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

Broekkamp, H.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
CHAPTER 5

ATTUNING TO PROCESSING DEMANDS

A comparison between students’ test expectations, teachers’ intended processing demands and the actual test questions

Abstract
History teachers (N = 22) and their 11th-grade students (N = 451) rated the degree in which four types of questions (verbatim, paraphrase, inference and skill items) were to be expected in an upcoming teacher-made test about a particular textbook chapter. In addition, teachers classified their actual tests according to these four question types. Ratings of individual students showed a low correspondence with those of their teacher and no correspondence with the actual test questions. However, teachers’ ratings also showed a considerable number of discrepancies with the actual test questions. These findings suggest that in study-test situations, more clarity is needed about processing demands.

1 INTRODUCTION

There has been increasing recognition in recent years that studying is contextually determined and varies with characteristics of the learning situation (Van Etten, Freebern, & Pressley, 1997, p. 193). An important category of contextual variables is formed by the processing demands that are associated with an upcoming test. It is assumed that strategic students flexibly adapt their studying to the test questions they anticipate (Pressley, Yokoi, Van Meter, Van Etten, & Freebern, 1997; Ramsden, 1988; Simpson & Nist, 2000; Thomas & Rohwer, 1986). For instance, when students expect questions that require memory of verbatim information, they may rehearse information or use keyword methods. In contrast, when students expect questions that demand integration or deep comprehension, they may self-generate questions, relate the learning materials to prior knowledge, reorganize these materials or recontextualize knowledge (cf. Hadwin, Winne, Stockley, Nesbit, & Woszczyna, 2001; Thomas & Rohwer, 1986). In order to effectively attune to processing demands, students need various skills, including identifying processing demands, selecting demand-appropriate learning strategies, executing these strategies, monitoring their effectiveness and, when necessary, adapting the strategies (Winne & Hadwin, 1998). Of these skills, ascertaining processing demands seems especially important, because students base all further efforts on the demands they perceive

In this study, we examined whether students who prepare for a classroom-test have a clear perception of processing demands.

A clear perception of processing demands is not a matter of course. Students’ perception of processing demands is dependent on both the information that is provided about the processing demands and the way students make use of this information (Winne & Hadwin, 1998). Because teachers vary according to the processing demands they set, students have to tune in to their particular teacher and the test that he or she is going to give (Nolen & Haladyna, 1990). However, because teachers do not want to give away the test questions, they are inclined to be sparing in giving test-related task information (Thomas & Rohwer, 1986, 1993). In addition to teacher information, students may rely on other sources of information on which to base their test expectations (cf. Doyle, 1983; Pressley et al., 1997; Van Etten et al., 1997). For example, the emphasis of study materials on a particular kind of processing may be indicative of the cognitive level of test questions that will be asked in the test. Similarly, students may infer from previous tests the type of test questions that their teachers are likely to ask in the upcoming test. However, the predictive value of these sources may differ across study-test situations. Although students indicate that a clear perception is important to them (Van Etten et al., 1997), it is uncertain whether they generally have a clear perception of processing demands when they prepare for a classroom-test.

Experimental research tends to show that students perform better on the type or format of test that they were led to expect (e.g., for reviews see Crooks, 1988; Lundeberg & Fox, 1991). This indirectly indicates that students’ perception is clear enough to adapt their studying according to these demands. However, most of the experimental studies have been conducted in laboratory settings (see Lundeberg & Fox, 1991; Crooks, 1988). Moreover, the experimental studies have been performed in classrooms involved conditions that differ from fully realistic classroom contexts in important ways (cf. Alvermann & Moore, 1991; Crooks, 1988; Lundeberg & Fox, 1991; Thomas & Rohwer, 1986). First, because studying was not embedded in the regular curriculum it was not accompanied by other activities such as following lectures or doing exercises. Second, the experimentally manipulated processing demands were typically more clearly defined than processing demands in realistic classroom contexts, where criteria of studying can be “shrouded in secrecy” (Thomas & Rohwer, 1986, p. 21). Third, demands in the experiments were mostly of one type or format (e.g., only lower order multiple choice items), whereas classroom tests often combine different types of questions and formats.

In the last two decades, experimental studies on studying have made room for an increasing number of descriptive classroom studies (Nist & Simpson, 2000; Thomas & Rohwer, 1993). Several of these studies have examined students’ self-reported study approaches across different contexts. A major finding is that study approaches are both consistent and variable across contexts (disciplines, departments, grade levels, courses, teachers, tasks) suggesting that students have stylistic preferences but also strategically adapt their studying to contexts (e.g., Hadwin et al., 2001; Trigwell, Prosser, & Waterhouse, 1999; Ramsden & Entwistle, 1983; Thomas & Rohwer, 1987; Thomas et al., 1993; Vermetten, Lodewijks, & Vermunt, 1999; see Vermetten, 1999).
Thomas and associates examined self-reported study activities in relation to test demands as a separate course characteristic. In one study, they examined social studies courses of various grade levels (Thomas & Rohwer, 1987). Across these grade levels, a general increase of higher level test demands was accompanied by an increase of higher level study strategies (e.g., relating information to previous knowledge). However, between courses within grade levels, no significant relations between test demands and study activities could be demonstrated. In a follow up study involving university biology courses (Thomas et al., 1993) the negative findings of Thomas and Rohwer (1987) were replicated. A possible explanation for these findings is that students had test expectations that were not in tune with the test demands (cf. Christopoulos, Rohwer, & Thomas, 1987, p. 322).

Thomas and associates did not include measures of student perceptions in their studies. Classroom research on studying that did include such measures, mostly focused on broad contextual characteristics, such as the perceived quality of instruction concerning departments (e.g., Ramsden & Entwistle, 1983), courses (e.g., Lawless & Richardson, 2002; Trigwell & Prosser, 1991) or study tasks given within courses (e.g., Eley, 1992). At a more specific level of course characteristics, Smith and Feathers (1983) conducted an interview study in which they compared the perceptions of three social studies teachers and their 8th-grade students concerning the relative importance of course objectives, including higher order thinking skills and factual knowledge. In addition, the tests that teachers gave to their students were analyzed according to the cognitive level of test questions. Teachers’ objectives showed large discrepancies with their test demands. Whereas according to their reports, teachers demanded higher order thinking, their tests mainly required reproduction of information. Such discrepancies have also been found in other studies that included teachers’ objectives and test demands – but not students’ perceptions (Ball, Doss, Dewalt, 1986; Bol & Strage, 1996). Interestingly, Smith and Feathers (1983) found that students’ perceptions of course objectives did correspond with the test demands instead of with their teachers’ objectives. This suggests that students did not rely on teacher information about processing demands but used other sources of task information, such as previous tests this teacher gave. These sources would have enabled them to “compensate” for discrepancies between teacher’s intended processing demands and the test demands.

As an exception among the classroom studies that included student perceptions, Simpson and Nist (1997) focused on processing demands of single study tasks. These researchers performed a case study in which they compared perceptions of university students who prepared for tests in a social study course with the demands that their teacher imposed upon them. Interviews with the teacher, lesson observations and analysis of the teacher’s tests showed that the teacher demanded especially higher level processing of his students, such as “thinking in conceptual terms”. The researchers were able to distinguish between three types among the ten students examined: (a) three students attuned to processing demands from the start, (b) three students learned to attune to processing demands during the course, and (c) four students were not attuned to processing demands.

An important contribution of the qualitatively oriented studies by Simpson and Nist (1997) and Smith and Feathers (1983) is that they demonstrated that discrepan-
cies exist between teachers' intended processing demands, students' test expectations and the cognitive level of test questions. However, given the small sample sizes of these studies, we cannot assume that these findings are representative for classrooms in general (cf. Wineburg, 1996, p. 432).

To examine the generalizability of the findings of Simpson and Nist (1997) and Smith and Feathers (1983), we conducted a quantitative oriented study. Our focus is restricted to processing demands that are associated with the cognitive level of test questions that students are to expect (e.g. the relative importance of reproductive questions versus higher order questions). However, our study is part of a larger investigation on (perceptions of) task demands that also includes content demands (i.e., the relative importance of textual topics; see Broekkamp, Van Hout-Wolters, Rijlaarsdam, & Van den Bergh, 2002; Broekkamp, Van den Bergh, Van Hout-Wolters, Rijlaarsdam, submitted). In the discussion section, we will discuss briefly findings pertaining to content demands in relation to findings pertaining to processing demands obtained in the present study.

We examined four questions (a) To what degree do teachers, who give a teacher-made test about the same learning materials, vary in their intended processing demands regarding students' test preparation? (b) To what degree do teachers' intended processing demands correspond with the actual cognitive level of test questions? (c) To what degree do individual students' test expectations correspond with their teachers' intended processing demands? (d) To what degree do students' test expectations regarding processing demands correspond with the actual cognitive level of test questions?

2 METHOD

2.1 Participants and context

Participants in this study were 22 history teachers and 451 11th-grade students. Teachers, who were from 17 schools in the Netherlands, each participated with one class of students. The average number of participating students in each class was 20.50 (SD = 6.86). Teachers, on average, had 17.6 years of teaching experience (SD = 8.45) and 16 of them were men. Students, who had a typical age of 17, were in the penultimate grade of the highest stream of Dutch secondary education that prepares for university. 253 of the students were female. The recruited teachers (and their classes) used a particular history textbook, which had the highest market share in the Netherlands and 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 processing demands effectively (cf. Simpson & Nist, 1997; Van Etten et al., 1997). The chapter had "The United States of America since 1945" as its theme and contained approximately 8,000 words.
2.2 The rating task

Teachers and students rated the degree in which four types of questions were to be expected in the upcoming test: verbatim, paraphrase, inference and skill items. The question categories and the corresponding labels were derived from a pilot study, in which they were identified as connecting most closely with the distinctions and terminology used by teachers and students. In the present study they were presented to the participants in the following way:

Below are four types of questions that might be posed in a test. The question types differ in the way you have to show what you have learnt.

1. Verbatim knowing
Show that you literally know what is in the learning materials

2. Telling it in your own words
Show that you understand what is in the learning materials by retelling it in your own words

3. Thinking through
Show that you can think through the information given in the learning materials by making connections and drawing conclusions

4. Applying a skill
Show that you can solve problems by taking the right steps.

The first three categories appear to be largely compatible with categorizations used by other researchers. For instance, Thomas and Rohwer (1987) distinguished between items requiring generation of verbatim, interpreted constructed information. Similarly, Doyle (1983) distinguished between verbatim, paraphrase and inference items. Although Doyle (1983) did not mention a category of skill test items, he distinguished procedural tasks, requiring domain specific skills, from comprehension and memory tasks, both requiring declarative knowledge. The difference between the question types is a matter of degree. For instance, whereas inference questions require content knowledge in combination with general reasoning skills, skill questions demand especially domain specific skills (like interpreting original historical texts or statistical information); content knowledge is secondary or provided in the question.

From talks with the students, it appeared that the importance of processing demands was mostly described in terms of the number of questions that were to be expected. Students, for instance, told us that they expected “many questions” that required thinking through. To connect with these descriptions, students in the present study were asked to indicate for each question type the number of questions they expected for the upcoming test. Similarly, teachers were asked to indicate for each category the number of questions students should in principle expect for the upcoming test. Both the students and the teachers used a 4-point scale that distinguished between “no questions” (1), “few questions” (2), “quite a lot of questions” (3) and “a great many questions” (4). To the teachers, we stressed that their ratings were to reflect intended processing demands (the questions that students are to expect), instead of actual test demands (the questions that will actually be asked in the test).
To check whether teachers and students understood the four categories as we intended them, a small “categorization test” was given before the rating task. The categorization test required participants to classify four sample questions according to the four categories (see Appendix A). The sample questions were about a preceding chapter (The United States during 1920–45) and were presented to the participants along with model answers and explanations that indicated the degree to which the required knowledge could be found in the learning materials. Sample questions were constructed in a way that classification was not too easy. Furthermore, by instructing participants to use each question category once, they were encouraged to think more deeply about the meaning of the categories. Thus, apart from a check of understanding, the classification task provided extra instruction to the participants. The outcomes of the categorization test indicated that 100% of the teachers and 73% of the students coded all four sample questions correctly. Incorrect responses mostly concerned distinctions between the adjacent categories verbatim and paraphrase items and between inference and skill items. When these responses are considered as correct, 90% of the students coded all four sample questions correctly.

The rating task was administered within two days before the actual classroom test, during a history lesson (16 classes) or during a separate session (6 classes). Students and their teachers, at the same time but independently, first made the short test with the four sample items and then rated the degree in which the four question types were to be expected. Both student and teacher ratings showed missing values in 3% of the cases (i.e., no rating was given).

2.3 The collection and content analysis of teacher-made tests

The categorization of the actual test questions according to the four question types was performed by the teachers themselves. We assumed that teachers were well equipped to categorize test questions. First, teachers had generally access to all information that was provided in classrooms (i.e., not only information provided by the textbook chapter, but also additional information such as that provided orally during lectures). Furthermore, teachers knew the kinds of answers that they would accept of their students. We were aware of a study by Bol and Strage (1996) that showed that biology teachers tend to underestimate the level of reproduction of their tests. The biology teachers were asked to give a global estimation of the number of test items students had seen or practiced in advance of the test. However, in our study, we asked teachers to classify single test items.

After the rating task, teachers placed their test materials on the table, which they were asked to take to the session. These materials included the test questions, the model answers to these questions and the distribution of test points. Teachers were asked to code their own test questions according to the four categories. Eventually 19 different teacher-made tests were collected and codings of 17 teachers could be included in the analysis.

As 4 teachers had not yet constructed their tests, they gave test materials to the researchers at a later stage. Of these 4 teachers only 2 responded to the request to code their test items. Furthermore, some teachers belonging to the same school shared tests. To increase the comprehensibility of the statistical
The importance or weight of question types was based on the maximum number of test points that students could obtain for individual questions. For each of the question types, the number of possible points for the corresponding questions was determined as a proportion of the total number of points in the test. For instance, in a certain test .20 of the possible test points could be earned with verbatim items, .10 with paraphrase items, .40 with inference items and .30 with skills items.

Unexpectedly, in 16% of the codings, teachers deviated from the instruction to assign the most suitable code per item and assigned two or more codes for a single test question. In 59% of the double codings, a combination of inference and skill items was assigned. Other double codings not only involved combinations of adjacent categories, like verbatim and paraphrase items but also combinations of remote categories such as verbatim and inference items. To determine the number of test points associated with double codings, both codings were considered as valid with the number of test points divided proportionally (e.g., a single inference-skill item with a maximum of eight test points yielded four points for inference demands and four points for skill demands).

3 RESULTS

3.1 Teachers' intended processing demands

According to the average teachers' ratings (see Table 1), teachers considered important especially questions that require inference from the learning materials. On average, students were to expect "quite a lot of questions" of this type ($M = 3.00$). Questions that required verbatim or paraphrased reproduction were deemed least important; on average, students could expect only "few questions" of this type (verbatim items: $M = 1.95$; paraphrase items: $M = 2.14$). Skill items were considered more important than verbatim and paraphrase items but less important than inference items ($M = 2.67$).

The individual scores (see Table 2) show that whereas according to teachers, students in all classes had to expect inference and skill questions, students should not always expect verbatim and paraphrase items. On the other hand, 12 of the 17 teachers indicated that students were to reckon with all four question types. Generally, there was variation among teachers especially in the number of test questions that their students were to expect for the four question types.

analyses involving test variables, we included only teachers (and their classes) who made or chose the test questions themselves. As a consequence, for each of the 3 schools that participated with 2 teachers, only 1 teacher was included in these analyses. This was the teacher that was known to have constructed or chosen the test questions. There was also one school that participated with 3 teachers. Because each teacher gave their own test, all 3 teachers were included in the analyses.

$^2$ A paired t test indicated that differences in mean ratings between the types of questions were all significant ($t \geq [2.94], df = 19, p \leq .008$), except for the difference between verbatim and paraphrase items ($t = [1.00], df = 20, p = .329$).
### Table 1. Descriptives for teachers’ intended processing demands, teachers’ actual test demands and students’ test expectations

<table>
<thead>
<tr>
<th>Processing demand variables</th>
<th>N</th>
<th>M</th>
<th>SD between teachers/classes</th>
<th>SD between students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers’ intended demands</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbatim items</td>
<td>22</td>
<td>1.95</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>Paraphrase items</td>
<td>21</td>
<td>2.14</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>Inference items</td>
<td>21</td>
<td>3.00</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>Skill items</td>
<td>21</td>
<td>2.67</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>Teachers’ actual test demands</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbatim items</td>
<td>17</td>
<td>.26</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Paraphrase items</td>
<td>17</td>
<td>.16</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>Inference items</td>
<td>17</td>
<td>.32</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Skill items</td>
<td>17</td>
<td>.26</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>Students’ test expectations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbatim items</td>
<td>445</td>
<td>2.12</td>
<td>0.12</td>
<td>0.48</td>
</tr>
<tr>
<td>Paraphrase items</td>
<td>447</td>
<td>2.68</td>
<td>0.18</td>
<td>0.57</td>
</tr>
<tr>
<td>Inference items</td>
<td>449</td>
<td>3.06</td>
<td>0.17</td>
<td>0.59</td>
</tr>
<tr>
<td>Skill items</td>
<td>448</td>
<td>2.72</td>
<td>0.30</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Note. Teachers’ intended processing demands and students’ test expectations are indicated by ratings that teachers and students gave to four types of questions, respectively. With these ratings, participants expressed the number of questions that were to be expected in the upcoming test (1 = “no questions, 2 = “few questions”, 3 = “quite a lot of questions”, 4 = “a great many of questions”). Teachers’ actual test demands indicate the weight that questions had in the test (i.e. the average proportion of test points that was associated with a question type). Teachers’ intended processing demands and actual test demands represent observed values. Students’ test expectations represent estimated values. Multilevel techniques were used to distinguish between the differences in ratings between classes and the differences in ratings between students (within classes). All variance estimates were significant at a .05 probability level. Statistical significance was determined by taking the ratio of the variance estimate and its standard error (which is t-distributed) and using a one-sided alpha level of .05, t > 1.658 (cf. Goldstein, 1995).

### 3.2 Teachers’ actual test demands compared with teachers’ intended processing demands

In addition to the average teachers’ ratings, indicating teachers’ intended processing demands, Table 1 also included average proportions based on their test codings, indicating actual test demands. Although the ratings performed on a 4-point scale cannot be directly compared with the test proportions, there seems to be a discrepancy between teachers’ ratings for verbatim items and the proportion of this question type in the actual test. Whereas teachers indicated that their students could expect few verbatim questions ($M = 1.95$), the actual tests indicate that students could earn .26 of the total of test points by answering these questions correctly. Such clear discrepancy cannot be demonstrated for the other question types. Paraphrase items, which were rated as relatively unimportant, took in .16 of the possible test points. Inference
items, rated as most important, had an average weight of .32. Finally, the proportion of skill items was .26.

Table 2. Individual scores of teachers’ intended processing demands and their actual test demands

<table>
<thead>
<tr>
<th>Nr</th>
<th>Verbatim</th>
<th>Paraphrase</th>
<th>Inference</th>
<th>Skill</th>
<th>Verbatim</th>
<th>Paraphrase</th>
<th>Inference</th>
<th>Skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>.22</td>
<td>.08</td>
<td>.55</td>
<td>.15</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>.23</td>
<td>.36</td>
<td>.13</td>
<td>.28</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>.00</td>
<td>.11</td>
<td>.67</td>
<td>.22</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>.33</td>
<td>.16</td>
<td>.22</td>
<td>.29</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>mis</td>
<td>mis</td>
<td>mis</td>
<td>mis</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>.23</td>
<td>.03</td>
<td>.26</td>
<td>.48</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>mis</td>
<td>mis</td>
<td>mis</td>
<td>mis</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>.62</td>
<td>.00</td>
<td>.23</td>
<td>.15</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>.43</td>
<td>.00</td>
<td>.38</td>
<td>.19</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>.36</td>
<td>.10</td>
<td>.32</td>
<td>.22</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>2</td>
<td>mis</td>
<td>mis</td>
<td>.36</td>
<td>.33</td>
<td>.12</td>
<td>.19</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>.08</td>
<td>.36</td>
<td>.18</td>
<td>.38</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>.13</td>
<td>.20</td>
<td>.34</td>
<td>.33</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>.46</td>
<td>.00</td>
<td>.44</td>
<td>.10</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>mis</td>
<td>mis</td>
<td>mis</td>
<td>mis</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>.40</td>
<td>.30</td>
<td>.15</td>
<td>.15</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>.30</td>
<td>.20</td>
<td>.25</td>
<td>.25</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>.12</td>
<td>.08</td>
<td>.36</td>
<td>.44</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>mis</td>
<td>4</td>
<td>3</td>
<td>.10</td>
<td>.20</td>
<td>.41</td>
<td>.29</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>mis</td>
<td>mis</td>
<td>mis</td>
<td>mis</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>.10</td>
<td>.19</td>
<td>.48</td>
<td>.23</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>mis</td>
<td>mis</td>
<td>mis</td>
<td>mis</td>
</tr>
</tbody>
</table>

Note. Teachers’ intended processing demands are indicated by ratings that teachers gave to four types of question (1 = “no questions”, 2 = “few questions”, 3 = “quite a lot of questions”, 4 = “a great many of questions”); teachers’ actual test demands is based on teachers’ categorization of their own test questions and the proportion of test points associated with these questions. Nr = number of respondent. Mis = missing value.

The individual scores (see Table 2) show that 13 of the 17 teachers included all four question types. However, as can be seen from both the individual scores (Table 2) and the standard deviations (Table 1), for each of the question types, there was con-

A paired t-test indicated that paraphrase items represented a significantly smaller proportion of the tests than inference items and skill items (t ≥ 2.62, df = 16, p ≤ .019); other comparisons yielded non-significant t-values (t ≤ 1.80, df = 16, p ≥ .090).
siderable variation among the teachers regarding the number of questions that students were to expect for these question types.

To examine the relations between individual teachers’ intended processing demands and their test demands, we calculated 16 Pearson correlations between teachers’ ratings and the proportions based on their test codings (4 ratings X 4 proportions; see Table 3, lower left quadrant). Teachers who indicated that their students were to reckon with a relatively large number of verbatim items, also showed more weight for this kind of question in the actual test \((r = .57, p = .017)\). Similarly, a positive relation was found between teachers’ codings for skill items and the weight of skill items in the test \((r = .49, p = .040)\). For the other comparisons between teachers’ ratings and their test demands, however, no significant correlations were found \((- .37 \leq r \leq .37, p \geq .082)\). These findings indicate that in many cases teachers’ intentions were discrepant with their actual test questions.

Table 3. Pearson correlations between teachers’ intended processing demands and their test demands

<table>
<thead>
<tr>
<th>Teachers’ intended demands</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Verbatim items</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Paraphrase items</td>
<td>-.33</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Inference items</td>
<td>-.37</td>
<td>.20</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Skill items</td>
<td>-.28</td>
<td>.01</td>
<td>.57**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teachers’ test demands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Verbatim items</td>
<td>.57*</td>
<td>-.09</td>
<td>-.32</td>
<td>.02</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Paraphrase items</td>
<td>-.37</td>
<td>.37</td>
<td>-.01</td>
<td>-.07</td>
<td>-.34</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Inference items</td>
<td>-.08</td>
<td>-.20</td>
<td>.20</td>
<td>-.29</td>
<td>-.41</td>
<td>-.53*</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>8 Skill items</td>
<td>-.34</td>
<td>-.01</td>
<td>.24</td>
<td>.49*</td>
<td>-.56</td>
<td>.13</td>
<td>-.19</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note. Teachers’ intended processing demands and teachers’ test demands are indicated by ratings and proportions for four question types, respectively. The correlations in the lower left quadrant are of primary interest: they represent comparisons between teachers’ intended processing demands and teachers’ test demands. The correlations in the upper left quadrant and the lower right quadrant represent comparisons between question types within teachers’ intended processing demands and teachers’ test demands, respectively. They provide indications for the discriminant validity of the four categories of question types. Because we expected positive relations between corresponding pairs (e.g. the correlation between teachers’ intended processing demands for verbatim items and teachers’ test demands for verbatim items), we determined the significance of correlations for these pairs by using a one-sided alpha level of .05. For non-corresponding pairs (e.g. the correlation between teachers’ intended processing demands for verbatim items and teachers’ test demands for paraphrase items) we used a two-sided alpha level of .05. * \(p < .05\), ** \(p < .01\).

3.3 Students’ test expectations compared with teachers’ intended processing demands

Whereas for analyses restricted to teacher variables, we could rely on observed scores, for analyses including students’ test expectations, we had to make use of multilevel estimates to take into account the fact that students were nested within
classes. These estimates allowed us to distinguish between the differences in ratings among classes and the differences in ratings among students (within classes) simultaneously. Furthermore, we could determine the significance of these differences by relating them to appropriate estimates of standard errors. Finally, estimated sample means gave an accurate representation of the sample because they were estimated simultaneously with the class and student level variation, taking into account, in an appropriate way, the number of students per class and the few missing values in the data (cf. Goldstein, 1995). To obtain multilevel estimates we used the software ML-WIN (Multi Level Project, 1999).

Table 1 includes the average (estimated) ratings of students for verbatim, paraphrase, inference and skill items. A comparison with the average teacher ratings (also included in Table 3) indicates the degree of correspondence between the two samples. Averaged over the four question types, students assigned somewhat higher ratings than teachers did \( (M = 2.65; M = 2.44; \text{not in Table}) \). Especially for items that demand paraphrasing the learning materials, students expected more questions than was intended by the teachers \( (M = 2.68; M = 2.14) \). On the other hand, the average ratings indicate that students agreed with their teachers that inference items were most important \( (\text{students}: M = 3.06; \text{teachers}: M = 3.00) \) and verbatim items were least important \( (\text{students}: M = 2.12; \text{teachers}: M = 1.95) \). Furthermore, skill items also received a rating that was quite similar to the teachers’ rating \( (\text{students}: M = 2.72, \text{teachers}: M = 2.67) \). In summary, the samples of teachers and students showed a reasonable correspondence in their rating pattern\(^4\).

### Table 4. Distribution of student rating values over the question categories

<table>
<thead>
<tr>
<th>Rating</th>
<th>Verbatim</th>
<th>Paraphrase</th>
<th>Inference</th>
<th>Skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>No questions</td>
<td>22</td>
<td>5</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Few questions</td>
<td>352</td>
<td>153</td>
<td>73</td>
<td>174</td>
</tr>
<tr>
<td>Many questions</td>
<td>64</td>
<td>261</td>
<td>277</td>
<td>187</td>
</tr>
<tr>
<td>A great many of questions</td>
<td>7</td>
<td>28</td>
<td>99</td>
<td>74</td>
</tr>
<tr>
<td>Missing</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Note. Student ratings express the number of questions that students \( (N=451) \) expected in the upcoming test for each of the four question categories (see Table 1).

The distribution of rating values over the four question categories (see Table 4) shows that the rating “no questions” was assigned in only 40 of the 1789 valid cases. This indicates that the large majority of the students reckoned with each of the four question types. However, the standard deviations (see Table 1) indicate that there were considerable differences among students regarding the number of question they expected for these four types. For all of the question types, the variation in ratings could in part be attributed to the class students belonged to \( (SD = 0.12, p = .040) \).

\(^4\)A paired t-test showed that all differences between students’ ratings for the question types were significant \( (t \geq 8.33, df = 446, p < .001) \) except for the difference between paraphrase and skill items \( (t \leq 0.38, df = 445, p \geq .703) \).
This finding is important as it suggests that, on average, students shaped their test expectations to the specific context they were in. Nevertheless, the major part of the variation in ratings was found at the student level. This means that within classes, ratings showed considerable differences between students ($SD = 0.48, p < 0.001$). Interestingly, compared with the other three question types, skill items showed both the highest variation at the class level ($SD = 0.30$) and the student level ($SD = 0.70$). Apparently, the processing demands that are associated with skill questions were very different between classes. These differences expressed themselves in student expectations despite the fact that students showed relatively low consensus about the importance of this question type.

To examine whether individual students shaped their expectations to the processing demands intended by the particular teacher, we determined the regression of student ratings on teacher ratings (see Table 5). In total, five of the 16 regression coefficients that were determined (4 teacher ratings X 4 student ratings) reached a significant value$^5$. Remarkably, only two significant regression coefficients concerned a relation between corresponding pairs of variables, the regression of student ratings for paraphrase items on teachers' ratings for paraphrase items ($\beta = 0.215, p < .001$) and the regression of student ratings for inference items on teachers' ratings for inference items ($\beta = 0.176, p = .030$). When teachers considered paraphrase items as relatively important, students of these teachers tended to consider this type of questions important as well. The same was true for inference items.

The other three significant regressions concerned students' ratings for paraphrase items and teachers' ratings for inference items ($\beta = -0.220, p = .014$), students' ratings for skill items and teachers' ratings for inference items ($\beta = 0.433, p < .001$) and students' ratings for inference items and teachers' ratings for skill items ($\beta = -0.197, p = .037$). The two negative regressions suggest that teachers' information about processing demands was interpreted by the students in a complementary way. For example, when the teachers stressed inference items, students might have learned that it was insufficient to paraphrase items from the learning materials and that they had to think more deeply about the learning materials. The positive regression suggests that students and teachers interpreted the adjacent categories of inference and skill differently: for instance, when teacher indicated to their students that they were to think more deeply about the learning materials, students might have thought that especially skills were required and that knowledge of the learning materials was secondary.

---

$^5$ Of the 16 regression values, we were primarily interested in the regressions that concern complementary pairs (i.e., verbatim-verbatim, paraphrase-paraphrase, inference-inference en skill-skill). Because we expected these regressions to have positive values, we tested them at a one-sided .05 probability level.
Table 5. Estimated regression of students’ test expectations on teachers’ intended processing demands

<table>
<thead>
<tr>
<th>Students’ test expectations</th>
<th>Teachers’ intended processing demands</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Verbatim</td>
</tr>
<tr>
<td>Verbatim items</td>
<td>0.082</td>
</tr>
<tr>
<td>Paraphrase items</td>
<td>0.043</td>
</tr>
<tr>
<td>Inference items</td>
<td>-0.113</td>
</tr>
<tr>
<td>Skill items</td>
<td>-0.003</td>
</tr>
</tbody>
</table>

Note. Students’ test expectations and teachers’ intended processing demands are indicated by ratings these participants gave to four types of questions (see Table 1). The regression of students’ ratings on teachers’ ratings were estimated in a multilevel way, accounting both for differences in ratings between classes and between students. Statistical significance of the regression estimates was determined by taking the ratio of the regression estimate and its standard error (which is t-distributed; cf. Goldstein, 1995). We used a one-sided alpha level of .05 (t > 1.658) for corresponding pairs (e.g. the regression of student ratings for verbatim items on teacher ratings for verbatim items) and a two-sided alpha level of .05 (t > 1.965) for non-corresponding pairs (e.g. the regression of student ratings for verbatim items on teacher ratings for paraphrase items) (cf. Goldstein, 1995). * p < .05. ** p < .01.

In general, the regression values were rather low. This becomes clear when effect sizes are calculated. Teacher ratings for skill items explain only 10% of the variation in student ratings for inference items. For the other significant relations, the explained variance is even lower. Thus, we could not demonstrate any strong relationship between students’ test expectations and their teachers’ intended processing demands.

3.4 Students’ test expectations compared with teachers’ actual test demands

The limited correspondence between ratings of individual students and their teachers suggests that teachers did not provide sufficient or clear information about processing demands to their students. Apart from information given by the teacher, however, students may also have relied on other sources of information upon which to base their test expectations (e.g., previous tests). This would make it possible that

* The size of effect of is calculated as the the square of the standardized regression coefficient. To calculate a standardized regression coefficient, the unstandardized coefficient is multiplied by the standard deviation of the independent variable, the teacher ratings, and then divided by the total standard deviation of the dependent variable (\( \beta * SD_{teacher} / SD_{student} \)). For instance, the unstandardized regression of students’ ratings for paraphrase items and teachers’ ratings for paraphrase items is 0.215 (Table 3). The standard deviation of teachers’ ratings for paraphrase items is 0.57 (Table 1). The standard deviation of students’ ratings for paraphrase items at the class and student level is 0.18 and 0.57, respectively (Table 1). The standardized regression coefficient is 0.205 (\( (0.215 * 0.57) / (0.18^2 + 0.57^2)^{1/2} \)). The effect size is .042 (0.205^2). So, only 4% of the variation in student ratings for paraphrase items is explained by teacher ratings for paraphrase items.
students' test expectations were better aligned with the test demands than with teachers' intended demands. To examine this possibility, we determined 16 regression coefficients expressing the relation between students' test expectations and teachers' test demands (4 ratings X 4 proportions). None of these values reached significance (p ≥ .076). Thus, we could not demonstrate any relationship between students' test expectations and the cognitive level of actual test questions.

4 DISCUSSION

This study focused on the relative importance of processing demands according to teachers' intentions, students' test expectations and actual teacher-made classroom tests. Importance was operationalized in terms of the number of questions or possible test points for different kinds of test items: verbatim, paraphrase, inference and skill items.

4.1 Teacher's intended processing demands and actual test demands

Although most teachers indicated that students could expect all four types of question, teachers varied considerably regarding the number of questions that students were to expect for these question types. Such differences in intended demands may be explained by various contextual characteristics (cf. Grossman & Stodolsky, 1994; Stark, 2000), including teachers' knowledge (Wilson & Wineburg, 1988, 1993) and their classroom context (Bol, Stephenson, Numery, & O’Conell, 1998; Fleming, Ross, Tollefson, & Green, 1998). For students, differences among teachers regarding processing demands underline the necessity to tune in to the information that their teachers provide about the processing demands. However, such attuning makes sense only when teachers' intended processing demands are consistent with the test demands.

Teachers' tests were consistent with teachers' intended processing demands in that they generally included each of the four question types. However, concerning the relative importance of question types, we found considerable inconsistencies. Whereas teachers, on average, indicated that their students were to reckon with "a few" verbatim questions, the actual test indicated that students could earn a quarter of the total of test points by answering these questions correctly. Moreover, a rather limited correspondence was found between teachers' individual ratings for the four question types and the weight that these questions had in this test.

The relatively large weight that verbatim question had in the test, when compared with teachers' intended processing demands, replicates previous research in which course objectives of social studies teachers (Ball et al., 1986; Smith & Feathers, 1983) and biology teachers (Bol & Strage, 1996) were of a higher cognitive level than their actual test questions. The present study suggests that such discrepancies may involve not only course objectives, but also objectives that pertain to a specific study task (i.e., intended processing demands).

To determine the regression of student ratings on teacher test codings, two multilevel models were needed for reasons of collinearity (test codings for the four question types sum up to 1.00).
The limited correspondence between teacher's intended demands and their actual test demands can be explained in at least three ways. First, teachers may not have had clearly defined intentions regarding processing demands. The rating task asked teachers to make their processing demands explicit. However, it is uncertain whether teachers had such an explicit representation of processing demands when teaching or when making the test items. Second, teachers could have failed to construct or choose a test that was representative of their processing demands, for example, because they did not consider carefully enough the processing activities required by their test items. Bol and Strage (1996) showed that when asked to give a global estimation of their test items, teachers tend to underestimate the level of reproduction. Third, teachers' intended processing demands may have reflected the formal curriculum and not the informal curriculum.

Concerning the third explanation, we asked teachers to express with their ratings both the likelihood that question types would appear in the test and the number of questions when tested. Although, with this instruction, we encouraged teachers to report realistic demands, we should take into account the possibility that teachers reported desired or formal demands instead of actual (i.e., informal) demands that they imposed upon students. For instance, Doyle (1983) described the situation in which teachers want to encourage higher level processing by their students but eventually choose for reproductive tests because of the pressure that students exert on the teacher to make the test not too difficult.

The distinction between formal and informal demands (cf. Doyle, 1992; Erickson & Shultz, 1992; Gehrke, Knapp, & Sirotnik, 1992) evokes the question of whether teachers' intended processing demands were more representative of the processing demands imposed upon students than the actual test demands. To get more insight into this question, we have to consider students' perceptions of processing demands (i.e., test expectations) as well.

4.2 Students' test expectations

Averaged across classes, students showed a similar rating pattern as the group of teachers, except for paraphrase items, which students considered clearly more important than was intended by the teachers. The systematic variation in student ratings that was found between classes suggests that students shaped their expectations to the specific context they were in. However, only a quite limited correspondence was found between ratings of individual students and those of their teachers. Moreover, we were not able to demonstrate a relation between ratings of individual students and the actual test questions that were given by their teachers.

The fact that we could not find a relation between students' test expectations and the test demands can be explained by the limited correspondence between teachers' intended processing demands and the test demands. When the teacher's intended processing demands are unrepresentative for the test demands, it becomes difficult for students to ascertain the demands of the upcoming test questions. Students could also obtain important data about processing demands from sources other than the teacher (e.g., previous tests, textbook exercises). However, the fact that we could not
demonstrate a relation between students’ test expectations and the test demands sug-
gests that these alternative sources did not fulfill a compensatory function.

The discrepancies that we found between teachers’ intended processing de-
mands, students’ test expectations and actual test demands leads us to conclude that
the relative importance of processing demands was generally unclear in the class-
rooms we examined. This does not mean that we exclude the possibility that some
students did have a clear perception of their teacher’s intended demands. However,
since we did not have a reliable criterion to ascertain what the more important proc-
essing demands really were in classrooms, we could not examine whether some stu-
dents had a clearer perception of processing demands than other students. Conse-
quently, we could not determine, in a meaningful way, the relationship between stu-
dents’ test expectations and students’ test performance (cf. Broekkamp, Van den
Bergh, Van Hout-Wolters, & Rijlaarsdam, 2002; Thomas & Rohwer, 1987)

4.3 Considerations for future research

Future research could optimize the methodology by which we obtained our findings. At least for secondary school history classes, these studies could make use of our four categories of test questions. The patterns of ratings and test proportions support the discriminant validity of the four categories. Next to differences at the aggregated level (e.g., inference items on average were considered more important than verba-
tim items), the categories did not show strong correlations⁸, indicating that all four
categories were needed to capture the relative importance of processing demands. Apart from skill items, our category system is quite similar to the three-category system used by Thomas and Rohwer (1987) to analyze social study tests. These re-
searchers found substantial proportions for question types requiring “verbatim, in-
terpreted and constructed information”, respectively. In contrast, studies that used
the category system of Bloom (1956) classified practically all questions in social
study classroom tests to the single category denoting “knowledge of terms or facts”
(Fleming & Chambers, 1983; Ball et al., 1986). Although these findings may reflect
the particular classroom contexts examined, the category system of Bloom seems
less suitable to distinguish between the diversity of test questions asked in social
study courses.

Although we suggest maintaining our categories, we feel that the scales that re-
late to these categories could be improved. Future studies could let participants rate,
separately, the likelihood that question types appear in the test and the weight that

---

⁸ Correlations between the four question types for individual teachers (both regarding intended demands
and test demands) are included in Table 3. To interpret the correlation values, it should be taken into
account that the four categories to some degree are inherently dependent on each other. For example,
teachers who indicate that students should expect a great many verbatim items are likely to expect fewer
items of the other question types. Moreover, this kind of dependency is even more inherent for the propor-
tions representing test demands, which necessarily sum up to 1.00. Still, the correlations were not very
strong ($r \leq 0.57$), which can be considered as support for the discriminant validity of the four categories.
For students, similar findings were obtained. Although at the class level, we found a correlation of $-0.90$
between ratings for verbatim items and ratings for inference items, the student level showed correlations
that were lower than $0.31$. 

questions would have when appearing in the test. Such ratings would allow for a more direct comparison between perceptions of processing demands and actual test demands.

Some of our teachers spontaneously gave double coding for single items despite our instruction to assign the most suitable category. These double codings could indicate that for some of the test items different types of processing received equal emphasis. However, double coding could also indicate that teachers had not adequately considered the cognitive level of test items.

To obtain more certainty about the quality of teachers' test codings, test questions could additionally be coded by the researcher. To this purpose, the researcher needs to collect oral and textual information that was provided in the classrooms as a supplement to information stated in the textbook. Furthermore, to know which answers teachers accept, the researcher cannot rely solely on teachers' model answers, but also has to collect students' test answers along with the teachers' evaluations of these answers. A less laborious possibility is that the researcher observes the test categorization process of teachers to examine whether teachers carefully consider the relation between information required by test questions and information stated in the text or supplementarily provided in classrooms. Moreover, the researcher could collaborate with teachers to ensure that such careful judgements take place. Double codings, then, could be reserved for items that indeed show equal emphasis in processing requirements.

Related to this, more certainty can be obtained about participants' conception of the four coding categories. The categorization test could include more sample questions to make students and teachers more accustomed with the four question categories. Furthermore, a distinction could be made between classification of questions for instructional purposes and classification as a means to assess participants' understanding of the question categories.

4.4 Relation with previous research and implications for educational practice

Our findings are consistent with previous classroom studies that demonstrated discrepancies between students' test expectations, teachers' intentions and/or actual test demands regarding the relative importance of processing demands (Simpson & Nist, 1997, Smith & Feathers, 1983a). Moreover, our findings are consistent with studies that failed to show relations between the variation in students' study strategies and the variation in the test demands of courses (Thomas & Rohwer, 1987; Thomas et al., Rohwer, 1993). Finally, in the introduction section, we mentioned that our study on the relative importance of processing demands is part of a larger investigation that includes students' and teacher's perceptions of content demands (i.e., the relative importance of textual topics). Regarding this type of task demand, our findings (see Broekkamp, Van Hout-Wolters, Rijlaarsdam, & Van den Bergh, 2002; Broekkamp, Van den Bergh, Van Hout-Wolters, Rijlaarsdam, submitted) and those of previous studies (e.g., Jetton & Alexander, 1997; Schellings & Van Hout-Wolters, 1995) indicated similar discrepancies between students' test expectations, teachers' intended task demands and/or actual test demands.
Taken together, research findings suggest that secondary school students generally do not have a clear perception of the relative importance of content and processing demands when they prepare for a test. Concerning the processing demands, a possible explanation is that teachers' tests include several types of question, and students find it difficult to ascertain the relative emphasis that their teacher assigns to these question types. Moreover, we should consider the possibility that students do not even attempt to attune their studying to the relative importance of question types. Instead, students may choose for a uniform processing strategy that allows them to answer reproductive questions and also higher level questions as long as the these questions do not require a too deep level of understanding. Our impression of the tests was that, next to factual knowledge, they generally demanded a basic but not very deep understanding of the learning materials. This could imply that higher level learning strategies such as relating learning materials to prior knowledge, concept mapping or self-generation of questions (see Hadwin et al., 2001) may have been helpful, but were not necessary to do well at the test. A rather reproductive study strategy combined with logical thinking when taking the test could have been sufficient to answer paraphrase, inference and even skill items. The explanation that social study tests at the secondary school level generally do not require students to use higher level learning strategies is in line with Thomas and Rohwer (1987), who found that students mostly reported using uniform and reproductive study strategies.

From this analysis, at least four instructional implications can be derived. First, teachers could raise their awareness for the cognitive levels of their test questions and the corresponding types of processing they demand from their students. Such awareness is necessary to align processing demands, test demands and the task information they provide to their students. Second, teachers could clearly emphasize one type of processing in their tests over other types of processing to encourage their students to attune study strategies to specific processing requirements. Third, when teachers want to encourage higher level study strategies, they have to ensure that these strategies are necessary to perform well at the test. Fourth, teachers could alternate their emphasis of question types for different tests. If teachers sufficiently prepare their students for the changing demands of tests, these tests will stimulate students to adopt study strategies that are attuned to the specific processing requirements. Experience in this kind of attuning will lead to students who flexibly adapt their studying to optimize their learning performance.
CATEGORIZATION TEST

Four sample questions are given below. The questions are about the section “From crisis to depression” of the chapter “The United States of America, 1929–1945”. Your task is to assign each sample question to one of the four categories. You can use each category once! So, choose the category that best fits. To make it easier, you can first read all questions, before you assign the categories.

Question A

**Question:** Describe why so many American people got high loans from the bank?

**Possible answer:** American banks wanted to make more profit and they encouraged people to borrow money from the bank. With this money, people bought shares. And with these shares, people could borrow more money from the bank to buy even more shares.

**Explanation:** The answer does not describe completely what is in the textbook, but it is more or less a summary of the relevant passage.

**Encircle the corresponding number:**
1) Verbatim knowing
2) Telling it in your own words
3) Thinking through
4) Applying a skill

Question B

**Question:** You have to use several sources to give an answer to the question of how severe was the crisis in the United States. One of the sources is a travel report of a diplomat from Europe. He visits New York in 1930 and is impressed by the wealth of Americans.

**Possible answer:** To answer the question you first have to examine whether the report is reliable and representative. You ask yourself questions like: Did the writer have access to the full reality of American life or only to small parts? Did he have own interests in what he saw? Could this have distorted his view? Eventually you draw the conclusion that you cannot use the diplomat’s report to answer the question, because his account is not representative.

**Encircle the corresponding number:**
1) Verbatim knowing
2) Telling it in your own words
3) Thinking through
4) Applying a skill
Question C

Question: Give four arguments that support the view that in 1933 the American depression was serious.

Possible answer: In 1933, a quarter of the population was unemployed; the national product was down fifty percent, compared with 1929; the income per head of the population was down fifty percent; and the wages down fifty percent.

Explanation: These answers are taken quite directly from the chapter text.

Encircle the corresponding number:
1) Verbatim knowing
2) Telling it in your own words
3) Thinking through
4) Applying a skill

Question D

Question: One of the causes of the crisis was unrealistic optimism (In 1929, a Democrat writes an article with the title "everybody must get rich"). In a later period, Roosevelt also wanted to let people share in prosperity. Was he also unrealistic in his optimism?

Possible answer: No, in contrast to the government in 1929, Roosevelt acknowledged the problems in different branches of the American economy; and he came up with a realistic plan to tackle these problems.

Explanation: For your answer, you can use the information from the passage about Roosevelt's policy in the textbook. However, the textbook does not provide a direct to the question.

Encircle the corresponding number:
1) Verbatim knowing
2) Telling it in your own words
3) Thinking through
4) Applying a skill