracy of the measures length, area, and capacity appears to increase with grade, which partly is in line with the literature (Corle, 1960; Desli & Giakoumi, 2017; Harel 2007; Huang, 2014; Ruwisch et al., 2015; Swan & Jones, 1980; Siegel et al., 1982). This indicates that the underlying concepts of measure, its measurement process, and the use of estimation strategies can be learned and are part of mathematics education. Forrester and Shire (1994) found higher estimation accuracy with increasing age, but surprisingly do not refer to grade and the skills learned with it as an explanation. The study presented here suggests that grade plays a significant role and may serve as an

explanation for increasing estimation accuracy with age. The primary influence of grade on estimation accuracy is also supported by the low effect size of the correlation between age and estimation accuracy (length and capacity) or the non-significant correlation (area and volume) when controlling for grade. The correlation coefficient is likely to be negative. The children who are particularly old compared to children of the same grade level may have repeated the grade and may be more likely to show weaker performance (e.g., even in the prerequisites for estimating or in estimating itself).

The surprisingly non-existent difference in estimation accuracy between 4th and 5th graders in this study may possibly be explained by the change of school from elementary to secondary school in conjunction with a neglect of the instructional treatment of measures, at least the estimation of measures. 4th graders presumably have had some experience, including instructional experience, with lengths and capacity, as they are more likely to be at the end of their elementary school years. 5th graders, on the other hand, are more likely to be at the beginning of their secondary school careers, and by then other content has presumably been focused on in mathematics instruction, so that no learning progress intended by mathematics instruction has been made on the topic of estimating measures. This result also fits with the observation of Desli and Giakoumi (2017), who found no difference between 5th graders and younger children in estimation tasks with the standardized unit cm (versus tasks with non-standardized units) and explained this with the lack of instruction on standardized units in elementary school.

Regarding the difference in estimation accuracy between female and male students, no general statement can be made; it is necessary to differentiate between the measures. Male students outperform female students in estimating lengths and capacity, but there is no difference in estimating accuracy of area and volume. However, according to Cohen (1988), the effect should be considered as small. The result therefore seems to primarily confirm the studies that also found no difference between female and male students (Desli & Giakoumi, 2017; Forrester & Shire, 1994; Harel, 2007; Hildreth, 1980;

Paull, 1971; Siegel et al., 1982). Nevertheless, it appears that again the male students estimated more accurately than the female students, as also shown by Swan and Jones (1980) for distance and height, i.e. also lengths, and Corle (1960) for several measures.

The ranking of measures in terms of accuracy is according to expectations. The highest accuracy in estimating lengths implies that children have the most benchmarks available for use in these estimation processes and/or they are most familiar with the measurement process of lengths and mentally comfortable with lining up or dividing length objects as units.

For the other measures, due to the low(er) score, it can be assumed that there are not enough benchmarks available for the estimation process, that the children are less familiar with the measurement process of these measures, or that they have difficulties with the concept of these measures. It was suspected that the different duration of experience with the measures (lengths are introduced first, followed by capacity. Area and volume are introduced later in school) might lead to differences in estimation accuracy. Especially for area and volume, difficulties with the concept of measures, their measurement process or their units (both recognizing an object as unit or the unit sign) might cause the low estimation accuracy. For example, only two-thirds to three-fourth of 8^{th} graders had a concept of area and showed additional difficulties with the formula, or confuse area, perimeter, and diameter (Kenney & Kouba, 1997; Lorenz, 1992), or failed in area measurement problems (Huang & Wirtz, 2011). Since in the study presented here the oldest students are 6^{th} graders, it could be assumed that they might have even more problems with area measurement and therefore with area estimation. Furthermore, structuring objects with a grid is not intuitive, both for areas in 2D (Battista et al., 1998; Outhred & Mitchelmore, 2000) and for volumes in 3D (Battista, 1999; Battista & Clements, 1996), but must be learned. However, this structuring can be regarded as a prerequisite for estimating if objects are to be mentally "applied" as a unit or if objects are to be mentally divided as benchmarks, since Reynolds and Wheatley (1996) describe it as prerequisite of measuring areas. The difficulty that for

volumes some of the units of the grid structure are not visible is probably added (Carpenter et al., 1975), especially since the estimation process or structuring is done only mentally. Furthermore, in the case of estimating volume, the recognition of the whole cube as a unit seems to be more difficult than recognizing a line as a unit for length (Carpenter et al., 1975). This matches the ranking of measures in this estimation study, as well as the ranking of measures (without capacity) in measurement tasks found by Carpenter et al. (1975), Hiebert (1981), and Tan Sisman & Aksu (2015): Thus, children find it easiest to estimate and measure lengths and hardest to estimate and measure volumes. Nevertheless, Tan Sisman and Aksu (2012, 2015) reported that even 6th and 7th graders had problems with basic measurement understanding of lengths, which possibly explains why lengths are also estimated with some deviation.

Operating in the rather higher number space, which may be less familiar, comes with increasing dimensions and thus larger dimension number as an additional difficulty. Beside lower mean accuracy, this might be an explanation for the greater range and standard deviation (the high range is also in line with literature; Axelrod & Millis, 1994; Jones et al., 2012) in measures with more dimensions: With higher dimension and therefore higher real values per object, the possibilities for dimension numbers increase. This can be illustrate by the following example: the height of a cube is 10cm, the front area's size is 100cm^2, its capacity is 1000ml, and its volume is 1000cm^3. In this consideration, however, the units used must also be included for comparisons between the measures; for example, the capacity of the cube could also be specified as 1l – and thus a very small measurement number. Since smaller (cm, cm^2, ml, cm^3) and larger units (m, m^2, l, m^3) are used equally frequently per measure in this study, presumably the explanation of dimensions can be resorted to.

Beside all differences between the measures, the expected correlation between all measures could be found. All correlation coefficients can be classified as moderate (Cohen, 1988). The fact that, in addition to the differences between the measures, strategy use may be similar (Weiher & Ruwisch, 2018) may be an explanation: For all four visible measures there is the option of determining the

respective size by comparing it with benchmarks, which is an object whose size is known, in a mental way (Bright, 1976; Heid, 2018; Hildreth, 1983; Siegel et al., 1982). For area and volume, there is the option of using the corresponding formula to calculate the area/volume (Hildreth, 1983), but these strategies need to be combined with a comparison strategy. The length of the sides/edges is then important for the estimation process, which is why benchmarks for lengths can be also used here for estimating area/volume. Since the strategies are similar, it is likely that the skills required to estimate the four measures are similar (Weiher & Ruwisch, 2018), which could also contribute to the correlation. Nevertheless, the correlation is not as high as expected. One reason for this could be that despite the same basic idea of measurement (Lehrer, 2003), the actual measurement process is different (Nunes et al., 1993). For example, an area or volume is determined by measuring and calculating the lengths of the sides or edges (which can also be the basis of an estimation strategy). However, even these formulas are difficult for children to understand (Kenney & Kouba, 1997), e.g., they do not recognize the connection between formula and grid structure (Huang, 2014) or between the measures themselves (Tan Sisman & Aksu, 2015). This leads to the assumption that when children use estimation strategies (e.g., length times width or length times width times height), they may not use, for example, the same support points and thus the mathematical relationship of the four measures.

It is interesting to note that in grade 6, the correlations between lengths and each of the other measures are lower than between area/volume, area/capacity, and volume/capacity. While the higher correlation between volume and capacity might be explained by the fact that they are two three-dimensional measures, this is not the case for the other two pairings. Nevertheless, it would be conceivable that the higher correlation is related to the recognition of any other dimension. This implies not only to recognize the measure with all dimensions at the object to be estimated, but also to imagine a corresponding unit (e.g. a benchmark).

The estimation accuracy of lengths can be used as a predictor of the estimation accuracy of the other measures with a moderate to high effect size. This is consistent with the view that understanding lengths serves as a basis for, and is a prerequisite for, understanding other measures, especially for measurement (Bragg & Outhred, 2001; Clarke et al., 2003; Joram et al., 2005; Nührenbörger, 2004). The present study can be therefore considered as a continuation of the theory by showing that this assumption also appears to hold for measurement estimation. The estimation accuracy of lengths share approximately between 10% and 16% variance with the estimation accuracy of the other measures, which is noticeable due to the high number of potential prerequisites for estimation accuracy. These could include the skills needed to estimate, which can be described in three groups (Weiher & Ruwisch, 2018): The first group is named “mental operating with benchmarks” and includes knowledge about measurement process and benchmarks, ability to comparisons, and spatial skills (Joram, 2003; Joram et al., 1998; Hildreth, 1983; Heid, 2018; Siegel et al., 1982; Sowder, 1992). The second group are executive functions, including strategy choice, application, and evaluation (as is investigated in psychological research; Axelrod & Millis, 1994; Brand et al., 2003; Bullard et al., 2004; D’Aniello et al., 2015; Della Sala, et al., 2003; Mendez et al., 1998, Shallice & Evans, 1978). The third group are general (mathematical) knowledge and basic skills, including geometrical and arithmetical knowledge as well as visual perception, as can be derived from the description of strategies. For example, calculations must be performed to obtain an estimated result for an area using the strategy length times width, or the corresponding measure must be detected on an object.

Nevertheless, it cannot be concluded from this study that it is sufficient for the estimation of area, capacity, and space to address only the estimation of lengths in school.

Limitations

The following limitations can be named:

1) Since the estimation accuracy was assessed using a written test, no statements can be made about the ways in which students arrived at an estimation result. In addition, it cannot be guaranteed that the students actually estimate the expected measure – it is also possible that they know or guess the size of the TBEO. It is also possible that students estimate a different measure than the one asked for (e.g., because they confuse perimeter and area, see above).

2) For two-thirds of the questions, the object to be estimated was not visible. Although efforts were made to be unambiguous (e.g., by using brand names or by describing the object precisely), it cannot be ensured that the students imagine the object in order to estimate its size. It can also occasionally happen that students do not know the object at all and still enter (guess) a size.

Implications and Outlook

The aim of this study was to contribute to the systematic assessment of the estimation accuracy of the visible measures length, area, capacity and volume and to examine their relation. As a particular strength of this study, the systematic, i.e. parallel, investigation of the four visible measures should be especially emphasized, so that the results are thus – at least for the visible measures – more generally valid and statements can be made for measurement estimation. The results of this study can be used for teaching mathematics as well as for further research on measurement estimation, as explained below.

With respect to estimation accuracy, the students seems to be rather heterogeneous. Individual attention to the taught class or the individual student as well as differentiated prior knowledge assessment seems to be indispensable. In addition, reducing gender differences in estimation accuracy should be a concern of mathematics lessons.

Since estimation accuracy improves with higher grades, measurement estimation seems to be learnable and should be a permanent part of mathematics lessons, along with the underlying skills, e.g.

measuring and mental use of benchmarks. When measuring, not only the procedural ability to measure should be taught (like putting on a measuring instrument and reading a scale), but also the basic ideas of measurement should be addressed, such as iterating a unit, and the concept of units in general, in order to create an understanding-based connection between estimation and measurement as well as the visible measures. The connection between the four measures could also be more strongly addressed in mathematics lessons. Thus, perhaps the benchmarks that can be used to estimate lengths could also be used to estimate the other measures and students are able to increase their estimation accuracy for area, capacity, and volume.

The moderate correlation between the measures shows that it can be useful to address different visual measures in one study/survey instrument. However, the reported differences also show that the selection has to be done carefully and justified – it is not useful to have "just any" measures estimated. This applies to both research and educational purposes.

Since the estimation accuracy of lengths predicts the estimation accuracy of the other measures at least moderately, the theory that the measurement concept of lengths serves as a prerequisite for the measurement concept of further measures can be extended for the estimation of these measures. Nevertheless, the present results also indicate that, despite higher accuracy in estimating lengths, there seems to be greater difficulties in estimating all the other measures studied. It should not be concluded that it is sufficient to train only the estimation of lengths.

Overall, against the background of the everyday relevance and the importance for the development of mathematical skills, the demand for greater attention to measurement estimation in mathematics education seems appropriate.

One possibility for further research is a longitudinal or interventional study to investigate the extent to which an improvement in the estimation accuracy of lengths also leads to an improvement in the estimation accuracy of the other visible measures. As a basis, a systematic empirical investigation of

skills related to measurement estimation seems to be necessary, as well as knowledge of students about the relationship between measures.

A second possibility for further research is to systematically examine teaching processes in order to clarify which learning paths contribute to increasing the estimation accuracy.

Last but not least, the investigation of the multitude of task characteristics is an approach for further research. This includes the influence of the real value as a number (in distinction to the size in sense of spatial extent of the TBEO), e.g., 1m vs. 100cm in addition to the "classical" characteristics like visibility and orientation of objects, and, in light of the results of this study, should also be done for each measure.

References

Axelrod, B. N., & Millis, S. R. (1994). Preliminary Standardization of the Cognitive Estimation Test. *Assessment, 1*(3), 269–274. https://doi.org/10.1177/107319119400100307

Battista, M. T. (1999). Fifth Graders' Enumeration of Cubes in 3D Arrays: Conceptual Progress in an Inquiry-Based Classroom. *Journal for Research in Mathematics Education, 30*(4), 417–448. https://doi.org/10.2307/749708

Battista, M. T., & Clements, D. H. (1996). Students' Understanding of Three-Dimensional Rectangular Arrays of Cubes. *Journal for Research in Mathematics Education, 27*(3), 258–292. https://doi.org/10.2307/749365

Battista, M. T., Clements, D. H., Arnoff, J., Battista, K., & Van Auken Borrow, C. (1998). Students' Spatial Structuring of 2D Arrays of Squares. *Journal for Research in Mathematics Education, 29*(5), 503–532. https://doi.org/10.2307/749731

Bragg, P., & Outhred, L. (2001): So That's What A Centimetre Look Like: Students' Understandings of Linear Units. In M. van den Heuvel-Panhuizen (Ed.), *Proceedings of the 25th Conference of the*

International Group for the Psychology of Mathematics Education (pp. 2-209–2-216). Utrecht, Netherlands: PME.

Brand, M., Fujiwara, E., Kalbe, E., Steingass, H.-P., Kessler, J., & Markowitsch, H. J. (2003). Cognitive Estimation and Affective Judgments in Alcoholic Korsakoff Patients. *Journal of Clinical and Experimental Neuropsychology, 25*(3), 324–334. https://doi.org/10.1076/jcen.25.3.324.13802

Brand, M., Kalbe, E., & Kessler, J. (2002). *Test zum Kognitiven Schätzen. Manual*. [Cognitive Estimation Test. Manual.] Göttingen: Beltz Test GmbH (in german).

Bright, G. W. (1976). Estimation as Part of Learning to Measure. In D. Nelson (Ed.), *Measurement in School* (pp. 87–104). Reston, VA.: NCTM.

Bright, G. W. (1979). Estimating Physical Measurements. *School Science and Mathematics, 79*(8), 581–586.

Bullard, S. E., Fein, D., Gleeson, M. K., Tischer, N., Mapou, R. L., & Kaplan, E. (2004). The Biber Cognitive Estimation Test. *Archives of Clinical Neuropsychology, 19*(6), 835–846. https://doi.org/10.1016/j.acn.2003.12.002

Carpenter, T. P., Coburn, T. G., Reys, R. E., & Wilson, J. W. (1975). Notes from National Assessment: Basic Concepts of Area and Volume. *The Arithmetic Teacher, 22*(6), 501–507. https://www.jstor.org/stable/41191275

Clarke, D., Cheeseman, J., McDonough, A., & Clarke, B. (2003). Assessing and Developing Measurement with Young Children. In D.H. Clements, & G. Bright (Eds.), *Learning and Teaching Measurement* (pp. 68–80). Reston, VA: NCTM.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum.

Corle, C. G. (1960). A Study of the Quantitative Values of Fifth and Sixth Grade Pupils. *The Arithmetic Teacher, 7*(7), 333–340. https://www.jstor.org/stable/41184340

Corle, C. G. (1963). Estimates of Quantity by Elementary Teachers and College Juniors. *The Arithmetic Teacher, 10*(6), 347–353. https://www.jstor.org/stable/41184817

D'Aniello, G. E., Castelnuovo, G., Scarpina, F. (2015). Could Cognitive Estimation Ability Be a Measure of Cognitive Reserve? *Frontiers in Psychology 6:608*. https://doi.org/10.3389/fpsyg.2015.00608

Della Sala, S., MacPherson, S. E., Phillips, L. H., Sacco, L., & Spinnler, H. (2003). How Many Camels Are There in Italy? Cognitive Estimates Standardized on the Italian Population. *Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 24*(1), S. 10–15. https://doi.org/10.3389/fpsyg.2015.00608

Desli, D., & Giakoumi, M. (2017). Children's Length Estimation Performance and Strategies in Standard and Nonstandard Units of Measurement. *International Journal of Research in Mathematics Education, 7*(3), 61–84.

Forrester, M. A., Latham, J., & Shire, B. (1990). Exploring Estimation in Young Primary School Children. *Educational Psychology, 10*(4), 283–300. https://doi.org/10.1080/0144341900100401

Forrester, M. A., & Shire, B. (1994). The Influence of Object Size, Dimension and Prior Context on Children's Estimation Abilities. *Educational Psychology, 14*(4), 451–465. https://doi.org/10.1080/0144341940140407

Gooya, Z., Khosroshahi, L. G., & Teppo, A. R. (2011). Iranian Students' Measurement Estimation Performance Involving Linear and Area Attributes of Real-World Objects. *ZDM Mathematics Education, 43*, 709–722. https://doi.org/10.1007/s11858-011-0338-1

Grassmann, M. (1999). Zur Entwicklung von Zahl- und Größenvorstellungen als wichtigem Anliegen des Sachrechnens [On the development of number and size concepts as an important concern of practical problems]. *Grundschulunterricht, 46*(4), 31–34 (in german).

Grund, K.-H. (1992). Größenvorstellungen – eine wesentliche Voraussetzung beim Anwenden von Mathematik. [Concepts of size – an essential prerequisite when applying mathematics.] *Grundschule, 12*, 42–44 (in german).

Harel, B. T., Cillessen, A. H., Fein, D. A., Bullard, S. E., & Aviv, A. (2007). It Takes Nine Days to Iron a Shirt: The Development of Cognitive Estimation Skills in School Age Children. *Child Neuropsychology, 13*(4), 309–318. https://doi.org/10.1080/09297040600837354

Heid, L.-M. (2018). *Das Schätzen von Längen und Fassungsvermögen: Eine Interviewstudie zu Strategien mit Kindern im 4. Schuljahr* [Estimation of Length and Capacity: Strategies of 4th-graders in an Interview Study]. Wiesbaden: Springer Spektrum (in German). https://doi.org/10.1007/978-3-658-18874-0

Hiebert, J. (1981). Units of Measure: Results and Implications from National Assessment. *The Arithmetic Teacher, 28*(6), 38–43. https://www.jstor.org/stable/41191805

Hildreth, D. J. (1980). *Estimation Strategy Uses in Length and Area Measurement Tasks by Fifth and Seventh Grade Students*. Dissertation at Ohio State University. Ann Arbor: University Microfilm International.

Hildreth, D. J. (1983). The Use of Strategies in Estimating Measurements. *Arithmetic Teacher, 30*(5), 50–54. https://www.jstor.org/stable/41192173

Hogan, T. P., & Brezinski, K. L. (2003). Quantitative Estimation: One, Two, or Three Abilities? *Mathematical Thinking and Learning, 5*(4), 259–280. https://doi.org/10.1207/S15327833MTL0504_02

Hoth, J., Heinze, A., Weiher, D. F., Ruwisch, S., Huang, H.-M. E. (2019). Primary School Students Length Estimation Competence – A Cross-Country Comparison Between Taiwan and Germany. In J. Novotná, & H. Moraová (Eds.), *Opportunities in Learning and Teaching Elementary Mathematics* (p. 201–211). Charles University, Faculty of Education. https://www.semt.cz/proceedings-19.pdf

Huang, H.-M. E. (2014). Investigating Children's Ability to Solve Measurement Estimation Problems. In S. Oesterle, P. Liljedahl, C. Nicol, & D. Allan (Eds.), *Proceedings of the Joint Meeting of PME 38 and PME-NA 36* (Vol. 3, pp. 353–360). Vancouver, Canada: PME. https://files.eric.ed.gov/fulltext/ED599830.pdf

Huang, H.-M. E., & Wirtz, K. G. (2011). Developing Children's Conceptual Understanding of Area Measurement: A Curriculum and Teaching Experiment. *Learning and Instruction, 21*(1), 1–13. https://doi.org/10.1016/j.learninstruc.2009.09.002

Jones, G., Forrester, J. H., Gardner, G. E., Andre, T., & Taylor, A. R. (2012). Students' Accuracy of Measurement Estimation: Context, Units, and Logical Thinking. *School Science and Mathematics, 112*(3), 171–178. https://doi.org/10.1111/j.1949-8594.2011.00130.x.

Joram, E. (2003). Benchmarks as Tools for Developing Measurement Sense. In D. H. Clements & G. Bright (Eds.), *Learning and Teaching Measurement* (pp. 57–67), Reston, VA: NCTM.

Joram, E., Gabriele, A. J., Bertheau, M., Gelman, R., & Subrahmanyam, K. (2005). Children's Use of the Reference Point Strategy for Measurement Estimation. *Journal for Research in Mathematics Education, 36*(1), 4–23. https://doi.org/10.2307/30034918.

Joram, E., Subrahmanyam, K., & Gelman, R. (1998). Measurement Estimation: Learning to Map the Route from Number to Quantity and Back. *Review of Educational Research, 68*(4), 413–449. https://www.jstor.org/stable/1170734

Kenney, P. A., & Kouba, V. L. (1997). What Do Students Know about Measurement? In P. A. Kenney, E. A. Silver (Eds.), *Results from the Sixth mathematics Assessment of the National Assessment of Education Progress* (pp. 141–163). Reston, VA: NCTM.

Lehrer, R., Jaslow, L., & Curtis, C. (2003): Developing an Understanding of Measurement in the Elementary Grades. In D. H. Clements, & G. Bright (Eds.), *Learning and Teaching Measurement* (pp. 100–121). Reston, VA: NCTM.

Lorenz, J. H. (1992). Größen und Maße in der Grundschule [Measures and Measurements in Primary School]. *Grundschule, 11*, 12–14 (in german).

Mendez, M. F., Doss, R. C., & Cherrier, M. M. (1998). Use of the Cognitive Estimations Test To Discriminate Frontotemporal Dementia from Alzheimer's Disease. *Journal of Geriatric Psychiatry and Neurology, 11*(1), 2–6. https://doi.org/10.1177/089198879801100102

Newcombe, N. (2014). The Origins and Development of Magnitude Estimation. *Ecological Psychology, 26(*1–2), 147–157. https://doi.org/10.1080/10407413.2014.875333

Nührenbörger, M. (2004). Childrens' Measurement Thinking in the Context of Length. In G. Törner, R. Bruder, A. Peter-Koop, N. Neill, H.-G. Weigand & B. Wollring (Eds.), *Developments in Mathematics Education in German-speaking Countries. Selected Papers from the Annual Conference on Didactics of Mathematics* (pp. 95–106). Ludwigsburg, Germany. http://webdoc.sub.gwdg.de/ebook/e/gdm/2001/Nuehrenboerger.pdf

Nunes, T., Light, P., & Mason, J. (1993). Tools for Thought: The Measurement of Length and Area. *Learning and Instruction, 3*(1), 39–54. https://doi.org/10.1016/S0959-4752(09)80004-2

O'Daffer, P. (1979). A Case and Techniques for Estimation: Estimation Experiences in Elementary School Mathematics – Essential, Not Extra! *Arithmetic Teacher, 26*(6), 46–51. https://doi.org/10.5951/AT.26.6.0046

Outhred, L. N., & Mitchelmore, M. C. (2000): Young Children's Intuitive Understanding of Rectangular Area Measurement. *Journal for Research in Mathematics Education, 31*(2), 144–167. https://doi.org/10.2307/749749

Paull, D. R. (1971). *The Ability to Estimate in Mathematics*. Dissertation, Columbia University. University Microfilms, A XEROX Company, Ann Arbor, Michigan.

Pike, C. D., & Forrester, M. A. (1997). The Influence of Number-sense on Children's Ability to Estimate Measures. *Educational Psychology, 17*(4), 483–500. https://doi.org/10.1080/0144341970170408

Reynolds, A., & Wheatley, G. H. (1996). Elementary Students' Construction and Coordination of Units in an Area Setting. *Journal for Research in Mathematics Education, 27*(5), 564–581. https://doi.org/10.2307/749848

Ruwisch, S., Heid, L.-M., & Weiher, D. F. (2015). Measurement Estimation in Primary School: Which Answer is Adequate? In K. Beswick, T. Muir, & J. Wells, (Eds.). *Proceedings of 39th Psychology of Mathematics Education conference* (Vol. 4, pp. 113–120), Hobart, Australia: PME.

Shallice, T., & Evans, M. E. (1978). The Involvement of the Frontal Lobes in Cognitive Estimation. *Cortex, 14*(2), 294–303. https://doi.org/10.1016/S0010-9452(78)80055-0

Siegel, A. W., Goldsmith, L. T., & Madson, C. R. (1982). Skill in Estimation Problems of Extent and Numerosity. *Journal for Research in Mathematics Education, 13*(3), 211–232. https://doi.org/10.2307/748557

Sowder, J. (1992). Estimation and Number Sense. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning. A project of the National Council of Teachers of Mathematics* (pp. 371–389). New York: Macmillan.

Swan, M., & Jones, O. (1971). Distance, Weight, Height, Area and Temperature Percepts of University Students. *Science Education, 55*(3), 353–360. https://doi.org/10.1002/sce.3730550315

Swan, M., & Jones, O. (1980). Comparison of Students' Percepts of Distance, Weight, Height, Area, and Temperature. *Science Education, 64*(3), 297–307. https://doi.org/10.1002/sce.3730640305

Tan Sisman, G., & Aksu, M. (2012). Seventh Grade Students' Understanding of Linear Measurement: What Has Been Learned about Linear Measurement in Seven Years of Schooling Process?

Procedia – Social and Behavioral Sciences, 46, 1905–1909. https://doi.org/10.1016/j.sbspro.2012.05.400

Tan Sisman, G., & Aksu, M. (2015). A Study on Sixth Grade Students' Misconceptions and Errors in Spatial Measurement: Length, Area, and Volume. *International Journal of Science and Mathematics Education, 14*, 1293–1319. https://doi.org/10.1007/s10763-015-9642-5

Weiher, D. F. & Ruwisch, S. (2018). Kognitives Schätzen aus Sicht der Mathematikdidaktik [Cognitive Estimation from a mathematic educational point of view: Estimation of visually perceptible measures and skills to do so] *mathematica didactica*, 41 (1), pp. 77–103 (in german). http://www.mathematica-didactica.com/Pub/md_2018/md_2018_Weiher_Ruwisch.pdf

Weiher, D. F., & Ruwisch, S. (2022): The Assessment of Measurement Estimation Results – a Discussion of Different Scorings [Manuscript submitted for publication].

Weiher, D. F. (2022): Measurement Estimation Accuracy: A Comparison of Different Approaches. [Manuscript submitted for publication].

Weiher, D. F. (2019). Framework for the Parallelized Development of Estimation Tasks for Length, Area, Capacity, and Volume in Primary School – A Pilot Study. *Journal of Research in Science, Mathematics and Technology Education, 2*(1), 9–28. https://doi.org/10.31756/jrsmte.212

B Übersicht weiterer Publikationen im Zusammenhang mit der Dissertation

Tabelle 3

Übersicht weiterer Publikationen

Kennzeichnung	**Literaturangabe**
A	Weiher, D. F. (2018a): Operationalisierung des Konstrukts „Schätzen von Längen, Flächeninhalten und Volumina“ für Grundschulkinder. In Fachgruppe Didaktik der Mathematik der Universität Paderborn (Hrsg.): *Beiträge zum Mathematikunterricht 2018* (S. 1939–1942). WTM. http://dx.doi.org/10.17877/DE290R-19764
B	Weiher, D. F. (2018b): Development of a measurement estimation test for length, area, and volume. In E. Bergqvist, M. Österholm, C. Granberg & L. Sumpter (Hrsg): *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education Vol. 5* (S. 307).
C	Weiher, D. F. (2018c): Schätzen von Längen, Flächeninhalten und Volumina mit verschiedenen Aufgabentypen. In A. S. Steinweg (Hrsg.): *Inhalte im Fokus – mathematische Strategien entwickeln. Tagungsband des AK Grundschule in der GDM 2018* (S. 105–108) University of Bamberg Press. https://fis.uni-bamberg.de/bitstream/uniba/44712/1/MDG8SteinwegGDM2018opusse_A3a.pdf
D	Weiher, D. F. (2019b): Merkmale von Schätzaufgaben zu Längen, Flächeninhalten, Fassungsvermögen und Rauminhalten. In A. Frank, S. Krauss & K. Binder (Hrsg.): *Beiträge zum Mathematikunterricht 2019* (S. 1419). WTM. http://dx.doi.org/10.17877/DE290R-20738
E	Weiher, D. F. (2020a): Der Zusammenhang zwischen Schätzgenauigkeit und Stützpunktausprägung bei Längen. In H.-S. Siller, W. Weigel & J. F. Wörler (Hrsg.). *Beiträge zum Mathematikunterricht 2020* (S. 1277–1280). WTM. http://dx.doi.org/10.17877/DE290R-21623
F	Weiher, D. F. (2020b): Die Vielfalt von Schätzaufgaben. Eine Darstellung verschiedener Merkmale von Schätzaufgaben zu visuell erfassbaren Größen. In *mathematik differenziert 3*, 16–21.
G	Weiher, D. F. (2022b). Measurement Estimation Accuracy: A Comparison of Different Approaches. In: IDMI-Primar Goethe Universität Frankfurt (Hrsg.), *Beiträge zum Mathematikunterricht 2022*. http://dx.doi.org/10.17877/DE290R-23806

C Testinstrumente

C.1 Schätztest Pilotstudie

Ich kann Längen schätzen!

Ich heiße:_____

Ich bin ein ☐ Mädchen

☐ Junge

Klasse: _____

1) Wie lang ist ein Fineliner ungefähr?

Ein Fineliner ist ungefähr _____ cm lang.

2) Wie hoch ist ein Schiffscontainer ungefähr?

Ein Schiffscontainer ist ungefähr _____ m hoch.

3) Wie breit ist ein Teebeutel ungefähr?

Ein Teebeutel ist ungefähr _____ cm breit.

4) Wie viele Fineliner sind ungefähr so lang wie die kurze Seite eines Donald Duck Hefts?

Ungefähr _____ Fineliner sind so lang wie die kurze Seite eines Donald Duck Hefts.

5) Wie viele Fineliner sind ungefähr so lang wie die lange Seite eines Zebrastreifen-Streifens?

Ungefähr _____ Fineliner sind so lang wie die lange Seite eines

Zebrastreifen-Streifens.

6) Wie viele Fineliner sind ungefähr so lang wie die lange Seite eines Zeichenblocks?

Ungefähr _____ Fineliner sind so lang wie die lange Seite eines Zeichenblocks.

7) Wie viele von den weißen Streifen an der Tafel sind ungefähr so lang wie die lange Seite von „Gregs Tagebuch“?

Ungefähr _____ weiße Streifen sind so lang wie die lange Seite von „Gregs Tagebuch“.

8) Wie viele von den weißen Streifen an der Tafel sind ungefähr so lang wie ein geschlossener Knirps-Regenschirm?

Ungefähr _____ weiße Streifen sind so lang wie ein geschlossener Knirps-Regenschirm.

9) Wie viele von den weißen Streifen an der Tafel sind ungefähr so lang wie ein Duplo-Schokoriegel?

Ungefähr _____ weiße Streifen sind so lang wie ein Duplo-Schokoriegel.

10) Wie lang ist der rote Streifen an der Tafel ungefähr?

Der rote Streifen an der Tafel ist ungefähr_____ cm lang.

11) Wie lang ist eine Seite der Ritter-Sport-Schokolade an der Tafel ungefähr?

Eine Seite der Ritter-Sport-Schokolade ist ungefähr _____ cm lang.

12) Wie lang ist die lange Seite des Posters an der Tafel ungefähr?

Die lange Seite des Posters ist ungefähr _____ cm lang.

13) Wie viele Fineliner sind ungefähr so lang wie der Ordner an der Tafel hoch ist?

Ungefähr _____ Fineliner sind so lang wie der Ordner an der Tafel hoch ist.

14) Wie viele Fineliner sind ungefähr so lang wie die kurze Seite des Handtuchs an der Tafel?

Ungefähr _____ Fineliner sind so lang wie die kurze Seite des Handtuchs an der Tafel.

15) Wie viele Fineliner sind ungefähr so lang wie der Löffel an der Tafel?

Ungefähr _____ Fineliner sind so lang wie der Löffel an der Tafel.

16) Wie viele von den weißen Streifen an der Tafel sind ungefähr so lang wie die kurze Seite des Posters an der Tafel?

Ungefähr _____ weiße Streifen sind so lang wie die kurze Seite des Posters an der Tafel.

17) Wie viele von den weißen Streifen an der Tafel sind ungefähr so lang wie der grüne Streifen an der Tafel?

Ungefähr _____ weiße Streifen sind so lang wie der grüne Streifen an der Tafel.

18) Wie viele von den weißen Streifen an der Tafel sind ungefähr so lang wie die längste Seite von dem gestreiften Karton an der Tafel?

Ungefähr _____ weiße Streifen sind so lang wie die längste Seite von dem gestreiften Karton an der Tafel.

19) Das blaue Band an der Tafel ist 1m lang.

Wie lang ist ein Rettungswagen ungefähr?

Ein Rettungswagen ist ungefähr _____ m lang.

20) Das blaue Band an der Tafel ist 1m lang.

Wie lang ist eine Seite einer Euro-Palette ungefähr?

Die Seite einer Euro-Palette ist ungefähr _____ m lang.

21) Das blaue Band an der Tafel ist 1m lang.

Wie hoch hängt eine Ampel über der Straße?

Eine Ampel hängt ungefähr _____ m über der Straße.

22) Das blaue Band an der Tafel ist 1m lang.

Wie lang ist das Seil A an der Tafel ungefähr?

Das Seil A an der Tafel ist ungefähr _____ m lang.

23) Das blaue Band an der Tafel ist 1m lang.

Wie lang ist das Seil B an der Tafel ungefähr?

Das Seil B an der Tafel ist ungefähr _____ m lang.

24) Das blaue Band an der Tafel ist 1m lang.

Wie lang ist das Seil C an der Tafel ungefähr?

Das Seil C an der Tafel ist ungefähr _____ m lang.

Danke!

Ich kann Flächeninhalte schätzen!

Ich heiße:_____

Ich bin ein ☐ Mädchen

☐ Junge

Klasse: _____

1) Wie groß ist ein Taschentuch ungefähr?

Ein Taschentuch ist ungefähr _____ cm^2 groß.

2) Wie groß ist die größte Fläche eines einzelnen Maoams ungefähr?

Die größte Fläche eines einzelnen Maoams ist ungefähr _____ cm^2 groß.

3) Wie groß ist ein Autokennzeichen ungefähr?

Ein Autokennzeichen ist ungefähr _____ cm^2 groß.

4) Wie viele Maoams haben ungefähr die gleiche Fläche wie eine CD-Hülle?

Ungefähr _____ Maoams haben die gleiche Fläche wie eine CD-Hülle.

5) Wie viele Maoams haben ungefähr die gleiche Fläche wie eine normale Postkarte?

Ungefähr _____ Maoams haben die gleiche Fläche wie eine normale Postkarte.

6) Wie viele Maoams haben ungefähr die gleiche Fläche wie eine Bankkarte?

Ungefähr _____ Maoams haben die gleiche Fläche wie eine Bankkarte.

7) Wie viele von den grünen Vierecken an der Tafel haben ungefähr die gleiche Fläche wie ein Zeichenblockblatt?

Ungefähr _____ grüne Vierecke haben die gleiche Fläche wie ein Zeichenblockblatt.

8) Wie viele von den grünen Vierecken an der Tafel haben ungefähr die gleiche Fläche wie ein Einbahnstraßen-Schild?

Ungefähr _____ grüne Vierecke haben die gleiche Fläche wie ein Einbahnstraßen-Schild.

9) Wie viele von den grünen Vierecken an der Tafel haben ungefähr die gleiche Fläche wie eine normale Postkarte?

Ungefähr _____ grüne Vierecke haben die gleiche Fläche wie eine normale Postkarte.

10) Wie groß ist das gelbe Viereck an der Tafel ungefähr?

Das gelbe Viereck an der Tafel ist ungefähr _____ cm^2 groß.

11) Wie groß ist der Kalender an der Tafel ungefähr?

Der Kalender an der Tafel ist ungefähr _____ cm^2 groß.

12) Wie groß ist die DVD-Hülle an der Tafel ungefähr?

Die DVD-Hülle an der Tafel ist ungefähr _____ cm^2 groß.

13) Wie viele einzelne Maoams haben die gleiche Fläche wie das grüne Viereck an der Tafel?

Ungefähr _____ einzelne Maoams haben die gleiche Fläche wie das grüne Viereck an der Tafel.

14) Wie viele einzelne Maoams haben die gleiche Fläche wie der Kalender an der Tafel?

Ungefähr _____ einzelne Maoams haben die gleiche Fläche wie der Kalender an der Tafel.

15) Wie viele einzelne Maoams haben die gleiche Fläche wie die Ritter-Sport-Schokolade an der Tafel?

Ungefähr _____ einzelne Maoams haben die gleiche Fläche wie die Ritter-Sport-Schokolade an der Tafel.

16) Wie viele von den grünen Vierecken an der Tafel haben ungefähr die gleiche Fläche wie das braune Viereck an der Tafel?

Ungefähr _____ grüne Vierecke haben die gleiche Fläche wie das braune Viereck an der Tafel.

17) Wie viele von den grünen Vierecken an der Tafel haben ungefähr die gleiche Fläche wie das Poster an der Tafel?

Ungefähr _____ grüne Vierecke haben die gleiche Fläche wie das Poster an der Tafel.

18) Wie viele von den grünen Vierecken an der Tafel haben ungefähr die gleiche Fläche wie das Handtuch an der Tafel?

Ungefähr _____ grüne Vierecke haben die gleiche Fläche wie das Handtuch an der Tafel.

19) Das braune Viereck an der Tafel ist $1m^2$ groß.

Wie groß ist ein Weichboden in der Sporthalle ungefähr?

Ein Weichboden in der Sporthalle ist ungefähr _____ m^2 groß.

20) Das braune Viereck an der Tafel ist $1m^2$ groß.

Wie groß ist eine normale Bettdecke ungefähr?

Eine normale Bettdecke ist ungefähr _____ m^2 groß.

21) Das braune Viereck an der Tafel ist 1m^2 groß.

Wie groß ist die Fläche eines Zebrastreifen-Streifens ungefähr?

Ein Zebrastreifen-Streifen ist ungefähr _____ m^2 groß.

22) Das braune Viereck an der Tafel ist 1m^2 groß.

Wie groß ist die rote Fläche an der Tafel ungefähr?

Die rote Fläche an der Tafel ist ungefähr _____ m^2 groß.

23) Das braune Viereck an der Tafel ist 1m^2 groß.

Wie groß ist das Poster an der Tafel ungefähr?

Das Poster an der Tafel ist ungefähr _____ m^2 groß.

24) Das braune Viereck an der Tafel ist 1m^2 groß.

Wie groß ist das Handtuch an der Tafel ungefähr?

Das Handtuch an der Tafel ist ungefähr _____ m^2 groß.

Danke!

Ich kann Fassungsvermögen schätzen!

Ich heiße:_____

Ich bin ein ☐ Mädchen

☐ Junge

Klasse: _____

1) Wie viel Wasser passt ungefähr in einen Fruchtzwerge-Becher?

In einen Fruchtzwerge-Becher passen ungefähr _____ ml Wasser.

2) Wie viel Wasser passt ungefähr in eine Badewanne?

In eine Badewanne passen ungefähr _____ l Wasser.

3) Wie viel Wasser passt ungefähr in eine Shampoo-Flasche?

In eine Shampoo-Flasche passen ungefähr _____ ml Wasser.

4) Wie oft muss man ungefähr von einem kleinen Putzeimer Wasser in eine Regentonne füllen, damit sie voll ist?

Man muss ungefähr _____ mal Wasser von einem kleinen Putzeimer in eine Regentonne füllen, damit sie voll ist.

5) Wie oft muss man ungefähr von einem kleinen Putzeimer Wasser in eine Badewanne füllen, damit sie voll ist?

Man muss ungefähr _____ mal Wasser von einem kleinen Putzeimer in eine Badewanne füllen, damit sie voll ist.

6) Wie oft muss man ungefähr Wasser von einem Fruchtzwerge-Becher in ein Honigglas füllen, damit es voll ist?

Man muss ungefähr _____ mal Wasser von einem Fruchtzwerge-Becher in ein Honigglas füllen, damit es voll ist.

7) Wie oft muss man ungefähr Wasser von dem weißen Gefäß an der Tafel in eine Badewanne füllen, damit sie voll ist?

Man muss ungefähr _____ mal Wasser von dem weißen Gefäß an der Tafel in eine Badewanne füllen, damit sie voll ist.

8) Wie oft muss man ungefähr Wasser von dem weißen Gefäß an der Tafel in einen kleinen Putzeimer füllen, damit er voll ist?

Man muss ungefähr _____ mal Wasser von dem weißen Gefäß an der Tafel in einen kleinen Putzeimer füllen, damit er voll ist.

9) Wie oft muss man ungefähr Wasser von dem weißen Gefäß an der Tafel in eine große Wasserflasche füllen, damit sie voll ist?

Man muss ungefähr _____ mal Wasser von dem weißen Gefäß an der Tafel in eine große Wasserflasche füllen, damit sie voll ist.

10) Wie viel Wasser passt ungefähr in die grüne Flasche an der Tafel?

In die grüne Flasche an der Tafel passen ungefähr _____ ml Wasser.

11) Wie viel Wasser passt ungefähr in die Gießkanne an der Tafel?

In die Gießkanne an der Tafel passen ungefähr _____ ml Wasser.

12) Wie viel Wasser passt ungefähr in den gestreiften Karton an der Tafel?

In den gestreiften Karton an der Tafel passen ungefähr _____ l Wasser.

13) Wie oft muss man ungefähr Wasser von einem Fruchtzwerge-Becher in die schwarze Flasche an der Tafel füllen, damit sie voll ist?

Man muss ungefähr _____ mal Wasser von einem Fruchtzwerge-Becher in die schwarze Flasche an der Tafel füllen, damit sie voll ist.

14) Wie oft muss man ungefähr Wasser von einem Fruchtzwerge-Becher in die Keksdose an der Tafel füllen, damit sie voll ist?

Man muss ungefähr _____ mal Wasser von einem Fruchtzwerge-Becher in die Keksdose an der Tafel füllen, damit sie voll ist.

15) Wie oft muss man ungefähr Wasser von einem kleinen Putzeimer in das Planschbecken vor der Tafel füllen, damit es voll ist?

Man muss ungefähr _____ mal Wasser von einem kleinen Putzeimer in das Planschbecken vor der Tafel füllen, damit es voll ist.

16) Wie oft muss man ungefähr Wasser von dem weißen Gefäß an der Tafel in den gestreiften Karton vor der Tafel füllen, damit er voll ist?

Man muss ungefähr _____ mal Wasser von dem weißen Gefäß an der Tafel in den gestreiften Karton vor der Tafel füllen, damit er voll ist.

17) Wie oft muss man ungefähr Wasser von dem weißen Gefäß an der Tafel in den Stiftebecher an der Tafel füllen, damit er voll ist?

Man muss ungefähr _____ mal Wasser von dem weißen Gefäß an der Tafel in den Stiftebecher an der Tafel füllen, damit er voll ist.

18) Wie oft muss man ungefähr Wasser von dem weißen Gefäß an der Tafel in die Keksdose an der Tafel füllen, damit sie voll ist?

Man muss ungefähr _____ mal Wasser von dem weißen Gefäß an der Tafel in die Keksdose an der Tafel füllen, damit sie voll ist.

19) In die blaue Flasche an der Tafel passen 1l Wasser.

Wie viel Wasser passt ungefähr in einen kleinen Putzeimer?

In einen kleinen Putzeimer passen ungefähr _____ l Wasser.

20) In die blaue Flasche an der Tafel passen 1l Wasser.

Wie viel Wasser passt ungefähr in eine Papiertonne?

In eine Papiertonne passen ungefähr _____ l Wasser.

21) In die blaue Flasche an der Tafel passen 1l Wasser.

Wie viel Wasser passt ungefähr in einen gelben Recycling-Sack?

In einen gelben Recycling-Sack passen ungefähr _____ l Wasser.

22) In die blaue Flasche an der Tafel passen 1l Wasser.

Wie viel Wasser passt ungefähr in die Chipsdose an der Tafel?

In die Chipsdose an der Tafel passen ungefähr _____ l Wasser.

23) In die blaue Flasche an der Tafel passen 1l Wasser.

Wie viel Wasser passt ungefähr in den Blumentopf an der Tafel?

In den Blumentopf an der Tafel passen ungefähr _____ l Wasser.

24) In die blaue Flasche an der Tafel passen 1l Wasser.

Wie viel Wasser passt ungefähr in die Keksdose an der Tafel?

In die Keksdose an der Tafel passen ungefähr _____ l Wasser.

Danke!

Ich kann Rauminhalte schätzen!

Ich heiße:_____

Ich bin ein ☐ Mädchen

☐ Junge

Klasse: _____

1) Wie groß ist ein 8er-Legostein ungefähr?

Ein 8er-Legostein ist ungefähr _____ cm^3 groß.

2) Wie groß ist ein Schiffscontainer ungefähr?

Ein Schiffscontainer ist ungefähr _____ cm^3 groß.

3) Wie groß ist eine Packung Butter ungefähr?

Eine Packung Butter ist ungefähr _____ cm^3 groß.

4) Wie viele 8er-Legosteine sind ungefähr genauso groß wie eine Kassettenhülle?

Ungefähr _____ 8er-Legosteine sind genauso groß wie eine Kassettenhülle.

5) Wie viele 8er-Legosteine sind ungefähr genauso groß wie eine Packung Butter?

Ungefähr _____ 8er-Legosteine sind genauso groß wie eine Packung Butter.

6) Wie viele 8er-Legosteine sind ungefähr genauso groß wie ein Milchkarton?

Ungefähr _____ 8er-Legosteine sind genauso groß wie ein Milchkarton.

7) Wie viele von den weißen Würfeln an der Tafel sind ungefähr genauso groß wie eine 10er-Eierpackung?

Ungefähr _____ weiße Würfel sind genauso groß wie eine 10er-Eierpackung.

8) Wie viele von den weißen Würfeln an der Tafel sind ungefähr genauso groß wie eine Packung Butter?

Ungefähr _____ weiße Würfel sind genauso groß wie eine Packung Butter.

9) Wie viele von den weißen Würfeln an der Tafel sind ungefähr genauso groß wie ein Paket Kopierpapier?

Ungefähr _____ weiße Würfel sind genauso groß wie ein Paket Kopierpapier.

10) Wie groß ist die Keksdose an der Tafel ungefähr?

Die Keksdose an der Tafel ist ungefähr _____ cm^3 groß.

11) Wie groß ist der gestreifte Karton an der Tafel ungefähr?

Der gestreifte Karton an der Tafel ist ungefähr _____ cm^3 groß.

12) Wie groß ist die Butterkeks-Packung an der Tafel ungefähr?

Die Butterkeks-Packung an der Tafel ist ungefähr _____ cm^3 groß.

13) Wie viele 8er-Legosteine sind ungefähr genauso groß wie die Keksdose an der Tafel?

Ungefähr _____ 8er-Legosteine sind genauso groß wie die Keksdose an der Tafel.

14) Wie viele 8er-Legosteine sind ungefähr genauso groß wie der Schwamm an der Tafel?

Ungefähr _____ 8er-Legosteinie sind genauso groß wie der Schwamm an der Tafel.

15) Wie viele 8er-Legosteine sind ungefähr genauso groß wie die Spielpackung an der Tafel?

Ungefähr _____ 8er-Legosteine sind genauso groß wie die Spielpackung an der Tafel.

16) Wie viele von den weißen Würfeln an der Tafel sind ungefähr genauso groß wie der grüne Karton an der Tafel?

Ungefähr _____ weiße Würfel sind genauso groß wie der grüne Karton an der Tafel.

17) Wie viele von den weißen Würfeln an der Tafel sind ungefähr genauso groß wie das Buch an der Tafel?

Ungefähr _____ weiße Würfel sind genauso groß wie das Buch an der Tafel.

18) Wie viele von den weißen Würfeln an der Tafel sind ungefähr genauso groß wie der gelbe Karton an der Tafel?

Ungefähr _____ weiße Würfel sind genauso groß wie der gelbe Karton an der Tafel.

19) Der rote Würfel an der Tafel ist 1 dm^3 groß.

Wie groß ist eine 10er-Eierpackung ungefähr?

Eine 10er-Eierpackung ist ungefähr _____ dm^3 groß.

20) Der rote Würfel an der Tafel ist 1 dm^3 groß.

Wie groß ist eine Papiertonne ungefähr?

Eine Papiertonne ist ungefähr _____ dm^3 groß.

21) Der rote Würfel an der Tafel ist 1 dm^3 groß.

Wie groß ist ein kleiner Sprungkasten in der Sporthalle ungefähr?

Ein kleiner Sprungkasten in der Sporthalle ist ungefähr _____ dm^3 groß.

22) Der rote Würfel an der Tafel ist 1 dm^3 groß.

Wie groß ist der Koffer an der Tafel ungefähr?

Der Koffer an der Tafel ist ungefähr _____ dm^3 groß.

23) Der rote Würfel an der Tafel ist 1 dm^3 groß.

Wie groß ist die Holzkiste an der Tafel ungefähr?

Die Holzkiste an der Tafel ist ungefähr _____ dm^3 groß.

24) Der rote Würfel an der Tafel ist 1 dm^3 groß.

Wie groß ist der schwarze Blumentopf an der Tafel ungefähr?

Der schwarze Blumentopf an der Tafel ist ungefähr _____ dm^3 groß.

Danke!

C.2 Schätzobjekte mit Realwerten (Pilotstudie)

Tabelle 4

Schätzobjekte für die Größe Längen mit Realwerten (Pilotstudie)

Schätzobjekt	**Realwert**
Fineliner Länge	16,5 cm
Schiffscontainer Höhe	2,5 m
Teebeutel Breite	4,5 cm
Donald Duck Heft kurze Seite	1 Fineliner
Zebrastreifen-Streifen lange Seite	18 Fineliner
Zeichenblock lange Seite	2,5 Fineliner
Gregs Tagebuch lange Seite	1,5 weiße Streifen
Knirps Regenschirm Länge	2 weiße Streifen
Duplo-Schokoriegel Länge	1 weißer Streifen
Roter Streifen Länge	22 cm
Ritter-Sport-Schokolade Länge	9 cm
Poster lange Seite	83 cm
Ordner Höhe	2 Fineliner
Handtuch kurze Seite	3,7 Fineliner
Löffel	1,3 Fineliner
Poster kurze Seite	5 weiße Streifen
Grüner Streifen	2 weiße Streifen
Karton längste Seite	4 weiße Streifen
Rettungswagen Länge	6 m
Euro-Palette	1,2 m
Höhe Ampel über der Straße	4,5 m
Seil A Länge	3 m
Seil B Länge	1,5 m
Seil C Länge	2 m

Tabelle 5

Schätzobjekte für die Größe Flächeninhalte mit Realwerten (Pilotstudie)

Schätzobjekt	**Realwert**
Taschentuch	441 cm^2
Einzelner Maoam größte Fläche	4 cm^2
Autokennzeichen	572 cm^2
CD-Hülle	35 Maoams
Postkarte	30 Maoams
Bankkarte	10 Maoams
Zeichenblockblatt	4 grüne Vierecke
Einbahnstraßen-Schild	7,6 grüne Vierecke
Postkarte	0,5 grüne Vierecke
Gelbes Viereck	500 cm^2
Kalender	387 cm^2
DVD-Hülle	257 cm^2
Grünes Viereck	67,5 cm^2
Kalender	75 Maoams
Ritter-Sport-Schokolade	15 Maoams
Braunes Viereck	31 grüne Vierecke
Poster	16 grüne Vierecke
Handtuch	26 grüne Vierecke
Weichboden	6 m^2
Normale Bettdecke	2,7 m^2
Zebrastreifen-Streifen	1,5 m^2
Rote Fläche	1,5 m^2
Poster	0,5 m^2
Handtuch	0,8 m^2

Tabelle 6

Schätzobjekte für die Größe Fassungsvermögen mit Realwerten (Pilotstudie)

Schätzobjekt	**Realwert**
Fruchtzwerge Becher	50 ml
Badewanne	200 l
Shampoo-Flasche	250 ml
Regentonne	110 kleine Putzeimer
Badewanne	40 kleine Putzeimer
Honigglas	7 kleine Fruchtzwerge Becher
Badewanne	400 weiße Gefäße
Kleiner Putzeimer	10 weiße Gefäße
Große Wasserflasche	3 weiße Gefäße
Grüne Flasche	750 ml
Gießkanne	1700 ml
Karton	41,7 l
Schwarze Flasche	2,5 kleine Fruchtzwerge Becher
Keksdose	40 kleine Fruchtzwerge Becher
Planschbecken	8,5 kleine Putzeimer
Karton	83,4 weiße Gefäße
Stiftebecher	0,8 weiße Gefäße
Keksdose	4 weiße Gefäße
Kleiner Putzeimer	5 l
Papiertonne	240 l
Gelber Recycling-Sack	90 l
Chipsdose	1 l
Blumentopf	2 l
Keksdose	1,8 l

Tabelle 7

Schätzobjekte für die Größe Rauminhalte mit Realwerten (Pilotstudie)

Schätzobjekt	**Realwert**
8er Legostein	4 cm^3
Schiffscontainer	37500000 cm^3
Packung Butter	245 cm^3
Kassettenhülle	28 8er-Legosteine
Packung Butter	60 8er-Legosteine
Milchkarton	160 8er-Legosteine
10er-Eierpackung	36 weiße Würfel
Packung Butter	9 weiße Würfel
Paket Kopierpapier	117 weiße Würfel
Keksdose	2588 cm^3
Karton	41760 cm^3
Butterkekspackung	819 cm^3
Keksdose	490 8er-Legosteine
Schwamm	48 8er-Legosteine
Spielpackung	48 8er-Legosteine
Grüner Karton	56 weiße Würfel
Buch	24 weiße Würfel
Gelber KArton	20 weiße Würfel
10er-Eierpackung	1,7 dm^3
Papiertonne	408 dm^3
Sprungkasten	140 dm^3
Koffer	10,9 dm^3
Holzkiste	6 dm^3
Blumentopf	1 dm^3

C.3 Manual (Pilotstudie)

Einführung

Hallo, mein Name ist ____ und ich arbeite an der Universität Lüneburg. Wisst ihr, was eine Universität ist?

Wenn man an einer Universität arbeitet, kann man Sachen erforschen. Ich erforsche, wie gut Kinder schätzen können. Schätzen bedeutet, dass man z.B. die Länge/den Flächeninhalt/das Fassungsvermögen/den Rauminhalt eines Gegenstandes sagen soll, ohne mit einem Lineal/einem Messbecher oder einem anderen Messinstrument nachzumessen.

Das wollen wir heute auch machen. Ich habe ein paar Aufgaben mitgebracht, in denen ihr Längen/Flächeninhalte/Fassungsvermögen/Rauminhalte schätzen sollt. Das sind ganz unterschiedliche Aufgaben: Manchmal sind die Gegenstände in echt da (an der Tafel), manchmal müsst ihr euch sie nur vorstellen. Und manchmal sollt ihr zwei Gegenstände miteinander vergleichen. Das steht dann alles genau in den Aufgaben.

Weil ihr kein Lineal oder ein anderes Messgerät benutzen dürft, ist es nicht schlimm, wenn ihr nicht ganz genau das richtige Ergebnis wisst. Ihr sollt nur so genau wie ihr könnt schätzen.

Die Aufgaben stehen alle in diesem Heft. Ich lege das gleich vor euch hin, und ihr dürft dann euren Namen und eure Klasse dort eintragen und ankreuzen, ob ihr ein Junge oder ein Mädchen seid. Bitte fangt noch nicht an, wir fangen alle gemeinsam an. Statt eurem echten Namen dürft ihr euch einen Fantasienamen ausdenken. Den müsst ihr euch dann merken, weil ich nochmal wiederkomme und ihr dann genau den gleichen Namen benutzen sollt wie heute.

HEFTE AUSTEILEN

Eder soll für sich in seinem eigenen Tempo arbeiten. Bitte schaut nicht, welche Ergebnisse bei euren Sitznachbarn stehen. Wenn ihr etwas nicht wisst oder versteht, könnt ihr die Frage einfach überspringen, das ist dann überhaupt nicht schlimm. Ihr werdet nicht benotet, sondern die Ergebnisse sind nur für mich.

Habt ihr noch Fragen?

Dann dürft ihr jetzt anfangen.

C.4 Beobachtungsbogen (Pilotstudie)

Evaluation

Abgabe erstes Heft: _____

Abgabe letztes Heft: _____

Fragen vor Beginn des Tests:

Fragen während des Tests:

Auffälligkeiten während der Testdurchführung:

Wann wurde Schätzen/Messen/die Größe das letzte Mal im Unterricht behandelt?

C.5 Schätztest Hauptstudie Version A

Ich kann schätzen!

Ich heiße _____

Ich bin ein ☐ Mädchen

☐ Junge

Ich habe am ____ ____ _____ Geburtstag.

Klasse: _____

Längen

1) Wie lang ist ein Duplo-Riegel ungefähr?

Ein Duplo-Riegel ist ungefähr _____ cm lang.

2) Wie lang ist ein Schwebebalken ungefähr?

Ein Schwebebalken ist ungefähr _____ m lang.

3) Wie breit ist ein Backofen ungefähr?

Ein Backofen ist ungefähr _____ cm breit.

4) Wie hoch ist das Brandenburger Tor (Berlin) ungefähr?

Das Brandenburger Tor ist ungefähr _____ m hoch.

5) Wie lang ist der rote Streifen an der Tafel ungefähr?

Der rote Streifen an der Tafel ist ungefähr _____ cm lang.

6) Wie lang ist das blaue Band an der Tafel ungefähr?

Das blaue Band an der Tafel ist ungefähr _____ m lang.

7) Wie lang ist die kurze Seite des Posters an der Tafel ungefähr?

Die kurze Seite des Posters an der Tafel ist ungefähr _____ cm lang.

8) Wie lang ist das Seil an der Tafel ungefähr?

Das Seil an der Tafel ist ungefähr _____ m lang.

9) Das blaue Band an der Tafel ist 1 m lang.

Wie lang ist das Schiff „Titanic“ ungefähr?

Das Schiff „Titanic“ ist ungefähr _____ m lang.

10) Das gelbe Band an der Tafel ist 1 cm lang.

Wie lang ist die lange Seite eines 5€-Scheins ungefähr?

Die lange Seite eines 5€-Scheins ist ungefähr _____ cm lang.

11) Das blaue Band an der Tafel ist 1 m lang.

Wie lang ist ein Tennis-Platz ungefähr?

Ein Tennis-Platz ist ungefähr _____ m lang.

12) Das gelbe Band an der Tafel ist 1 cm lang.

Wie lang ist eine Barbie-Puppe ungefähr?

Eine Barbie-Puppe ist ungefähr _____ cm lang.

Flächeninhalte

13) Wie groß ist ein auseinander gefaltetes Taschentuch ungefähr?

Ein auseinander gefaltetes Taschentuch ist ungefähr _____ cm^2 groß.

14) Wie groß ist eine Rettungsdecke ungefähr?

Eine Rettungsdecke ist ungefähr _____ m^2 groß.

15) Wie groß ist ein gelber Klebe-Notiz-Zettel ungefähr?

Ein gelber Klebe-Notiz-Zettel ist ungefähr _____ cm^2 groß.

16) Wie groß ist ein Boxring ungefähr?

Ein Boxring ist ungefähr _____ m^2 groß.

17) Wie groß ist das gelbe Viereck an der Tafel ungefähr?

Das gelbe Viereck an der Tafel ist ungefähr _____ cm^2 groß.

18) Wie groß ist das braune Viereck an der Tafel ungefähr?

Das braune Viereck an der Tafel ist ungefähr _____ m^2 groß.

19) Wie groß ist der Kalender an der Tafel ungefähr?

Der Kalender an der Tafel ist ungefähr _____ cm^2 groß.

20) Wie groß ist der rote Stoff an der Tafel ungefähr?

Der rote Stoff an der Tafel ist ungefähr _____ m^2 groß.

21) Das braune Viereck an der Tafel ist 1 m^2 groß.

Wie groß ist ein Weichboden in der Sporthalle ungefähr?

Ein Weichboden in der Sporthalle ist ungefähr _____ m^2 groß.

22) Das grüne Viereck an der Tafel ist 1 cm^2 groß.

Wie groß ist eine Bank-Karte ungefähr?

Eine Bankkarte ist ungefähr _____ cm^2 groß.

23) Das braune Viereck an der Tafel ist 1 m^2 groß.

Wie groß ist eine Tischtennis-Platte ungefähr?

Eine Tischtennis-Platte ist ungefähr _____ m^2 groß.

24) Das grüne Viereck an der Tafel ist 1 cm^2 groß.

Wie groß ist ein „Was-ist-was“-Buch ungefähr?

Ein „Was-ist-was“-Buch ist ungefähr _____ cm^2 groß.

Fassungsvermögen

25) Wie viel Wasser passt ungefähr in einen kleinen Fruchtzwerge-Becher?

In einen kleinen Fruchtzwerge-Becher passen ungefähr _____ ml Wasser.

26) Wie viel Wasser passt ungefähr in einen gelben Müllsack für Plastik?

In einen gelben Müllsack für Plastik passen ungefähr _____ l Wasser.

27) Wie viel Wasser passt ungefähr in eine Capri-Sonne?

In eine Capri-Sonne passen ungefähr _____ ml Wasser.

28) Wie viel Wasser passt ungefähr in einen Turnbeutel?

In einen Turnbeutel passen ungefähr _____ l Wasser.

29) Wie viel Wasser passt ungefähr in die blaue Flasche an der Tafel?

In die blaue Flasche an der Tafel passen ungefähr _____ l Wasser.

30) Wie viel Wasser passt ungefähr in die Duschgel-Flasche an der Tafel?

In die Duschgel-Flasche an der Tafel passen ungefähr _____ ml Wasser.

31) Wie viel Wasser passt ungefähr in das Planschbecken an der Tafel?

In das Planschbecken an der Tafel passen ungefähr _____ l Wasser.

32) Wie viel Wasser passt ungefähr in die schwarze Flasche an der Tafel?

In die schwarze Flasche an der Tafel passen ungefähr _____ ml Wasser.

33) In die blaue Flasche an der Tafel passt 1 l Wasser.

Wie viel Wasser passt ungefähr in eine Papiertonne (vor einem Haus)?

In eine Papiertonne passen ungefähr _____ l Wasser.

34) In das die grüne Flasche an der Tafel passt 1 ml Wasser.

Wie viel Wasser passt ungefähr in ein kleines Teelicht?

In ein kleines Teelicht passen ungefähr _____ ml Wasser.

35) In die blaue Flasche an der Tafel passt 1 l Wasser.

Wie viel Wasser passt ungefähr in einen Kinderschuhkarton?

In einen Kinderschuhkarton passen ungefähr _____ l Wasser.

36) In die grüne Flasche an der Tafel passt 1 ml Wasser.

Wie viel Wasser passt ungefähr in ein Wasserglas?

In ein Wasserglas passen ungefähr _____ ml Wasser.

Rauminhalte

37) Wie groß ist ein 8er-Legostein ungefähr?

Ein 8er-Legostein ist ungefähr _____ cm^3 groß.

38) Wie groß ist ein Schiffscontainer ungefähr?

Ein Schiffscontainer ist ungefähr _____ m^3 groß.

39) Wie groß ist ein Paket Butter ungefähr?

Ein Paket Butter ist ungefähr _____ cm^3 groß.

40) Wie groß ist der hintere Teil eines Rettungswagens ungefähr?

Der hintere Teil eines Rettungswagens ist ungefähr _____ m^3 groß.

41) Wie groß ist die Spielpackung an der Tafel ungefähr?

Die Spielpackung an der Tafel ist ungefähr _____cm^3 groß.

42) Wie groß ist der Klassenraum ungefähr?

Der Klassenraum ist ungefähr _____m^3 groß.

43) Wie groß ist der gelbe Karton an der Tafel ungefähr?

Der gelbe Karton an der Tafel ist ungefähr _____cm^3 groß.

44) Wie groß ist das Würfelmodell an der Tafel ungefähr?

Das Würfelmodell an der Tafel ist ungefähr _____m^3 groß.

45) Der rote Würfel an der Tafel ist ungefähr 1 cm^3 groß.

Wie groß ist eine Butterkekspackung ungefähr?

Eine Butterkekspackung ist ungefähr _____ cm^3 groß.

46) Das Würfelmodell an der Tafel ist ungefähr 1 m^3 groß.

Wie groß ist ein normaler Stadtbus ungefähr?

Ein normaler Stadtbus ist ungefähr _____ m^3 groß.

47) Der rote Würfel an der Tafel ist ungefähr 1 cm^3 groß.

Wie groß ist eine einzelne eckige Smarties-Packung ungefähr?

Eine einzelne eckige Smarties-Packung ist ungefähr _____ cm^3 groß.

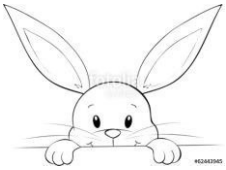
48) Das Würfelmodell an der Tafel ist ungefähr 1 m^3 groß.

Wie groß ist ein Sammelbehälter für Altkleider ungefähr?

Ein Sammelbehälter für Altkleider ist ungefähr _____ m^3 groß.

Danke!

C.6 Schätztest Hauptstudie Version B



Ich kann schätzen!

Ich heiße _____

Ich bin ein ☐ Mädchen

☐ Junge

Ich habe am ____ ____ _____ Geburtstag.

Klasse: _____

Längen

1) Wie lang ist die lange Seite eines 5€-Scheins ungefähr?

Die lange Seite eines 5€-Scheins ist ungefähr _____ cm lang.

2) Wie lang ist ein Tennis-Platz ungefähr?

Ein Tennis-Platz ist ungefähr _____ m lang.

3) Wie lang ist eine Barbie-Puppe ungefähr?

Eine Barbie-Puppe ist ungefähr _____ cm lang.

4) Wie lang ist das Schiff „Titanic“ ungefähr?

Das Schiff „Titanic“ ist ungefähr _____ m lang.

5) Wie lang ist der rote Streifen an der Tafel ungefähr?

Der rote Streifen an der Tafel ist ungefähr _____ cm lang.

6) Wie lang ist das blaue Band an der Tafel ungefähr?

Das blaue Band an der Tafel ist ungefähr _____ m lang.

7) Wie lang ist die kurze Seite des Posters an der Tafel ungefähr?

Die kurze Seite des Posters an der Tafel ist ungefähr _____ cm lang.

8) Wie lang ist das Seil an der Tafel ungefähr?

Das Seil an der Tafel ist ungefähr _____ m lang.

9) Das blaue Band an der Tafel ist 1 m lang.

Wie hoch ist das Brandenburger Tor (Berlin) ungefähr?

Das Brandenburger Tor (Berlin) ist ungefähr _____ m hoch.

10) Das gelbe Band an der Tafel ist 1 cm lang.

Wie lang ist ein Duplo-Riegel ungefähr?

Ein Duplo-Riegel ist ungefähr _____ cm lang.

11) Das blaue Band an der Tafel ist 1 m lang.

Wie lang ist ein Schwebebalken ungefähr?

Ein Schwebebalken ist ungefähr _____ m lang.

12) Das gelbe Band an der Tafel ist 1 cm lang.

Wie breit ist ein Backofen ungefähr?

Ein Backofen ist ungefähr _____ cm breit.

Flächeninhalte

13) Wie groß ist eine Bank-Karte ungefähr?

Eine Bank-Karte ist ungefähr _____ cm^2 groß.

14) Wie groß ist ein Weichboden in der Sporthalle ungefähr?

Ein Weichboden in der Sporthalle ist ungefähr _____ m^2 groß.

15) Wie groß ist ein „Was-ist-was“-Buch ungefähr?

Ein „Was-ist-was“-Buch ist ungefähr _____ cm^2 groß.

16) Wie groß ist eine Tischtennis-Platte ungefähr?

Eine Tischtennis-Platte ist ungefähr _____ m^2 groß.

17) Wie groß ist das gelbe Viereck an der Tafel ungefähr?

Das gelbe Viereck an der Tafel ist ungefähr _____ cm^2 groß.

18) Wie groß ist das braune Viereck an der Tafel ungefähr?

Das braune Viereck an der Tafel ist ungefähr _____ m^2 groß.

19) Wie groß ist der Kalender an der Tafel ungefähr?

Der Kalender an der Tafel ist ungefähr _____ cm^2 groß.

20) Wie groß ist der rote Stoff an der Tafel ungefähr?

Der rote Stoff an der Tafel ist ungefähr _____ m^2 groß.

21) Das braune Viereck an der Tafel ist 1 m^2 groß.

Wie groß ist eine Rettungsdecke ungefähr?

Eine Rettungsdecke ist ungefähr _____ m^2 groß.

22) Das grüne Viereck an der Tafel ist 1 cm^2 groß.

Wie groß ist ein auseinander gefaltetes Taschentuch ungefähr?

Ein auseinander gefaltetes Taschentuch ist ungefähr _____ cm^2 groß.

23) Das braune Viereck an der Tafel ist 1 m^2 groß.

Wie groß ist ein Boxring ungefähr?

Ein Boxring ist ungefähr _____ m^2 groß.

24) Das grüne Viereck an der Tafel ist 1 cm^2 groß.

Wie groß ist ein gelber Klebe-Notiz-Zettel ungefähr?

Ein gelber Klebe-Notiz-Zettel ist ungefähr _____ cm^2 groß.

Fassungsvermögen

25) Wie viel Wasser passt ungefähr in ein kleines Teelicht?

In ein kleines Teelicht passen ungefähr _____ ml Wasser.

26) Wie viel Wasser passt ungefähr in eine Papiertonne (vor einem Haus)?

In eine Papiertonne passen ungefähr _____ l Wasser.

27) Wie viel Wasser passt ungefähr in ein Wasserglas?

In ein Wasserglas passen ungefähr _____ ml Wasser.

28) Wie viel Wasser passt ungefähr in einen Kinderschuhkarton?

In einen Kinderschuhkarton passen ungefähr _____ l Wasser.

29) Wie viel Wasser passt ungefähr in die blaue Flasche an der Tafel?

In die blaue Flasche an der Tafel passen ungefähr _____ l Wasser.

30) Wie viel Wasser passt ungefähr in die Duschgel-Flasche an der Tafel?

In die Duschgel-Flasche an der Tafel passen ungefähr _____ ml Wasser.

31) Wie viel Wasser passt ungefähr in das Planschbecken an der Tafel?

In das Planschbecken an der Tafel passen ungefähr _____ l Wasser.

32) Wie viel Wasser passt ungefähr in die schwarze Flasche an der Tafel?

In die schwarze Flasche an der Tafel passen ungefähr _____ ml Wasser.

33) In die blaue Flasche an der Tafel passt 1 l Wasser.
Wie viel Wasser passt ungefähr in einen gelben Müllsack für Plastik?

In einen gelben Müllsack für Plastik passen ungefähr _____ l Wasser.

34) In die grüne Flasche an der Tafel passt 1 ml Wasser.

Wie viel Wasser passt ungefähr in einen kleinen Fruchtzwerge-Becher?

In einen kleinen Fruchtzwerge-Becher passen ungefähr _____ ml Wasser.

35) In die blaue Flasche an der Tafel passt 1 l Wasser.

Wie viel Wasser passt ungefähr in einen Turnbeutel?

In einen Turnbeutel passen ungefähr _____ l Wasser.

36) In die grüne Flasche an der Tafel passt 1 ml Wasser.

Wie viel Wasser passt ungefähr in eine Capri-Sonne?

In eine Capri-Sonne passen ungefähr _____ ml Wasser.

Rauminhalte

37) Wie groß ist eine Butterkekspackung ungefähr?

Eine Butterkekspackung ist ungefähr _____ cm^3 groß.

38) Wie groß ist ein normaler Stadtbus ungefähr?

Ein normaler Stadtbus ist ungefähr _____ m^3 groß.

39) Wie groß ist eine einzelne eckige Smarties-Packung ungefähr?

Eine einzelne eckige Smarties-Packung ist ungefähr _____ cm^3 groß.

40) Wie groß ist ein Sammelbehälter für Altkleider ungefähr?

Ein Sammelbehälter für Altkleider ist ungefähr _____ m^3 groß.

41) Wie groß ist die Spielpackung an der Tafel ungefähr?

Die Spielpackung an der Tafel ist ungefähr _____cm^3 groß.

42) Wie groß ist der Klassenraum ungefähr?

Der Klassenraum ist ungefähr _____m^3 groß.

43) Wie groß ist der gelbe Karton an der Tafel ungefähr?

Der gelbe Karton an der Tafel ist ungefähr _____cm^3 groß.

44) Wie groß ist das Würfelmodell an der Tafel ungefähr?

Das Würfelmodell an der Tafel ist ungefähr _____m^3 groß.

45) Der rote Würfel an der Tafel ist ungefähr 1 cm^3 groß.

Wie groß ist ein 8er-Legostein ungefähr?

Ein 8er-Legostein ist ungefähr _____ cm^3 groß.

46) Das Würfelmodell an der Tafel ist ungefähr 1 m^3 groß.

Wie groß ist ein Schiffscontainer ungefähr?

Ein Schiffscontainer ist ungefähr _____ m^3 groß.

47) Der rote Würfel an der Tafel ist ungefähr 1 cm^3 groß.

Wie groß ist ein Paket Butter ungefähr?

Ein Paket Butter ist ungefähr _____ cm^3 groß.

48) Das Würfelmodell an der Tafel ist ungefähr 1 m^3 groß.

Wie groß ist der hintere Teil eines Rettungswagens ungefähr?

Der hintere Teil eines Rettungswagens ist ungefähr _____ m^3 groß.

Danke!

C.7 Schätzobjekte mit Realwerten (Hauptstudie)

Tabelle 8

Schätzobjekte für die Größe Längen mit Realwerten (Hauptstudie)

Schätzobjekt	**Realwert**
Duplo	11 cm
Schwebebalken	5 m
Backofen	59 m
Brandenburger Tor	20 m
roter Streifen	22 cm
blaues Band	1 m
Poster	59 cm
Seil	5 m
Titanic	269 m
5 Euro	12 cm
Tennisplatz	24 m
Barbie	29 cm

Tabelle 9

Schätzobjekte für die Größe Flächeninhalte mit Realwerten (Hauptstudie)

Schätzobjekt	**Realwert**
Taschentuch	441 cm^2
Rettungsdecke	3 m^2
Post-It	58 cm^2
Boxring	38 m^2
gelbes Viereck	500 cm^2
braunes Viereck	1 m^2
Kalender	390 cm^2
roter Stoff	3 m^2
Weichboden	6 m^2
Bank-Karte	46 cm^2
Tischtennisplatte	4 m^2
Was-ist-was-Buch	644 cm^2

Tabelle 10

Schätzobjekte für die Größe Fassungsvermögen mit Realwerten (Hauptstudie)

Schätzobjekt	**Realwert**
Fruchtzwerge	50 ml
gelber Sack	90 l
Capri-Sonne	200 ml
Turnbeutel	12 l
blaue Flasche	1 l
Duschgel	200 ml
Planschbecken	42 l
schwarze Flasche	125 ml
Papiertonne	240 l
Teelicht	9 ml
Kinderschuhkarton	5 l
Wasserglas	250 ml

Tabelle 11

Schätzobjekte für die Größe Rauminhalte mit Realwerten (Hauptstudie)

Schätzobjekt	**Realwert**
8er Legostein	4 cm^3
Schiffscontainer	88 m^3
Butter	245 cm^3
Rettungswagen	21 m^3
Spielpackung	486 cm^3
Klassenraum	140-260 m^3
gelber Karton	192 cm^3
Würfelmodell	1 m^3
Butterkekspackung	819 cm^3
Stadtbus	90 m^3
Smartiespackung	20 cm^3
Altkleider	2,6 m^3

C.8 Manual (Hauptstudie)

Schätztest

Längen, Flächeninhalte, Fassungsvermögen & Rauminhalte

MANUAL

Einleitung

Um ungefähr die gleichen Testbedingungen in allen Klassen und Testgruppen zu ermöglichen, ist es wichtig, dass einige Aspekte beachtet werden. Bitte folgen Sie den Empfehlungen bezüglich des Materials und den Anweisungen während des Tests. Weisen Sie bitte auch die Lehrkraft daraufhin!

Bitte behalten Sie im Hinterkopf, dass das Ziel des Tests die Erhebung der aktuellen Schätzfähigkeit ist. Bitte vermeiden Sie zu viele Hilfestellungen und Erklärungen – die Test-Situation ist nicht als Lern-Situation konzipiert.

Für den Test werden die folgenden Materialien benötigt:

Längen:	*Roter Streifen, blaues Band, Poster, Seil, gelber Streifen*
Flächeninhalte:	*Gelbes Viereck, Kalender, braunes Viereck, roter Stoff, grünes Viereck,*
Fassungsvermögen:	*Blaue Flasche, Duschgel, Planschbecken, schwarze Flasche, grüne Flasche*
Rauminhalte:	*Spielpackung Ligretto, gelber Karton, großer Würfel, (Klassenraum), roter Würfel*

Die Materialien stehen nach Absprache zur Verfügung. Außerdem benötigen Sie einen *Zollstock*, um Höhe, Breite und Länge des Klassenraumes abzumessen – diese Angaben werden für ein Item der Größe Rauminhalte benötigt.

Für die 4. Klasse sind die Größen Längen und Fassungsvermögen vorgesehen. Die 5. und 6. Klassen bearbeiten Aufgaben zu allen Größen.

Bitte entfernen/verdecken Sie vor Beginn des Tests, möglichst ohne die Aufmerksamkeit der Kinder, sämtliche Messinstrumente im Klassenraum (Tafellineale, Plakate und Poster, Lineale auf den Tischen, …).

Testvorbereitung und Testdurchführung

Es ist wichtig, dass die Materialien in allen Testgruppen in der gleichen oder zumindest in einer sehr ähnlichen Weise angeordnet werden. Es ist darauf zu achten, dass nichts bündig abschließt, sondern dass die Materialien immer leicht versetzt angeordnet werden (auch zur Tafel).

Wenn möglich, arrangieren Sie die Materialien in der folgenden Weise.

Das Seil und den roten Stoff legen Sie dort, wo Platz ist, auf den Fußboden. Sollte der Platz für das Seil nicht reichen, können Sie es wie auf dem Foto ausrichten. Der Stoff sollte jedoch ausgebreitet werden.

Der Kubikmeter sollte *nicht* unter/in der Nähe des Quadratmeters stehen. Beides (Kubik- und Quadratmeter) können Sie auch woanders im Raum aufstellen, wenn der Platz an/bei der Tafel nicht reicht.

Achten Sie darauf, dass alle Kinder einen uneingeschränkten Blick auf die Materialien haben. Es sollte nicht erforderlich sein, dass die Kinder durch die Klasse gehen müssen, um das Material ansehen zu können.

Zur Einführung des Tests benutzen Sie bitte Sie Anleitung vom Beobachtungsbogen. Achten Sie darauf, keine Strategien zu erklären oder anzudeuten. Vermeiden Sie außerdem Beschreibungen von Schätzen als „raten“ oder ähnliches. Abweichungen sollen auf dem Beobachtungsbogen notiert werden.

Während der Durchführung ist auf eine leise Arbeitsatmosphäre zu achten. Jedes Kind arbeitet für sich. Bitte achten Sie darauf, dass die Kinder nichts aus den Testheften der Sitznachbarn abschreiben.

Es ist nicht erlaubt, Messhandlungen vorzunehmen. Bitte achten Sie darauf, dass keine Lineale sichtbar sind (weder auf den Tischen noch das Tafellineal). Auch Messhandlungen mit den Fingern sind nicht erlaubt, wenngleich diese schwer zu entdecken bzw. vermeiden sind.

Die Testzeit beträgt 45 Minuten inklusive der Testeinführung. Es sollte daher nicht zu viel Zeit (nur ca. 5 Minuten) für die Einführung aufgewendet werden. Fragen zu einzelnen Items können während des Tests individuell beantwortet werden. Auch hierbei ist zu beachten:

- Keine Größen mit den Händen zeigen
- Keine Strategien erklären
- Keine Vergleichsobjekte nennen

Wenn ein Kind ein Item auch nach Erklärung nicht versteht, darf es das Item überspringen.

Bitte stellen Sie sicher, dass jedes Testheft mit Namen, Klasse, Geschlecht und Geburtsdatum (Jahr!) versehen ist.

Eintragen der Rohwerte

Für die Auswertung steht eine Excel-Tabelle zur Verfügung, in die die Rohwerte eingetragen werden. Dazu gehören auch die Daten der teilnehmenden Kinder. Bitte tragen sie den Namen, das Geschlecht (m/w) und die Klassenstufe (3/4/5/6) ein. Bilden Sie außerdem für jedes Kind eine ID nach folgendem Muster:

Schulform-Schule_Jahrgang_Klasse_Geschlecht_Alter-Jahr_Alter-Monat_Laufende-Nummer

Für die Bildung der ID gilt:

- Bitte genau vier Buchstaben für Schulform-Schule auswählen (wird vorgegeben)
- Bitte genau eine Zahl und einen Buchstaben für Jahrgang und Klasse auswählen
- Bitte genau einen Buchstaben (M oder W) für Geschlecht auswählen (U, wenn nicht ausgefüllt)
- Bitte genau vier Zahlen für das Alter in Jahren und Monaten auswählen (fehlende Zehner durch 0 ersetzen)

Ein Mädchen aus der Grundschule Deutsch Evern, 3c, welches am 3.12.2009 Geburtstag hat, hat also zunächst folgende ID (für einen Testzeitpunkt Mitte Januar):

GSDE3CW0901

Dieser ID wird eine laufende Nummer hinzugefügt. Diese gehört zu ihrer Schule und ist vierstellig. Die wird Ihnen gesondert mitgeteilt.

Nach diesen Richtlinien hat eine ID immer genau 15 Stellen.

Für das Eintragen der Rohwerte gelten die folgenden Richtlinien:

- Alle Rohwerte werden ohne Einheit eingetragen
- Wenn eine Größenspanne gegeben ist, wird der Mittelwert der beiden Zahlen eingetragen.
- Wenn zwei Werte gegeben sind, wird der Mittelwert der beiden Zahlen eingetragen.

Es kann vorkommen, dass Werte nicht eindeutig zu lesen sind. Bitte folgen Sie in diesen Fällen den folgenden Richtlinien:

- Antworten, die nicht sinnvoll zu interpretieren sind (z.B. wegen einer ergänzten unpassenden Einheit, unleserliche Schrift), werden mit einem “n” eingetragen. Diese werden mit “0” kodiert.
- Felder für Fragen, die nicht beantwortet wurden, werden mit einem „m“ gekennzeichnet. Auch diese werden mit „0“ kodiert.
- Geschriebene Antworten mit Erklärungen wie “1 einhalb” werden in numerische Antworten übersetzt, in diesem Fall “1,5”. Felder, in denen solche Änderungen vorgenommen werden, bitte mit einer pinken Farbhinterlegung kennzeichnen.
- Wenn eine Antwort eine ergänzte Einheit enthält, die zur Größe passt, wird die Antwort in die erfragte Einheit umgerechnet und wie eine standardmäßige Antwort behandelt. Felder, in denen solche Änderungen vorgenommen werden, bitte mit einer grünen Farbhinterlegung kennzeichnen.
- Felder, die eine Größenspanne oder zwei Werte enthalten, müssen in einen Wert umgerechnet werden. Dazu wird das arithmetische Mittel gebildet (z.B. für „5-10“ oder „5, 10“ wird 7,5 eingetragen). Felder, in denen solche Änderungen vorgenommen wurden, bitte mit einer gelben Farbhinterlegung kennzeichnen.

Wenn eine uneindeutige Antwort hier nicht aufgeführt ist, wählen Sie bitte den Fall aus, der der Antwort am nächsten kommt. Insgesamt ist es wichtig, dass für gleiche Fälle immer gleich entschieden wird. Wenn Sie einen neuen Fall einführen, beschreiben Sie ihn bitte in der Datei und kennzeichnen Sie die Felder in einer einheitlichen Farbe.

C.9 Beobachtungsbogen (Hauptstudie)

Text und Anmerkungen für die testleitende Person	**Anmerkungen**
Hallo, mein Name ist ____ und ich arbeite an der Universität Lüneburg. Wisst ihr, was eine Universität ist? Oder etwas ähnliches – kurzer Eisbrecher	
Wenn man an einer Universität arbeitet, kann man Sachen erforschen. Ich erforsche, wie gut Kinder schätzen können. Schätzen bedeutet, dass man z.B. die Länge/den Flächeninhalt/das Fassungsvermögen/den Rauminhalt eines Gegenstandes sagen soll, ohne mit einem Lineal/einem Messbecher oder einem anderen Messinstrument nachzumessen. Oder ähnlich – Hinführung zum heutigen Thema Wörter wie „raten“, „ungenau“ bitte nicht benutzten, auch keine Hinweise auf Strategien oder mögliche Stützpunkte geben	
Das wollen wir heute auch machen. Ich habe ein paar Aufgaben mitgebracht, in denen ihr Längen/Flächeninhalte/Fassungsvermögen/Rauminhalte schätzen sollt. Weil ihr kein Lineal oder ein anderes Messgerät benutzen dürft, ist es nicht schlimm, wenn ihr nicht ganz genau das richtige Ergebnis wisst. Ihr sollt nur so genau wie ihr könnt schätzen. Bitte so nah wie möglich am Text bleiben – auf keinen Fall Strategien oder ein Beispiel erklären!	
Für die Aufgaben habe ich ein bisschen Material mitgebracht, das seht ihr hier vorne. [jedes Material einmal benennen und darauf zeigen]. *Die Aufgaben stehen alle in diesem Heft. Ich lege das gleich vor euch hin, und ihr dürft dann euren Namen euren Geburtstag und eure Klasse dort eintragen und ankreuzen, ob ihr ein Junge oder ein Mädchen seid. Statt eurem echten Namen dürft ihr euch einen Fantasienamen ausdenken.* *Bitte fangt noch nicht an, wir fangen alle gemeinsam an.* Relativ zügig wieder zu den heutigen Aufgaben kommen. Auch hier, falls Nachfragen kommen, keine Strategien erklären, keine Größen mit den Händen zeigen etc.	
Hefte austeilen, darauf achten, dass noch niemand anfängt.	
Jeder soll für sich in seinem eigenen Tempo arbeiten. Bitte schaut nicht, welche Ergebnisse bei euren Sitznachbarn stehen. Bitte bearbeitet die Aufgaben der Reihe nach. Ihr werdet nicht benotet, sondern die Ergebnisse sind nur für mich. Wenn ihr eine Frage habt, könnt ihr euch melden, dann komme ich zu euch. Bitte möglichst nah am Text bleiben. Es soll darum gehen, dass die Kinder nicht zu nervös sind, aber sie sollen auch nicht verleitet werden, nicht mehr nachzudenken (daher nicht laut ansagen, dass sie Fragen überspringen dürfen).	
Habt ihr noch Fragen? … Dann dürft ihr jetzt anfangen. Bitte auf die Uhr sehen.	

Rahmendaten

Schule:

Klasse: Größe des Klassenraumes: _____m^3

Das erste Heft wurde nach _____ Minuten abgegeben.

Das letzte Heft wurde nach _____ Minuten abgegeben.

.

Fragen vor Beginn des Tests:

Fragen während des Tests:

Auffälligkeiten während der Testdurchführung:

Wann wurden Schätzen/Messen/Größen das letzte Mal im Unterricht behandelt? Welches Schulbuch wird verwendet?

Die Erhebung ist ordnungsgemäß und gemäß dem Protokoll durchgeführt worden.

Datum: _____ Unterschrift: _____