Task demands and test expectations. Theory and empirical research on students' preparation for a teacher-made test

Broekkamp, H.

Citation for published version (APA):
CHAPTER 3

WILL THAT BE ON THE TEST?

Perceived task demands and test performance in a classroom context

Abstract
Examing a realistic classroom context, this study focuses on the relationship between students’ perception of task demands and their learning performance. History teachers (N = 20) and their 11th-grade students (N = 375) rated the relative importance of sections of an instructional text on which teachers would be giving a test. In addition, test grades were collected. It was hypothesized that the degree of correspondence between ratings of individual students and those of their teacher would reflect the accurateness of students’ perception of task demands and therefore would be positively related with test grade. Multilevel analysis revealed partly support for this hypothesis. Students with higher grades more consistently showed a relative close correspondence with their teacher than students with lower grades. Nonetheless, a substantial number of students who showed a relatively close correspondence with their teacher, obtained a low test grade. This suggests that a more accurate perception of task demands did not guarantee a higher test performance.

1 INTRODUCTION

"Will that be on the test?" Without doubt this is one of the most frequently asked questions by students in secondary education. Even so, teachers will rarely answer it directly, because they do not want to give away the test questions. Students know this and still try to obtain the most specific information possible about the forthcoming test (Doyle, 1983; Miller & Parlett, 1974; Ramsden, 1988; Van Eten, Freebern, & Pressley, 1997). This “game” is sometimes seen as an artifact in education, which has more to do with students’ test-orientation than with learning. However, considering contemporary theories of learning, the outcome of this game has important consequences. Clear knowledge about task demands is an important prerequisite for students to learn in a strategic, self regulated and effective way (Brown, Bransford, Ferrara, & Campione, 1983; Butler & Winne, 1995; Nist & Simpson, 2000; Thomas & Rohwer, 1986). With respect to text studying, a clear perception of task demands allows students to focus their attention on task-relevant text elements and to choose appropriate strategies to learn these elements (R.C. Anderson, 1982; Anderson & Armbruster, 1984; Hidi, 1995; Marton & Säljö, 1976; Reynolds, 1992).
As is stressed by several researchers, the *task relevance* or the *instructional importance* of text elements does not necessarily coincide with the *structural importance* of those elements (Alexander & Jetton, 1996; Goldman, 1997; Schraw, Wade & Kardash, 1993; Van Dijk, 1979; Van Hout-Wolters, 1986). For example, teachers may assign higher importance to text elements that describe specific content and have low structural importance than to elements that provide global information and are closely related to the gist. Similarly, teachers may differ with regard to the importance they assign to text elements, even when these elements have similar structural importance. Such differences between teachers have important instructional consequences. In classrooms, teachers generally determine the task demands and also test (mostly a selection) of these demands (Doyle, 1992). Differences between teachers imply that students have to tune in to the particular task demands that their own teacher sets (Alexander & Jetton, 1996). Moreover, it seems that students can increase their test performance by knowing more accurately which text elements their teacher deems instructionally important.

Several studies have examined the relative importance that students and teachers assign to text units when a test has been set (Alexander, Jetton, Kulikowich & Woehler, 1994; Broekkamp, Van Hout-Wolters, Rijlaarsdam, & Van den Bergh, 2002; Jetton & Alexander, 1997; Schellings & Van Hout-Wolters, 1994, 1995; Van Hout-Wolters, 1990a, 1990b/1997). In all of these studies, only a limited correspondence was found between teachers, between students, and/or between teachers and their students with respect to their importance assignment. Moreover, these findings were obtained for different text levels (sentence, paragraph, section), different measures (rating, question writing, selection, notetaking), different subjects (science, biology, history) and different correspondence measures including multilevel estimates (Broekkamp, Van Hout-Wolters, Rijlaarsdam & Van den Bergh, 2002). These studies underscore that teachers differ in the task demands they define for their students and that students differ in their perception of these demands. However, as these studies did not include measures for test performance in their analyses, they cannot answer the question whether a more accurate perception of task demands indeed helps the student to obtain a higher test performance. To learn more about the relation between students’ perception of task demands and their test performance, we have to rely on experimental test expectancy research.

1.1 *Experimental research on the effect of test expectancy*

There is a long tradition of experimental research that examines the effect of expected testing demands on learning performance (for reviews, see R.C. Anderson & Biddle, 1975; Anderson & Armbruster, 1984; Crooks, 1988; Hamaker, 1986; McConkie, 1977; Reynolds, 1992). This research consistently shows that text elements that are made important by task manipulations (e.g., inserted questions, and objectives) are better learned than “non-targeted” text elements. With regard to the non-targeted text elements, on the contrary, effects are inconsistent. “In various studies acquisition of the non-targeted information by students have been facilitated, inhibited, or unaffected” (McConkie, 1977, p. 24).
In interpreting the heterogeneous results, at least three factors should be taken into account. The first factor is the clarity of task demands. “It is likely that if the targeted information is easily identifiable, and the subjects are convinced that other information is irrelevant to the task, they will be selective in their learning. Increased retention of the targeted information will then be at the expense of the other information presented in the text” (McConkie, 1977, p. 24). Thus, when task demands are unclear or do not distinguish between important and unimportant information, students are discouraged to process the text in a selective way. Note that the clarity of task demands depends on the available task information as well as the way the student makes use of this information.

The second factor is the study time that students have available. In his meta-analysis on the effects of inserted questions, Hamaker (1986), similar to McConkie (1977), found heterogeneous results concerning the learning of non-targeted information. However, Hamaker was able to identify “time control” as an explaining variable. When study time was controlled, the task manipulation led to a decrease in the learning of non-targeted information. On the contrary, when study time was unrestricted, a small positive effect was found. It seems, then, that when time pressure increases, students are more selective in their learning.

The third factor explaining the heterogeneous results concerns the way students enact upon their perception of task demands. As Anderson and Armbruster (1984) stressed “knowledge of the criterion task may be a necessary condition for optimal studying, but is obviously not a sufficient condition” (p. 660). They pointed out that students also have to focus their attention on the task relevant information and engage in strategies, which ensure that this information is well understood, and likely to be remembered (p. 660). Plainly, the effect of task perception on learning is dependent on the way these activities are carried out.

The importance of the first and third factor is underscored by several studies of Reynolds and associates on text studying (Reynolds, 1992). In these studies, content related task demands were implicitly presented by means of preparatory questions inserted in the text. Students’ “awareness of task demands” was examined by administering retrospective interviews. In addition, the degree of selective processing was examined by measuring the proportion of attention for targeted and non-targeted text. Regression analyses showed positive relationships between reading ability, the level of task awareness, degree of selective processing and test performance for both college students and tenth-grade students. However, whereas selective processing and test performance were causally related for college students, this was not the case for tenth-grade students. The researchers concluded that only the college students were able to use the selective attention strategy in an effective way.

1.2 Test expectations and test performance in classroom contexts

When students study in preparation for a test, classroom contexts are likely to differ from experimental contexts in a great number of factors, including preparation time, students’ motivation, nature of task demands, availability of task information, and authority defining task demands. In general, naturalistic classroom contexts pose...
additional complexities for students (cf. Hacker, Bol, Horgan, & Rakow, 2000). Therefore, in classroom contexts, students’ perception of task demands may be less clear than in laboratory settings. An important question is whether task demands in classroom contexts are clear enough for students to use the selective attention strategy in an effective way.

In classroom contexts, study time is rarely restricted in a direct way like in many of the test expectancy experiments. However, the time that students have at their disposal will certainly be constrained by factors like study loads, fixed testing moments, competing goals, and flaws in students’ planning (Van Etten et al., 1997). Depending on the nature of these constraints, time pressure will be higher or lower. Given that time pressure is low and expected test demands are not fully clear, studying will be most effective if students use a nonselective attention strategy and study intensively all information if it has any chance to be tested. On the contrary, when time pressure increases, a more selective strategy is called for. Under these conditions, an accurate perception of the task demands becomes more important to learning performance.

The experimental studies of Reynolds and associates (Reynolds, 1992) suggest that secondary students do not effectively use the selective attention strategy. However, classroom conditions might be more favorable for using the selective attention strategy than the conditions of the experimental studies were. Moreover, even when this is not the case, there is still a good reason to assume a positive relationship between students’ perception of task demands and their learning performance. Several researchers have argued that understanding task demands involves metacognitive skills, which distinguish strategic learners from their less strategic counterparts (e.g., Reynolds, 1992; Winne & Hadwin, 1998). More specifically, these skills include paying attention to task cues and analyzing and interpreting these cues (Butler, 1999). Even when more strategic students eventually do not selectively attend to task relevant text elements, other attributes than a greater task awareness will help them to learn more effectively (e.g., more developed skills with regard to time planning, regulating motivation or decoding textual information). These associated attributes seem to explain why Reynolds and colleagues found a relationship for tenth-graders between reading ability, task awareness and test performance, despite the fact that students did not effectively used the selective attention strategy.

1.3 Perceptions as indicators for task demands

So far, task demands in classroom contexts were implicitly equated with teachers’ task demands, i.e., task demands that are intended/perceived by teachers. Task demands, however, can also be considered from students’ perspective.

From the perspective of the student, task demands foremost coincide with their test expectations. As is stressed by Doyle (1983), students only will take those demands seriously for which they are held accountable, i.e., demands that have a reasonable probability of being tested. Since, within classrooms, teachers generally know best which information is likely to be tested, they are naturally a crucial source of information for students to base their test expectations upon. However,
information that teachers give about task demands may be indirect, incomplete, unclear and even misleading. For these reasons, students may prefer to rely on other sources that provide information about the task demands, such as the learning materials or previous tests (see Doyle, 1983; Thomas & Rohwer, 1986; Van Etten et al., 1997).

Taking into account the student perspective, the teacher’s task demands constitute a valid indicator of the factual task demands only when teacher’s task demands are sufficiently accessible to students and when students are convinced that these task demands correspond with the upcoming test. If this is not the case, the perceptions of students themselves could provide a better indicator of the factual task demands than the teacher’s perceptions of task demands. Obviously, as students’ perceptions vary across individuals, they cannot directly be taken as an objective measure. However, when students’ perceptions are aggregated, they can be considered as representing the consensus of students within a class. As such they could provide a useful measure for the task demands in a particular classroom context. (For a similar approach, see Ryan, Gheen & Midgley, 1998; these researchers used student aggregates as measures for the motivational goal structures within classroom contexts.)

In previous studies on instructional importance, only a limited correspondence was found between individual students and their teacher regarding their perception of the task demands (e.g., Broekkamp, Van Hout-Wolters, Rijlaarsdam & Van den Bergh, 2002; Jetton & Alexander, 1997). Given the discussion above, this finding evokes questions about the validity of the teachers’ perceptions as an indication of the factual task demands. It is possible, then, that those task demands that really mattered in the classrooms were disparate from those perceived/intended by the teachers. Moreover, in order to obtain a high test performance it could have been more important for students to show a close correspondence with the “class perception” than with the teacher’s perception.

1.4 The present study

In this study, students’ and teachers’ perceptions of task demands were indicated by the instructional importance participants assigned to text elements of a chapter that would be tested in the short term. We built on a study on instructional importance of Broekkamp, Van Hout-Wolters, Rijlaarsdam, and Van den Bergh (2002). However, whereas this previous study focused on the degree of correspondence among participants’ task perceptions, the present study focuses on the relationship between task perceptions and test performance. This relationship was examined in three different ways.

First of all, we examined whether there is a direct relation between perceived instructional importance and test performance. More specifically, we examined whether students who, on average, assign more importance to text elements also have a higher test performance. A positive relationship can be expected when students have sufficient time available and allocate their attention in a nonselective way. Given these conditions, students who, on average, perceive the text as more
important, can be expected to spend more overall attention to the text and also learn more from it.

Secondly, we examined whether the correspondence in perceived instructional importance between individual students and their teacher is related with test performance. A positive relationship can be expected for both causal and noncausal reasons. A causal reason is that when students are under time pressure and are able to use the selective attention strategy in an effective way, a more accurate perception of the task demands will increase the learning of task-relevant information and hence will yield a better test performance. A noncausal reason is that abilities concerning the identification of task demands are related to other characteristics that may contribute to effective learning and test performance. Note that both reasons assume that an accurate perception of task demands is indicated by the correspondence a student’s perception shows with the teacher’s perception.

Thirdly, we examined whether the correspondence in perceived instructional importance between individual students and their class is related with test performance. A positive relationship can be expected for similar causal and noncausal reasons as those pertaining to the student-teacher correspondence. Now, however, an accurate perception of task demands is indicated by the correspondence a student’s perception shows with the average perception across students within a class.

As we used correlational methods and did not collect data on the instructional-learning process, it was beforehand clear that we could not demonstrate a causal relationship between the quality of students’ perception of task demands and their test performance. However, we could examine an important condition for inferring causality, the existence of a relationship. Hereto we tested three – partly conflicting – hypotheses:
1) The average importance that individual students assign across text elements is positively related to their test performance.
2) The correspondence in perceived importance between individual students and their teacher is positively related to test performance.
3) The correspondence in perceived importance between individual students and their class is positively related to test performance.

2 METHOD

2.1 Participants

Participants in this study were 20 history teachers and 375 11th-grade students. Teachers, who were from 16 schools in the Netherlands, each participated with one class of students. The average number of students in a class was 20.45 ($SD = 6.73$). Teachers, on average, had 18.1 years of experience ($SD = 7.87$) and 14 of them were

---

1 Initially, 22 teachers and 451 students participated in the study. As two teachers did not provide data on students’ test performance, they were excluded from the study, together with their classes (42 students). Also, of the remaining 20 classes, 34 students were excluded because they were not present at the time that the test was made by the majority of the classroom.

2 Of the 16 schools, 13 schools participated with one teacher and 3 schools with two teachers. Sample size did not allow us to examine school effects in a reliable way.
male (70%). Students had a typical age of 17 and were in the penultimate year of pre-university education. 214 of the students (57%) were female.

In order to recruit participants, we first approached schools all over the country that trained students of the level mentioned, and also used a particular history textbook, which had the highest market share. From a first orientation it appeared that teachers widely differed with regard to the chapters they selected for instruction and the sequence in which they taught them. In addition, teachers differed in the number of lessons spent on chapters and in the way they constructed tests – as no standard chapter tests were provided by the textbook, the vast majority of teachers made or assembled the test questions themselves. As we intended to study a realistic classroom context, we did not persuade teachers to adapt their curriculum. Instead, we chose a chapter most frequently taught at grade 11 and recruited teachers who tested their 11th-grade students on this chapter in the second half of the school year. At this point, students would be more used to their teachers and would be more likely to attune to task demands effectively (Doyle, 1983; Van Etten et al., 1997).

2.2 Materials and procedure

The chapter had “the United States of America since 1945” as its theme and contained approximately 8,000 words. It was structured in 16 sections of differing length that described the presidential terms from Roosevelt to Clinton. These descriptions were chronological, but not proportional; the text on certain presidents took up two or three sections, while other presidents were described within a single section.

The chapter’s section structure formed the basis of a rating task that was performed by the teachers and the students. To obtain text units of similar length, longer sections were subdivided (by reducing differences in text length ratings) in order to more purely reflect differences in instructional importance, instead of differences in structural importance; see Alexander & Jetton, 1996.) This revised division contained 26 “text sections”, with each section encompassing a sequence of three or four coherent paragraphs.

To help students to consider the text at a global level and to compare topics of broad units in a balanced way, sections were presented to the participants by means of a short title and a brief summary, for example:

“Truman, Roosevelt’s successor
About Truman’s internal policy and the resistance that it met with. About his unexpected victory in the elections and the subsequent reaction of the Republicans.”

Titles were taken from the book verbatim, or formulated in a similar style.

The rating task was performed within two days before the actual classroom test during a history lesson (15 classes) or during a separate session (5 classes). Students and their teacher, at the same time but independently, rated the relative importance of the 26 sections. Respondents indicated for each section whether it was important to the upcoming test, and was thus likely to be tested. It was explained to teachers that their ratings should express the degree to which a section was important for their students preparing for the test, taking into account that the student did not
know the test questions beforehand. Importance was rated on a four-point scale (unimportant, less important, important, very important) (cf. Brown & Smiley, 1977).

During the rating task, students and teachers were asked not to refer to the textbook or their notes. With this constraint we tried to ensure a common ground for text ratings and to avoid that some students would focus on additional test preparation, as was the case in a pilot study we conducted. Moreover, we could more easily present our revised division with text units of similar length.

We assumed that section titles and explanations would generally be familiar to the students. As a check, however, students could indicate if they were unfamiliar with a section. In the event, this option was used in 2% of the ratings. In addition, for 1% of the total number of student observations, no rating was given. For teachers, this percentage was also 1%. In the analysis, both kinds of data points were considered as missing values.

After teachers scored the test, they informed the experimenter about the grades students received. In the Dutch educational system, grades typically range from 1 (worst) to 10 (best). As we examined a realistic context, no attempt was made to influence the grading procedure.

2.3 Analysis

The three hypotheses that were tested in this study all assume a relation between the quality of students’ task perception and their test performance, but differ with regard to the measure for this quality. For the first hypothesis the quality of a student’s task perception is operationalized as the value of a student’s importance rating; for the second hypothesis as the correspondence – or complementary, as the difference – between a student’s rating and the rating of his/her teacher; and for the third hypothesis as the correspondence – or the difference – between the rating of the student and the average rating of his/her class. The fact that the three hypotheses differ with regard to only one variable allowed us to test them with three variants of the same basic multilevel model. Multilevel modeling was chosen for, because this method would take into account the nested structure of our data (cf. Goldstein, 1995; Holt, Scott, & Hewings, 1980).

Although a multivariate model might, in principle, be preferable over three univariate models, we explicitly chose for the latter because of simplicity of the models and hence interpretation. In a multivariate model, restrictions have to be imposed on the covariance matrices, depending on the levels and operationalizations. For instance, “student rating – class rating” is the only operationalization that does not vary between classes. Similarly, the correlation between “student rating – class rating” and “student rating” equals unity and both have equal variance at the student level; this is because the rating of students is the same for both measures. In a multivariate model a diversity of such restrictions have to be imposed on the parameter estimates, each making the interpretation of the model less transparent. For the sake of completeness, the correlation between “student rating” and “student rating – teacher rating” equals .31 at the class level, unity at the student level and .69 at the section level. So, the operationalizations behave differently at the three levels.
2.4 The basic multilevel model

The basic model takes into account that the quality of students’ perceptions may differ between classes, students and sections. So, the parameter estimates of this model include the mean quality score, the intercept, as well as the variance in the intercept between classes, students and sections. In addition to these four parameters, the basic model includes six parameters that describe the relation between the quality measure and test grade. In the first place, a regression coefficient is estimated, indicating the fixed effect of test grade on the quality measure. Furthermore, two variance components are estimated that allow the effect of test grade to vary between classes and between sections. In addition, the model includes two parameters that describe the covariation between differences in regression coefficients and differences in intercepts. At the class level, a positive covariance coefficient would indicate that in classes that higher quality score, there is a stronger relationship with test grade than in classes with a lower quality score. At the section level, a positive covariance value would indicate that sections with a higher quality score show a stronger relationship with test grade than sections with a lower quality score. Finally, the model includes a covariance parameter at the student level, which indicates the relationship between differences in grades and differences in the average quality score across ratings. A positive covariance value would indicate that students with a higher grade show more variation in their average quality score across ratings than students with a lower grade.

For the first quality measure, analysis with this ten parameter-model answers our questions. For the second and third operationalization there still remains a (technical) problem to be tackled, as the correspondence between student and teacher ratings, as well as the correspondence between student and class ratings, may come about in different ways. For instance, students who follow their teacher’s rating pattern closely have a low deviation from their teacher across sections, indicating an accurate perception of task demands. However, this may hold as well for students who give all sections a same rating that approximates the teacher’s average rating across sections. Therefore, we have to add a second explanatory variable to take into account differences between a student’s section ratings, e.g., the within student standard deviation.

---

4 By considering the regression of test grade on the quality measure, we could determine the relationship between these two variables for the 26 sections simultaneously. Data limitations did not allow us to use a multivariate model with 26 instances of the quality measure predicting test grade.

5 A positive covariance coefficient could, in principle, also be interpreted as indicating that in classes that on average have higher grades, there is a stronger relationship between grades and quality scores than in classes with lower grades. This interpretation, however, is not convincing as teachers typically gave grades on a norm referenced base. Hence, differences in grades between classes could not be interpreted in a meaningful way.

6 The covariances at the class and section level could also be interpreted as the relationships between differences in grades and differences in the quality scores at the class and section level, respectively. As our model allows us to estimate the variance between sections and classes at each point of the grading scale, we don’t have to make the assumption of homoscedasticity. The conditional covariance of the quality scores can be written as: \[ \text{VAR} \{ \text{Quality Score} | \text{Grade} \} = \text{VAR} \{ \text{Quality Score} \}^2 + 2 \times \text{COV} \{ \text{Quality Score} \times \text{Grade} \} + \text{VAR} \{ \text{Grade} \}^2 \] (cf. Goldstein, 1995).
CHAPTER 3

2.5 Expectations pertaining to the first quality measure (student rating)

Based on hypothesis 1, we formulated two expectations. First of all, a positive effect of test grade on student ratings was expected, whereby a high test grade would be associated with a high importance rating. Secondly, at the student level, a negative covariance was expected. That is, students with higher test grades would more consistently have a high importance rating than students with lower test grades. As a consequence, students with higher test grades would show a lower variation in their average rating across sections.

2.6 Expectations pertaining to the second quality measure (student rating – teacher rating) and the third quality measure (student rating – class rating)

As expectations of the second and the third quality measure were similar, we only discuss those of the second quality measure. Departing from hypothesis 2, a negative effect of test grade on the quality measure was expected, whereby a higher test grade would be associated with a smaller deviation from the teacher’s rating. Furthermore, we expected that students with higher test grades would more consistently have a relative accurate perception of task demands than students with lower test scores. As a consequence, the covariance at the student level would have a negative value.

The variance and co-variance components at the different hierarchical levels were estimated with the software ML-wiN (Multi Level Models Project, 1999). Analyses were performed on normalized scores. Statistical significance of the parameter estimates can easily be determined as the ratio of the parameter estimate and its standard error is t-distributed. For variance estimates we used a one-sided alpha level of .05 ($t > 1.658$), for covariance estimates a two-sided alpha level of .05 ($t > 1.965$) (see Goldstein, 1995).

3 RESULTS

Table 1 presents the observed mean, the standard deviation and the range for variables included in the analysis. As can be seen, both students and teachers showed substantial variation in their ratings on the four-point scale ($SD$ for student ratings = 0.797; $SD$ for teacher ratings = 0.664). For individual students, the standard deviation for sections ranged from 0.421 to 1.184. For teachers this range was 0.484 – 0.823. All participants, then, distinguished between the importance of sections, with some showing a larger spread among their ratings than others. Sections were deemed rather important than unimportant, whereby the average teacher rating ($M = 2.892$) was somewhat higher than the average student rating ($M = 2.775$). As a consequence, the average student-teacher correspondence across sections has a negative

---

Footnote: In fact, two analyses have been performed. The first on the observed ratings, and for the second analyses, which are reported in this paper, a normalizing transformation is applied to the data, after which they were transposed to the original scale values. Both types of analyses show practically the same results.
value \((M = -0.117)\). The corresponding standard deviation indicates that students and their teachers varied considerably in their ratings for separate sections \((SD = 0.960)\). Apparently, students and teachers generally had quite different perceptions of the task demands. However, student perceptions were not the same. The standard deviation of the student–class correspondence shows that separate ratings of individual students often differed from ratings that students of the corresponding class, on average, gave to these sections \((M = 0.00; SD = 0.789)\). Students within a class, then, often showed differences in the text elements they considered important for the test. Finally, test grades, ranging from 1 to 10, had a mean value of 6.296 and a standard variation of 1.336.

Results from the multilevel analysis are organized according to our three quality measures. For each quality measure, the influence of test grade on the quality of students’ perceptions is estimated, taking into account the influence of the within student standard deviation (Table 2). (In the Appendix, an unconditional model is presented, which shows the distribution of the variance in the quality measures at the different hierarchical levels.)

**Table 1. Descriptives of observed values of variables included in the study**

<table>
<thead>
<tr>
<th></th>
<th>(n)</th>
<th>(M)</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Students</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating</td>
<td>6429</td>
<td>2.775</td>
<td>0.797</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>(M) of ratings across sections</td>
<td>375</td>
<td>2.775</td>
<td>0.302</td>
<td>2.039</td>
<td>3.875</td>
</tr>
<tr>
<td>(SD) of ratings across sections</td>
<td>375</td>
<td>0.797</td>
<td>0.147</td>
<td>0.421</td>
<td>1.184</td>
</tr>
<tr>
<td>Test grade</td>
<td>375</td>
<td>6.296</td>
<td>1.336</td>
<td>1.000</td>
<td>10.000</td>
</tr>
<tr>
<td><strong>Teachers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating</td>
<td>515</td>
<td>2.892</td>
<td>0.664</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>(M) of ratings across sections</td>
<td>20</td>
<td>2.892</td>
<td>0.269</td>
<td>2.577</td>
<td>3.500</td>
</tr>
<tr>
<td>(SD) of ratings across sections</td>
<td>20</td>
<td>0.664</td>
<td>0.101</td>
<td>0.484</td>
<td>0.823</td>
</tr>
<tr>
<td>Student rating – teacher rating</td>
<td>6429</td>
<td>-0.117</td>
<td>0.960</td>
<td>-3.000</td>
<td>3.000</td>
</tr>
<tr>
<td>Student rating – class rating</td>
<td>6429</td>
<td>0.000</td>
<td>0.789</td>
<td>-2.073</td>
<td>1.401</td>
</tr>
</tbody>
</table>

*Note. For both teachers and students descriptives are given for importance ratings of individual sections, the average of a participant's ratings across sections, and the standard deviation of a participant's ratings across sections. Furthermore, descriptives are given for students' test grade and two difference scores: "student rating – teacher rating" and "student rating – class rating".*
3.1 Student rating

The expectations derived from hypothesis 1 could not be confirmed. First, the regression of test grades on students’ ratings was nonsignificant ($\beta = 0.000, p > .05$). Second, although the covariance at the student level showed a negative value, this value did not reach significance ($S^2 = -0.004, p > .05$). These results, then, provide no indication that the height of students’ importance ratings is directly related to test grade.

The within student standard deviation did have an influence on the student-teacher correspondence ($\beta = -0.284$, $p < .05$). Its negative value suggests that students with more spreading in their ratings, on average, had a closer correspondence with their teacher.

Table 2. Estimated quality of students’ importance ratings, conditional model

<table>
<thead>
<tr>
<th></th>
<th>Student rating</th>
<th>Student rating – teacher rating</th>
<th>Student rating – class rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
</tr>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (intercept)</td>
<td>2.782</td>
<td>0.034</td>
<td>0.110</td>
</tr>
<tr>
<td>Regression test grade ($\beta_1$)</td>
<td>0.000</td>
<td>0.014</td>
<td>0.004</td>
</tr>
<tr>
<td>Regression within student SD ($\beta_2$)</td>
<td></td>
<td></td>
<td>-0.284*</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3 (class/teacher)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.010</td>
<td>0.006</td>
<td>0.070</td>
</tr>
<tr>
<td>Regression ($\beta_1$)</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>Covariance</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>Level 2 (student)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.057*</td>
<td>0.007</td>
<td>0.047</td>
</tr>
<tr>
<td>Regression ($\beta_1$)</td>
<td>-0.004</td>
<td>0.003</td>
<td>-0.006*</td>
</tr>
<tr>
<td>Covariance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1 (section)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.563*</td>
<td>0.012</td>
<td>0.804*</td>
</tr>
<tr>
<td>Regression ($\beta_1$)</td>
<td>0.004</td>
<td>0.004</td>
<td>0.002</td>
</tr>
<tr>
<td>Covariance</td>
<td>0.006</td>
<td>0.004</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Note. Estimates (Est) are based on three separate models that correspond to the different operationalizations for the quality of student ratings. The three models can be considered as variants of a general conditional model with test grade ($\tilde{\beta}$) and the within students standard deviation ($\beta_2$) as explaining variables. Note that for the first quality measure (student rating) the within student standard deviation could not be included in the model. Similarly, the model for the third quality measure (student rating – class rating) only includes variance estimates for the student and section level. Formula description: (quality of student rating) = $\beta_{0jk} + \beta_{1jk}(test\ grade)_{ik} + \beta_2(\text{within\ student\ SD}) + e_{ijk}$. $\beta_{0jk} = \beta_{000} + v_{00k} + u_{0jk} + e_{ijk}$. * $p < .05$. 
3.2 Student rating – class rating

The expectations derived from hypothesis 3 could not be confirmed. First, the regression of test grade on the student-class correspondence was nonsignificant ($\beta = 0.002$, $p > .05$). Second, although the covariance at the student level had a negative value, this value did not reach significance ($COV = -0.004$, $p > .05$). Interestingly, we did find a significant effect of the within student standard deviation, showing that students with more spreading in their ratings, on average, had a closer correspondence with their class ($\beta = -0.2742$, $p < .05$). It seems, then, that the results for the third quality measure converged with those of the second quality measure but were less pronounced.

![Figure 1. Student-teacher correspondence as a function of test grade.](image)

*Note.* The correspondence score pertains to the difference in importance ratings between individual students and their teacher. Negative and positive deviations are aggregated across the 26 sections that were rated by the participants. The horizontal line is the average correspondence across the 375 students. Variance values were predicted by substituting the parameter estimates of the conditional model in the formula of note 6 and by using a 90 percent confidence interval ($\pm 1.65 \times SD$).
Contemporary theories on learning assume that the quality of students' perceptions of task demands is related to their learning performance. We have focused on a well-known instance of this assumption that involved students' expectations regarding test content (indicated by the relative importance they assigned to text sections) and their test performance (indicated by test grades).

4.1 Are perceived task demands and test performance related?

The first hypothesis was that a positive relationship exists between the average importance students assign across sections and their test performance. This hypothesis was not supported, because the height of importance ratings showed no relationship with test grades.

The second hypothesis was that the correspondence in perceived importance between individual students and their teachers is positively related with students' test performance. This hypothesis was in part supported. For separate sections, a relationship between the student-teacher correspondence and test grade could not be demonstrated; the regression value was nonsignificant. This suggests that a relative accurate perception of task demands did not guarantee a good test performance. However, across sections, the covariance at the student level indicates that students with higher test grades were a more homogeneous group than students with lower test grades: i.e., they more often showed a relatively close correspondence with their teacher. This suggests that students with higher test grades had a more accurate perception of task demands.

The third hypothesis was that the correspondence in perceived importance between individual students and their class is positively related with students' test performance. This hypothesis was not supported, because both the regression and the covariance values did not reach significance.

We interpret these outcomes as partly support for the assumption that the quality of students' perception of task demands is related to their learning performance. This brings us to a crucial question: does this relation have a causal nature?

4.2 Does a more accurate perception of task demands contribute to test performance?

To answer the question of causality, at least three factors should be taken into account. The first factor is the clarity of task demands. The differences among students in their rating pattern suggest that some students had a more accurate perception of task demands than other students. Moreover, students who demonstrated less spreading in their ratings showed more discrepancy with their teacher and class. The lower spreading could mean that these students refrained from making sharp distinctions in their ratings because they were uncertain about the instructional importance of these sections. This finding, then, confirms that some students had a more clear perception
of task demands than other students did. Nonetheless, even for students with a more clear perception, the perception might still have been insufficiently clear to use (effectively) a selection attention strategy. Therefore, clarity of task demands remains an important explanation why we did not find significant regression coefficients concerning the second and third hypothesis.

The second factor is the study time that students had available. In the theoretical account of our hypotheses, we distinguished between study conditions with high and low time pressure. Whereas high time pressure would encourage selective processing, low time pressure would induce a nonselective attentional strategy. It was reasoned that accurate knowledge of the task demands was important for the selective strategy but not for the nonselective strategy. For the nonselective strategy, we hypothesized that the more importance students assigned to text sections the more attention they would pay to these sections — resulting in a higher test grade. The relation between higher ratings and an increase in the amount of attention, however, may not be that clear-cut. For instance, students may have differed in their interpretation of the rating scale. Moreover, as we did not collect data on the available study time and the studying process, our findings still allow for the explanation that student had sufficient time to learn all section equally intensive and that the accuracy of their task perception was of minor importance.

The third factor is the way students acted on their perception of task demands. It is uncertain whether the eleventh grade-students in our study, possessed the abilities to use the selective attention strategy in a fully effective way. Reynolds and associates (Reynolds, 1992) found that this strategy had a causal effect on test performance for college students, but not for tenth-grade students. The eleventh-grade students of our study could be considered as falling in between the tenth grade- and the college students. A direct comparison, however, is not possible, since the laboratory context of the studies of Reynolds and associates are quite different from the classroom contexts we examined. The studies of Reynolds and associates, however, make clear that even when task perception and task performance are related, we cannot assume that this relationship has a causal nature.

Considering these three factors, we conclude that for students with the highest test grades, their relative accurate perception of task demands may have helped them to obtain these grades. However, their relative accurate perception of task demands could also have been an epiphenomenon related to learning abilities that caused the higher test performance. For the students who had a lower test grade, there is no indication for a causal relationship at all.

Given our research methods, it was beforehand clear that we could not distinguish between a causal and noncausal relationship concerning the quality of students’ perception of task demands and test performance. This study, however, was meant as a first attempt to investigate the nature of this relationship in a classroom context. As our findings and our discussion of the three factors make clear, the nature of this relation is complex. To examine it more thoroughly, our study could be extended by including data on students’ attentional strategies when studying text as well as on the study time available and the efficiency of the learning performance.
4.3 Measuring (perceptions of) task demands

Importance ratings are only an indirect indicator of participants’ perception of task demands. The meaningful relations that were found among variables at different levels can be seen as a confirmation of the validity of the ratings. However, since no alternative measures were included in the study (e.g., measures derived from retrospective interviews), reservations should be made regarding an absolute interpretation of the parameter estimates that we found. Similarly, the fact that we could not find some of the relations that were hypothesized cannot be taken as a conclusive proof that these relations in fact were absent.

The finding of previous studies that importance the importance assignment of individual students and their teachers showed only a moderate correspondence (e.g., Broekkamp, Van Hout-Wolters, Rijlaarsdam & Van den Bergh, 2002) gave reason to question the validity of teachers’ perceptions as a measure for the factual task demands. Therefore, we examined the average ratings across students within classes as an additional measure. This additional measure appeared to yield partly convergent findings. That is, students who had less spreading in their ratings showed more discrepancy both with their teacher and with their class. Furthermore, as for the student-teacher correspondence, the student-class correspondence showed a negative covariance with test grade at the student level. However, for the student-class correspondence this value did not reach significance. We therefore conclude that, in our study, the class perceptions constituted a convergent but less powerful measure of the factual task demands than the teacher’s perceptions.

How can we explain this finding? We decided that it was appropriate to use student aggregates because rating patterns of students of the same class were clearly more similar than rating patterns of students of different classes (cf. Ryan et al., 1998, p. 533). Nonetheless, rating patterns within classes still showed considerable variation. It seems, then, that in order to use student aggregates as a powerful measure more consensus is needed within classes. The required level of this consensus, however, remains an unresolved question (cf. Ryan et al., 1998, p. 533).

Perceptions of task demands are not the only indications of task demands. Case studies demonstrate that observing the instructional process or analyzing tests may help to gain additional insights into task demands that are imposed upon students (cf. Jetton & Alexander, 1997; Simpson & Nist, 1997). Moreover, analysis of teachers’ tests could additionally provide insight into the validity of test grades as a measure for students’ test performance.

4.4 Implications

Given the lack of other classroom studies on test expectations and test grades, special precaution is needed in generalizing our results to other populations, other academic tasks and other kinds of task demands. Generalizations should take into account, that our study concentrated on relatively advanced secondary students studying an instructional history text in preparation for a test. Moreover, we assessed task perceptions at a single occasion. Concerning the nature of task demands, we focused on the relative importance of topics involving rather broad units of text (a section
included three or four paragraphs). In doing so, we did not include at least six other kinds of task demands that are involved in text studying in classroom contexts. Task demands may also pertain to (a) the instructional importance of text elements at lower text levels (e.g., the sentence and paragraph level; Jetton & Alexander, 1997; Schellings & Van Hout-Wolters, 1995) (b) the type of the required information (e.g., broad concepts versus factual information; Thomas, Bol, Warkentin, Wilson, Strage, & Rohwer, 1993); (c) the cognitive level at which students have to demonstrate that information has been mastered (e.g., memorization or application; Marton & Säljö, 1976; Thomas et al., 1993); (d) the social and academic expectations of lesson participation (e.g., Green & Weade, 1987; Winne & Marx, 1982) (e) the required level of effort and the motivational goals that are encouraged in classrooms (e.g., Brookhart, 1997; Ryan et al. 1998); (f) the epistemological beliefs and study orientations that are appropriate in academic contexts (e.g., Ramsden, 1988; Simpson & Nist, 1997). Concerning all these kinds of task demands, students use task information to construe a perception of these demands. Research has just begun to unravel the different kinds of task information that are used by students to self-regulate their learning and to increase their learning performance (Butler & Winne, 1995).

In this study, we concentrated on a single type of task information – the instructional or task relevance of text sections. Experimental research has consistently showed that this kind of task information, provided in the form of objectives and inserted questions, increases the attention for task relevant text elements and results in better learning of these text elements (e.g., Reynolds, 1992). The present study, however, is conducted in a realistic classroom context. Moreover, the effect was examined of individual differences in the perception of task demands on test performance. The fact that our findings provide only limited support for a causal relationship between perceived task demands and test performance, points to crucial differences between laboratorial and classroom contexts regarding the nature of task demands and information about these demands. Several phenomenographic classroom studies make clear that the emergence and communication of task demands – including instructional importance – are very complex processes (Green & Weade, 1987; Jetton & Alexander, 1997; Simpson & Nist, 1997; Wade & Moje, 2000; Wilson & Wineburg, 1993). As shown in the present study, as well as several other studies, these processes may result in large differences in perceived task demands between teachers and between teachers and their students. In our opinion, especially the latter kind of discrepancies is important, because it seems to set boundaries for students to study in strategic ways.
### Estimated quality of students’ importance ratings, unconditional model

|                      | Student rating | Student rating | Student rating |
|----------------------|----------------|----------------|----------------|----------------|
|                      | Estimate | $SE$ | Estimate | $SE$ | Estimate | $SE$ |
| **Fixed**             |          |      |          |      |          |      |
| Mean (intercept)      | 2.782   | 0.034 | -0.096  | 0.079 | -0.001  | 0.017 |
| **Random**            |          |      |          |      |          |      |
| Level 3 (teacher/class)| 0.010*  | 0.005 | 0.070*  | 0.030 |          |      |
| Level 2 (student)     | 0.057*  | 0.007 | 0.048*  | 0.007 | 0.053*  | 0.007*|
| Level 1 (section)     | 0.570*  | 0.010 | 0.807*  | 0.015 | 0.570*  | 0.010*|

Note. Estimates are based on a general unconditional model, which is the same for the three operationalizations of the quality of student’s ratings. Note that, as the average student-class correspondence could not vary between classes, the model for the third quality measure only includes estimates for the student and section level. Formula description: (quality of student rating) = $\beta_0 + \alpha_k + \eta_{ijk} + \epsilon_{ijk}$. * $p < .05$. 