Teachers’ task demands, students’ test expectations, and actual test content

Hein Broekkamp¹*, Bernadette H. A. M. Van Hout-Wolters¹, Huub Van den Bergh¹,² and Gert Rijlaarsdam²
¹ Graduate School of Teaching and Learning University of Amsterdam, The Netherlands
² Utrecht Institute for Linguistics, Utrecht University, The Netherlands

Background. Previous studies on instructional importance show that individual students and their teachers differ in the topics that they consider important in the context of an upcoming teacher-made test.

Aims. This study aimed to examine whether such differences between students’ test expectations and teachers’ intended task demands can be explained by the actual test content.

Samples. Participants were history teachers (N=19) and their 11th-grade students (N=388).

Method. Teachers and students rated the importance of text sections that would be tested in the near future. By means of multilevel analysis, ratings were compared with the occurrence of sections in the tests.

Results. Although teachers considered a majority of sections as important and tested only a minority of the sections, their tests still included sections rated as unimportant. The number of such discrepancies, however, was relatively small. Sections the teachers rated important had a much higher probability of being included in the test than sections rated unimportant. For students, a similar but lower degree of correspondence between ratings and test content was found. Interestingly, for sections that teachers considered important, students more often gave a higher rating when these sections appeared in the test than when they did not. The same holds for sections that teachers considered unimportant.

Conclusions. For both teachers and students there is a limited correspondence between perceived task demands and test content. Furthermore, students’ perceptions of task demands show a compensation for some of the differences between their teachers’ intended task demands and the test demands.

* Correspondence should be addressed to Hein Broekkamp, Graduate School of Teaching and Learning, University of Amsterdam, Wibautstraat 4, 1091 GM, Amsterdam, The Netherlands (e-mail: broekkamp@ilo.uva.nl).
When students have clear and realistic expectations about a test, they know the goals and criteria towards which they should direct their studying. They have, in other words, an adequate perception of the requirements of the study task. In classrooms, it is generally the teacher who is responsible for defining the task demands, and how these demands are tested. Because teachers will differ in their task demands, students try to attune to the task and test demands that their teacher sets (cf. Doyle, 1983; Nolen & Haladyna, 1990; Ramsden, 1988; Simpson & Nist, 1997; Thomas & Rohwer, 1986; Van Etten, Freebern, & Pressley, 1997; Winne & Hadwin, 1998). The present study focuses on a specific kind of demand that concerns the ‘instructional importance’ of textual topics in the context of an upcoming test (cf. Alexander & Jetton, 1996). A clear perception of this kind of demand may help students to focus their attention especially on important text parts and as a consequence to increase their learning performance (Anderson & Armbruster, 1984; Reynolds, 1992).

Several studies have examined the instructional importance that students and teachers assign to textual topics 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 these studies, a limited correspondence was found between teachers, between students, and/or between teachers and their students with respect to their importance assignment. These findings were obtained for different text levels (sentence, paragraph, section), different measures (rating, question writing, underlining, note-taking), different subjects (science, biology, language arts, history) and different correspondence measures, including multilevel estimates. Furthermore, the student-teacher correspondence varied between teachers, students and textual topics. These studies suggest the following conclusions: (a) teachers differ in the task demands they set, (b) teachers differ in the way they convey task demands to their students, (c) students, generally, do not have a very accurate perception of their teachers’ task demands, (d) some students have a more accurate perception of the task demands than other students.1

Jetton and Alexander (1997) analyzed the tests teachers constructed and gave to their students. Students’ importance assignments showed only a limited correspondence with the required contents of the test. For instance, some topics that students considered relatively unimportant were addressed by more than one test item while topics that they considered relatively important were not addressed at all. Even more interestingly, such discrepancies also existed for the teachers’ ratings, with one of the three participating teachers clearly showing a lower correspondence with her own test questions than the other two teachers did. The qualitative character of the study does not allow for generalizations. Nonetheless, the findings suggest that differences exist between teachers’ (intended) task demands (the demands towards which students

---

1 It should be noted that the perspective by which participants assigned importance to the text units showed differences among the studies (see also Broekkamp et al., 2002). For instance, in the study of Broekkamp et al. (2002), students’ importance ratings expressed both the emphasis of topics in the class and the likelihood that these topics would appear in the test. The rating task in the study of Jetton and Alexander (1997) was also followed by a test, but students received the more general instruction to indicate content they deemed important to know. In the study of Schellings and Van Hout-Wolters (1995) students were asked to pretend study for a regular test given by their teacher and to underline text units that were likely to be asked in such a test. In this study, no actual test was given to the students. Despite such differences, it can be argued that in all of the studies that we mention, the – actual or imaginary – upcoming test exerted influence on the assignment of importance. For this reason we allow ourselves to describe the findings in terms of task demands for all of these studies. In doing so, we deviate from Alexander and associates, who did not use the term task demands overtly and used the more general term of instructional importance to describe their findings.
should direct their learning when preparing for the test) and the actual *test demands*
(the contents required by the test items).

When students study in preparation for a test, the test demands at best represent a
sample of teachers’ task demands. That is, the test generally cannot cover all
information that students are required to learn. Depending on the number of topics that
have to be studied, available testing time, the kinds of test questions or arbitrary factors
(e.g., spending little time on test construction) teachers can select a smaller or larger
part of the learning content for the test. Discrepancies between teachers’ task demands
and test demands occur when less important topics are included in the test while more
important topics are excluded or when less important topics in the test receive more
weight than more important topics. Such discrepancies could emerge when teachers
have only vague intentions regarding task demands. Another possibility is that teachers
fail to construct test items that are representative of their task demands.

What do differences between teachers’ task demands and the test demands mean for
the student? Students may base their test expectations on at least five different sources
(cf. Alexander & Jetton, 1996; Doyle, 1983; Miller & Parlett, 1974; Schraw, Wade, &
Kardash, 1993; Snyder, 1971; Van Etten et al., 1997). First, students may rely on the
explicit information that teachers provide about their intended task demands. For
instance, during a teacher-led classroom discussion, the teacher says that a certain topic
is important and will probably be tested. Second, students can make use of implicit
information that teachers provide. For instance, when the teacher dwells upon a
specific part of the learning content, this is a signal to the students that this part might
be very important in the test. Third, students can base their expectations on their
experiences with previous study tasks and tests. For instance, in previous tests their
history teacher may have stressed economic explanations for events over cultural
explanations (cf. Wilson & Wineburg, 1988). Fourth, students can rely on the learning
materials. For instance, a topic that is signalled by the textbook author as crucial,
probably will receive more attention in the end test. Finally, students can obtain task
information indirectly from peers.

When students sense that the teacher’s explicit information about his intended task
demands is unrepresentative of the test demands or incomplete, they may disregard this
information or supplement it by information derived from other sources. By doing so,
they may ‘compensate’ for differences between the teachers’ task demands and the test
demands. Such compensation may explain why in one of the three classes studied by
Jetton and Alexander (1997), students’ importance ratings showed a reasonable
correspondence with the test questions despite the fact that students’ ratings were
unrelated to the teacher’s importance ratings.

The present study extends the study of Jetton and Alexander (1997) by examining
the relations between teachers’ (intended) task demands, students’ test expectations
and test demands in a quantitative way. We focused on three questions: (a) What is the
degree of correspondence between teachers’ task demands and their test demands? (b)
What is the degree of correspondence between individual students’ test expectations
and the test demands? (c) To what degree are individual students able to compensate
for differences between teachers’ task demands and the test demands?

The three questions were investigated in the first place to shed light on previous
studies on instructional importance. Furthermore, these questions are important from
the perspective of assessment in classroom contexts (cf. Airasian, 1996; Brookhart,
1999; Crooks, 1988; Dochy & Moerkerke, 1997; Fleming & Chambers, 1983). Previous
investigations did not allow us to formulate firm expectations concerning our research
questions. However, they gave us reason to expect that test demands would show a limited correspondence with both teachers’ task demands and students’ test expectations, and that students would compensate for differences between teachers’ task demands and the test demands.

Method

Participants
Participants in this study were 19 history teachers and 388 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 18.4 years of experience (SD = 7.60) and 15 of them were male. Students had a typical age of 17 and were in the penultimate year of pre-university education. Of the students 210 were female. The recruited teachers (and their classes) used a particular history textbook, which had the highest market share in the Netherlands and which did not provide standard chapter tests. 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 gave a teacher-made test to their 11th-grade students on this chapter in the second half of the school year. At this point, students would be more accustomed to their teachers and would be more likely to attune to task demands effectively (cf. Van Etten et al., 1997).

Materials and procedure
The chapter had ‘The United States of America since 1945’ as its theme and contained approximately 8,000 words. Originally, the chapter was structured in 16 sections of different length. For the purpose of this study, longer sections were subdivided. The new division, which followed quite naturally from the text, consisted of 26 text ‘sections’ of similar length, each encompassing three to four paragraphs. The 26 sections formed the objects of a rating task. Participants rated the importance of sections based on short summaries of these sections. 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.

By asking students to assign importance to summaries of text units, we deviated from previous studies in which instructional importance was assigned to the actual text (e.g., Jetton & Alexander, 1997; Schellings & Van Hout-Wolters, 1995). For our purpose, the ‘summary rating task’ had both practical and theoretical advantages over a ‘text rating task’. A practical advantage was that a summary rating task allowed us to collect ratings concerning the whole text in a short time period. This was important, because the text in our study was considerably longer than the text used in previous studies to examine

---

2 Initially, 22 teachers from 17 schools each participated with one class of students (451 students in total). For the purpose of this study, however, we examined only teachers (and their classes) who made or chose the test questions themselves. Consequently, for each of the three schools that participated with two teachers, only one teacher was examined. This was the teacher that was known to have constructed or chosen the test questions; the teacher that was left out of our analyses shared the test with his or her colleague but was not responsible for writing or choosing the test questions. There was also one school that participated with three teachers. Because each teacher gave their own test, all three teachers were included in the analyses.
Concerning the theoretical advantages, we assumed that the summary rating task exerts less influence on participants' perceptions of task demands than a text rating task because it prevents them from reading the text. Reading the text with the explicit purpose to assign importance to text elements would interfere with the usual way in which task demands and test expectations develop in classrooms. In this regard, a summary task is less obtrusive, because it encourages participants to operate from memory in forming their importance judgments.

To ensure that none of the students engaged in additional reading, they were asked not to review the textbook or their notes during the rating task. Similarly, teachers were asked not to examine the text or their actual test materials at this point. With this latter constraint, we tried to encourage teachers to report intended task demands, instead of actual test demands.

The rating task was performed within two days before the actual classroom test, during a history lesson (16 classes) or during a separate session (6 classes). Students and their teacher, at the same time but independently, rated the importance of the 26 sections. Participants were instructed to indicate with their importance ratings both the emphasis that section topics received in the classroom and the likelihood that these topics would be included in the upcoming test. To the teachers, we stressed that their ratings were to reflect intended task demands, instead of actual test demands. Importance was rated on a 4-point scale (1 = unimportant, 2 = less important, 3 = important, 4 = very important; cf. Brown & Smiley, 1977).

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. 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 the session, teachers handed over to the researcher the test they would give to their students. In addition, they gave on paper their model answers to the test questions and the allocation of test points over the test questions. These three kinds of test information formed the basis for the test analysis.

**Analysis of teachers' tests**

For each test item, we identified chapter sections that were associated with the required information. Sections were not only identified when they provided explicit information required to answer a test question (cf. Jetton & Alexander, 1997) but also when they offered implicit information (the required knowledge was intended by the textbook author), indirect information (the required knowledge could be inferred by making use of information provided by the section) or related information (the required knowledge was clearly associated with the topic of the section, but the section was not of help in answering the question) (cf. Bean, 1985; Doyle, 1983). The identification of
sections was performed by two judges, who showed an agreement of 89%. Discrepant codings were resolved by discussion.

After a section was identified, we determined the ‘weight’ that the section had in the test by taking the number of points that could be obtained for the corresponding questions as a proportion of the total number of points in the test. When a question involved more than one section, the number of points was proportionally divided.

Based on the test analysis, three variables were derived that were associated with test demands: (a) section test occurrence: whether a section was addressed by the test or not (values: 1, 0); (b) section test weight: the average weight of a section when addressed by a test question (values: 0.00 – 1.00); (c) section test coverage: the number of sections addressed in the test (values: 0 – 26).

Statistical analysis

Data were analyzed in a multilevel way, which allowed us to take into account their nested structure. For instance, ratings were nested within students, students nested within classes. To facilitate the interpretation of the multilevel models, the four values of the importance ratings were collapsed into two values (ratings 1 and 2: unimportant, ratings 3 and 4: important). The parameters of the multilevel models were estimated with the software MLwiN (Multi Level Models Project, 1999). Statistical significance of a parameter estimate can easily be determined as the ratio of the estimate and its standard error is $t$-distributed. For variance estimates, we used a one-sided alpha level of .05 ($t > 1.658$) (cf. Goldstein, 1995).

Results

Descriptives of observed ratings and test characteristics

Table 1 shows that sections more often were deemed important (ratings 3 and 4) than unimportant (rating values 1 and 2). Teachers, on average, considered .77 of the sections as important. This proportion ranged between .54 and 1.00 ($SD = .35$). Students, on average, considered .67 of the sections as ‘important’. This proportion ranged between .19 and 1.00 ($SD = .15$). The ranges of variation show that some of the students and some of the teachers considered all sections of the textbook chapter as instructionally important. However, the standard deviations indicate that most participants distinguished between important and unimportant sections. Appendix B, which includes descriptives per section, shows that averaged across teachers, the proportion of ratings with the value ‘important’ varied between sections from .18 to .91. Averaged across students, this range was .50 to 1.00. This indicates that some sections had more ‘general instructional salience’ than other sections (cf. Alexander & Jetton, 1996). Nevertheless, most of the sections showed considerable variation in assigned ratings. Apparently, the importance of sections did not only depend on the text but also on individual and/or contextual factors (see Reynolds, 1992).

Table 1 includes descriptives for the three test variables. The average occurrence of sections in the test was .37 ($SD = 0.48$). This means that teachers’ tests, in general, were rather selective. As shown by the section test coverage variable, on average, only 9.53

---

5 In Appendix A, descriptives are given of variables based on four rating values. These descriptives show that although some participants adhered to the higher points of the scale, each of the participants showed a substantial variation in their ratings.
sections were covered in the test. Although there were large differences between teachers, with one teacher covering only two sections and another teacher covering 17 sections, there was no teacher covering all 26 sections. The weight of sections in the test, on average, was .09, and ranged between section from .01 and .45. Thus, in some cases, sections were only minimally addressed, whereas in other cases, a single section took in almost half of the possible test points. Appendix B shows that section test occurrence and section test weight varied between sections. Whereas one section was addressed by .74 of the tests, another section was not addressed at all. Similarly, for the sections that appeared in one or more tests, the average weight ranged from .02 to .15.

**Importance ratings and section test occurrence**

Table 2 shows the multilevel estimates pertaining to a 2 x 2 x 2 matrix of teachers’ ratings by students’ ratings by section test occurrence. Instances of each cell were estimated as a proportion of the total number of the combinations observed. In the following discussion, it is gradually shown how the matrix can be used to determine the kinds of correspondences that are of interest to our research questions. For all comparisons that are discussed, a significant difference was found (p < .01).

<table>
<thead>
<tr>
<th>Teacher Student</th>
<th>Non-occurrence of topic on test</th>
<th>Occurrence of topic on test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unimp.</td>
<td>.09</td>
<td>.01</td>
</tr>
<tr>
<td>Imp.</td>
<td>.10</td>
<td>.03</td>
</tr>
</tbody>
</table>

Note. Instances of each of the 2 x 2 x 2 categories of section test occurrence (non-occurrence, occurrence) by teachers’ ratings (unimportant, important) by students’ ratings (unimportant, important) are estimated as a proportion of the total number of 9,731 instances of the eight categories. Column and row totals of the proportions are presented in the margins.
The correspondence between teachers’ and students’ ratings

Of the total of sections that teachers considered important, 70 were considered important by the students as well; 30 of these sections were considered unimportant by the students: \( \frac{.29 + .25}{.44 + .33} = .70 \); \( \frac{.15 + .08}{.44 + .33} = .30 \). Furthermore, of the total of sections that teachers considered unimportant, \( .43 : \frac{.09 + .01}{.19 + .04} \) were considered unimportant by the students as well; \( .57 \) were considered important. These proportions indicate that students were more likely to overestimate the importance of an unimportant section than they were to underestimate the importance of an important section. The ratio of students’ ratings (important/unimportant) is lower for sections that teachers considered unimportant than for sections that teachers considered important: \( \frac{.10 + .03}{.09 + .01} = 1.3 \); \( \frac{.29 + .25}{.15 + .08} = 2.3 \). These proportions indicate that teachers and students showed a correspondence in their importance ratings. However, \( .36 \) of the ratings that students gave to sections were discrepant with the ratings that teachers gave to these sections \( (.10 + .15 + .03 + .08) \).

We will now examine whether these discrepancies can be explained by the occurrence of sections in the test.

The correspondence between teachers’ ratings and section test occurrence

Teachers considered \( .77 \) of the sections important \( (.44 + .33) \). However, only \( .37 \) of the sections were included in the test. Given this selective way of testing, a test being fully congruent with teachers’ ratings would include no sections considered unimportant by the teachers. However, Table 2 shows that \( .11 \) of the sections in the test were considered unimportant by the teachers \( (.04 / .37) \). This indicates that there was a substantial number of discrepancies between teachers’ ratings and section test occurrence. On the other hand, sections that were deemed important had a much higher probability of occurring in the test than sections that were deemed unimportant. Whereas \( .17 \) of the sections that teachers considered unimportant did occur in the test \( .43 \) of the sections that teachers considered important occurred in the test: \( .04/[.19 + .04] = .17 \); \( .33/[.44 + .33] = .43 \). In sum, teachers’ ratings show a limited correspondence with the occurrence of sections in the test.

The correspondence between students’ ratings and section test occurrence

Of the sections included in the test \( .24 \) were considered unimportant by students \( (.09 / .37) \). This means that students showed more discrepancies with section test occurrence than their teacher did \( (.24 \text{ vs. } .11) \). Nevertheless, sections that students considered important had a higher probability to occur in the test than sections that they deemed unimportant. Whereas \( .27 \) of the sections that students considered unimportant did occur in the test: \( .42 \) of the sections that students considered important occurred in the test: \( .09/[.24 + .09] = .27 \); \( .28/[.39 + .28] = .42 \). Thus, students’ ratings showed a limited correspondence with the occurrence of sections in the test.

The correspondence between teacher ratings, student ratings and section test occurrence

To learn whether students compensated for a difference between teachers’ task demands and the test demands, we can compare for each of the four combinations of teachers’ ratings and students’ ratings the probability that a section was included in the test. For sections that were considered important by both the teacher and the student, this probability was \( .46 : .25 / [.29 + .25] \). For sections that were considered important by the teacher and unimportant by the student, this chance was \( .35 : .08 / [.15 + .08] \). The difference between these two proportions indicates that for some topics the
student was able to know that these topics would not appear on the test, even though
the teacher considered them important. Similarly, for sections that were considered
unimportant by both the teacher and the student the chance of appearing in the test
was $0.10 / [0.09 + 0.01]$, whereas for sections that were considered unimportant by
the teacher and important by the student this chance was $0.23 / [0.10 + 0.03]$. The
difference between these two proportions indicates that for some topics the student
was able to know that these topics would appear in the test, although the teacher
considered them unimportant. These comparisons indicate that students compensated
for some of the differences between teachers’ task demands and the test demands.

**Variance between classes**

Appendix C shows the estimated variance of the eight categories. The variance
estimates indicate whether there were differences between classes regarding the four
kinds of correspondence we have just described. All eight variance estimates had a
significant value. For instance, the probability of sections appearing in the test and rated
unimportant by the teacher and important by the student varied between classes from
$0.00$ to $0.27$. This range indicates that whereas students showed compensation for
discrepancies between teachers’ task demands and the test demands in one class such
compensation was absent in another class.

**Importance ratings and section test weight**

To examine whether the weight of sections in the test could explain differences
between students’ ratings, teachers’ ratings and section test occurrence, a multilevel
analysis was performed. For each of the four categories of teachers’ and students’
importance ratings involving sections occurring in the test, the section test weight was
estimated (Table 3). Sections that were rated important by teachers clearly had more
weight in the test than sections that were rated as unimportant ($0.11$ vs. $0.05$, $\chi^2 > 23.55$,
$p < .01$). This means that the discrepancy of teachers considering sections in the test as
unimportant was partly countered by a relatively low test weight. For students such a
counter-effect could not be determined. The weight of sections that were rated
unimportant by students did not differ from the weight for sections that were rated as
important ($0.10$ vs. $0.11$, $\chi^2 > 3.11$, $p = .08$).

The weight of sections per category was estimated along with the variance between
classes (see Appendix D). For each of the four categories of sections appearing in the

| Table 3. Estimated section test weight for 2 x 2 categories of teachers’ ratings by students’ ratings |
|--------------------------------------------------|------------------|------------------|
| Student Rating | Teacher rating |                |
|                |                | Unimportant | Important |
| Unimportant    | $.05$          | $.10$       | $.10$      |
| Important      | $.05$          | $.12$       | $.11$      |
|                | $.05$          | $.11$       | $.09$      |

Note. Section test weight is estimated for 2 x 2 categories of teachers’ ratings by students’ ratings
(unimportant, important) pertaining to sections that occurred in the test. Section test weight refers to
the number of test points associated with a section. The margins represent the average values for the
column and row totals.
test, there were differences between classes in the average weight of the sections. For instance, the average weight for sections that were rated important by the teacher and unimportant by the student ranged from .01 to .19. This means that in some classes, discrepancies between students’ and the teacher’s perceptions of task demands might have followed from students’ anticipation of a low test weight.  

Importance ratings and section test coverage
The low test weight for sections in the test that teachers considered unimportant could have been typical for teachers that included many sections in the test. To examine this possibility, regression coefficients were estimated that expressed the influence of the number of sections included in the test for four categories of teachers’ and students’ ratings pertaining to sections that appeared in the test. The estimated regression was non-significant for all four categories ($t < 1.13, p > .13$). This means that the distribution of sections in the test over the four categories is similar for teachers. Section test coverage, then, cannot explain discrepancies between teachers’ ratings, students’ ratings and section test occurrence.

Discussion
As expected, we found a limited correspondence between perceptions of task demands and test demands both for teachers and for students. In addition, students’ perceptions of task demands showed compensation for some of the differences between teachers’ (intended) task demands and the test demands.

On average, teachers’ tests were quite selective. Teachers considered the majority of the sections as important but included less than half of these sections in the test. Still, the tests included sections that teachers considered unimportant. Such discrepancies, however, were quite rare. The probability that unimportant sections appeared in the test was much smaller than the probability that important sections appeared in the test. Moreover, when a relatively unimportant section was addressed in the test, this section had a relatively low weight. Especially because of a low content test coverage, we conclude that the correspondence between teachers’ content demands and the test demands was limited.

Like the teachers’ task demands, students’ test expectations showed a limited correspondence with the test demands. Sections that were rated important had a higher probability of appearing in the test, but the majority of the sections that students deemed important was not included in the test. Furthermore, compared to the teachers’ task demands, students’ test expectations showed more discrepancies with the test demands. Moreover, when students rated sections as unimportant these sections did not have a lower weight in the test. Plainly, students did not know as well as their teacher which sections were important in the upcoming test.

On the other hand, it was found that for sections that teachers considered important, students more often gave high importance ratings when these sections appeared in the test than when they did not appear in the test. The same was true for sections that

---

Note that, averaged across classes, test weight does not explain differences between teachers’ and students’ ratings. Concerning the sections that were rated important by teachers, the weight did not significantly differ for sections considered important by students and those that were considered unimportant by the students ($12 \text{ vs. } .10, \chi^2 \geq 3.25, p = .07$). The same was true for sections that were rated unimportant by teachers ($0.5 \text{ vs. } .05, \chi^2 \geq 0.31, p = .58$).
teachers considered unimportant. This suggests that students compensate for some of
the differences between teachers’ task demands and the test demands. Apart from
teacher’s information about the task demands, they seem to derive task cues from other
sources, such as the textbook or previous tests.

We should also consider the possibility that students may use teacher information
about task demands that is not in tune with the teacher’s own intentions. For instance,
when a teacher dwells on a topic during class discussion, students may infer that this
topic is important for the test, even though the teacher considers it as a sidetrack.
Nevertheless, when this topic appears on the test, students would have been right in
using the implicit teacher information. In future studies, an attempt could be made to
assess both explicit and implicit task information that resides in classrooms and to
systematically compare this information with students’ and teachers’ perceptions of
task demands as well as with the test demands (cf. Green & Weade, 1987; Jetton &
Alexander, 1997; Simpson & Nist, 1997; Winne & Marx, 1982).

So far, we have discussed the general picture for classes. However, considerable
variation among classes was found regarding the number of sections covered in the test
as well as in the correspondence between importance ratings and test content. The
latter variation indicates that differences between teachers’ task demands and the test
demands are more apparent in some classes than in other classes. Similarly,
compensation for such differences varies across classes.

This quantitative study confirms the findings of a qualitative study of Jetton and
Alexander (1997), which indicated that teachers’ intended task demands may deviate
from their actual test demands and that students may compensate for such differences.
Moreover, our study provides information about the generalizability of these findings.
Although the tests generally showed a low content coverage, discrepancies between
teachers’ task demands and the test demands did not occur very frequently.
Furthermore, we found that students compensated for only some of the differences
between teachers’ task demands and the test demands. Test content, then, appears to
fall short in explaining the considerable differences that were found between students’
test expectations and their teachers’ task demands. Such differences seem to be better
explained by limitations in the ways teachers convey task demands to their students
and students’ use of task information (see Jetton & Alexander, 1997; Broekkamp et al.,
2002).

It should be noted that the average degree of correspondence between students’
perceptions of task demands on the one hand and teachers’ intended task demands and
actual test demands on the other hand appeared to be lower in our study than in the
However, whereas we used a rating task based on summaries of a relatively long text, in
the studies of Alexander and associates participants assigned importance to an actual
text, which was considerably shorter. Another methodological explanation is the way
researchers aggregated ratings to obtain a correspondence measure. Whereas
Alexander and associates compared teacher’s ratings with the average ratings of their
class, we compared teacher’s ratings with individual students’ ratings. Our findings are
in line with previous studies that examined instructional importance at the individual
level (e.g., Schellings & Van Hout-Wolters, 1995).

Our findings could possibly be explained by the particular instructional context. For
instance, the low content coverage of tests that we found could be typical for the
school subject that we examined. History can be characterized by relatively long texts
and high levels of autonomy for teachers with respect to defining what they teach and
assess (Stodolsky & Grossman, 1995). Concerning the grade level, our study as well as previous studies on instructional importance have focused on secondary schools. Future studies could examine the relations between perceptions of instructional importance and actual test content at the university level.

The present study focused on a specific kind of task demand, which concerned the relative importance of section topics encompassing three or four paragraphs. For a more complete picture, other kinds of task demands involved in study-test situations are of interest as well. For instance, task demands may pertain to the instructional importance of content elements at lower text levels (e.g., the sentence and paragraph level; cf. Jetton & Alexander, 1997; Schellings & Van Hout-Wolters, 1995), or the processing demands that are associated with the cognitive level of test questions (e.g., memorization versus application of information; see Thomas & Rohwer, 1986). Concerning the cognitive level of test questions, discrepancies between teachers' course objectives and their test demands have been found (Bol & Strage, 1996). Future studies could examine to what degree students' perceptions show a compensation for such discrepancies (cf. Simpson & Nist, 1997).

The present study has implications for classroom testing as well as for the task demands that teachers formulate in relation to test preparation study tasks. It is generally accepted that the congruence of task demands and test demands is a crucial condition for the validity of a test (e.g., see Airasian, 1996; Brookhart, 1999; Dochy & Moerkerke, 1997). In this light, the low content coverage that we found is worrying. It seems that teacher-made tests need to cover higher proportions of the content that teachers deem important in order to obtain valid test scores. Furthermore, as this study and other studies make clear, task demands according to teachers' intentions may differ from those perceived by their students. In order to obtain a valid test, then, teachers have to align their intended task demands both with the students and with the test. Hereby, teachers should take into account that students base their test expectations both on explicit and implicit task information and point their students to possible inconsistencies among task information sources. Finally, teachers should regularly inform themselves whether students' perceptions still match with their own perceptions of the study task.

References


Received 8 January, 2002; revised version received 19 December, 2002
## Appendix A

### Descriptives based on four values of importance ratings

<table>
<thead>
<tr>
<th></th>
<th>( n )</th>
<th>( M )</th>
<th>( SD )</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Students’ importance ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Separate ratings</td>
<td>9732</td>
<td>2.81</td>
<td>0.80</td>
<td>1.00</td>
<td>4.00</td>
</tr>
<tr>
<td>( M ) of ratings across sections</td>
<td>387</td>
<td>2.81</td>
<td>0.31</td>
<td>1.77</td>
<td>3.88</td>
</tr>
<tr>
<td>SD of ratings across sections</td>
<td>387</td>
<td>0.80</td>
<td>0.72</td>
<td>0.15</td>
<td>0.27</td>
</tr>
<tr>
<td><strong>Teachers’ importance ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Separate ratings</td>
<td>489</td>
<td>3.03</td>
<td>0.73</td>
<td>1.00</td>
<td>4.00</td>
</tr>
<tr>
<td>( M ) of ratings across sections</td>
<td>19</td>
<td>3.03</td>
<td>0.27</td>
<td>2.62</td>
<td>3.50</td>
</tr>
<tr>
<td>SD of ratings across sections</td>
<td>19</td>
<td>0.67</td>
<td>0.10</td>
<td>0.50</td>
<td>0.85</td>
</tr>
</tbody>
</table>

*Note.* For both students and teachers descriptives are given for importance ratings of separate sections, the average of a participant’s ratings across sections, and the standard deviation of a participant’s ratings across sections. Average ratings for separate sections (not in table) range from 1.87 and 3.39 for students (\( M = 2.81, SD = 0.34 \)) and from 2.47 to 3.79 for teachers (\( M = 3.03; SD = 0.37 \)).

## Appendix B

### Descriptives of importance ratings, section test occurrence and section test weight per section

<table>
<thead>
<tr>
<th>Section</th>
<th>Student rating</th>
<th>Teacher rating</th>
<th>Section test occurrence</th>
<th>Section test weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( M )</td>
<td>SD</td>
<td>( M )</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>.81</td>
<td>.39</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>2</td>
<td>.86</td>
<td>.35</td>
<td>.68</td>
<td>.47</td>
</tr>
<tr>
<td>3</td>
<td>.76</td>
<td>.43</td>
<td>.80</td>
<td>.40</td>
</tr>
<tr>
<td>4</td>
<td>.44</td>
<td>.50</td>
<td>.75</td>
<td>.43</td>
</tr>
<tr>
<td>5</td>
<td>.75</td>
<td>.43</td>
<td>.83</td>
<td>.38</td>
</tr>
<tr>
<td>6</td>
<td>.91</td>
<td>.28</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>7</td>
<td>.69</td>
<td>.46</td>
<td>.88</td>
<td>.33</td>
</tr>
<tr>
<td>8</td>
<td>.18</td>
<td>.39</td>
<td>.70</td>
<td>.46</td>
</tr>
<tr>
<td>9</td>
<td>.68</td>
<td>.47</td>
<td>.67</td>
<td>.47</td>
</tr>
<tr>
<td>10</td>
<td>.85</td>
<td>.36</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>11</td>
<td>.62</td>
<td>.49</td>
<td>.92</td>
<td>.27</td>
</tr>
<tr>
<td>12</td>
<td>.24</td>
<td>.43</td>
<td>.61</td>
<td>.49</td>
</tr>
<tr>
<td>13</td>
<td>.64</td>
<td>.48</td>
<td>.69</td>
<td>.47</td>
</tr>
<tr>
<td>14</td>
<td>.88</td>
<td>.32</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>15</td>
<td>.77</td>
<td>.42</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>16</td>
<td>.72</td>
<td>.45</td>
<td>.75</td>
<td>.43</td>
</tr>
<tr>
<td>17</td>
<td>.63</td>
<td>.48</td>
<td>.77</td>
<td>.42</td>
</tr>
<tr>
<td>18</td>
<td>.66</td>
<td>.47</td>
<td>.72</td>
<td>.45</td>
</tr>
<tr>
<td>19</td>
<td>.60</td>
<td>.49</td>
<td>.50</td>
<td>.50</td>
</tr>
<tr>
<td>20</td>
<td>.53</td>
<td>.50</td>
<td>.63</td>
<td>.48</td>
</tr>
<tr>
<td>21</td>
<td>.90</td>
<td>.30</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>22</td>
<td>.57</td>
<td>.50</td>
<td>.61</td>
<td>.49</td>
</tr>
<tr>
<td>23</td>
<td>.78</td>
<td>.41</td>
<td>.47</td>
<td>.44</td>
</tr>
<tr>
<td>24</td>
<td>.65</td>
<td>.48</td>
<td>.53</td>
<td>.50</td>
</tr>
<tr>
<td>25</td>
<td>.56</td>
<td>.50</td>
<td>.51</td>
<td>.50</td>
</tr>
<tr>
<td>26</td>
<td>.60</td>
<td>.49</td>
<td>.69</td>
<td>.47</td>
</tr>
</tbody>
</table>

*Note.* Descriptives of importance ratings are based on two collapsed values: rating 1 and 2 = unimportant (0), rating 3 and 4 = important (1). Note that section test weight could be determined only for sections that occurred in the test.
**Appendix C**

*Estimated probability and variance for 2 x 2 x 2 categories of section test occurrence by teachers’ ratings by students’ ratings*

<table>
<thead>
<tr>
<th>Test</th>
<th>Teacher</th>
<th>Student</th>
<th>Probability (log)</th>
<th>SE</th>
<th>Var class SE</th>
<th>Probability (prop)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>–1.06</td>
<td>0.12</td>
<td>0.28*</td>
<td>0.09</td>
<td>.25</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>–</td>
<td>–2.48</td>
<td>0.13</td>
<td>0.31*</td>
<td>0.11</td>
<td>.08</td>
</tr>
<tr>
<td>+</td>
<td>–</td>
<td>+</td>
<td>–3.60</td>
<td>0.37</td>
<td>2.50*</td>
<td>0.84</td>
<td>.03</td>
</tr>
<tr>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–4.60</td>
<td>0.28</td>
<td>1.31*</td>
<td>0.50</td>
<td>.01</td>
</tr>
<tr>
<td>–</td>
<td>+</td>
<td>+</td>
<td>–0.89</td>
<td>0.11</td>
<td>0.22*</td>
<td>0.07</td>
<td>.29</td>
</tr>
<tr>
<td>–</td>
<td>+</td>
<td>–</td>
<td>–1.72</td>
<td>0.10</td>
<td>0.16*</td>
<td>0.06</td>
<td>.10</td>
</tr>
<tr>
<td>–</td>
<td>–</td>
<td>+</td>
<td>–2.21</td>
<td>0.19</td>
<td>0.64*</td>
<td>0.21</td>
<td>.15</td>
</tr>
<tr>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–2.34</td>
<td>0.16</td>
<td>0.44*</td>
<td>0.15</td>
<td>.09</td>
</tr>
</tbody>
</table>

*Note.* Instances of 2 x 2 x 2 categories of section test occurrence (– = non-occurrence, + = occurrence) x teachers’ ratings (– = unimportant, + = important) x students’ ratings (– = unimportant, + = important) are estimated as a proportion of the total number of 9,731 combinations that could be determined. Probabilities are presented both in logits (log) and proportions (prop). For proportions, the range of variance is presented at a 90% probability level.

* Appendix D

*Estimated section test weight and variance for 2 x 2 categories of teacher by students’ ratings*

<table>
<thead>
<tr>
<th>Teacher rating</th>
<th>Student Rating</th>
<th>Weight</th>
<th>SE</th>
<th>Var class SE</th>
<th>SE</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>0.12</td>
<td>0.017</td>
<td>0.0052*</td>
<td>0.0017</td>
<td>.01 – .24</td>
</tr>
<tr>
<td>+</td>
<td>–</td>
<td>0.10</td>
<td>0.014</td>
<td>0.0033*</td>
<td>0.0011</td>
<td>.01 – .19</td>
</tr>
<tr>
<td>–</td>
<td>+</td>
<td>0.05</td>
<td>0.006</td>
<td>0.0004*</td>
<td>0.0002</td>
<td>.02 – .08</td>
</tr>
<tr>
<td>–</td>
<td>–</td>
<td>0.05</td>
<td>0.006</td>
<td>0.0005*</td>
<td>0.0001</td>
<td>.01 – .09</td>
</tr>
</tbody>
</table>

*Note.* The estimated section test weight (the number of test points associated with a section) is presented for 2 x 2 categories of teachers’ ratings and students’ ratings (– = unimportant, + = important) pertaining to sections that occurred in the test. In addition, the estimated between-class variance is given along with its range at a 90 percent probability level.

* p < .05