Who Makes More Accurate Judgments of Comprehension: Students or Their Teachers?

A comparison study of students' and teachers' judgment accuracy and their accompanying use of cues

E. Kramer (4145518)
Bachelor thesis educational sciences
Utrecht University
Supervisor: Janneke van de Pol

Utrecht University

Abstract

The importance of accurate judgments of comprehension for both students and their teachers is well-acknowledged throughout the literature. However, there has not been comparative research about which actor is more skilled in making accurate judgments. Especially in view of the current developments in educational practices, which place greater emphasis on metacognitive skills of students, this comparison is of great importance. This explorative study, based on the cue-utilization framework, is the first to examine the comparison between teachers' and students' judgments accuracy and their accompanying use of cues. To this end, secondary school students read several texts, completed causal diagrams, and were tested for comprehension of the causal relations from those texts. Both students themselves and their teachers judged the students' extent of comprehension. A key finding of this study is that, overall, teachers make significantly more accurate judgments of students' comprehension than students themselves, although the effect was small. This difference in judgment accuracy between teachers and students could partly be explained by significant differences in cue-utilization. Practical implications and suggestions for further research are discussed.

Keywords: judgments of comprehension, cue-utilization, judgment accuracy, causal relations, diagram

Who Makes More Accurate Judgments of Comprehension: Students or Their Teachers?

In the last decades, the focus of education has shifted from purely transferring knowledge to teaching students to guide their own learning processes (Delfino & Persico, 2009; Thomas & Brown, 2011). This view of education places more emphasis on metacognitive skills, which concern the procedural knowledge and executive skills that are required for regulation and monitoring of one's learning activities (Brown & DeLoache, 1978; Flavell, 1992). Currently, new forms of education are rising which presume metacognitive and self-regulated learning skills of students. However, there has not been much comparative research concerning which actor is more skilled to give direction to learning processes: students themselves or their teachers? An essential part of this skill involves making accurate judgments about learning (Nelson & Narens, 1990; Schneider, 2008), which is the main focus of this study.

Research on making judgments builds on the well-acknowledged cue-utilization framework of Koriat (1997), which states that to judge their learning, people use cues that are accessed prior to making a judgment. Examples of such cues are the perceived relative difficulty of the study items or the ease with which information normally comes to mind (Thiede, Griffin, Wiley, & Anderson, 2010). Because cues are used to make a judgment about test performance, the accuracy of the judgment will be determined by how well those cues predict test performance, i.e., cue-diagnosticity (Brunswik, 1956; Koriat, 1997). When cues are used that are more diagnostic of subsequent test performance, judgment accuracy will improve (Thiede et al., 2010).

The primary goal of this study is to compare students' and teachers' judgment accuracy and their accompanying use of cues. First, a description of the process and importance of making accurate judgments for both students and teachers will be presented. Subsequently, the necessity of this comparison is described, together with the theoretical predictions of this study.

Student Judgments of Comprehension

Importance of students' judgment accuracy. Previous studies have used various definitions and formulations for the concept of student judgments. In this study, students' judgments of comprehension are defined as judgments made by students concerning their ability of recalling and applying information from a text on a subsequent test (partly based on Koriat, 1997). Those judgments are said to be accurate when they are consistent with objective assessments of the same skill (Ready & Wright, 2011). Students who can accurately judge their level of understanding are able to learn more from textual information (Dunlosky & Rawson, 2012; Thiede, Anderson, & Therriault, 2003). In particular, if students can judge which materials they have understood well and which they have not, they can focus their attention solely on not-

understood information (Dunlosky, Hertzog, Kennedy, & Thiede, 2005). As a result, students' judgment accuracy is critical for continued strategy use, study decisions and consequently learning efforts (Metcalfe & Finn, 2008; Thiede, Anderson, & Therriault, 2003). Unfortunately, students' judgments of text comprehension are often inaccurate (De Bruin, Thiede, Camp, & Redford, 2011; Dunlosky & Lipko, 2007). As a result, students will not be able to use their judgments to appropriately guide their learning processes.

Improving students' judgment accuracy: delayed diagram completion task. Multiple studies have shown that to improve judgment accuracy when learning from text, learners need to base their judgments on cues that arise from processing information about the gist of the text (e.g., Thiede et al., 2010; Rawson, Dunlosky, & Thiede, 2000). With respect to expository texts, this gist comprehension depends largely on a reader's ability to connect and understand the causal relations in a text (Graesser, Singer, & Trabasso, 1994).

An intervention that helps students focus on causal relations is the diagram completion task, as used by Van Loon, De Bruin, Van Gog, Van Merriënboer, and Dunlosky (2014). In this task, students had to depict the steps in causal chains of several texts in pre-structured and partly-filled in causal diagrams. The results of this study showed that this task provides learners with cues that indicate whether they have understood the causal relations within the text (Van Loon et al., 2014). Because these cues are diagnostic of subsequent performance, judgment accuracy is improved. Especially when the diagram completion task was delayed, i.e., completed some time after studying the texts, higher judgment accuracy was supported (Van Loon et al., 2014). This is in line with prior research, suggesting that cues produced by a task vary in the degree of diagnosticity when the task is performed immediately versus at a delay, due to the difference in the level of mental representation involved in performing the task (e.g., Thiede, Dunlosky, Griffin, & Wiley, 2005). For complete text comprehension, learners should go beyond the processing of factual information and establish a coherent mental representation of the gist of the text (Kintsch, 1998). By means of the delay, the diagram completion task helps readers focus on the quality of their mental model, which yields diagnostic cues (Van Loon et al., 2014).

However, these specific findings of Van Loon and colleagues (2014) have not been verified yet. Since replication studies are valuable for the reliability of results (John, Loewenstein, & Prelec, 2012; Lakens, Haans, & Koole, 2012), a secondary aim of this study is to verify whether the delayed diagram completion task indeed supports higher students' judgment accuracy.

Cue-utilization. Based on the diagram completion task, Van Loon et al. (2014), established the existence of three diagnostic cues: the extent to which correct causal relations

were provided in the diagram responses (correct relations), the extent to which provided answers were not based on the information of the text (commission errors) and the extent to which no response was given (omissions). Their study is one of the first in this specific area yielding quantitative cues, and since those cues are found to be helpful in analysing judgment accuracy and cue-utilization, the same cues are taken into account in this study. In other words, when making judgments about their comprehension, students are expected to use the three cues described above to a greater or lesser extent.

Teacher Judgments of Comprehension

Importance of teachers' judgment accuracy. Teachers also make ongoing judgments about students' understanding (e.g., Alvidrez & Weinstein, 1999). In this study, teachers' judgments of comprehension are defined as judgments made by a teacher concerning a students' ability of recalling and applying information from a text on a subsequent test. The ability to accurately assess students' comprehension is considered to be an important aspect of teachers' professional competence (Ready & Wright, 2011). Teachers' judgments guide instructional decisions that may affect student performance (Südkamp, Kaiser, & Möller, 2012). Specifically, more accurate judgments can lead to better differentiation of instruction, which produces greater gains in student learning (Thiede et al., 2015). Moreover, teachers' judgments influence their expectations about students' abilities (Brophy & Good, 1986), they influence students' academic self-concept (Möller, Pohlmann, Köller, & Marsh, 2009), and they identify struggling students (Bailey & Drummond, 2006).

Teachers' judgment accuracy: room for improvement. Recently, a meta-analysis of 75 articles about teachers' judgment accuracy is conducted by Südkamp, Kaiser, and Möller (2012). Their results show that teacher judgments are far from perfect and that there is plenty of room for improvement. Remarkably, a lot of variation in teachers' judgment accuracy could not be explained, suggesting that teachers vary widely in their judgment accuracy. Understanding these different levels of accuracy is complicated by the fact that researchers have used a variety of approaches to compute the correlation between predicted and actual performance, i.e., judgment accuracy (Thiede et al., 2015). As in prior research, the focus of this study is placed on relative accuracy, which is the degree to which predictions discriminate between the different levels of test-performance for one text relative to another (Van Loon et al., 2014).

Cue-utilization. There are a variety of cues available to teachers to judge students' comprehension, for example the former achievements of the students or the global characteristics of the texts studied (Thiede et al., 2015). As described above, this study builds upon the cues presented by Van Loon et al. (2014): correct causal relations, commission errors and omissions,

based upon the diagram response categories. These cues are also expected to be used by teachers, to a greater or lesser extent, to make judgments about the degree of comprehension of their students.

The Present Study

In the present study, judgment accuracy and cue-utilization of both students and teachers are examined. As described above, different studies have used a variety of approaches to measure students' and teachers' judgment accuracy, which makes comparison challenging. A study is needed which measures those constructs in the same way, so that results can be compared and subsequently explained. This study is the first to make this comparison. Especially in view of the described current developments in educational practices, this comparison is of great importance. The main research questions of this study are: 1. To what extent are there differences between students and teachers regarding judgment accuracy? 2. Can potential differences be explained by differences in cue-utilization? Previous studies have shown differing results of judgment accuracy for both students and teachers (De Bruin et al., 2011; Dunlosky & Lipko, 2007; Südkamp et al., 2012; Thiede, 2015; Van Loon; 2014). Since there has not been enough univocal evidence to give direction to the hypotheses, this study will take an exploratory perspective.

Because this study relies much on that of Van Loon et al. (2014), a secondary research question concerns replicating their findings: 3. To what extent does the delayed diagram completion task lead to more accurate students' judgments in comparison to omission of this task? Based on the results of Van Loon et al. (2014), judgment accuracy is hypothesized to be higher when students complete the delayed diagram completion task, relative to when they do not.

Method

Participants

Both students and teachers participated in this study. By means of a convenience sample, fifteen teachers from six secondary schools across the Netherlands were recruited, of which 10 women and 5 men. The teachers were recruited through contact with the rectors of the schools. The average age of the teachers was 40.40 years (*SD* = 10.90, range 24-58 years). Teachers from various subjects participated; the inclusion criterion was that reading and studying explanatory texts was an essential part of the subjects' curriculum. The average years of teaching experience of the teachers was 14.88 (*SD* = 9.36, range 2.7-35 years), with on average 8.87 months of exposure to the students they judged in this study (*SD* = 3.38).

Each teacher participated with one of their third year classes. In this way, 261 high school students participated, of which 60% were female and 40% were male. All students were between

11 and 16 years old (M = 14.59, SD = 0.64) and followed one of the two educational programs that lead to higher education (this was the inclusion criterion for the classes): 60% students followed the higher general secondary education program (HAVO, middle level of secondary education), and 40% students followed the pre-university program (VWO, highest level of secondary education).

Materials

Table 1 provides a schematic overview of the various phases of this study. The materials that were presented during each phase will be described below.

Text study. Since part of this study comprises replication of the results of Van Loon et al. (2014), the same six explanatory texts were used as in their study. In those texts were both causal and factual relations presented (see Appendix A for an example of a text). Each texts contained five causal relations. The topics of the texts were 'Sinking of metro cars', 'Botox', 'The Suez Canal', 'Music makes smart', 'Money does not bring happiness' and 'Renovation of concrete constructions'.

Diagram completion task. In the diagram completion task, students were provided with a pre-structured diagram for each text. Each diagram consisted of five textboxes representing either serial or simultaneous causal relations, of which one textbox was already filled in (see Figure 1 for a completely filled-in example diagram).

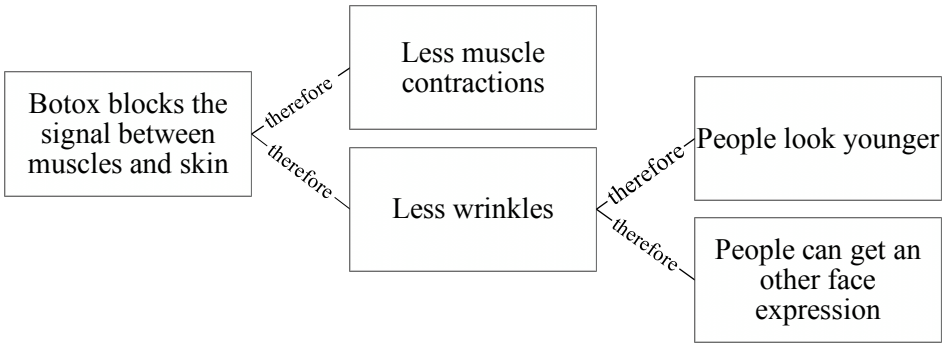

Figure 1. A correctly completed diagram for the text 'Botox'.

Scoring of diagrams. The students' responses in the boxes of the diagrams from the diagram completion task were classified into three categories, corresponding with the cues compiled by Van Loon et al. (2014). A response was scored as correct (category 1) when answers literally showed the causal relations or showed gist understanding of the text. A response was scored as a commission error (category 2) when incorrect causal relations were established or vague answers were given. Lastly, when students had not filled in a textbox, this was scored as an omission error (category 3). Based on these categories, a score was compiled concerning the number of correct relations mentioned, ranging from 0 to 4 since students had to identify four

Table 1

Schematic Overview of the various Phases of this Study

Students	
Delayed-diagram condition	No-diagram condition
Practice session	
Text study: Read text 1	Text study: Read text 1
Text study: Read text 2	Text study: Read text 2
…	…
Text study: Read text 6	Text study: Read text 6
Diagram completion task: Diagram 1	Filler task
Diagram completion task: Diagram 2	
…	
Diagram completion task: Diagram 6	
Student Judgments of Comprehension	
Test	
Teachers	
Practice session with student materials	
Practice session Teacher Judgments of Comprehension	
Teacher Judgments of Comprehension	

causal relations from each text. Three raters independently coded part of the diagrams and inter-rater agreement was sufficient, Kappa = .74 (Cohen, 1960; McHugh, 2012).

Judgments of Comprehension. Students were for each text they had read asked to provide a judgment of their comprehension concerning the causal relations of that text. The title of the text was presented to them, accompanied by the following question: How many questions concerning the causal relations of this text do you expect to answer correctly on the test? The response scale for this question ranged from 0 to 4. Teachers were also asked to provide a judgment for each text the students had read. They were presented with the same question and response scale, but then about the students' understanding of causal relations from the texts.

Test. Students were tested for their understanding and remembrance of causal relations from each text with a test. This test included for each text one question about the causal relations, in which students were asked to identify four causal relations from the corresponding text.

Scoring of test performances. For each question on the test, a score was computed that indicated the number of correctly stated causal relations. This score could range from 0 to 4, since students had to identify four causal relations from each text. Because comprehension was emphasized, responses were also scored correct when the students did not respond with what was literally stated in the text, but instead responded with an answer indicating gist comprehension of causal relations from the text. Again three raters independently coded part of the test performances. Inter-rater reliability analysis yielded a Kappa of .61, which indicates a moderate but substantial level of agreement (Cohen, 1960; McHugh, 2012).

Design and Procedure

A between-subjects design was used to compare students and teachers regarding their judgment accuracy and cue-utilization (research questions 1 and 2). To confirm whether or not the delayed diagram completion task was indeed valuable for making accurate judgments (research question 3), an experimental between-subjects design was used. To address this question, students were randomly assigned to either the delayed-diagram condition or the no-diagram condition. All data collected in the experiment was processed anonymously and interpreted with care and precision, which guarantees the confidentiality of the participants.

Table 1 broadly depicts the procedure of this study for both students and teachers. In advance, informed consent of both students, their parents and teachers was formally taken care of. In the student practice session, students were instructed about the type of texts they were going to read, the diagram completion task, judgments of comprehension, and lastly the test format. Students were informed that they would study six texts for a later performance test with questions about the causal relations from those texts. Students in the delayed-diagram condition first read all six texts and then started with the diagram completion task (see also Table 1). Students in the no-diagram condition completed a filler task after reading all six texts, in which they needed to spot differences between images. Thereafter, students were asked to judge their comprehension of each text by predicting future test performance and subsequently they took the test. Students completed all the tasks in one session that lasted for approximately one and a half hour, which took place in their own classroom. The texts were presented according to the Latin Square Design, whereas the sequence of the other materials presented was random. All experimental tasks were self-paced by the students.

Teachers also started with a practice session about the student materials. Subsequently, teachers practiced with predicting students' performances by estimating the performances of three random selected students. The teachers were provided with the students' completed diagrams and were instructed to base their predictions of student performance on those diagrams.

After this practice session, all teachers provided judgments of comprehension for each text for fifteen random selected students. The teacher sessions also took place in classrooms of their schools and lasted approximately one hour.

Analyses

Judgment accuracy is operationalized as the intra-individual correlation between test performance and judgment of comprehension, and is measured using the gamma correlation. This non-parametric statistic has been considered one of the most appropriate measures of relative accuracy (Nelson, 1984; Van Loon et al., 2014). The value of the gamma indicates the strength of the association between judgments of comprehension and test performance, ranging from –1 to +1, with a stronger positive correlation indicating greater accuracy. To examine the differences between students and teachers regarding judgment accuracy (RQ 1), a one-way between groups ANOVA is performed. An ANOVA avoids the problem of an inflated experimentwise alpha level, that could occur using an independent samples *t* test (Gravetter & Walnau, 2013), which is why an ANOVA was chosen. The factor being studied is 'person' (students or teachers), and the dependent variable is judgment accuracy. This ANOVA tests whether in general students can more accurately judge themselves or whether teachers can more accurately judge them.

The second main research question, whether potential differences in judgment accuracy can be explained by differences in cue-utilization, consists of two parts. Firstly, differences in cue-utilization are examined. Cue-utilization is measured using a within-participant correlation between the diagram responses and judgments of comprehension. Intra-individual gamma correlations between the number of the response types (i.e., cues) and judgments of comprehension were calculated for each text. A correlation involving a particular response type that is greater than 0 indicates the cue is used for making judgments, with increasingly higher correlations (closer to +1.0) indicating greater utilization. Potential differences between students and teachers concerning cue-utilization (RQ 2) are examined with a MANOVA. Again the independent variable is 'person' (students or teachers), and the dependent variables are the cue-utilization of correct relations, commission errors and omissions. Secondly, the relation between judgment accuracy and person was tested for influence by cue-utilization (RQ 2). Because of the exploratory perspective this study takes, the choice was made for moderation-analyses to examine whether, and how, cue-utilization of each cue influences this relationship.

Lastly, to examine whether the delayed diagram completion task leads to more accurate student judgments in comparison to omission of this task (RQ 3), a one-way between-groups ANOVA is performed. Again an ANOVA was chosen to avoid the problem of an inflated experimentwise alpha level (Gravetter & Walnau, 2013). The dependent variable of this analysis

is students' judgment accuracy and the groups are formed by the two conditions (delayed-diagram condition versus no-diagram condition).

All analyses are performed with IBM SPSS Statistics, version 24. Due to aggregating problems, data is missing at random.

Results

Below, the results of the performed analyses for each research question are presented. Effects are reported as significant at $p < .05$.

Differences in Judgment Accuracy (RQ 1)

The mean judgment accuracy of the students was .16 ($N = 1187$, $SD = .66$), and the median was .11, whereas the mean judgment accuracy of the teachers was .25 ($N = 1254$, $SD = .65$), with a median of .33. The statistical significance of these differences was tested with a one-way ANOVA. Both the assumption of normality and the assumption of homogeneity of variance were violated. However, since the ANOVA is quite robust for these violations (Glass, Peckham, & Sanders, 1972; Schmider et al., 2010), the analysis was continued. The analysis showed that there were significant differences between students and teachers regarding judgment accuracy, $F(1, 2439) = 10.90$, $p = .001$, $\eta^2 = .004$, indicating that teachers judge the level of comprehension of students more accurately than students themselves.

Differences in Cue-Utilization (Part One RQ 2)

Table 2 displays descriptive statistics of both students and teachers regarding cue-utilization. The table presents the means of the intra-individual gamma correlations for each response type (i.e., cue), along with the standard deviations and the median correlations.

Table 2

Descriptive Statistics of Cue-Utilization for both Students and Teachers

	Students			Teachers		
Cues	Mean[a]	*SD*	Median	Mean[a]	*SD*	Median
Correct relations	.20	.57	.20	.39	.63	.56
Commission errors	-.07	.61	.00	-.14	.61	-.13
Omissions	-.22	.64	-.33	-.28	.65	-.33

[a] All correlations differ significantly from zero, $p < .01$.

A MANOVA was conducted to statistically test these differences between cue-utilization of students ($N = 516$) and teachers ($N = 540$) for the three cues. The assumptions of normality, linearity and homogeneity of variance-covariance matrices were not met, yet the analysis was continued since robustness of the MANOVA for these violations has been demonstrated (Olson,

1974; Kariya, 1981). Findings showed that there were significant differences between students and teachers for cue-utilization of the three cues combined, F (3, 1052) = 9.87, p = <.001, partial η^2 = .027. Follow-up analyses conducted to examine the differences between students and teachers for each cue are described below.

Correct relations. Significant differences were found between students and teachers for cue-utilization of correct relations, F (1, 1054) = 27.84, p = < .001, partial η^2 = .026. This indicates that teachers made significant more use of the cue correct relations than students (see also Table 2).

Commission errors. The cue-utilization of the cue commission errors was also significantly different between students and teachers, F (1, 1054) = 4.91, p = .027, partial η^2 = .005. These results indicate that teachers made significant less use of the cue commission errors in comparison to students (see also Table 2).

Omissions. Both students and teachers made little use of omissions as a cue, as can be inferred from the negative values for cue-utilization in Table 2. The analysis showed no significant differences between the levels of cue-utilization of students and teachers for this cue, F (1, 1054) = 2.09, p = .149.

The Relation Between Judgment Accuracy and Cue-Utilization (Part Two RQ 2)

To examine whether, and how, the relation between judgment accuracy and person (students or teachers) was influenced by cue-utilization, moderation-analyses were performed. To test each cue-utilization variable for moderation, three multiple regression analyses were conducted. All required assumptions were met for each analysis. Only information about the interaction effects is provided, since those results are most relevant to the research question of the current study.

Correct relations. The interaction effect of person and cue-utilization of correct relations was not significant, B = .007 [-.031, .045], t(1070) = .37, p = .713. Therefore, the use of this cue did not influence the relationship between person and judgment accuracy.

Commission errors. The interaction effect of person and cue-utilization of commission errors was significant, B = .045 [.007, .083], t(1070) = 2.34, p = .020. This indicates that cue-utilization of commission errors moderates the relationship between person and judgment accuracy, meaning that the effect of person on judgment accuracy is dependent on the extent of the use of commission errors as a cue. In combination, the three predictor variables explained about 2% of the variance in judgment accuracy, R^2 = .017, adjusted R^2 = .014. A visual representation of the interaction effect is shown in Figure 2. For students, the high use of

commission errors as a cue results in more accurate judgments, compared to teachers. On the other hand, low use of this cue results for students in lower judgment accuracy than for teachers. In this way, the extent of cue-utilization of commission errors influences the relationship between person and judgment accuracy.

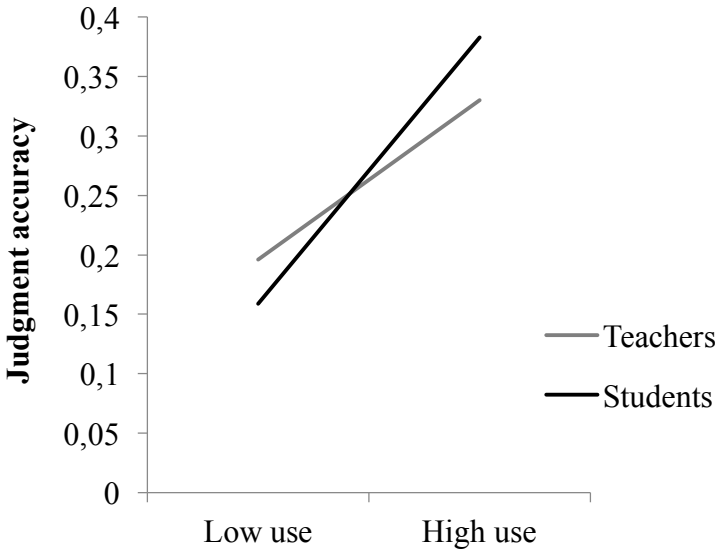

Figure 2. A visual representation of the moderating influence of the use of commission errors on the relationship between person (teachers or students) and judgment accuracy.

Omissions. No moderation analysis was conducted with the use of omissions as a cue, since no significant differences were found between students and teachers in cue-utilization of this cue. Therefore, it makes no theoretical sense to conduct a moderation analysis, which examines whether differences in judgment accuracy of students and teachers are dependent on differences in cue-utilization influence.

Effects of the diagram-completion task (RQ 3)

The effect of the diagram-completion task on the judgment-accuracy is examined with a one-way ANOVA. To this end, two conditions were specified in this study: delayed-diagram condition and no-diagram condition. Students were however very disproportionately distributed over these conditions, which could result in distortion of the results of the ANOVA. Therefore, a random sample was compiled and used for this analysis, with a more equal number of students in the delayed-diagram condition (N = 130, M = .19, SD = .67) and the no-diagram condition (N = 153, M = .20, SD = .67). Note that in this random sample the means of both groups are higher than in the overall sample of students, as used in research question 1 (N = 1187, M = *.16, SD* = .66). Assumptions of normality and homogeneity of variance were met. The ANOVA

demonstrated no significant differences between the two groups, F (1, 281) = .051, p = .822, indicating that the delayed diagram completion task does not lead to more accurate student judgments in comparison to omission of this task.

Discussion

The main goal of this study was to compare students' and teachers' judgment accuracy and their corresponding use of cues. The main research questions were: 1. To what extent are there differences between students and teachers regarding judgment accuracy? 2. Can potential differences be explained by differences in cue-utilization? Because of the exploratory perspective, no hypotheses were formulated for these questions.

Different judgment accuracy levels: implications and a possible explanation

A key finding is that, overall, teachers make more accurate judgments of students' comprehension than students themselves. Since this study was the first to make this comparison, this finding cannot directly be related as consistent or inconsistent with previous studies. These results might have important practical implications for contemporary teaching practices. As described, shifts in education require increasingly more advanced metacognitive skills of students, of which the ability to make accurate judgments is an important part (Delfino & Persico, 2009; Thomas & Brown, 2011; Nelson & Narens, 1990; Schneider, 2008). Important tasks of teachers are transferred to the responsibility of students. However, since the judgment accuracy of students was found to be lower, maybe students are not (yet) ready for this increased responsibility. Therefore, based on the results of this study, the desirability of current developments in educational practices can be questioned.

However, the remark must be made that only small effects were found and that sometimes multiple assumptions required for the statistical analyses could not be met. Therefore, we must interpret the results of this study with care. Nevertheless, it raises important questions.

The age of the students involved might be an important variable at play in explaining these results. The target group of this study were students from around 15 years old. However, children might become more accurate in their academic self-perceptions when they approach the adolescence (Harter, 1985). In this stage of life, they develop more understanding of academic tasks (Brown & Smiley, 1977). As a result, their monitoring of the differences in effectiveness of their cognitive strategies for learning grows (Pressley, Joel, & Ghatala, 1984). It would be interesting to explore whether the results found in this study are also found when an older target group is used. Differences between students and teachers regarding judgment accuracy might in that case be smaller. Therefore, future research comparing the judgment accuracy of students and teachers might focus on older students.

Differences partly explained by cue-utilization

Teachers made significantly more use of the cue correct relations, and at the same time significantly less use of the cue commission errors than students did. These differences in cue-utilization could partly explain the observed differences in judgment accuracy, by means of the conducted moderation analyses. More specifically, cue-utilization of commission errors has been found to moderate the relationship between judgment accuracy and person. In this way, the lower cue-utilization of commission errors by teachers was linked to their higher judgment accuracy.

However, this moderating role was not found for the use of correct relations as a cue. As a result, part of the differences between students' and teachers' judgment accuracy could not be explained by the cues taken into account in this study. An important limitation of this study is therefore the restricted number of cues taken into account. Students and teachers probably use several more cues besides the number of correct relations, commission errors and omission. When more cues would be taken into account, possibly more differences between students and teachers regarding judgment accuracy could be explained. Therefore, studies examining which cues students and teachers actually use in the process of making judgments are of value and should be carried out.

Unexplained variance in both students' and teachers' judgment accuracy

Another notable finding is the variation within both groups, students as well as teachers, with regard to judgment accuracy. In their meta-analysis, Südkamp et al. (2012) also found high levels of variance in teachers' judgment accuracy, suggesting that teachers vary widely in their judgment accuracy. This suggestion is confirmed by this study. In addition, also high levels of variance in students' judgment accuracy are found. Südkamp et al. (2012) could not explain the high levels of variance. The variances were not addressed in this study because of the between-groups design focusing on the differences between, and not within, groups of students and teachers. As a results, the problem of unexplained variance remains unsolved and requires further research. However, Südkamp et al. (2012) advocated for more studies assessing how teacher characteristics relate to teacher judgment accuracy, suggesting these characteristics might be an explanatory factor. In this study, the average years of teaching experience of the teachers had a range of more than 32 years and the average age of the teachers had a range of 34 years. These wide variations in characteristics may indeed be related to the variations in judgment accuracy. It is important to know where these variances come from, in order to design adequate interventions that help improving judgment accuracy. Future research should therefore focus on explanations of these differences with possibly more qualitative research designs to figure out these sources differences. For example, in-depth interviews could be valuable in which teachers explain how

they come to their judgments. Also, observations can be used to study the natural setting in which teachers make everyday judgments, i.e. the classroom, which can yield valuable insights in this matter.

Effect of the delayed diagram completion task

The secondary goal of this study was to verify the results of Van Loon et al. (2014) concerning the effects of delayed-diagram completion on judgment accuracy. Based on their findings, judgment accuracy was hypothesized to be higher when students completed the delayed diagram completion task, relative to when they did not. The results of this study, however, do not support this hypothesis, since no significant differences were found between the judgment accuracy of the students in the delayed-diagram condition and the no-diagram condition.

Since the same materials were used as in the study of Van Loon et al. (2014), this cannot be an explanatory factor. However, differences in results can possibly be caused by differences in explanation and guidance given to students during the data collection. This explanation and guidance could have influenced the way students performed at the experimental tasks. Also external circumstances might have played a role, such as the timing of the experiment in the week and year schedule in combination with the motivation of the students for completing the tasks. Again the age of the students might also be an important variable in explaining this incongruence. Van Loon et al. (2014) included students of the age between 14 and 16. However, in the current study also younger students participated (age 11, 12 and 13). The distribution of these younger students over the two groups might play a role in explaining the results.

These results are not only inconsistent with the specific findings of Van Loon et al. (2014), but also with broader studies suggesting that the delayed timing of a task can lead to more accurate judgments due to the level of mental representation involved (e.g., Thiede, Dunlosky, Griffin, & Wiley, 2005; Kintsch, 1998). This emphasizes the importance of replication studies once again (John, Loewenstein, & Prelec, 2012; Lakens, Haans, & Koole, 2012), since it indicates that results of studies should not too easily be considered as truth.

Importance of improving judgments

A last notable finding of this study is that the accuracy of both students' and teachers' judgment is quite low. In line with De Bruin et al., 2011, Dunlosky & Lipko, 2007, Südkamp et al. (2012), and Van Loon et al. (2014), I can state that both teachers' and students' judgment accuracy are far from perfect and that there is plenty of room, and need, for improvement. As described, judgments of comprehension have important consequences in the educational practice and therefore low levels of both students' and teachers' judgment accuracy can have a detrimental effect on student learning and development (Metcalfe & Finn, 2008; Südkamp,

Kaiser, & Möller, 2012; Thiede, 2015; Thiede, Anderson, & Therriault, 2003). Research has shown that relative judgment accuracy can be improved by means of practicing with monitoring and making judgments (Hacker, Dunlosky, & Graesser, 2009; Vesonder & Vos, 1985). In addition, Nelson and Dunlosky (1991), have found a correlation as high as .93. This shows there is no need for pessimism, although there is work to be done. Given the importance of accurate judgments of both teachers and students for the learning processes and academic careers of students, future studies should focus on more interventions that could be used in the educational practice to help both actors improving their judgment accuracy. This is especially of great importance considering the advanced knowledge society students nowadays grow up in.

References

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology*, *91*(4), 731-746. doi:10.1037/0022-0663.91.4.731

Bailey, A. L., & Drummond, K. V. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment*, *11*(3), 149-178. doi:10.1207/s15326977ea1103&4_2

Brown, A. L., & Smiley, S. S. (1977). Rating the importance of structural units of prose passages: A problem of metacognitive development. *Child Development, 48*(1), 1-8. doi:10.2307/1128873

Bruin, A. B. de, Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology*, *109*(3), 294-310. doi:10.1016/j.jecp.2011.02.005

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. Berkeley: University of California Press.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46. doi:10.1177/001316446002000104

Delfino, M., & Persico, D. (2009). Self-regulated learning: issues and challenges for initial teacher training. In W. H. L. Tan & R. Subramaniam. *Handbook of research on new media literacy at the K-12 level: Issues and challenges (pp.* 839-845). Hershey: IGI Global.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*(4), 228-232. doi:10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271-280. doi:10.1016/j.learninstruc.2011.08.003

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. doi:10.3102/00346543042003237

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371-395. doi:10.1037/0033-295x.101.3.371

Gravetter, F. J., & Wallnau, L. B. (2013). *Statistics for the Behavioral Sciences (9th ed.)*. Belmont, CA: Wadsworth Cengage Learning.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*(1), 93-103. doi:10.3758/mc.36.1.93

Hacker, J. D., Dunlosky, J., & Graesser, A. C. (2009). Handbook of Metacognition in Education. New York. NY: Routledge.

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. doi: 10.1177/0956797611430953

Kariya, T. (1981). Robustness of Multivariate Tests. *The Annals of Statistics, 9*(6), 1267-1275. doi:10.1214/aos/1176345643

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349-370. doi:10.1037/0096-3445.126.4.349

Lakens, D., Haans, A, & Koole, S. (2012). Eén onderzoek is geen onderzoek. *De Psycholoog, 9,* 10-18.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic**.** *Biochemia Medica, 22*(3), 276 82. doi:10.11613/bm.2012.031

Metcalfe, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1084-1097. doi:10.1037/a0012580

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research*, *79*(3), 1129-1167. doi:10.3102/0034654309337522

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109-133. doi:10.1037/0033-2909.95.1.109

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation.* New York, USA: Academic Press.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association, 69*(348), 894-908. doi:10.1080/01621459.1974.10480224

Pressley, M., Joel, R. L., & Ghatala, E. S. (1984). Memory strategy monitoring in adults and children. *Journal of Verbal Learning and Verbal Behavior, 23*(2), 270-288. doi:10.1016/S0022-5371(84)90181-6

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, *28*(6), 1004-1010. doi:10.3758/bf03209348

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities the role of child background and classroom context. *American Educational Research Journal*, *48*(2), 335-360. doi:10.3102/0002831210374874

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology, 6*(4)*,* 147-151. doi:10.1027/1614-2241/a000016

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education, 2*(3), 114-121. doi:10.1111/j.1751-228x.2008.00041.x

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. doi:10.1037/a0027627.supp

Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, *28*(2), 129-160. doi:10.1016/s0361-476x(02)00011-5

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66-73. doi:10.1037/0022-0663.95.1.66

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., & Jesse, D. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education*, *49*, 36-44. doi:10.1016/j.tate.2015.01.012

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1267-1280. doi:10.1037/0278-7393.31.6.1267

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331-362.

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring

during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Greasser (Eds.), *Handbook of metacognition in education* (pp. 85-106). Mahwah, NJ: Erlbaum.

Thomas, D., & Brown, J. S. (2011). *A new culture of learning: Cultivating the imagination for a world of constant change.* Lexington, KY: CreateSpace.

Van Loon, M. H., Bruin, A. B. de, Gog, T. van, Merriënboer, J. J. van, & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143-154. doi:10.1016/j.actpsy.2014.06.007

Appendix A

Text “The Suez Canal”

The Suez Canal, which connects the Indian Ocean and the Mediterranean Sea with each other, is of great importance to the world. Originally, there was no natural water connection between the Atlantic and the Indian Ocean. Between these two seas is a desert. This meant that trading ships that travelled from the harbour city Jeddah in Saudi Arabia to Europe had to make a long journey around the whole African continent. It was therefore decided that a shorter waterway was needed that would connect the two oceans with each other. For this reason, the Suez Canal, which was designed by the Austrian engineer Alois Negrelli, was dug. For years, workers were digging; the canal was finally opened in 1869 for shipping. By the digging of the Suez Canal, the distance from the harbour city of Jeddah to the harbour city of Rotterdam has been reduced by 40%. Through the Suez Canal, the distance between these cities is 6,337 nautical miles, when ships sail around the African continent this distance is 10,743 nautical miles.