TEACHERS’ JUDGEMENT ACCURACY OF WORD PROBLEMS AND INFLUENCING TASK FEATURES

Sara Becker[1], Tobias Dörfler[2]
[1]Freiburg University of Education, [2]Heidelberg University of Education

The ability to judge accurately the difficulty of mathematical tasks is considered as a central facet of the diagnostic competence of mathematics teachers. An underlying reason is that the accurate judgement of task difficulty is the basis for achieving an optimal level of instruction for the learning group. Although a lot of studies have already investigated the judgement accuracy and the influence of additional factors, like teacher knowledge, there is a lack of a detailed look at the task features as possible influencing factors. Therefore, in the present study, we first investigated the judgement accuracy of word problems with fractions. Afterwards, by means of theoretical varied task features and an empirical study with 153 6th graders as well as 64 prospective teachers, we explored differences in the tasks regarding the judgement accuracy.

THEORETICAL BACKGROUND

Accurate diagnostic judgments are considered to be crucial for adaptive teaching (e.g., Hardy, Decristan, & Klieme, 2019). In mathematics teaching, the task diagnoses play a key role for shaping teaching and influencing learning processes (e.g., Sullivan, Clarke, & Clarke, 2013). In this context, to estimate the task difficulty is one method of adapting the teaching to a learning group (e.g., Hammer, 2016) and for achieving an optimal level of challenge for the learning group (e.g., Urhane, & Wijnia, 2020). Anders et al. (2010) were able to show in a study that students are more cognitively stimulated during instructional activities when a teacher can make an adequate judgement of task difficulty. Thus, accurate judgements of task difficulty have also been identified as one of the core tasks of mathematics teachers.

In a large part of the empirical studies on teachers' diagnostic competence conducted so far, the quality of diagnostic judgments is regarded as judgment accuracy. There is knowledge from over 40 years of research on the accuracy of teacher judgement (e.g., Urhane, & Wijnia, 2020). Judgment accuracy describes the degree to which teachers' judgments on task solution agree with empirically collected solution rates (Hoge, & Coladarci, 2016; Südkamp, Kaiser, & Möller, 2012). It is important to distinguish more precisely between person-related, task-related and person-specific teacher judgement accuracy (e.g., McElvany et al., 2009). In this contribution, we focus on the task-related judgement accuracy. Empirical studies on the judgement of task difficulty could show that the task difficulties are mostly underestimated and are judged with a too low of variance (e.g., Urhane, & Wijnia, 2020). Although the rank component is judged well on average, it shows considerable variance. Furthermore, teacher overestimate the level of their classes up to 45.5 % (e.g., Hosenfeld, Helmke, &

2022. In C. Fernández, S. Llinares, A. Gutiérrez, & N. Planas (Eds.), *Proceedings of the 45th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 67-74). PME.

Schrader, 2002). Only a few studies considered the three components of judgment accuracy in their analysis of teachers' diagnostic competence. However, these few studies reported low positive associations between the three components of accuracy (e.g., Schrader & Praetorius, 2018). Teacher judgment accuracy in a given content area differs across and within studies and show high inter-individual variances in the teacher judgment accuracy (e.g., Südkamp, Kaiser, & Möller, 2012; Urhane, & Wijnia, 2020). Research has identified several moderators that can determine the degree of judgment accuracy (e.g., Südkamp, Kaiser, & Möller, 2012). The accuracy can be influenced by teacher characteristics, judgment characteristics, student characteristics, class-level characteristics as well as test and task characteristics. Test and task characteristics refer to features of the tests or the tasks that have been used to measure student achievement. For example, Südkamp et al. (2012) examined the role of subject matter and the domain specificity of the achievement test as moderators of the judgement accuracy. But, none of these moderators affected teacher accuracy of judging and similary, Machts et al. (2016) found no evidence that test standardization moderated the judgement accuracy.

Despite the long period of research, we can look back on, and despite the numerous factors that have been investigated as influencing factors on the judgment accuracy, there has been little research on the association between teachers' judgment accuracy, the empirical solution frequency and the features of a task.

OBJECTIVE

According to the need for research pointed out in the previous section, we investigate first, whether the judgement accuracy and the interindividual variances regarding the judgment accuracy of student solutions reported in meta-analyses and task difficulty could also be shown in the judgement accuracy of mathematical word problems with fractions. Afterwards, as the focus of the study, we explicitly look more closely at the difficulty-generating task features with regard to the judgement accuracy and to the empirical solution frequency with the aim to analyze task features as influencing factors.

SAMPLE AND METHODS

The tasks that we focus on in this contribution is part of a larger study in which the influence of stress on the cognitive processes underlying diagnostic judgements on tasks (Becker et al., 2020) and the resulting judgement accuracy was examined. The difficulty of the tasks, eight mathematical word problems with fractions, were theoretically determined and empirically verified. For this reason, we designed fraction word problems with varied task features based on tasks frequently found in mathematics textbooks. The difficulty of word problems and fraction tasks has already been investigated in a number of studies. The task features chosen in the previous study were deduced from a review of those studies. In the present study, we considered two mathematical as well as two linguistic task features. First, we differed the relationship between the denominators (like or unlike fractions) (Padberg, & Wartha, 2017). It has been shown that like fractions have a lower requirement for the solution of a task in comparison to arithmetic tasks with unlike fractions and that tasks containing like fractions are easier to solve by students because of the analogy to the familiar natural numbers (e.g., Padberg, & Warta, 2017). Second, we distinguished between the number of calculation steps that have to be executed until the task is solved (one or two steps) (e.g., Jordan et al., 2006). It has been shown that tasks including one calculation step based on one mental model of operation and are therefore easier to solve by students than tasks that require two steps, because they include a further mental model of operation (Jordan et al., 2006). Furthermore, in word problems, the mathematical operation is part of the semantic structure of the text, which can also influence the difficulty of tasks (e.g., Verschaffel et al., 2020). For example, passive constructions can cause a change of the subject and the object of a sentence and can therefore be a further difficulty for students (e.g., Wessel, Büchter, & Prediger, 2018). Therefore, we varied as the third difficultly, the sentence structure of the tasks by using a passive construction in the task or not. Fourth, we distinguish between the use of words that can be considered as unfamiliar to 6th graders and the abandonment of those words, because it has been repeatedly shown that the use of those words can influence the solution of tasks and therefore the difficulty (e.g., Gürsoy et al., 2013). The number of the four difficulty-generating task features determines the theoretical difficulty in the present study.

The theoretically defined difficulty of the tasks has been proven in an empirical study with $N = 153$ 6th graders at various secondary schools in Germany. For this purpose, the students edited the word problems during their lessons in a randomized order to prevent sequence effects. Correctly solved tasks were subsequently coded with 1, incorrectly solved or unsolved tasks with 0. The students had sufficient time to solve the tasks.

Based on the solution frequency of each task, an empirical difficulty was determined by assigning a corresponding difficulty to the task on a ten-point scale (e.g.: 100 % - 90.1 % solution frequency corresponds to difficulty level 1, 90 - 80.1 % solution frequency corresponds to difficulty level 2, etc.; see task difficulty – students in table 1, 2 and 3). Furthermore, based on the empirical solution frequency, a ranking of the tasks was created.

Afterwards, in the main study, $N = 64$ prospective teachers of the educational university of Heidelberg judged the difficulty of the mathematical word problems in fractions for 6th graders on a ten-point scale. In a previous questionnaire, the semesters of the participants and whether any courses regarding the difficulty of fraction tasks had already been attended, but could be excluded as influencing factors in subsequent calculations. The mean of the participants' judgements is referred to as task difficulty – prospective teacher in the tables below (see task difficulty – p. teachers; table 1, 2 and

3). In the divisions of the means, the values were rounded down when the non-whole number is less than .5, higher than that the values were rounded up.

RESULTS

In all three components of judgment accuracy, the teachers' judgements deviated from the empirically determined difficulties of the tasks. On average, the task difficulty and the variance of task difficulty was underestimated. The rank component showed low positive correlations between prospective teachers' judgments and the empirically solution frequency. Furthermore, the results indicated high inter-individual variances in the teachers' judgments. No correlations were found between the individual components of judgment accuracy (between -0.001 and -0.180; averaged correlation is 0.000).

In view of the aim to identify task features that could influence the judgement accuracy of tasks, in the following, the varied task features of the eight word problems, the judgements of the prospective teachers and the empirical solution frequency of the 6th graders are examined in more detail with regard to each word problem.

The prospective teachers estimated the tasks mostly accurately, that are solved correctly by the students to a large extent (close to 50 % or more) (see table 1). This includes task 1, that is correctly solved by 89 % of the 6th graders, task 5, that is correctly solved by 42 % of the 6th graders, and task 7, that is correctly solved by 48 % of the 6th graders. Taking a closer look at the tasks, that are mostly judged accurately by the prospective teachers and are correctly solved by the 6th graders at a rate of almost 50 %, it is noticeable that task 1, 5 and 7 include only mathematical difficulty-generating task features. The lower the theoretical difficulty of the task, the more often the task is solved correctly. The theoretical task difficulty of task 1 is two, of task 5 it is five and of task 7 it is four.

	task difficulty		task features			
task	p. teachers	students	fraction	steps	lexicology	syntax
1	2	2 (89 %)	1	1	0	0
5	5	6 (42 %)	3	2	0	0
7	5	6 (48 %)	2	2	0	0

Table 1: Task difficulty of task 1, 5 and 7, judged by prospective teachers and derived from the solution frequency of the empirical survey, and task features

Task 2 and task 6 don't fit into the previously recognized structure, although both include exclusively mathematical difficulty-generating task features and would therefore have to be assigned to table 1. The theoretical task difficulty of task 2 is four and of task 6 it is also four.

	task difficulty		task features			
task	p. teachers	students	fraction	steps	lexicology	syntax
2	5	4 (25 %)	2	2	0	0
6	6	4 (35 %)	2	2	0	0

Table 2: Task difficulty of task 2 and 6, judged by prospective teachers and derived from the solution frequency of the empirical survey, and task features

But the teachers' judgements are not accurate and the empirical solution frequency is in the lower third. If we take a closer look at task 6, it is noticeable that this is not a classic fraction calculation task. The solution is already given in the word problem. The word problem was therefore less about mathematical calculation and more about understanding the text of the task. Analyzing the verbal protocols of the participants, it is noticeable that some participants noticed this and therefore classified it as easy and other participants classified it as difficult for 6th graders. Some participants did not recognize the given solution in the task and analyzed the mathematical calculation with regard to the difficulty for the 6th graders. No verbal protocols are available for the solutions of the 6th graders. But the solutions of the 6th graders showed that some pupils recognized and noted the solution in the text of the task, other pupils tried to solve the word problem by calculating.

Finally, we will take a closer look at those tasks that are mostly judged accurately by the prospective teachers with regard to the theoretical task difficulty, but that are not accompanied by the empirical solution frequency and thus the empirical difficulty of the tasks (see table 3). Task 3, 4 and 8 include mathematical difficulty-generating task features as well as semantic and linguistic difficulty-generating task-features. The theoretical task difficulty of task 3 is five, of task 4 it is also five and of task 8 it is six.

	task difficulty		task features			
task	p. teachers	students	fraction	steps	lexicology	syntax
3	5	8 (30 %)	2	1	1	1
4	5	9 (18 %)	2	2	1	0
8	6	9 (12 %)	2	2	1	1

Table 3: Task difficulty of task 3, 4 and 8, judged by prospective teachers and derived from the solution frequency of the empirical survey, and task features

DISCUSSION AND CONCLUSION

The present study investigated the judgement accuracy of task difficulty by prospective teachers. The results of the present study for the domain of word problems with fractions are consistent with the research findings reported in the literature presented.

Because these results were consistent with our assumptions, we investigated the task features with regard to the teachers' judgement accuracy and the empirical solution frequency. It is noticeable that in particular such tasks are accurately judged, that include only mathematical difficulty-generating task features and that are solved correctly by a large part of the students (task 1, 5 and 7). The theoretical difficulty, the teachers' judgement and the empirical solution frequency largely coincide for these three tasks. This is consistent with previous research showing that teachers can accurately judge those tasks in particular, that are easier to solve for students in particular (e.g., Urhane, & Wijnia, 2020). Tasks that contain linguistic difficulty-generating task features in addition, may be accurately judged by the prospective teachers with regard to the theoretical task difficulty (task 3, 4 and 8). However, the theoretical difficulty and the judgement do not concur with the empirical solution frequency. Students seem to find the linguistic difficulties more challenging than judged by the teachers. Two tasks were included in the test, where the empirical solution frequency largely correspond with the theoretical difficulty (task 2 and 6). However, it seems that it was difficult for teachers to judge these tasks accurately. The reason could be, for example in task 6, that the solution was already obtained and the task was, insofar as one recognized this as a student, very easy. This was sometimes not recognized by the teachers or was listed as a point of discussion.

Before discussing possible implication for international research on teachers' judgement accuracy, we would like to recall the limitations of this research, which suggest interpreting the evidence with care. First, it must be pointed out that prospective teacher may not yet be familiar with judging task difficulty for students. However, in order to exclude further influencing factors, such as experience, we first conducted the study with prospective teacher. An important further research approach would therefore be to replicate the results also through studies with in-service teachers. Furthermore, the results of this research report are based on only eight tasks, precisely word problems with fractions. It would be crucial to transfer the results to other content areas and task frameworks. Moreover, further research should complement these findings by means of different methodological approaches, especially quantitative data.

However, the findings of the present study provide a first, explorative insight into the influence of task features as influencing factors on teachers' judgment accuracy. Since the interindividual variances of teachers' judgments have still not been satisfactorily elucidated, despite over 40 years of research, the study offers a starting point for further investigation of the influence of task features on the teachers' judgement accuracy.

Acknowledgment

This research was funded by the Ministry of Science, Research and Arts of Baden-Wuerttemberg within the Research Training Group "Diagnostic Competences of Teachers".

References

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

Becker, S., Spinath, B., Ditzen, B., & Dörfler, T. (2020). Der Einfluss von Stress auf Prozesse beim diagnostischen Urteilen – eine Eye Tracking-Studie mit mathematischen Textaufgaben. *Unterrichtswissenschaft, 48(4)*, 531-550.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*, 177-182.

Gürsoy, E., Benholz, C., Renk, N., Prediger, S., & Büchter, A. (2013). Erlös = Erlösung? Sprachliche und konzeptuelle Hürden in Prüfungsaufgaben zur Mathematik. *Deutsch als Zweitsprache, 13*, 14-24.

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online, 11(2)*, 169-191.

Hammer, S. (2016). *Professionelle Kompetenz von Mathematiklehrkraften im Umgang mit Aufgaben in der Unterrichtsplanung: Theoretische Grundlegung und empirische Untersuchung*. München: Ludwig-Maximilians-Universitat.

Hoge, R. D., & Coladarci, T. (2016). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59(3)*, 297-313.

Hosenfeld, I., Helmke, A., & Schrader, F.-W. (2002). Diagnostische Kompetenz. Unterrichts- und lernrelevante Schülermerkmale und deren Einschätzung durch Lehrkräfte in der Unterrichtsstudie SALVE. *Zeitschrift für Pädagogik, Beiheft 45*, 65-82.

Jordan, A., Ross, N., Krauss, S., Baumert, J., Blum, W., Neubrand, M., Löwen, K., Brunner, M., & Kunter, M. (2006). *Klassifikationsschema für Mathematikaufgaben. Dokumentation der Aufgabenkategorisierung im COACTIV-Projekt*. Max-Planck-Inst. für Bildungsforschung.

Machts, N., Kaiser, J., Schmidt, F. T. C., & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review, 19*, 85-103.

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., & Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschatzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern. *Zeitschrift für Pädagogische Psychologie, 23(34)*, 223-235.

Padberg, F., & Wartha, S. (2017). *Didaktik der Bruchrechnung*. (5. Auflage). Mathematik Primarstufe und Sekundarstufe I & II. Berlin: Springer Spektrum.

Schrader, F.-W., & Praetorius, A.-K. (2018). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost, J. R. Sparfeldt, & S. Buch (Eds.), Beltz Psychologie 2018. *Handwörterbuch Pädadagogische Psychologie* (5th ed., S. 92-98). Weinheim, Basel: Beltz.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104(3),* 743-762.

Sullivan, P., Clarke, D., & Clarke, B. (2013). *Teaching with tasks for effective mathematics learning*. New York, NY: Springer.

Urhane, D., & Wijnia, L. (2020). A review on the accuracy of teacher judgments. *Educational Research Review, 32*, 100374.

Verschaffel, L., Schukajlow, S., Star, J., & van Dooren, W. (2020). Word problems in mathematics education: A survey. *ZDM, 52(1),* 1-16.

Wessel, L., Büchter, A., & Prediger, S. (2018). Weil Sprache zählt - sprachsensibel Mathematikunterricht planen, durchführen und auswerten. *Mathematik lehren*, *206*, 2-7.