Contextuality of Application Tasks in Large-Scale Summative Assessments at Lower Competence Levels for Lower Secondary Education

Sven Basendowski[1] and Gilbert Greefrath[2]

[1]University of Rostock, Germany; sven.basendowski@uni-rostock.de

[2]University of Münster, Germany; greefrath@uni-muenster.de

This article investigates the item pool of measurements of mathematical competencies in students at or below the lowest proficiency level through a qualitative analysis of Germany's 2018 IQB-Trends in Students Achievement. Contrary to conventional views, this article emphasizes a nuanced understanding of the difficulty of application tasks in large-scale assessments, with a focus on contextualization. By introducing a novel task pool designed for students with special educational needs, the study addresses previous limitations in accurately assessing their performance. Qualitative analysis reveals improved authenticity and relevance in the new tasks, particularly in private contexts. The findings highlight the importance of refining authenticity and relevance criteria for application tasks at lower proficiency levels, and provide valuable insights for inclusive education contexts.

Keywords: *Applications, test items, authenticity, relevance, special education.*

Introduction

The results of large-scale summative assessments (LSAs) in mathematics are crucial for the further development of the education system. In recent years, there has been research-based debate about whether and how LSAs can accurately measure the mathematical competencies of students at or below the lowest competency level, as found in assessments such as the Organization for Economic Cooperation and Development’s (OECD) Programme for International Student Assessment (PISA). This article presents the results of a qualitative analysis on the contextualization of application tasks in the lower competence range from a German national LSA, the 2018 Trends in Student Achievement, conducted by the Institute for Quality Development in Education (IQB). Contrary to the often-discussed general assessment in the literature (Knoche & Lind, 2004), which suggests that application tasks are inherently more difficult than the same mathematical tasks without an application reference, the analysis by Mahler et al. (2020) concludes that the contextualization of application tasks must be considered in a more differentiated way. This finding is crucial for the further development of application tasks in LSAs for an inclusive education system aligning with the principles of the United Nations Convention on the Rights of Persons with Disabilities (UN CRPD). Application-related tasks in LSAs serve an educational goal in an inclusive education system. For this purpose, we analysed characteristic contextual features within the current 2018 IQB-Trends’ task pool in Germany.

Relevance of the application reference in LSA

Application reference is of particular importance in many LSAs. International LSAs focus on real-world references when assessing mathematical competencies in adolescence and adulthood. In the PISA study, mathematical competencies are assessed "in authentic application situations" (Baumert et al., 2001, p. 19). This applies equally to the PISA framework for 2022, which aims to emphasize the relevance of mathematics for students and continues to set tasks in authentic contexts. The application reference is a fundamental feature of tasks in PISA studies and other internationally known LSAs, representing a central educational goal for all students. However, there are indications that application-related tasks are empirically more difficult than comparable inner-mathematical tasks Knoche & Lind, 2004). Contrary to the relevance of application tasks as an educational goal for all students, there is a tendency in Germany that the lower the level of completion of an educational program at lower secondary level, the less often modelling tasks are set (Neubrand, 2007).

Operationalization of the application reference in LSAs

The extent to which the applied tasks vary for different target groups in LSAs has not been analyzed to date. This applies, for example, to PISA tasks, which can be divided into contexts relevant to young people according to different areas of life (OECD, 2023). A distinction is made between contexts that address a personal, professional, social and scientific domain (Reinhold et al., 2019). In addition to the range of relevant contexts to be considered, the LSA frameworks emphasize the authenticity of the test items as a requirement (OECD, 2001). Authenticity refers to an extra-mathematical context that needs to be addressed in the situation using mathematical means. The extra-mathematical context should be authentic and not just constructed for the particular mathematical task. Thus, the use of mathematics in this situation should not be limited to mathematics lessons. Authenticity of tasks is included in various classification schemes and descriptions of modelling tasks (e.g. Maaß, 2010).

Authenticity can be operationalized in terms of different dimensions such as the situation, the question or the information and tools provided (Palm, 2007). When multiple dimensions of a task are authentic, students are more likely to make the necessary real-world considerations for the solution (Palm, 2008). Palm (2008) therefore considers a situation to be authentic if it represents a real task situation and if important aspects of that situation are simulated to an appropriate degree. An authentic task is therefore credible to the learner and realistic in terms of the environment (Palm, 2007). A focus on the authenticity of key dimensions of tasks, such as situation, question and methods, is useful (Turner et al., 2022) and will be used in this article.

Tasks in LSAs at the lower competence levels of the 2018 IQB-Trends

In Germany, there is an institutionalized category of special educational needs for students with learning difficulties and disadvantages (SEN ldd students) as distinct from SEN for students with disabilities (OECD, 2007). The special educational needs of students are "considered to arise primarily from problems in the interaction between the student and the educational context" (OECD, 2007, p. 20) and are manifested by a general and persistent failure to achieve school standards, such as in mathematics. As a result, students with learning difficulties perform in the lower proficiency levels of PISA: 61.5% below proficiency level I, 27.9% at proficiency level I and 10.6% above proficiency level I (Müller et al., 2017). Due to this institutionalization in the German education system, it is possible to specifically examine a group of students that perform within the lower competence range among application tasks.

As part of the 2018 IQB-Trends, a target group-specific task pool for SEN ldd students was developed (Mahler et al., 2020). The IQB-Trends examines a representative sample of all students in grades 4 and 9 without SEN every six years. In 2018, the representative sample also included SEN ldd students and was intended to provide information about the extent to which learners in grade 9 have achieved the competence expectations formulated in the national educational standards (Stanat et al., 2019). Competence level Ib is associated with the minimum requirements to gain the lowest formal secondary school diploma (Kölm & Mahler, 2019).

The new target group-specific development in the 2018 IQB-Trends was deemed necessary because the task pools and competence structure models used nationally (e.g., Südkamp et al., 2015) and internationally (Müller et al., 2017) had not proven to be empirically suitable for adequately recording the performance of SEN ldd students. In the national studies cited, items from pools for the 4th or 6th grade level were used without taking contextual features into account. As a result, there was no task pool or competence structure model that could measure the performance of SEN ldd students in grade 9 with precision (Müller et al., 2017). The new development was able to empirically demonstrate a satisfactorily valid instrument for comparisons between all learners in secondary I programs except for the highest level (Mahler et al., 2020). In doing so, it focused conceptually on the lower competence areas. Further adaptations focused on processing times and the reduction of language and context barriers in the test material, i.e. by gaining a high degree of authenticity (Mahler et al., 2020). The results suggest that the general valuation that application tasks are empirically more difficult per se needs to be differentiated. The closeness to the real world and the authenticity of the context plays a decisive role here (Mahler et al., 2020). However, no analyses have been conducted to determine how these adaptations were reflected in the final task pool or how they influenced performance. Nevertheless, the results are relevant because modelling competencies are central to the subject of mathematics (KMK, 2004), and this area of competence in LSAs must be ensured for each student in an inclusive education system (UN CRPD).

Research question

The state of the art has highlighted the conceptual, didactic and participatory relevance of application-related items in LSAs in mathematics. The new development of an LSA in the educational trend 2018 provides important indications that, considering specific adaptations to authenticity and the relevance of the application reference, it is possible to develop a valid and reliable task pool that ensures the recording of the performance of young people from different educational pathways in the lower competence area. Given the documented test quality of the newly designed task pool for young people from almost all lower secondary programmes except the highest level, it can be assumed that the adaptations of the newly designed task pool can be applied to any student performing at the lower levels of competence in the IQB Trends studies. Similarly, there is no consensus in the research on whether, due to their complexity, application-related tasks in mathematical LSAs should be eliminated at the lower proficiency levels for students in lower secondary I.

The aim of this analysis is therefore to characterize the features of context embedding in a task pool that can demonstrably and validly measure performance in application tasks at the lowest competence levels. For this purpose, the items of the newly designed task pool from the 2018 IQB-Trends are compared with those of the previous task pool in terms of the authenticity and relevance of the application reference for each lower competence level. This is of interest to be able to provide indications as to which features of the context are suitable for the improvement of future task pools for LSAs. The research question arises:

To what extent do empirically suitable LSA items with an application reference at the lower competence levels differ from previously used LSA items in terms of authenticity and relevance of the application reference per competence level?

Study Design and Material

A structured qualitative content analysis is recommended for the intended qualitative analysis of the selected didactic criteria and characteristics of context embedded in application tasks (Gläser-Zikuda et al., 2020). For the present study, the standard procedure based on Mayring (2010) was adapted. The deductive category system was scrutinized and modified after the first run of analysis of the material. For this first run through the material, approximately half of the items were randomly selected and prepared for coding with the initial deductive category system in MAXQDA. Coding was carried out independently by two coders.

In this analysis of the 2018 IQB-Trends, the entire task pool, including information on item difficulty and assigned competence level, was provided by the IQB (Basendowski & Greefrath, 2024). However, this data can only be used on the condition that no items are published.

The subject of this structured qualitative content analysis is the set of test items used in the 2018 IQB-Trends (Stanat et al., 2019), which includes both the items and the coding guide. The items of the 2018 IQB-Trends were developed by teachers under the guidance of the IQB. Some items were reused from the 2012 test, while others were supplemented with newly designed items specifically for SEN LD students. The redesign was prompted by the insufficient measurement accuracy for SEN LD students identified in previous item pools, as explained above. Before being used in the 2018 IQB-Trends, the items were tested with several hundred SEN ldd students in both general and special schools and then selected. In the 2018 IQB-Trends itself, the 188 newly developed items were used as part of the task pool of a total of 521 items for all students, regardless of the level of the secondary I programme they attended (Mahler, Schipolowski & Weirich, 2019).

In accordance with the given research interest, only items with an application reference were selected. As the redesigned item pool is exclusively at levels Ia, Ib and II, only items at these levels were selected. The total of 128 items selected consists of 88 items from the redesigned pool and 40 items from the existing pool.

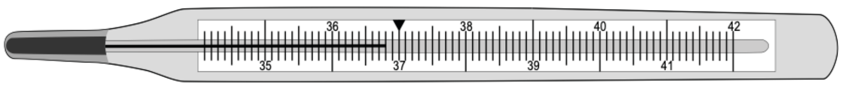
Revision of the category system and final material pass including intercoder reliability check

Gläser-Zikuda et al. (2020) identify testing through intercoder agreement as a common quality criterion for qualitative content analyses and specify a target value of Kappa = 0.70 as a sufficient indicator. The kappa values in the reported study for each supercategory in the 2nd run are between 0.69 and 0.91, determined by MAXQDA. As a result, after the 2nd run, there is an outcome that does not necessitate any further revisions of the deductive-inductive category system. All three authenticity categories, i.e. question, situation, and method (Turner et al., 2022), could be coded in "at least simulated authentic" (= credible and realistic), "simulated" (= untrustworthy) and "not assessable". The relevance categories (private, social, professional, scientific) were coded once per item. The final deductive-inductive category system can be found in Basendowski & Greefrath (2024). For an exemplary explanation of the authenticity and relevance criteria, see Table 1 and Figure 1: "Indicate how many degrees C the measured body temperature in the figure is."

Table 1: exemplary coding

Authenticity	question	at least simulated authentic	the problem of measuring a body temperature by reading a clinical thermometer is not implausible
	situation	simulated	reading of a thermometer is not credible without a person being present; in a problem situation, the measured body temperature may have very different consequences
	tools	simulated	the clinical thermometer is virtually available; in a problem situation

Figure 1: Example for previous concept pool items



Grafik: © IQB

Teilaufgabe 1

Gib an, wie viel °C die gemessene Körpertemperatur in der Abbildung beträgt.

____________ °C

(source: https://www.iqb.hu-berlin.de/vera/aufgaben/ma1/)

Results

A comparison between the previous concept and the new concept reveals the following striking differences: The authenticity of the tasks in both pools is generally low. However, there appears to be a trend that the newly designed tasks in the situation category are more consistently simulated authentic than the tasks used in the previous concept. No differences were found in the other authenticity categories. There are major differences in terms of the relevance of the tasks. In both groups, the relevance is essentially limited to scientific contexts, i.e. exemplary questions are described that are not relevant to the students' private, social and future professional lives. However, there are significantly more items among the newly developed tasks that can be classified as relevant in a private context (see Table 2).

Table 2: Number of items in the previous concept and the new concept

		Competence Level Ia		Competence Level Ib		Competence Level II	
		Previous Concept	New Concept SEN ldd	Previous Concept	New Concept SEN ldd	Previous Concept	New Concept SEN ldd

Authenticity: question / situation / tools	not assessable	11 / 0 / 13	36 / 0 / 44	0 / 0 / 0	33 / 0 / 31	25 / 0 / 31	15 / 0 / 14
	simulated	11 / 14 / 8	32 / 41 / 26	2 / 0 / 2	23 / 26 / 22	21 / 31 / 19	6 / 22 / 8
	at least simulated authentic	1 / 7 /1	8 / 35 / 4	0 / 2 / 0	15 / 35 / 18	9 / 22 / 6	6 / 5 / 5
relevance	private	0	18	0	21	0	1
	social	0	2	0	4	0	1
	professional	0	2	0	2	0	0
	scientific	22	54	2	47	50	25

Reading note for lines 2-4: The first number refers to the authenticity category of the question, the second to the authenticity category of the situation and the third to the authenticity category of the tool.

Discussion

The study of Mahler et al. (2020) has revealed that it is indeed possible to develop suitable and valid application tasks at the lowest competence levels. The study indicated that the newly designed task pool allows for a valid assessment of the performance of adolescents at almost all levels of secondary I programs, including the one for SEN ldd students.

It was found that there are clear differences between the concepts. This suggests that the new conceptualisation (Mahler et al., 2020) can provide new insights for the development of LSA items with practical application relevance at lower proficiency levels. In comparison, the new conceptualisation of items for SEN students shows that the novel characteristics 'relevance to everyday life' and 'authenticity of question, situation and tools' of application-related items at lower proficiency levels Ia and Ib were not as prominent in the previous conceptualization of IQB-Trends. It is generally important to enhance the credible problem character of the situation to make it more authentic – even though the new item test pool deviate from the degree of authenticity and relevance intended by the theoretical concept (Mahler et al., 2020).

The absence of authentic contexts on all levels (question, situation, and tools) needs to be discussed in relation to the PISA framework as well (OECD, 2023). There may be a need for improvement. If the demand of the tasks was to increase through authentic and relevant contexts, differences could potentially be identified by comparing the lower competence level with those above. However, in this investigation, no significant differences could be found. This might imply that authentic and relevant contexts can be constructed across all competence levels. Nevertheless, there are indications that authenticity is somewhat less frequent at the lowest competence level, Ia. Therefore, these concerns cannot be entirely dismissed. Other limitations concern items at higher levels of proficiency and more detailed categories.

References

Basendowski, S., & Greefrath, G. (2024). Anwendungsbezug in Mathematik-Large-Scale-Assessments im Bildungsmonitoring für den Sekundarstufe I – Bildungsgang des sonderpädagogischen Schwerpunkts Lernen. Journal für Mathematik-Didaktik. https://doi.org/10.1007/s13138-024-00230-y

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (Eds.). (2001). PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Leske + Budrich. https://doi.org/10.1007/978-3-322-83412-6

Gläser-Zikuda, M., Hagenauer, G., & Stephan, M. (2020). The Potential of Qualitative Content Analysis for Empirical Educational Research. Forum: Qualitative Social Research, 21(1), 17. https://doi.org/10.17169/fqs-21.1.3443

KMK (Ed.). (2004). Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss. Beschluss vom 4.12.2003. Luchterhand.

Knoche, N., & Lind, D. (2004). Bedingungsanalysen mathematischer Leistung: Leistungen in den anderen Domänen, Interesse, Selbstkonzept und Computernutzung. In M. Neubrand (Ed.), Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland: Vertiefende Analysen im Rahmen von PISA 2000 (pp. 205–226). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-80661-1_11

Kölm, J., & Mahler, N. (2019). Kompetenzstufenbesetzungen im Ländervergleich. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich, & S. Henschel (Eds.), IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich (pp. 157–168). Waxmann.

Lehmann, R. H., & Hoffmann, E. (Eds.). (2009). BELLA: Berliner Erhebung arbeitsrelevanter Basiskompetenzen von Schülerinnen und Schülern mit Förderbedarf 'Lernen'. Waxmann.

Maaß, K. (2010). Classification Scheme for Modelling Tasks. Journal Für Mathematik-Didaktik, 31(2), 285–311. https://doi.org/10.1007/s13138-010-0010-2

Mahler, N., Kölm, J., & Werner, B. (2020). Entwicklung von Mathematiktestaufgaben für Schüler*innen mit einem sonderpädagogischen Förderbedarf im Lernen – Konzeption und empirische Ergebnisse. In C. Gresch, P. Kuhl, M. Grosche, C. Sälzer, & P. Stanat (Eds.), Schüler*innen mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen (pp. 109–146). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-27608-9_5

Mahler, N., Schipolowski, S., & Weirich, S. (2019). Anlage, Durchführung und Auswertung des IQB-Bildungstrends 2018. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich, & S. Henschel (Eds.), IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich (pp. 99–124). Waxmann.

Mahler, N., Weirich, S., & Becker, B. (2019). Auswertung, Trendschätzung und Ergebnisdarstellung. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich, & S. Henschel (Eds.), IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich (pp. 125–130). Waxmann.

Mayring, P. (2010). Qualitative Inhaltsanalyse. In G. Mey & K. Mruck (Eds.), Handbuch Qualitative Forschung in der Psychologie (pp. 601–613). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92052-8_42

Müller, K., Prenzel, M., Sälzer, C., Mang, J., Heine, J.-H., & Gebhardt, M. (2017). Wie schneiden Schülerinnen und Schüler an Sonder- und Förderschulen bei PISA ab? Analysen aus der PISA 2012-Zusatzerhebung zu Jugendlichen mit sonderpädagogischem Förderbedarf. Unterrichtswissenschaft, 2, 194–211. https://doi.org/10.3262/UW1702175

Neubrand, M. (2007). Professionelles Wissen von Mathematik-Lehrerinnen und Lehrern: Konzepte und Ergebnisse aus der PISA- und der COACTIV-Studie und Konsequenzen für die Lehrerausbildung. In: F. Kostrzewa (Ed.), Lehrerbildung im Diskurs (pp. 53–72). gata-Verlag.

OECD (Ed.). (2001). Knowledge and Skills for Life. First Results from the OECD Programme for international Student Assessment. OECD.

OECD. (2007). Students with Disabilities, Learning Difficulties and Disadvantages: Policies, Statistics and Indicators. OECD. https://doi.org/10.1787/9789264027619-en

OECD (2023). PISA 2022 Assessment and Analytical Framework. OECD. https://doi.org/10.1787/dfe0bf9c-en

Palm, T. (2007). Features and Impact of the Authenticity of Applied Mathematical School Tasks. In W. Blum, P. L. Galbraith, H.-W. Henn, & M. Niss (Eds.), Modelling and Applications in Mathematics Education. The 14th ICMI Study (Vol. 10, pp. 201–208). Springer US. https://doi.org/10.1007/978-0-387-29822-1_20

Palm, T. (2008). Impact of authenticity on sense making in word problem solving. Educational Studies in Mathematics, 67(1), 37–58. https://doi.org/10.1007/s10649-007-9083-3

Reinhold, F., Reiss, K., Diedrich, J., Hofer, S. I., & Heinze, A. (2019). Mathematische Kompetenz in PISA 2018 – aktueller Stand und Entwicklung. In K. Reiss, M. Weis, E. Klieme, & O. Köller (Eds.), PISA 2018: Grundbildung im internationalen Vergleich (pp. 187–209). Waxmann. https://doi.org/10.31244/9783830991007

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S. (Eds.). (2019). IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich. Waxmann.

Südkamp, A., Pohl, S., Hardt, K., Jordan, A.-K., & Duchhardt, C. (2015). Kompetenzmessung in den Bereichen Lesen und Mathematik bei Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf. In P. Kuhl, P. Stanat, B. Lütje-Klose, C. Gresch, H. A. Pant, & M. Prenzel (Eds.), Inklusion von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen (pp. 243–272). Springer.

Turner, E. E., Bennett, A. B., Granillo, M., Ponnuru, N., Roth Mcduffie, A., Foote, M. Q., Aguirre, J. M., & McVicar, E. (2022). Authenticity of elementary teacher designed and implemented mathematical modeling tasks. Mathematical Thinking and Learning, 26(1), 1–24. https://doi.org/10.1080/10986065.2022.2028225