Comparing global judgments and specific judgments of teachers about students' knowledge: Is the whole the sum of its parts?

Karina Karst, Stefanie Dotzel, Oliver Dickhäuser

University of Mannheim

Author Note:

Prof. Dr. Karina Karst, Assistant Professor of Quality of Instruction in Heterogeneous Contexts, Department of Psychology, University of Mannheim; Stefanie Dotzel, M.A., Department of Psychology, University of Mannheim; Prof. Dr. Oliver Dickhäuser, Full Professor of Educational Psychology, Department of Psychology, University of Mannheim

This research was supported in part by grants (PLI 3026A/ Kassel: Prof. Frank Lipowsky & PLI 3026B/ Bamberg: Prof. Gabriele Faust) from the German Federal Ministry of Education and Research (BMBF).

Correspondence concerning this article should be addressed to Prof. Dr. Karina Karst, Assistant Professor of Quality of Instruction in Heterogeneous Contexts, Department of Psychology, University of Mannheim, A5,6, 68131 Mannheim.

Email: karst@uni-mannheim.de

Abstract

Teachers' judgments about students' knowledge and skills can be global or specific depending on the diagnostic situation during teaching. We test the relationship between these judgments, their accuracy, and whether global judgment (GJ) accuracy can be measured by aggregating specific judgments (SJ). Judgments of 52 primary school teachers about their students' achievement in a standardized mathematics test were assessed. SJs and GJs correlated high. However, SJs were slightly more accurate than GJs. Additionally, teachers' GJ accuracy is not similar to the accuracy of aggregated SJs. We conclude that teachers use different judgment strategies for GJs and SJs.

Highlights

Teachers are quite consistent in their judgments about their students.

Judgment accuracy of teachers depends on the kind of judgment and on the used accuracy measure

Student-specific judgments (SJ) were slightly more accurate than student-global judgments (GJ).

The accuracy measures of aggregated SJs do not represent teachers' GJ accuracy.

The whole is not the sum of its parts

Introduction

Students' learning is more likely to be successful the more closely the learning process is related to their existing knowledge (Baumert, Lüdtke, Trautwein, & Brunner, 2009, Harwell et al., 2007; Weinert & Helmke, 1995). This is why it is helpful for teachers to have insight into students' prior knowledge in order to initiate adaptive learning opportunities for their students in the classroom (Brühwiler & Blatchford, 2011). Teachers receive this insight during various diagnostic situations in or outside of the classroom. Such situations require the teacher to make judgments about students' knowledge and skills. Thus, making diagnostic judgments about students seems to be a central and necessary teacher task (Burns, 1984; Herppich et al., in press).

In a recently suggested model of teachers' assessment competencies, Herppich et al. (in press) assumed that the kind of diagnostic judgments about the student *and* the accuracy of the judgments can differ with respect to the degree of specification of the judgment. This variation in specification arises from different aims of diagnostic situations during teaching that determine a particular perception of the teacher about the student. Some teaching situations may require a global judgment (e.g., when a teacher has to prepare test-tasks for an exam that are not too easy or difficult, and therefore reflects on the mean level of the students' competencies), whereas some situations may require a more specific judgment (e.g., when a teacher prepares specific individual support for a student).

The differentiation of either global or specific teacher judgments has only recently been made within the literature. A review of the research literature in 2012 suggested that a systematic analysis of similarities and differences between these two kinds of judgment is lacking (Südkamp, Kaiser, & Möller, 2012). Therefore, it seems relevant to investigate how student-specific judgments (SJs) correspond to student-global judgments (GJs), and whether SJ and GJ differ with regard to accuracy. Additionally, there is little educational research on processes underlying these kinds of judgment (Crisp, 2013; Herppich et al., in press). It is unclear whether teachers make GJ based on summing up many single SJs or not. If teachers in fact make GJ by aggregating many single SJs, the initial judgments and the judgment processes would be the same and we would neither assume any differences between the teacher's GJ and SJs, nor between the corresponding accuracy measures.

Thus, the whole (GJ) would be the sum of its parts (SJs). This would also reduce the time needed to investigate teachers' judgment accuracy. Assessing only one kind of teachers' judgment would suffice for gaining insight into overall teachers' judgment accuracy.

Diagnostic Situations During Teaching

Diagnostic situations emerge during and beyond teaching and involve judgments about individual students regarding characteristics that are relevant for learning- (e.g., prior knowledge, self-concept, and emotional stability; Herppich et al., in press). The aim of the teacher within these situations is to gain information for an upcoming pedagogical decision (Author, 2001).

One can imagine a huge variability between those situations. Author (2001) postulated four independent dimensions that describe the variability of these situations. These dimensions are: Purpose of the judgment (assessment for learning/formative assessment or assessment of learning/summative assessment), the option that the situation can be planned in advance, the binding character and consequence of the judgment for the student, and the degree of specificity of the teachers' perception of the student. We will focus on the last dimension "specificity of the teachers' perception of the student". Concerning this dimension, it is possible to differentiate between two endpoints so that the perception of the student could be more global or more specific.

The following example illustrates the importance of student-global perceptions of a teacher. The teacher plans a lesson with the aim to strengthen and consolidate the students' newly acquired knowledge. For this aim, he/she realizes internal differentiation by adopting methods supporting an internal differentiation resulting in groups of learners with different achievement levels. In order to manage this situation, the teacher should consider the students' different competence levels and group their students according to their overall competence levels in the learning unit (Shavelson & Stern, 1981). This is the step where GJ accuracy becomes relevant. The judgmental task for the teacher is to answer the question: "How good is my student at solving such exercises that I'd like to implement during collaborative learning?" This requires a rather global (comprehensive) judgment about a student (Author, 2012).

By contrast, SJs probably more strongly influence teachers' behavior in situations where individual support takes place (Helmke & Schrader, 1987). This assistance could be realized by, for

example, specific explanations or targeted use of teaching material or individual feedback. The underlying question the teacher has to answer is: "How good do I think my student would be at solving this specific exercise? Does he or she need a specific explanation?" Thus, the perception of the student and the judgmental task are more specific. It is a specific judgment of an individual student. The contrast to the student-global diagnostic situation is that the teacher focuses on an individual student's prerequisites of a single student that are needed to master certain exercises.

Because of these different situations, the perception of the teachers about their students varies, as do the kinds of judgments, which can change from more or less specific judgments to global judgments (Author, 2001). The question, therefore, is to identify whether there are differences in the judgments and in the accuracy of the judgments depending on the kinds of judgment.

Judgment Processes

It is not only important to consider the differences in the results of these kinds of judgment but one should also pay attention to the information processing that underlies the judgments. Theories of social judgments (Chaiken, 1980; Fiske & Neuberg, 1990) essentially state that people (e.g., teachers) can apply two different strategies to process information for judging other people (e.g., students). Individuals can use either a heuristic strategy or a controlled strategy. The heuristic or short-cut strategy is assumed to be relatively automatic and to require little cognitive effort. Using this strategy leads to a simplification of the information in order to handle the complexity of the judgment more easily. Teachers may use the strategy as a result of (adverse) external conditions during teaching such as less time or energy of the teacher for competing tasks. Using the controlled strategy, the individual collects more information and integrates it into a judgment.

Regarding the two kinds of judgments (GJ vs. SJ), it seems likely that for a GJ, the teacher uses a heuristic strategy, whereas a controlled strategy is more likely to be applied when making the SJ. A SJ about students requires and facilitates the consideration of individualizing information about the student and about characteristics of the exercise (task characteristics). This, in turn, possibly requires more elaborated cognitive processes than a GJ because the teacher has to compare the ability of the students with the concrete task characteristics. GJs, by contrast, are probably easier to make because the teacher can use heuristics as a base for judgments. The idea that GJs and SJs are

associated with different types of processing is based on the reasoning developed by Sherman, Beike and Ryalls (1999). The authors argue, that general judgments are more likely to be influenced by heuristic processing as heuristics are more likely to exist for general cases. Consistent with this idea it has been found that abstract entities are more likely to be judged based on stereotypes (Fiske & Neuberg, 1990).

However, an assumed lower difficulty in making GJs does not necessarily mean that these judgments are more accurate than SJs. As SJs more likely require a systematic and controlled approach, one may assume that the accuracy for SJ is higher than for GJ. However, there are also reasons to assume higher accuracy of GJ compared with SJ. Brighton and Gigerenzer (2012) could show that heuristics do not necessarily lead to lower judgment accuracy than more complex procedures of information integration. Thus, information resulting from heuristic strategies might be a better predictor (Brighton & Gingerenzer, 2015) and thus lead to higher judgment accuracy. With regard to teachers, this could mean that they only consider information that is actually relevant for a global judgment, which in turn leads to a higher level of judgment accuracy.

Thus, we cannot make a clear prediction regarding if GJs or SJs lead to higher judgment accuracy. Nevertheless, we assume that different judgment processes underlie two different kinds of judgments. If teachers are asked to judge a student more globally, they will not do this based on many specific judgments. Low or no correspondence between GJ and aggregated SJs and its analogous accuracy measures would indicate those different processes. Thus, the sum of many SJs would not result in a GJ.

Kinds of Judgment and Judgment Accuracy − State of Research

As mentioned before, the differentiation between GJ and SJ is relatively new. However, this does not mean that previous research did not analyze differences in judgment accuracy for different kinds of judgment. The following section outlines the state of research concerning the accuracy of different kinds of judgments.

The level-, differentiation-, and rank component (Schrader, 1989, Helmke & Schrader, 1987) are the most widely used accuracy measures of teachers' judgments about their students' skills and knowledge. The level component shows whether the teacher is a good judge of the students' mean

achievement level. It derives from the difference between the average teacher judgment and the average criterion value (e.g., average student achievement in a standardized test). Values above 0 show that the teacher overestimated the students' achievement, whereas levels below 0 indicate underestimation. The differentiation component is the quotient of the variance of teacher judgments and the variance of the student's test scores. Values above 1 indicate that the variance of the students' achievement was overestimated by the teacher, whereas values below 1 indicate underestimation. The rank component is the product-moment correlation between teacher judgments and students' test scores (criteria), and illustrates whether the teacher is aware of the rank order of his/her student concerning the judged characteristic/criterion.

Overall, it can be summarized that teachers' judgment accuracy is moderate. In a meta-analysis, Südkamp et al. (2012) reported a mean rank-component of .63 with a high variability (-0.03 up to 1.18; Fisher *z*-transformed correlations). The results are even more heterogeneous for the level and differentiation component. Whereas some studies reported an overestimation of students' achievement level (Bates & Nettelback, 2001; Südkamp, Möller, & Pohlmann, 2008), some studies reported an underestimation or over- and underestimation within one study for two measurement points (Feinberg & Shapiro, 2009; Stang & Urhahne, 2016). The same heterogeneity concerning the results emerges from the differentiation component. Furthermore, there were only weak correlations between the three different judgment accuracy measures (Lorenz & Artelt, 2009; Praetorius, Karst, Dickhäuser, & Lipowsky, 2011; Südkamp, Möller, & Pohlmann, 2008; Spinath, 2005; Schrader, 1989). These weak correlations indicate independency between the judgment accuracy measures and make a separate analysis of each accuracy measure necessary to gain a comprehensive picture about teacher's judgment accuracy.

Within their meta-analysis and by conceptualizing a heuristic model of judgment accuracy, Südkamp et al. (2012) attempted to find reasons for this heterogeneity of results. They postulated that different kinds of judgments would be associated with different levels of judgment accuracy. According to Südkamp et al. (2012) the following factors (a-d) characterize the kind of judgment: (a) informed versus uninformed; (b) congruence of domain specificity, (c) direct versus indirect, and (d) specificity of the judgment.

An (a) informed teacher judgment means that the teacher judges their students' test performance in an achievement test, which is also simultaneously administered to the students. In contrast to this, an uninformed judgment means that the teacher judges the competence of their students in any domain using a Likert-type rating scale. Here, teachers are uninformed about the criterion their judgment will be compared with (e.g., "How do you assess the language skills of student x?"). The latter kind of judgment is usually more comprehensive and global about the student. Südkamp et al. (2012) found that the judgment accuracy is significantly higher for informed judgments than for uninformed judgments ($\beta = .15$; $p = .045$).

The factor (b) congruence of domain specificity refers to the accordance of teacher judgments and the achievement test. If the teacher judgment is specific (judgment of an academic ability in one subject) *and* the achievement test is specific (covering a single subject), Südkamp et al. (2012) called it a high correspondence of domain specificity. In their meta-analysis, they found higher judgment accuracy for teachers within studies that measured judgment accuracy with a high congruence of domain specificity ($\beta = -.13$; $p = .009$).

A (c) direct judgment is an informed judgment of the teacher with more or less correspondence of domain specificity, whereas an indirect judgment is defined as an uninformed judgment with low correspondence. For this factor, Südkamp et al. (2012) found no significant differences for judgment accuracy. Additionally, Hoge and Colardarci (1989)—authors of an earlier meta-analysis—found slightly higher judgment accuracy (rank components) for direct judgments, but the differences were not significant neither. In addition, in a narrative review, Harlen (2005) reported a higher accuracy for direct judgments as compared to indirect judgments.

Lastly, teachers' judgments about students' knowledge and skills can be differentiated according to the factor (d) specificity. Based on suggestions by Hoge and Coladarci (1989), Südkamp et al. (2012) differentiated between five levels of specification. The most specific judgment is needed when a teacher is asked "Will your student x solve the exercise y?" (Level 5). Less specific and thus more global is the judgment when a teacher is asked "How many of these seven exercises will your student x solve?" (Level 4). The lowest degree of specificity (Level 1) is realized by ratings when the teacher rates their students' competence on a rating-scale (e.g., poor to excellent). Judgment accuracy

was significantly higher for judgments with a higher level of specificity in the Hoge and Coladarci meta-analysis (1989), but not in the meta-analysis of Südkamp et al. (2012).

Looking at recently published studies that had not been included in the reported meta-analyses, the findings are heterogeneous. Karing, Matthäi, and Artelt (2011) analyzed direct (specific) and indirect less specific judgments and compared the rank components. For 64 primary school teachers, they reported a significantly higher rank component for the indirect judgment than for the direct judgment ($t = 1.70$, $p < .05$, $d = 0.27$). They concluded that direct judgment accuracy was lower because direct judgments were more difficult for teachers to make due to their cognitive complexity.

In a study published by Zhu and Urhahne (2014) 16 primary school teachers were asked to make direct (multiple-item) judgments and indirect (single-item) judgments about their students' learning relevant motivation and emotions. The authors computed two different types of accuracy measures: the rank component (for the direct and indirect judgment) and percent agreement (only for the direct judgment). They reported no significant difference between the accuracy of indirect judgments and the accuracy of direct judgments. Thus, the authors regarded an indirect measure as a simple, efficient and valid method to evaluate teacher's judgment accuracy about students' motivation and emotions.

Considering the described state of the research, we conclude that the kinds of judgment teachers were asked to conduct in educational studies probably influences the judgment accuracy. However, it is not possible to decide whether teachers are better judges when they are prompted to make direct, or informed, or specific, or indirect, or uninformed, or less specific judgments.

The comparisons of direct and indirect judgments and their corresponding rank component (as one possible accuracy measure) dominated previous research. So far, no analyses exist that measure differences and relationships in judgment accuracy for the other two components, namely level and differentiation component. In most of the cases, such analyses were simply not possible because these components could not be calculated for single-item/indirect judgments. As these two components are also important elements of judgment accuracy, they should be integrated to investigate the question of whether judgment accuracy varies depending on the kind of judgment. However, a few studies exist, which compared various judgment accuracy measures (e.g., Zhu & Urhahne, 2014). Nevertheless,

those studies focused on the correlations between the accuracy measures and ignored to calculate the mean differences of the accuracy measures.

As seen from previous considerations, there is a clear overlap between the factors characterizing the kind of judgment and the present differentiation of GJ and SJ (Table 1).

Table 1

However, these factors characterizing the kinds of judgment are not entirely identical. An SJ is always a direct measure, but a GJ could be direct or indirect. In the manuscript, we analyze GJ on a direct level, which corresponds to the fourth category of specification according to Hoge and Coladarci (1989). Even though GJ and SJ are not identical to the factors of former research, they can, however, be integrated into the heuristic model of judgment accuracy (Südkamp et al., 2012). Both judgments (GJ and SJ) are informed judgments with a high congruence of domain specificity. In the present study, these two factors will be kept constant in order to avoid the confounding of more or less specific judgments with other judgment characteristics influencing judgment accuracy, such as direct or indirect measures. In almost all previous studies, the indirect measure was operationalized by a global judgment and the direct measure by a specific one. Thus, the effect of differences to the extent of judgment accuracy because of variations in the kind of judgment could not be ascribed to a single factor. Furthermore, the distinction between GJ and SJ takes the teaching activities and diagnostic tasks of teachers into account. Thus, the theoretical innovation and benefit of this distinction is the derived link to teaching situations. This theoretical derivation provides insight into when teachers need to make a GJ or SJ.

Research Questions

Based on this previously discussed research and the existing gaps, the present study intended to empirically investigate the following research questions:

(1) Does the accuracy of judgments differ with respect to the kind of judgments?

(2) Is it possible to express GJ-accuracy by aggregating single SJs of the teacher?

Method

Participants

To answer the above stated research questions empirically, we used the data set of a German primary school study. This study was conducted in three federal states of Germany: Saxony, Berlin, and Mecklenburg-Western Pomerania.

The teacher sample comprised 52 teachers working in 37 classes in 20 primary schools: the higher number of teachers compared to the number of classes is explained by a particular type of classroom organization in several schools, in which two teachers work in the same classroom with the same students for Mathematics and German (15 of the 37 classes in this case had two teachers). Of these teachers, 59% were in the age range 35 to 45, 98% (all but one) were female, and average time in teaching was 18.9 years (in general, on this basis, the teacher sample represented the primary teacher population in Germany (Tarelli, Lankes, Drossel, & Gegenfurtner 2012). The roughly gender-balanced student sample comprised 700 first graders (mean age 7 years, 4 months, 52% girls). The average class size included 18.9 students.

Materials and Procedure

Students completed a standardized math test at the end of grade 1. The *student's math test* was developed for this purpose and was based on a standardized achievement test (DEMAT 1 and 2; Krajewski, Küspert, & Schneider, 2002; Krajewski, Liehm, & Schneider, 2004). The test assessed students' arithmetic competence. The 45-minute test consisted of 49 test-tasks (items; example: Do the math of this addition! 7+21 = ___). All students received the same test (Author, 2012).

While the students filled out their test, the teacher judgments about a subset of the whole math test were collected via questionnaire. First, in order to assess teachers' *student-global judgment* (GJ), each teacher judged the number of each of their students' correct responses to items that represented two item-types (Example: "How many of these items will your student solve?"). The item-types were addition of two whole numbers (7 items) and reading numbers (8 items). For the *student-specific judgments (SJ)*, the teachers had to judge 12 students of their class that were randomly chosen. The basis was six arithmetic items concerning basic arithmetic operations (e.g. addition with unknowns in all positions, basic word problems, and double a number). For each of these six items, the teachers reported whether each of the twelve students would be able to solve it. In comparison to the GJ, this

judgment asked the teacher to decide, for each item separately, whether a specific student would be able to solve the item (Author, 2012).

To achieve an informed test situation, the teachers received all test-items of the students' math test and the instruction of the administered math test. We used the same scale for the teacher judgments and student achievement. This enabled the direct comparison of the student responses with the two kinds of judgments and resulted in a high congruence of domain specificity.

The analyses for the GJ were conducted with 49 teachers and 639 students because three teachers did not completely fill in the GJ. For the SJ, the data were complete for all teachers.

The eight teacher judgments (two GJs and six SJs) had a high reliability, Cronbach's $\alpha = .86$. In addition, the subset of the students' math-test items (21 items: 7 addition items, 8 number reading items and 6 basic arithmetic operation items), which were the basis for the teachers' judgments, had a high reliability, Cronbach's $\alpha = .88$.

Computing Accuracy Measures

Before computing the accuracy measures, we transformed the teacher judgments and the students' test scores from absolute to relative numbers. This was necessary since we intended to compare the different kinds of judgments with the student achievement test scores directly.

Following the approach of Schrader (1989), the accuracy components were calculated for each kind of judgment (student-global and student-specific) and for each teacher. The components are the level-, differentiation-, and ranking component. Additionally, we computed level- and differentiation errors for both kinds of judgments. The *level-component* derives from the difference between the average teacher judgment and the average students' criteria value. The *differentiation component* is defined by the quotient of the variance of the teacher judgment and the variance of the students' values. The optimum value of the level component is 0 and the optimum value of the differentiation component is 1. Deviations from those optimum values can be positive or negative. Values above the optimum value indicate an overestimation of the level or the variance of the students' knowledge and skills. Values beneath these optimums indicate an underestimation of the level or the variance. Both the level and the differentiation components lacked an absolute zero point. This makes it impossible to analyze directed relationships of the different kinds of judgments of each component. Therefore, we

also computed the corresponding errors of the level and the differentiation component. The level error is the absolute value of the difference between the average judgment and the average students' score. The differentiation error is the absolute value of the difference between the variance of the judgments and the variance of the students' scores. The higher these absolute values are, the higher is the corresponding error and the lower is the judgment accuracy of the teacher. The *rank component* is the the product-moment correlation between teacher judgments and students' test scores (criteria).[1] To report the mean of the rank components the teacher-specific correlations were transformed using a Fisher-*Z* transformation. This converts correlations into an almost normally distributed measure and allowed us to calculate the mean rank component. Then, this mean *Z*-value was recalculated to *r*. The level of significance was set at 5% for all inferential statistical analyses. For effect sizes we refer to Cohen (1992).

Results

First, we present the descriptive results for the teacher judgments and the pertaining student test scores. Then we continue with the results for the two research questions.

Table 2 shows the descriptive results for the teacher judgments as well as the achievement test scores.

Table 2

Overall, the teachers expected their students to perform better on the math items they assessed for the specific judgments than for the items they assessed for the global judgments. This assessment proves to be correct because students performed better in the six specific items than in the global items. Moreover, the teacher judgments show a broader range than the students' test scores.

---

[1] The Pearson's correlation coefficient is rated as a stricter criterion in contrast to Spearman's rank correlation coefficient. The reason for using the Pearson's correlation coefficient is that the calculation not only refers to a teacher's capability to estimate the ranking position of individual pupils within the class, but also to a teacher's capability to estimate the relative performance gap between individual pupils (cf. Schrader,1989).

Research Question 1 – Comparing Accuracy of GJ and SJ. Figure 1 displays the descriptive statistics of the accuracy measures depending on kind of judgment.

Figure 1

Differences between these accuracy measures were calculated by a 2 × 4 ANOVA, with repeated measures (kind of judgment [global vs. specific] × accuracy measure [level, differentiation, level error, differentiation error]). To compare the two rank components, we computed a *Z*-test for the equality of two dependent correlations following Steiger (1980).

The 2×4 ANOVA with repeated measures demonstrates two statistically significant main effects: The main effect for kind of judgment, $F(1,48) = 86.05$, $p < .001$, indicates that the accuracy differs with respect to the two conducted judgments (specific and global). The main effect for the accuracy measures, $F(3,46) = 924.71$, $p < .001$, indicates that the means of the four accuracy measures (level and differentiation) are unequal. The interaction between the two factors "kind of judgment" and "accuracy measure" is significant, too, $F(3,46) = 15.44$, $p < .001$. Table 3 shows the pairwise comparisons with Bonferroni correction and the result of the *Z*-Test (Steiger, 1980).

Table 3

The level component of the GJ indicates that the teachers underestimate the students' achievement level ($M = -0.07$), whereas the level component of the SJ indicates an overestimation ($M = 0.03$). This difference between the means is significant (Table 3). However, the deviation from the optimum value zero is only significant for the global level component, $t(48) = -5.47$, $p < .001$.

The levels of the differentiation components also show significant differences between the two kinds of judgment ($p = .001$) with a medium effect size ($d = .75$). Teachers tend to underestimate the heterogeneity of their students, but only in GJ, $t(48) = -6.98$, $p < .001$, because the differentiation component of the SJ does not deviate from the optimum value of 1.

The corresponding error components do not differ significantly between the kinds of judgment. Thus, the accuracy of the two kinds of judgment does not differ in absolute discrepancy of the criterion, but in the direction of discrepancy.

The mean rank component of GJ is $r = .64$, whereas the rank component of the SJ is $r = .36$. The comparison of means shows that the correlations are significantly different; $Z = 2.25$, $p = .024$.

The effect size with $q = .38$ illustrates that this difference is a moderate effect (Cohen, 1992). Thus, the rank component of the global judgment is significantly higher than the rank component of the specific judgment.

Research Question 2– Transferring Specific into Global Judgments. To analyze the second research question, we had to aggregate the six specific judgments by summing these judgments up for each student. After that, we calculated the accuracy measures as described above.

At first, we had a look at the correlation between the two kinds of judgments. We computed for each teacher semi-partial product-moment correlations between the GJ and aggregated GJ. Semi-partial correlation was necessary to control for the two different item-types on which the teachers judged in the GJ (addition vs. reading numbers). We also conducted class-specific product-moment correlation between the averaged students' test scores underlying the GJs and the averaged test scores underlying the SJs. Subsequently, we transformed all correlations using Fisher-*Z* transformation as described above. This allowed us to compare the correlations to each other. Secondly, we calculated, as before by a 2×4 repeated measure ANOVA, if the accuracy measures of each kind of judgment (original global and aggregated global) differ in their value. Thirdly, and finally, product-moment correlations measure the relationship between the same accuracy measures of the original global and the aggregated global judgment.

(1) The semi-partial product-moment correlation between original GJ and aggregated GJ revealed a strong relationship. The mean correlation is $r = .78$ ranging from .41 to .93 between the teachers. The mean correlation between the students' test scores is significantly smaller with $r = .55$ ($p < .001$) than the correlation between original GJ and aggregated GJ ($z = -10.22$; $p < .001$). In addition, the correlations between the student test scores show a larger range (from .01 to .80).

(2) The results of the 2×4 ANOVA are as follows: The main effect for original versus aggregated judgment, $F(1,48) = 68.96$, $p < .001$ indicates that the accuracy differs with respect to this factor. The main effect for the accuracy measure, $F(3,46) = 415.11$, $p < .001$ indicates that the means of the four accuracy measures (level and differentiation) are unequal. The interaction effect is also significant, $F(3,46) = 20.24$, $p < .001$. Figure 2 shows the descriptive statistics of the aggregated and global accuracy measures in comparison with each other.

Figure 2

To test differences of the means between the GJ-accuracy measure and aggregated GJ-accuracy measures, we conducted pairwise comparison corrected by Bonferroni adjustment and the Z-test referring to Steiger (1980). The results are shown in Table 4. The results for the level component and the level error are equivalent to the results in Table 3 because the level component and the level error do not differ between original SJs and aggregated GJ .[2] As mentioned above the level component for the original GJ indicates underestimation of the students, whereas the level component of the aggregated GJ is highly accurate in the mean. However, there is no significant difference between the two level errors. The aggregated differentiation component is, on average, significantly higher than the original global differentiation component (d = 1.19). This can also be shown for the differentiation error (d = 0.46). Due to aggregating the specific judgments, there is no difference between the two rank components (p = .84). Thus, measuring the rank component by global judgments shows approximately the same results as the mean rank component measured by aggregating specific judgments.

Table 4

(3) For the correlations between the accuracy measures, the findings suggest that there is no significant relationship between the original global level error and the aggregated one (Table 5). The relationship between the two differentiation errors is low and non-significant, but is negative. The relationship between the two rank components is slightly positive, but also not significant (r = .21). These findings suggest that teachers' assessment accuracy is not stable over different kinds of judgments, and that it is not possible to make a statement about the global judgment accuracy of a teacher by summing up specific judgments.

[2] The aggregated level component and the level error are identical to their non-aggregated original components (specific judgment). The reason for this is that the level component and the level error are theoretically defined as aggregated components because the mean students' test scores are subtracted from the mean teacher judgments.

Table 5

Discussion

In the scope of this study, two research questions have been investigated that focused on the comparison of teachers' GJs and SJs about their students' knowledge and skills. These judgments are part of everyday school life when the teacher has to adapt instruction to learning requirements, and go together with a certain perception of the students, namely a global versus specific perception.

Looking at teachers' judgment accuracy, our data showed the following: On average, the teachers do very well for both kinds of judgment. For example, the level error of 0.09 for the GJs indicates that teachers, on average, only overestimated or underestimated the students' performance by approximately one and a half correctly solved items (test-task) out of a total of 15 items (test-tasks). Also, the global rank component (r = .64) is fairly high and was found to be similar to the reported mean rank component (r = .63) in the meta-analysis of Südkamp et al. (2012).

The low rank component of the specific judgment is rather surprising. A first conclusion could be that the specific judgments are more difficult than global judgments. In specific judgments, the focus is on individuating characteristics of the students (e.g. prerequisites needed to master an exercise) and on characteristics of the exercise. The teachers have to decide whether a student will solve the exercise or not. In this case, false positive and false negative deviations have a more substantial impact on the calculation of the product-moment correlation (Bortz, 2005, p. 227), than it is the case for the GJ. In line with this assumption is the meta-analysis of Machts, Kaiser, Schmidt, and Möller (2016), in which the authors compared, among others, the judgment accuracy between dichotomous to non-dichotomous judgment scales. The judgment accuracy (measured as rank component) for non-dichotomous judgments (more than two response categories) was significantly higher than for dichotomous response scales (r = .45 vs. r = .36).

Concerning the first research question, it is not possible to clearly state whether one kind of judgment is more accurate than the other. Whereas the value of the level error and the value of the differentiation error do not depend on the kind of judgment, the level and differentiation components show better values for SJ than for GJ. Furthermore, the level component of SJ does not significantly deviate from the optimum value zero. Thus, the teachers overall reach an accurate assessment of their

students' achievement level when they make specific judgments about their students. The same picture emerges concerning the differentiation component. This component does not significantly differ from the optimum value of one. Based on many specific judgments about their students, teachers reach an accurate assessment of overall heterogeneity in the class. The differences in accuracy between the kinds of judgments relate to the direction of the judgments in the sense of over- or underestimation of the students' knowledge and skills, and not to the absolute difference of students' tests scores and teacher judgments. By contrast, the rank component is significantly higher and thus better for GJ than for SJ. This leads to the conclusion that it might be easier for teachers to get an accurate estimation of the achievement level of their students and of the amount of heterogeneity of their students when they make SJs about the students. However, when it comes to estimating the rank order and the relative performance gap of their students, in terms of the rank component, the more accurate judgments are GJ.

The focus of the second research question was the comparability of original GJ and aggregated GJ and their corresponding accuracy measures.

At first, the results suggest a strong relationship between GJ and aggregated GJ. This means, that the teachers' assessments of the students is quite consistent and that the specific type of math exercise (e.g., addition of whole numbers; double a number) makes no difference to their assessment. The relationship between teachers' judgments for individual students is considerably closer than the relationship between the individual students' performances. The closer relationship may also indicate that the teachers expect more stable student achievement across the judged test-tasks than is actually present. This shows that teachers almost use an overall assessment benchmark to judge the performance of their students, and that the teachers do not differentiate between the difficulties of the exercises.

Concerning the judgment accuracy, we found mean differences between the aggregated and original global accuracy measures. The aggregated level component is closer to 0 and thus significantly better than the original global level component. The level errors do not differ significantly. The aggregated differentiation component is higher than the original one and significantly above the optimum value of 1. Whereas original GJ leads to an underestimation of the

amount of heterogeneity, the aggregated GJ leads to an overestimation of the amount of students' heterogeneity. This finding suggests that by the integration of individuating information—as could be assumed for SJ—differences between students were unconsciously accentuated. Hence, this accentuation does not become obvious until the SJs are aggregated to GJ. By aggregating and thereby averaging the test scores, intra-individual differences are eliminated, which would otherwise be part of the computation of heterogeneity and thus of the differentiation component. The misjudgment of heterogeneity by aggregating the SJ is also obvious for the comparison of the differentiation errors. The error is larger for the aggregated differentiation error than for the original global differentiation error. Finally, the results show that the aggregated specific rank component is almost as high as the original global rank component. Further, correlation analysis shows no substantial and significant relationships between the same accuracy measures of the aggregated global and original global variants. This means that a GJ is not the sum of many single SJs about one student, and indicates that the judgment processes differ between the kinds of judgments.

Theories of social information processing (Chaiken, 1980; Fiske & Neuberg, 1990) postulate that different (social) judgments depend of different types of information, different information integration processes, which may result in different outcomes. Applied to the SJ assessed in the present study, one may assume that an SJ fosters a controlled and systematic information processing. SJ means that a teacher judges a student with regard to the likelihood to master an exercise. In doing so, the focus of the teacher is on individuating information about the student *and* the exercise. The way of information integration is guided by these specific characteristics of the judgment. Concerning GJ, one may assume that other factors besides individuating information influence information processing and thus also influence the GJ. Such factors may be characteristics that result from social comparisons of the teacher between the students. Through these social comparisons, the teachers draw attention to general differences between their students' knowledge and skills and thus neglect specific characteristics of individual students, which might be necessary for a high level of judgment accuracy. These social comparisons might then result in lower accuracy of the level and differentiation component for the GJ than is the case for these accuracy measures resulting from SJ. We consider this finding as a hint that GJ is based in a rather heuristic processing of information. The results for the

rank component can be interpreted similarly, since the rank component indicates the knowledge of the teacher about the relative positions of his students in the class regarding achievement. The lower rank component emerges when the teachers make SJ. By aggregating these judgments, the rank component increased in the mean. There are no differences in the mean between original global and aggregated rank component. By aggregation, intra-individual differences are averaged and neglected. This indicates different judgment processes. If individuating information is ignorable, then the rank order of students can be better assessed.

Implications

The results show that the kind of judging has an impact on the accuracy of the judgment. Within the scope of teaching and teacher education, we can derive the following implications: GJ are more appropriate to assess the rank order of the students' skills and knowledge. On the contrary, global judgments (which may be assumed to be more strongly based on heuristic information processing, Scherman et al., 1999) would be less suitable to assess the level and differentiation of the students as accurately as possible. For the latter purpose, specific judgments (which are assumed to be subject to controlled information processing) seem particularly appropriate. As in previous studies, there is no clear statement possible about which judgment (in our case: specific vs. global judgment) leads to higher accuracy. One could say that it depends on the component of accuracy that is considered and perhaps on the teachers' information processing mode. Related to teachers' professional development and teaching practice, we can conclude that teachers should use *those* judgments which are more accurate for pedagogical decisions. If the teacher aims to gain information about the achievement level of the students and about the amount of heterogeneity, results of this study would advise the teacher to mostly judge the students specifically. If the teacher seeks information about the rank order of the students, global judgments would be more accurate. Following this assumption, this would mean that the global assessment of the students leads to a more appropriate internal differentiation in the class because the teacher is more able to determine the rank order of the students.

These findings also indicate that it will be necessary in further research to provide different kinds of judgments because it is impossible to draw inferences about one accuracy measure (e.g., GJ)

given the other measure (e.g., SJ). Teachers' GJ are not a result of aggregated single SJs. The differentiation between kinds of judgment, and thus between diagnostic situations during teaching with respect to the perspective of judges (teachers) about the judged target (student), seems necessary from an empirical and theoretical point of view. This means that for professional teacher development, teachers should be instructed to use such judgments that fit best in the underlying diagnostic situation.

Limitations of the Present Study

Based on the results of our study, we see initial evidence supporting the assumption that the information processing underlying global and specific judgments differs. However, further research is needed. Additionally, considering that judgment accuracy is not a uniform construct (Praetorius et al., 2011), it seems important to gain more insight into judgment processes that can lead to different outcomes in judgment accuracy.

Limits of our study are due to the different types of arithmetic items, which form the basis for the two kinds of judgment. Colardarci (1986) showed that the accuracy of teachers' judgments differs with respect to the judged item-types within one subject, namely math. In his study, the difference between the items, however, concerned the complexity of the cognitive processes, which were necessary to solve the test-items.). Such a difference is not applicable for the item-types used in our study. All test-items showed a rather similar level concerning the complexity of cognitive processes, which were relevant for solving. Moreover, all items, which were assessed by teachers, are scalable in accordance with the 1PL Rasch model (Wu, Adams, & Wilson, 1998). In line with Adams and Khoo (1996), the measured fit statistics of this scale are satisfactory (EAP/PV reliability = .95, WMNSQ: 0.75-1.26). To avoid this limitation, it might be possible in future studies to prompt teachers to judge the same items for each kind of judgment. However, this might seem very strange and artificial to the teachers, so such an approach would need other research designs. Additionally, it would be positive to expand the thematic areas of the judged items. This would raise the validity and generalizability of the results.

The fact that each teacher was required only to offer SJs for each of 12 of their students and, even more importantly, for just six separate test items is another weak spot of our study. A more

robust basis for the SJ data – that is more students judged and more SJs provided per student – would strengthen the validity of the findings, and therefore in turn its interpretational value.

Further limit occurs with the operationalization of the GJs. Referring to the GJ the situation for the teachers in the study and the situation they experience in ordinary work are not directly comparable. Teachers usually will make even more global judgments about their students' knowledge and skills, for example about a larger domain. Thus, the global judgments in our study would be closer to the endpoint of specificity than to the global endpoint if one thinks about these kinds of judgments as a dimension with two endpoints (cf. Hoge & Coladarci, 1989). However, empirical research about the kinds of judgments that are part of teachers' daily work is still missing. Until evidence based research exists the specificity of judgments remains a question of definition.

Our teacher-sample is a non-representative convenience sample of primary school teachers in Germany. Although it is not a random sample, it is comparable to the whole population of primary school teachers in Germany, namely with regard to gender ratio and mean age (Tarelli et al.; 2012; www.destatis.de). However, the results are not transferable to the whole population. The analyses are limited to the subject of math in primary school classes. Additionally, a larger sample would be desirable to verify the results. A small sample size could lead to an underestimation of correlations and their significance because the standard errors of the correlation coefficients become greater with a decreasing sample size (Bortz, 2005).

The reported accuracy measures in this study are manifest measures. However, some recently published studies measured the accuracy of teachers' judgments via latent variable modeling in the scope of multilevel analyses (Johansson, Myrberg & Rosén, 2012; Leucht, Tiffin-Richards, Vock, Pant, & Köller, 2012; Author, 2017). This strategy would also be an elaborated and desirable opportunity to compare SJ and GJ in future research. However, it is not possible to measure the differentiation component and differentiation error via linear mixed models. Thus, we decided to use the manifest variant because we wanted to compare all accuracy measures.

Conclusion

Considering this state of knowledge, we can summarize as follows: SJs lead to higher judgment accuracy concerning the assessment of the achievement level and heterogeneity of the class,

whereas GJs lead to higher judgment accuracy concerning the assessment of the students' rank order of the students in the class. Thus, teachers and researchers should prefer a specific judgment to receiving accurate insight into the achievement level of the students and the amount of heterogeneity in a class. To get information about the rank order and relative gap between the students, teachers and researchers should refer to global judgments. However, if only specific judgments are available, results show that the aggregated specific judgments reveal almost the same mean rank component as is obvious for the global judgments.

References

Adams, R. J., & Khoo, S. T. (1996). *Quest*. Melbourne: Australian Council for Educational Research.

Author, 2001. [details removed for peer review]

Author, 2012. [details removed for peer review]

Author, 2017. [details removed for peer review]

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology*, *21*, 177–187. doi:10.1080/01443410020043878

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review*, *4*, 165–176. doi:10.1016/j.edurev.2009.04.002.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler* (Vol. 6). Heidelberg: Springer.

Brighton, H., & Gigerenzer, G. (2012). How heuristics handle uncertainty. In P. M. Todd, G. Gigerenzer, & ABC Research Group (Eds.), *Ecological rationality. Intelligence in the world* (pp. 33–60). New York: Oxford University Press.

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research*, *68*, 1772–1784. doi:10.1016/j.jbusres.2015.01.061

Brühwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction*, *21*, 95–108. doi:10.1016/j.learninstruc.2009.11.004

Burns, R. B. (1984). The process and context of teaching: A conceptual framework. *Evaluation in Education*, *8*, 95–112.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*, 752–766. doi:10.1037/0022-3514.39.5.752

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

Coladarci, T. (1986). Accuracy of teacher judgements of student responses to standardized test items. *Journal of Educational Psychology*, *78*, 141–146. doi:10.1037/0022-0663.78.2.141

Crisp, V. (2013). Criteria, comparison and past experiences: How do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice*, *20*, 127–144. doi:10.1080/0969594X.2012.741059

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of Educational Research*, *102*, 453–462. doi:10.3200/JOER.102.6.453-462

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). New York: Academic Press.

Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*, 245–270. doi:10.1080/02671520500193744

Harwell, M. R., Post, T. R., Maeda, Y., Davis, J. D., Cutler, A. L., Andersen, E., & Kahan, J. A. (2007). Standards-based mathematics curricula and secondary students' performance on standardized achievement tests. *Journal for Research in Mathematics Education*, *38*, 71–101.

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgment accuracy on achievement. *Teaching and Teacher Education*, *3*, 91–98. doi:10.1016/0742-051X(87)90010-2

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., ... Südkamp, A. (in press). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Journal of Teaching and Teacher Education.* Advance online publication. https://doi.org/10.1016/j.tate.2017.12.001

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297–313. doi:10.2307/1170184

Johansson, S., Myrberg, E., & Rosén, M. (2012). Teachers and tests: Assessing pupils' reading achievement in primary schools. *Educational Research and Evaluation*, *18*, 693–711. doi:10.1080/13803611.2012.718491

Karing, C., Matthäi, J., & Artelt, C. (2011). Hängt die diagnostische Kompetenz von Sekundarstufenlehrkräften mit der Entwicklung der Lesekompetenz und der mathematischen Kompetenz ihrer Schülerinnen und Schüler zusammen? [Is there a relationship between lower secondary school teacher judgment accuracy and the development of students' reading and mathematical competence?] *Journal for Educational Research Online*, *3*, 119–147.

Krajewski, K., Küspert, P., & Schneider, W. (2002). *Deutscher Mathematiktest für erste Klassen (DEMAT 1+)*. Göttingen: Hogrefe.

Krajewski, K., Liehm, S., & Schneider, W. (2004). *Deutscher Mathematiktest für zweite Klassen (DEMAT 2+)*. Göttingen: Hogrefe.

Leucht, M., Tiffin-Richards, S., Vock, M., Pant, H. A., & Köller, O. (2012). Diagnostische kompetenz von Englischlehrkräften bei der bewertung von schülerleistungen mit hilfe des Gemeinsamen Europäischen Referenzrahmens für Sprachen. [English teachers' diagnostic skills in judging their students' competencies on the basis of the Common European Framework of Reference.]. *Zeitschrift Für Entwicklungspsychologie und Pädagogische Psychologie*, *44*, 163–177. doi:10.1026/0049-8637/a000071

Lorenz, C., & Artelt, C. (2009). Fachspezifität und Stabilität Diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. [Domain specificity and stability of diagnostic competence among primary school teachers in the school subjects of German and mathematics.]. *Zeitschrift Für Pädagogische Psychologie*, *23*, 211–222. doi:10.1024/1010-0652.23.34.211

Machts, N., Kaiser, J., Schmidt, F., T. C., & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Preview*, *19*, 85–103. doi:10.1016/j.edurev.2016.06.003

Praetorius, A., Karst, K., Dickhäuser, O., & Lipowsky, F. (2011). Wie gut schätzen Lehrer die Fähigkeitsselbstkonzepte ihrer Schüler ein? Zur diagnostischen Kompetenz von Lehrkräften.

[How teachers rate their students: On teachers' diagnostic competence regarding the academic self-concept.]. *Psychologie In Erziehung Und Unterricht*, *58*, 81–91. doi:10.2378/peu2011.art30d

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts* [Diagnostic competencies of teachers and their relevance for composition and effectiveness of teaching] (Vol. 289). Frankfurt am Main, Bern, NY, Paris: Peter Lang.

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research*, *51*, 455–498. doi:10.2307/1170362

Sherman, S. J., Beike, D. R., & Ryalls, K. R. (1999). Dual-processing accounts of inconsistencies in responses to general versus specific cases. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 203-227). New York: Guilford Press.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. [Accuracy of Teacher Judgments on Student Characteristics and the Construct of Diagnostic Competence.]. *Zeitschrift für Pädagogische Psychologie*, *19*, 85–95. doi:10.1024/1010-0652.19.12.85

Stang, J., & Urhahne, D. (2016). Wie gut schätzen Lehrkräfte Leistung, Konzentration, Arbeits- und Sozialverhalten ihrer Schülerinnen und Schüler ein? Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften. [How teachers rate students achievement, attention, work habits and social behavior? A contribution to the diagnostic competence of teachers.]. *Psychologie In Erziehung Und Unterricht*, *63*, 204–219. doi:10.2378/peu2016.art18d

Steiger, J. H. (1980). Tests for Comparing Elements of a Correlation Matrix. *Psychological Bulletin*, *87*, 245–251. doi:10.1037/0033-2909.87.2.245

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgements of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*, 743–762. doi:10.1037/a0027627

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz. [The simulated classroom: An experimental

study on diagnostic competence.]. *Zeitschrift für Pädagogische Psychologie*, *22*, 261–276. doi:10.1024/1010-0652.22.34.261

Tarelli, I., Lankes, E.M., Drossel, K., & Gegenfurtner, A. (2012). Lehr- und Lernbedingungen an Grundschulen im internationalen Vergleich [Teaching and learning conditions at primary schools in international comparison]. In W. Bos, I. Tarelli, A. Bremerich-Voss, & K. Schwippert (Eds.), *IGLU 2011—Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* [IGLU 2011—Reading competencies of primary schoolchildren in Germany in international comparison] (pp. 137–173). Münster: Waxmann.

Weinert, F. E., & Helmke, A. (1995). Interclassroom differences in instructional quality and interindividual differences in cognitive development. *Educational Psychologist*, *30*, 15–20. doi:10.1207/s15326985ep3001_2

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalized item response modelling software.* Melbourne: Australian Council for Educational Research.

Zhu, M., & Urhahne, D. (2014). Assessing teachers' judgements of students' academic motivation and emotions across two rating methods. *Educational Research and Evaluation*, *20*, 411–427. doi:10.1080/13803611.2014.964261

Table 1

Characteristics of SJ and GJ integrated into earlier definitions of kinds of judgments

	Direct vs. Indirect	Informed vs. Uninformed	Congruence of Domain Specificity	Specificity[a]
Specific Judgment	Direct	Informed	High	Level 5
Global Judgment	Direct	Informed	High	Level 4

Notes. [a] Specificity according to the definition of Hoge and Coladarci (1989)

Table 2

Descriptive results for teacher judgments and student achievement test scores (relative frequencies)

	n	*M*	*SD*	Range
Teacher judgment				
Global	49	.57	.26	.11-.80
Specific	52	.69	.29	.38-.96
Student test score				
Global	700	.64	.29	.35-.76
Specific	434	.66	.25	.41-.80

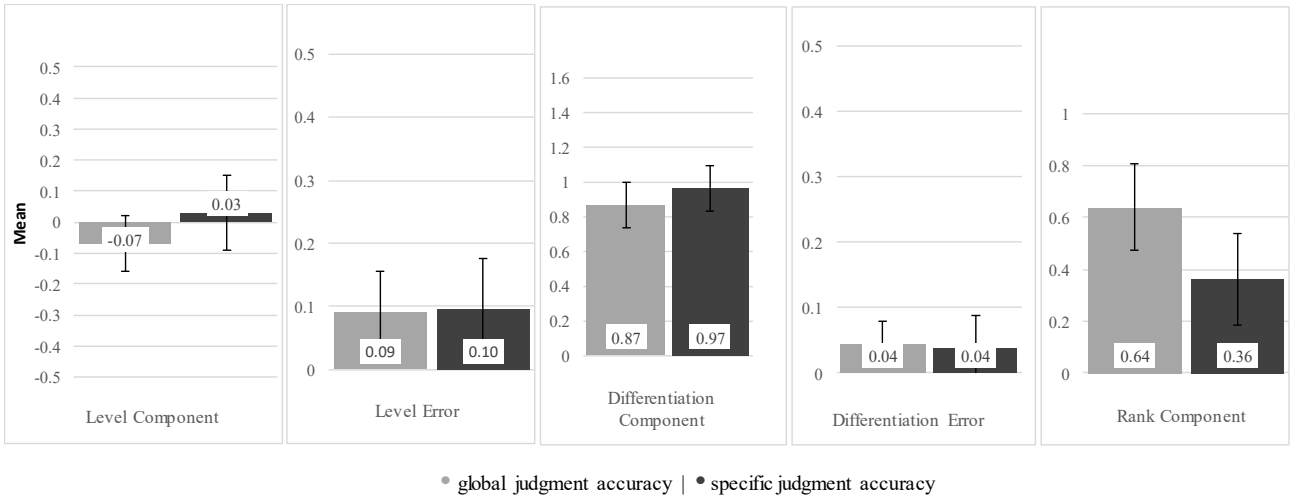

Figure 1. Descriptive statistics (*M*, *SD*) of the accuracy components depending on kinds of judgment.

Table 3

Pairwise comparisons for the accuracy measures depending on kind of judgment corrected using Bonferroni adjustment

Accuracy measure	Δ	*p*	*d*
Level component	-0.099	<.001	.926
Level error	-0.005	.743	.067
Differentiation component	0.100	.001	.748
Differentiation error	0.007	.472	.157
	Z	*p*	*q*
Rank component	2.254	.024	.376

Notes. Δ = difference between GJ-accuracy and SJ-accuracy; *d* = Cohen's *d*; *q* = Cohen's *q*.

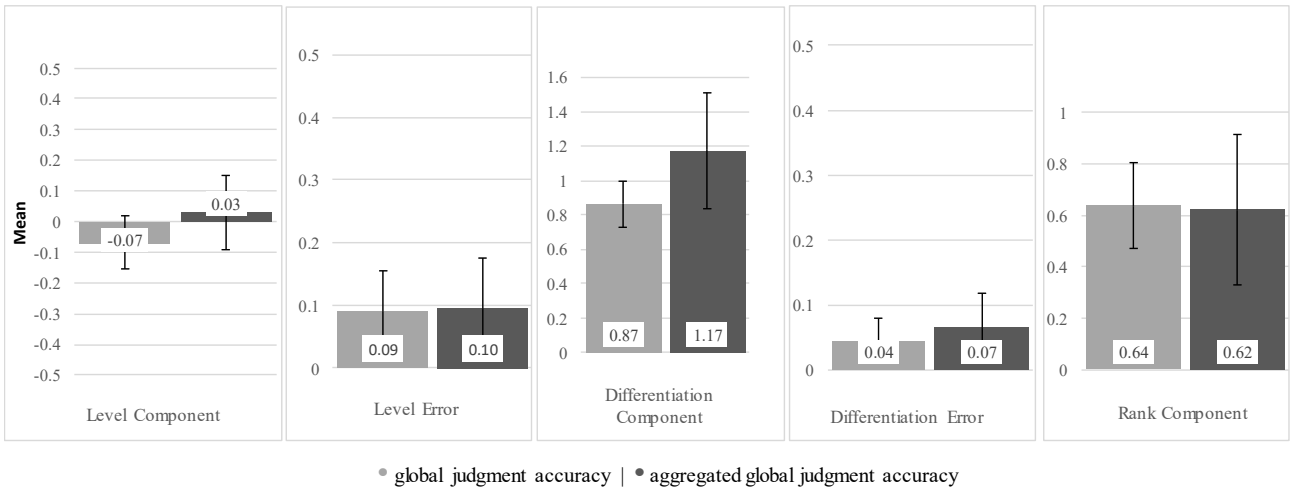

Figure 2. Descriptive statistics (*M*, *SD*) of the accuracy measures depending on GJ and aggregated GJ.

Table 4

Pairwise comparisons for the accuracy measures depending on original GJ and aggregated GJ corrected using Bonferroni adjustment

Accuracy measure	Δ	*p*	*d*
Level component	-0.099	<.001	0.926
Level error	-0.005	.743	0.067
Differentiation component	0.308	<.001	1.199
Differentiation error	-0.020	.044	0.456
	Z	*p*	*q*
Rank component	0.201	.841	.026

Notes. Δ = difference between GJ-accuracy and aggregated GJ-accuracy; *d* = Cohen's *d*; *q* = Cohen's *q*.

Table 5

Correlation between the accuracy of aggregated GJ and original GJ (N = 49)

	Original (Global)		
	Level	Differentiation	Rank
Aggregated (Global)			
Level Error	.00		
Differentiation Error		-.21	
Rank Component			.21