

Evidenter Mangel an Evidenz

Reinhard Oldenburg und Timo von Oertzen

Einleitung

Die Forderung, didaktische Entscheidungen auf allen Ebenen – von der einzelnen Handlung einer Lehrkraft bis zur Strukturierung von Curricula – auf solide Evidenz zu gründen, ist ebenso wichtig wie herausfordernd. In dem 2025 vorgestellten „Gutachten der Expertinnen- und Expertenkommission zur Weiterentwicklung der Lehrkräftebildung in Bayern“ wird nicht ein- oder zweimal, sondern 27-mal Evidenzbasierung gefordert. Für viele praxisrelevante Fragestellungen existieren allerdings gar keine Studien, und wenn doch, stellen sich sofort Fragen nach deren Verallgemeinerbarkeit und Anwendbarkeit in konkreten Situationen. Darüberhinaus hat Ioannidis (2005) grundsätzlich argumentiert, dass viele veröffentlichte Forschungsergebnisse falsch sein könnten, und die Replikationskrise in der Psychologie (Pashler & Wagenmakers, 2012) mahnt zur Vorsicht bei der Interpretation empirischer Daten.

Das Forschungsinteresse dieser kurzen Arbeit liegt darin, zu ergründen, inwieweit die bisher erbrachte Evidenz ausreichend ist, und wichtige praktische Entscheidungen treffen zu können. Um dies zu konkretisieren, wird exemplarisch untersucht, inwieweit Forschungsergebnisse zur Bedeutung des Interesses an Mathematik für den Lernfortschritt systematischem Zweifel standhalten. Das Thema wurde ausgewählt, weil teilweise evidenzbasiert argumentiert wird, Interesse werde überschätzt. Folglich solle eine evidenzbasiert arbeitende Lehrkraft keine Zeit in der Unterrichtsvorbereitung darauf verwenden, interessante Inhalte auszuwählen, sondern z. B. lieber mehr Zeit auf Diagnose und Förderung. Das erkenntnisleitende Interesse dieser kleinen Untersuchung ergibt sich daraus, dass dies aller Erfahrung praktizierender Lehrkräfte widerspricht und deswegen ist zu klären, ob die empirische Evidenz stark genug ist, um das Bauchgefühl zu widerlegen. Vorab sei festgehalten, dass den Studien, die hier kritisch hinterfragt werden, keineswegs mangelnde wissenschaftliche Sorgfalt unterstellt werden soll, im Gegenteil: Die Studien haben ein methodisch hohes, einige sogar exzellentes Niveau. Das Argument lautet vielmehr, dass trotz dieser wissenschaftlichen Qualität das Evidenzniveau – im Sinne von Belastbarkeit, Generalisierbarkeit und Relevanz für didaktisches Handeln – nicht überschätzt werden sollte.

Die konkreten Forschungsfragen lauten:

1. Könnten publizierte Ergebnisse zur Bedeutung des Interesses auch durch andere (simulierte) Daten zustande kommen, obwohl der Prozess der Datengenerierung den publizierten Schlussfolgerungen widerspricht?
2. Gibt es in den Rohdaten publizierter Studien Strukturen, die durch die Auswertungsmethode übersehen werden?
3. Kommen alternative Zusammenhangsmaße zu einheitlichen Ergebnissen?

Interesse

Die Rolle des Interesses für das Lernen ist ein klassisches Thema der pädagogischen Psychologie. Dennoch sind die Befunde erstaunlich vielfältig. Einige Beispiele: Ahmad et al. (2024) fanden eine Korrelation von 0.38 zwischen Leistung und Interesse an Mathematik, Salifu und Bakari (2022) fanden eine Korrelation von 0.51. Heinze et al. (2005) berichten, dass diese Korrelation von der Note abhängt und zwischen 0.42 und 0.30 variiert. Okafor (2020) fand sogar heraus, dass das Interesse allein 57 % der Varianz in der Physikleistung erklärt. Das mag zwar beeindruckend erscheinen, könnte aber daran liegen, dass das Vorwissen in diesen Studien nicht kontrolliert wurde. Tatsächlich ändern sich die Ergebnisse drastisch, wenn man das Vorwissen berücksichtigt. So fanden Köller et al. (2001) in einem linearen Pfadmodell ein standardisiertes Beta für den Pfad von der Leistung in der 7. Klasse zur Leistung in der 10. Klasse von 0.44, während der Pfad vom Interesse in der 7. Klasse zur Leistung in der 10. Klasse mit einer Gewichtung von nur 0.09 nicht signifikant war. Interessanterweise war in der gleichen Studie der Pfad vom Interesse in der 10. Klasse zum Erfolg in der 12. Klasse mit einem Beta-Gewicht von 0.30 signifikant. Jansen et al. (2016, Tab. 2) bestimmten bei Kontrolle kognitiver Fähigkeiten den Einfluss von Interesse an Mathematik zu $\beta = 0.14$ (und damit war der Einfluss bei Mathematik am stärksten, im Vergleich zu anderen Schulfächern). Rianto et al. (2022) fanden die umgekehrte Reihenfolge der Relevanz: In ihrer (eher kleinen) Studie mit Gymnasiastinnen und Gymnasiasten fanden sie ein Beta für das Vorwissen von 0.33

und für Interesse von 0.57. Kosiol et al. (2019) unterschieden zwischen verschiedenen Interessen- und Leistungsbereichen im ersten Semester der Universitätsmathematik und fanden heraus, dass das allgemeine Interesse an Mathematik ein standardisiertes Gewicht bei der Erklärung der Leistung zwischen 0.12 und 0.30 hatte, je nachdem, welche anderen Variablen kontrolliert wurden. In einem Modell, das zwischen dem Interesse an Schulmathematik und dem Interesse an Hochschulmathematik zusammen mit dem allgemeinen mathematischen Interesse unterschied, hatte letzteres ein Gewicht von 0.30, während die spezifischen Formen ein negatives Gewicht hatten, wobei das Gewicht für das Interesse an Schulmathematik ein Gewicht von -0.28 hatte (nahe an Signifikanz). Man kann spekulieren, dass sich Interesse, das nicht mit den Lehrinhalten übereinstimmt, negativ auswirken kann. Dagegen fanden Lazarides und Ittel (2013) ein Beta-Gewicht von Interesse auf Leistung von 0.21, was in ihrer Studie nicht signifikant war.

Unterschiedliche Daten, gleiche Ergebnisse

In diesem Abschnitt soll die erste Forschungsfrage beantwortet werden. Inwiefern können unterschiedliche Rohdaten zu den gleichen Schlussfolgerungen führen, obwohl die Struktur der Rohdaten diese Schlussfolgerung nicht legitimiert. Anders gesagt: Es wird geprüft, ob eine unpassende Auswertungsmethodik zu falschen Ergebnissen führen könnte. Dazu sollen exemplarisch die Ergebnisse von Köller et al. (2001) zum Zusammenhang zwischen den Leistungen in den Jahrgangsstufen 7 und 10 (A_7 , A_{10}) und dem Interesse in der 7. Klasse (I_7) analysiert werden. Die Autoren berichten über deskriptive Statistiken, die nach Geschlecht und Art der Lehrveranstaltung differenziert sind, die die Teilnehmenden besucht haben. Die Unterschiede zwischen diesen Bedingungen waren nicht sehr groß, sodass für die vorliegende Untersuchung über diese Gruppen gemittelt wurde. Die Ergebnisse sind in Tabelle 1 aufgeführt.

Der Vollständigkeit halber mag es auch interessant sein, dass das Interesse in Jahrgang 10 deutlich geringer ist als in Jahrgang 7: 2.59.

Dies könnte erklären, warum Interesse in dieser Studie nur ein moderater Prädiktor ist: Die Inhalte, die zwischen diesen beiden Messpunkten vermittelt werden, entsprechen möglicherweise nicht dem Interesse, und daher wirkt das Interesse geringer als potentiell möglich. Wir werden diesen Weg der Kritik jedoch nicht weiter verfolgen (ebensowenig wie den pädagogischen, nach dem für die unterrichtende Lehrkraft das Vorwissen eine Konstante ist, während das Interesse

zumindest eine gewisse Elastizität aufweist), sondern einen Blick auf das von Köller et al. (2001) entwickelte Pfadmodell werfen. Wie oben berichtet, fanden sie standardisierte Beta-Koeffizienten $A_7 \rightarrow A_{10}$ von 0.44 und von $I_7 \rightarrow A_{10}$ von 0.09 (was in ihrer Stichprobe nicht signifikant war). Köller et al. schlussfolgern für die Entwicklung von Jahrgangsstufe 7–10:

Interests are usually considered to be important antecedents of successful academic learning (e.g., Krapp, 1999; U. Schiefele et al., 1992). Our findings, which are based on longitudinal analyses, suggest that the empirical support for this assumption is weak, at least at the lower secondary school level. After controlling for prior knowledge, there was no significant effect of interest on achievement.

Dass sie eine höhere Bedeutung des Interesses in der gymnasialen Oberstufe festgestellt haben, erklären die Autoren mit den geänderten Lehrformen in dieser Altersstufe, in der die Lehre nicht mehr so stark strukturiert ist.

Die erste Einsicht, die hier diskutiert wird, ist trivial: Auswertungen mit (multipler) linearer Regression, Pfad- oder Strukturgleichungsmodellen analysieren letztlich nur die Kovarianz-Matrix der Daten, d. h. alle nichtlinearen Strukturen in den Daten werden wegprojiziert. Es ist deswegen problemlos möglich, in einer Simulation einen relevanten nichtlinearen Einfluss von I_7 auf A_{10} zu simulieren, ohne dass sich an den Auswertungen mit einem Pfadmodell etwas ändert. Konkret simuliert (Realisierung in Simulation1.R, wie auch die anderen Simulationen dieses Beitrags erhältlich unter myweb.rz.uni-augsburg.de/~oldenbre/EvidenzR.zip) man z. B. N Werte von A_7 und I_7 mit der Verteilung

$$(A_7, I_7) \sim \mathcal{N}\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right) \quad (1)$$

Sei dann $I_7^2 = (I_{7,1}^2, \dots, I_{7,N}^2)$, dann ist der zu I_7 orthogonale Anteil von I_7^2 :

$$Q = I_7^2 - I_7 \cdot \frac{I_7^T I_7^2}{I_7^T I_7} \quad (2)$$

Setzt man nun A_{10} fest als

$$A_{10} := 0.44A_7 + 0.09I_7 + b_3Q + \varepsilon \quad (3)$$

mit (z. B.) normalverteilterm Fehler ε , dann ergibt sich unabhängig von b_3 das gleiche Ergebnis wie bei Köller et al. (2001), d. h. man würde schließen, dass die Bedeutung des Interesses gering sei, obwohl man das Gewicht b_3 und damit die Bedeutung des Interesses beliebig groß wählen kann.

Selbstverständlich ist dies keine neue Einsicht: Wenn man Daten mit einem unpassenden Modell auswertet, kann man falsche Ergebnisse erhalten, das ist

Tabelle 1. Mittelwerte und Standardabweichungen der Stichprobe. Gepoolte deskriptive Statistik aus Köller et al. (2001).

Variable	Mittelwert	SD	A_7	I_7	A_{10}
Leistung 7. Klasse (A_7)	101.1	29.8	1	0.20*	0.48*
Interesse 7. Klasse (I_7)	2.73	0.67		1	0.20*
Leistung 10. Klasse (A_{10})	164.1	32.2			1

trivial. Heikel an dieser Erkenntnis ist aber, dass sich die Unangemessenheit eines linearen Modells bei multipler linearer Regression zwar mit dem Ramsey-Test effektiv nachweisen lässt (Christodoulou-Volos & Tserkezos, 2023), dass aber für Strukturgleichungsmodelle kein entsprechender Test existiert – und selbst dort, wo es möglich wäre, wird der Ramsey-Test längst nicht in allen Studien angewendet. Selbstverständlich kann man die Unangemessenheit des linearen Modells sehen, wenn man das richtige Modell vermutet und im Vergleich rechnet, aber dies ist in der Praxis selten der Fall.

Diese erste Überlegung hat gezeigt, dass es möglich ist, dass in den Daten nichtlineare Strukturen existieren, die durch eine lineare Auswertungsmethodik "übersehen" werden, sodass letztlich unangemessene Schlussfolgerungen gezogen werden könnten. Allerdings werden die Daten durch die Orthogonalisierung in hohem Maße künstlich erzeugt: Dies ist kein plausibler Mechanismus, wie die Daten in der zu simulierenden Realität entstanden sein könnten. Der folgende vorgeschlagene nichtlinearer Mechanismus hat allerdings eine gewisse psychologische Plausibilität.

Wäre ein lineares Modell wirklich eine treue Abbildung des datenerzeugenden Prozesses in der Realität, würde ein hohes Interesse selbst bei fehlendem Vorwissen einen guten Lernzuwachs prognostizieren lassen. Dies ist unplausibel, denn Vorwissen ist in fast allen Lernprozessen wichtig, wie z. B. Bransford et al. (2000, S. 97) feststellen. Viel naheliegender ist es, dass es einen Interaktionseffekt gibt, bei dem Vorwissen wirksamer ist, wenn es mit Interesse gepaart wird und umgekehrt. Ein symmetrischer Interaktionseffekt wie I_7A_7 kann allerdings noch keine Asymmetrie erklären zwischen der Bedeutung der beiden Variablen. Plausibel scheint allerdings, dass ein hohes Interesse bei mäßigem Vorwissen zu Frust führt, was zu besonders niedrigen Lernergebnissen führt, oder umgekehrt gesprochen, dass nur die Kombination aus hohem Interesse und sehr hohem Fachwissen sehr stark wirkt, und das kann durch $I_7^2A_7$ beschrieben werden.

Die Realisierung in Simulation2.R simuliert zunächst korrelierte Variablen I_7 und A_7 mit Werten im Einheitsintervall, die sich aus zwei uniform¹ verteilten Komponenten zusammensetzen, und berechnet dann mit den zentrierten Variablen $I_{70} := I_7 - \bar{I}_7$, $A_{70} := A_7 - \bar{A}_7$ die simulierte Leistung A_{10} als

$$A_{10} = b_1I_7 + b_2A_7 + b_3I_{70}^2A_{70} + \varepsilon \quad (4)$$

Durch Setzungen wie etwa $b_1 = 0.1$, $b_2 = 0.3$, $b_3 = 0.7$ wird dann simuliert, dass das Interesse von großer Bedeutung für A_{10} ist. Trotzdem liefert die lineare Auswertung mit einem Strukturgleichungsmodell die umgekehrte Aussage, insbesondere die Regression $A_{10} \sim 0.28 + 0.28 \cdot A_7 + 0.10I_7$. Alle üblichen Fit-Indizes sind dabei im unproblematischen Bereich, wie die Simulation2.R zeigt. Es ist leicht zu erklären, warum das lineare Modell den Koeffizienten von A_7 groß macht: Es addiert zu b_2 den mittleren Effekt aus dem Interaktionsterm. Dieser nichtlineare Effekt lässt sich in diesem Fall problemlos diagnostizieren, wenn man ihn vermutet. Werden die Konstrukte aber als latente Variablen in einem Strukturgleichungsmodell geschätzt, muss man sich schon viel Mühe geben, denn viele nichtlineare Strukturgleichungsprogramme erlauben nur quadratische Interaktionsterme, während in diesem Modell ein kubischer Term relevant ist. Für den Effekt ist es im Übrigen unerheblich, ob man mit Gleichverteilung oder Normalverteilung simuliert, wie Simulation2S.R zeigt.

Gleiche Daten, unterschiedliche Ergebnisse

Bisher wurde untersucht, ob die gleichen Ergebnisse statistischer Auswertungen durch unterschiedliche Daten erzeugt sein könnten. In diesem Abschnitt wird die Perspektive umgekehrt und damit die zweite Forschungsfrage adressiert: Ausgehend von realen Datensätzen wird aufgezeigt, dass unterschiedliche Auswertungsverfahren zu unterschiedlichen Antworten kommen können. Auch diese Erkenntnis ist nicht neu, die

¹ Die Wahl einer Uniformverteilung ist von geringer Bedeutung, motiviert ist sie, weil bei Erhebung mit Likertitems die alternative Normalverteilung in der Regel auch nicht sehr gut getroffen wird. Es gibt aber auch eine Variante der Simulation, die mit normalverteilten Zufallsvariablen arbeitet.

Konkretisierung an einem für die Didaktik relevanten Beispiel sollte dennoch von Interesse sein.

Der erste verwendete Datensatz stammt aus der High School Longitudinal Study HSLS:09, deren Daten über nces.ed.gov/surveys/hsls09/ frei verfügbar sind. Die Auswertungsdatei `analyse_HSLS.R` geht davon aus, dass die Datei `hsls_17_student_pets_sr_v1_0.rdata` von dort heruntergeladen wurde. Der Datensatz enthält u. a. die Variablen `X1TXMTSCOR`, `X2TXMTSCOR`, `X1MTHINT`, und `X2MTHINT_R`, die die Mathematikscores und das Interesse an Mathematik zu zwei Zeitpunkten angeben. Im Folgenden werden die zu Mittelwert 0 zentrierten Variablen als M_1, M_2, I_1, I_2 bezeichnet. Nach der Elimination von Fällen mit fehlenden Daten bleibt ein Datensatz mit 17427 Fällen. Zunächst werden lineare Regressionen berechnet, wobei als Prädiktoren aber auch Produkte der zentrierten Variablen verwendet werden: $I_1 M_1 := I_1 \cdot M_1, I_2 M_1 := I_2 \cdot M_1$. Die Ergebnisse für das lineare Modell (standardisierte Koeffizienten), wenn man I_1 resp. I_2 als Prädiktor für M_2 verwendet, lauten

$$M_2 \sim -0.46I_1 + 0.55M_1 \quad (5)$$

$$M_2 \sim 0.90I_2 + 0.12M_1 \quad (6)$$

und für das nichtlineare Modell

$$M_2 \sim 0.01I_1 + 0.74M_1 + 0.71I_1M_1 \quad (7)$$

$$M_2 \sim 0.90I_2 + 0.11M_1 + 0.03I_2M_1 \quad (8)$$

Diese Ergebnisse stehen im Einklang mit den Befunden von Köller et al. (2001): Bei Kontrolle des Vorwissens ist das Interesse zu Beginn kein Prädiktor für Lernerfolg. Hochgradig interessant ist aber, dass das Interesse am Ende des Lernprozesses von größter Bedeutung ist. Als Erklärung kann man vermuten, dass der Unterricht in der Zwischenzeit nicht das Interesse bedient, das am Anfang vorhanden war. Ein hohes Kompetenzniveau erreichen am Ende nur die Lernenden, die in der Lernzeit Interesse hatten und retrospektiv Interesse äußern. Dann erklärt dieses spätere Interesse so viel, dass auch der Interaktionsterm irrelevant wird. Wenn aber das Interesse zum ersten Messzeitpunkt herangezogen werden soll, dann zeigt sich der Interaktionsterm als ebenso wichtig wie das Vorwissen.

Erfreulicherweise wurden uns von Köller et al. (2001) auch die Rohdaten zur Verfügung gestellt. Damit werden jetzt die eben dargestellten Analysen wiederholt, wobei nur der Übergang von der siebten zur zehnten Jahrgangsstufe betrachtet wird und die entsprechenden Interesse- und Leistungsvariablen wie oben bezeichnet werden. Für die lineare Regression

ergeben sich dann folgende Ergebnisse:

$$M_2 \sim 0.10I_1 + 0.46M_1 \quad (9)$$

$$M_2 \sim 0.19I_2 + 0.43M_1 \quad (10)$$

Mit Interaktionstermen ergibt sich:

$$M_2 \sim 0.11I_1 + 0.46M_1 + 0.11I_1M_1 \quad (11)$$

$$M_2 \sim 0.19I_2 + 0.43M_1 + 0.07I_1M_1 \quad (12)$$

In diesem Datensatz ist also sowohl der nichtlineare Effekt geringer als bei HSLS als auch der Unterschied zwischen der Bedeutung des Interesses zu beiden Zeitpunkten. Die Ergebnisse der Studie werden damit im Wesentlichen bestätigt: Es gibt in diesem Datensatz als in dem der HSLS-Studie keine Hinweise auf eine substantielle Verzerrung durch nichtlineare Effekte.

Alternative Zusammenhangsmaße

Regressionsanalysen sind extrem weit verbreitet, aber es gibt Alternativen, und dies motiviert die dritte Forschungsfrage. Eine Implikationsanalyse in der entropischen Variante nach Gras et al. (2008) der HSLS-Daten zeigt folgende Implikationsgewichte:

$$I_1 \cdot M_1 \rightarrow M_2 : 0.61 \quad (13)$$

$$M_1 \rightarrow M_2 : 0.45 \quad (14)$$

$$I_2 \rightarrow M_2 : 0.42 \quad (15)$$

$$I_1 \rightarrow M_2 : 0.41 \quad (16)$$

Die Ergebnisse zeigen, dass der Interaktionsterm am besten hohe Leistung impliziert, die anderen drei Variablen im Wesentlichen vergleichbar sind. Ähnliche Ergebnisse ergeben sich auch, wenn man die Implikation mit dem Maß ι nach Noel (2021) bestimmt. Offensichtlich hängt die Frage, was man als guten Prädiktor betrachten will, von weitergehenden Überlegungen ab, und es ist keineswegs klar, dass lineare Regression die optimale Entscheidung darstellt.

Bei den Daten von Köller et al. (2001) zeigt sich folgendes Bild:

$$M_1 \rightarrow M_2 : 0.24 \quad (17)$$

$$I_2 M_1 \rightarrow M_2 : 0.23 \quad (18)$$

$$I_1 M_1 \rightarrow M_2 : 0.20 \quad (19)$$

$$I_1 \rightarrow M_2 : 0.16 \quad (20)$$

Die Implikationsgewichte liegen also relativ nahe beieinander und sind allesamt recht gering. Gemessen mit ι zeigt $I_2 M_1$ das höchste Implikationsgewicht (0.32 gegenüber 0.24 für M_1).

Implikationsanalysen geben also keine Stütze ab für das Argument, Interesse sei verhältnismäßig unwichtig.

Schlussfolgerung

Die obigen Überlegungen und Untersuchungen zeigen, dass bei Auswertung von Studien mit üblichen linearen Modellen substantielle, nichtlineare Effekte unentdeckt bleiben können (HSLs-Studie), dass aber in tendenziell ähnlich angelegten Studien nicht notwendig ein solcher nichtlinearer Effekt in stärkerem Maße auftritt. Ob man aus den statistischen Ergebnissen schlussfolgern kann, dass Interesse gegenüber Vorwissen von sehr geringer Bedeutung ist, erscheint aber fraglich. Insbesondere zeigen Implikationsanalysen teilweise ein anderes Bild als Regressionsanalysen.

Zusammengenommen ist die Arbeit ein Plädoyer dafür, auch nichtlineare Modelle und alternative Zusammenhangsmaße stärker in Betracht zu ziehen, um zu prüfen, ob relevante, nicht lineare Effekte vorhanden sein könnten, die vor allem bei Strukturgleichungsmodellierung leicht übersehen werden können. Insbesondere ist bei der Interpretation von Ergebnissen, die nur aus linearen Analysen gewonnen wurden, Vorsicht geboten.

Es sollte deutlich geworden sein, dass der aktuelle Stand der empirischen Forschung zwar wichtige Beiträge liefern kann, aber noch nicht in jedem Fall ein Evidenzniveau garantieren kann, das ausreicht, um wichtige Entscheidungen darauf zu basieren. Es bedarf also noch viel mehr methodischer und empirischer Forschung, bevor die berechnete Forderung nach evidenzbasierter Didaktik auch tatsächlich umgesetzt werden kann. Aktuell bleibt ein evidenter Mangel an Evidenz.

Literatur

- Ahmad, M. U., Isa, A. D., & Zakariyya, A. A. (2024). Interest as predictor of mathematics achievement among senior secondary school students in Bida educational zone, Niger State, Nigeria. *Journal of Instructional Mathematics*, 5(1), 13–23.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academy Press.
- Christodoulou-Volos, C., & Tserkezos, D. (2023). Sensitivity of the Ramsey's regression specification error term test on the degree of nonlinearity of the functional form. *Journal of Applied Economic Sciences*, 18(1[79]), 5–10. DOI: 10.57017/jaes.v18.1(79).01.
- Gras, R., Suzuki, E., Guillet, F., & Spagnolo, F. (Hrsg.). (2008). *Statistical implicative analysis: Theory and applications*. Springer.
- Heinze, A., Reiss, K., & Rudolph, F. (2005). Mathematics achievement and interest in mathematics from a differential perspective. *ZDM – The International Journal on Mathematics Education*, 37(3), 212–220. DOI: 10.1007/s11858-005-0011-7.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, 2(8), e124. DOI: 10.1371/journal.pmed.0020124.
- Jansen, M., Lüdtke, O., & Schroeders, U. (2016). Evidence for a positive relation between interest and achievement: Examining between-person and within-person variation in five domains. *Contemporary Educational Psychology*, 46, 116–127. DOI: 10.1016/j.cedpsych.2016.05.004.
- Köller, O., Baumert, J., & Schnabel, K. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education*, 32(5), 448–470. DOI: 10.2307/749801.
- Kosiol, T., Rach, S., & Ufer, S. (2019). (Which) mathematics interest is important for a successful transition to a university study program? *International Journal of Science and Mathematics Education*, 17, 1359–1380. DOI: 10.1007/s10763-018-9925-8.
- Lazarides, R., & Ittel, A. (2013). Mathematics interest and achievement: What role do perceived parent and teacher support play? A longitudinal analysis. *International Journal of Gender, Science and Technology*, 5(3), 207–231.
- Noel, Y. (2021). Statistical implication analysis and the Rasch model. *Proceedings of the 86th International Meeting of the Psychometric Society (IMPS 2021)*. <http://yvonnick.noel.free.fr/papiers/implication-rasch-imps2021-ynoel-2x2.pdf>
- Okafor, O. S. (2020). Interest as predictor of academic achievement of secondary school students in physics. *British Journal of Education, Learning and Development Psychology*, 3(3), 1–9.
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528–530. DOI: 10.1177/1745691612465253.
- Rianto, B., Lutfi, A., Muhamad Naser, A. D., & Dasari, D. (2022). The effect of prior knowledge and learning motivation on mathematics learning outcomes of junior high school students. *Jurnal Pendidikan MIPA*, 23(4), 1876–1886. DOI: 10.23960/jpmipa/v23i4.pp1876-1886.
- Salifu, A. S., & Bakari, A. (2022). Exploring the relationship between students' perception, interest and mathematics achievement. *Mediterranean Journal of Social & Behavioral Research*, 6(1), 13–20. DOI: 10.30935/mjosbr/11491.

Reinhard Oldenburg, Universität Augsburg
reinhard.oldenburg@math.uni-augsburg.de

Timo von Oertzen, Thomas Bayes Institut
timo.vonoertzen@thomasbayes.de