Exam analysis in Remindo: *Learning Analytics*

Exam analysis in Remindo

Remindo has a functionality in student view called *Learning Analytics*, which provides insight into the quality of the exam as a whole and the quality of specific questions. This handout explains the meaning of the analysis' results and how they can be used to gain+ insight into the quality of the exam, improve questions, and implement substantively based changes to the exam and/or the results, if necessary.

The importance of exam analysis

First of all, the results of the analysis provide the lecturer with substantive feedback. In which sections did candidates perform well, and in which less good? Could this be fully accounted to knowledge of the material, or also to the way in which a question has been formulated?

Analyses can contain clues in order to understand deviating answer patterns and, if necessary, editing the answer keys. The information is also important to be able to decide whether a question is suitable for entry into the item bank or it should be edited.

No exam is flawless, and there will always be a number of candidates who, based on a specific exam, unjustly fail or succeed. Hence, we need to be careful with editing afterwards. The results of an analysis give indications.

The last part of this handout focuses on how to use the analysis and/or the consequences coming from the conclusion on the quality of the exam. An analysis is mainly of importance as a source of information when the success rate and/or the reliability of the exam as a whole is falling short and adjustment of the results is needed. There are several possibilities to make alterations. It can be decided, for example, to approve multiple alternatives for specific questions, delete questions from the exam or edit the cut score. The lecturer bases these decisions on substantive grounds and on information coming from the analysis.

It is wise to keep this in mind beforehand, as the meaning of the results depends on multiple factors.

Questions in advance

Is the exam a first take or a re-sit? And is the number of participating candidates>60?

The guidelines for a good and distinguishing exam, as mentioned later on, subordinate to the size and composition of the group of participants of the exam.

The exam's task is to express the knowledge of the learning material. The larger and more heterogenic a group of candidates is, the higher the chances of differences are, and the more reliable the analyses' results will be. As a rule of thumb, for an analysis to be of value, the minimum number of participants for an exam is 60.

If there are less differences between candidates, regarding their knowledge of the learning material, the distinction, as expressed in several parameters in the analyses' results, is expected to be small, with the consequence that the reliability of the test will fall short. In the event of a re-sit, the candidate population is expected to be smaller and more homogeneous than at a first take. The reliability is expected to be lower than in a first take, because there is less distinction to be made between candidates.

Do candidates already have knowledge about the standard?

Have expectations already been raised concerning the standard (cut score)? In case of using the results of the analysis to adjust the exam or cut score afterwards, it is important that the adjustments accord to the information that has been given to candidates beforehand, and the adjustments need to have a similar or better candidate performance as a result.

Is this the only exam on which the grade is based, or are there more test components?

The explanation of the interpretation of the analyses' results further on in this handout is based on the fact that the decision on passing or failing is fully or partially based on the results of this specific exam. Usually, however, more candidate performances need to be carried out, in a varied way of test-taking, that decide the final grade. In those cases, the used standards can be edited pro rata.

Step-by-step plan for using the results of the analysis

Step 1: The first impression

Is the success rate in accordance with expectations? And, is the reliability of the text sufficient? NB: is this the only exam on which the grade is based? If yes, a Quick Scan (1A) will be sufficient! If not, please continue with step 2.

> ***Step 1A: Quick Scan****: Quality of the questions (adjustments for entry into the item bank)*
>
> Use the item analysis to mark possibly dubious questions (P' and Rir < 0,1). Review the distractors that seem attractive to many candidates (a-value) and to candidates who are performing well (Rar-value). Search for an explanation for the deviating score pattern by reviewing the content of the question. Can the phrasing of the question or of the alternatives be improved?

Step 2: In case of a disappointing success rate

Question beforehand: Can the success rate be accounted for (small group, poorly performing cohort)? Or could there be a chance the disappointing results can be attributed to the exam itself? Before you go into depth (inspection of the question quality), it is wise to reflect on whether a marginal adjustment could lead to an increase of the success rate. If that is the case, it takes off much pressure already. You could ask the following questions: With which cut score would the success rate be as expected (acceptable)? How much does that differ from the used cut score, and could a great effect be realized by a small adjustment?

Step 3: Deep Scan*: Make use of the explanation, interpretation and results of analysis*

- In case of a low success rate, focus on hard questions: P'-value < 0,15;
- In case of a low success rate and a low reliability, focus on questions with a (too) low Rir-value < 0,10, and a low P'-value (< 0,10).

Step 4: Look for deviating score patterns (always on the basis of substantive arguments!): In case of a low P'-value: look for the attractive distractor (often chosen, so with a relatively high a-value). In both cases, the question is whether or not something could be said for approving an alternative as a correct answer, or even all alternatives.

On the following pages, an explanation of the parameters Remindo shows at exam level and at question level is provided, first summarized and then more elaborately, on the basis of examples - screenshots- as presented in Remindo.

Summary *Learning Analytics* in Remindo

Full overview	
Pie chart success %:	❶ The success and fail rate is the share of candidates meeting the appointed standard for this test, and a result of the appointed success and fail rate (cut score) and the way of creating a standard. A standard can be edited in three ways: the score for the lowest grade (10); the lowest grade 1 or 0; and the score that is a close pass (grade 5,5 or 6).
Curve diagram	❷ This diagram shows how the scores are divided. Is the score pattern divided as usual? Are there peaks concerning the success and fail rate? What are the highest and lowest scores?

Cronbach's α: **SEM:**	❸*Reliability and consistency of the test*. Alpha is a standard for reliability. That means the internal consistency of the test, and whether the questions measure the same (knowledge of the material). If the reliability (α) is low, it could indicate something about the low precision of the test results and the degree of coincidence. Including more questions can lead to a higher reliability. ❸*Standard Error of Measurement (SEM).* The standard error of measurement is a standard for the average (un)accuracy of measured test scores. In practice, a standard error of measurement lower than 10% of the maximum score is usual and acceptable.
Analyze questions	
Max. score: **Results:**	❹ The maximum amount of points that can be obtained for a question. *Response.* The actual number of candidates that has answered the question. If that number is low, it could say something about the level of difficulty of the question (see P'-value) or about lack of time.
P':	❺ *Level of difficulty*. In Remindo, the P'-value is the average score %: the total amount of scored points on the question by all candidates is divided by the maximum amount of points that can be obtained. The P'-value is an indication of the level of difficulty of the question. The higher the P'-value, the better candidates have performed on that question. If the P'-value is low, it means the question was hard.
Rir (Rar):	❻ *Distinctive capacity.* Rir-values and Rar-values are indicators of the distinctive capacity of the correct answer (Rir) and distractors (Rar) to a question. They indicate to which extent the alternatives distinguish high and low scoring candidates. Rir is the item rest correlation: the correlation between the item score and the total score of all remaining items of the same test. The Rar is the correlation between the score and the specific distractor and remaining items. A well discriminating item provides a high positive Rir-value and negative Rar-values.
STD:	❼ The standard deviation provides spreading of the attributed scores in relation to the average score (P'-value).
Duration:	❽ The average time that is needed to answer a question provides information about the level of difficulty of that question and the time that is needed for reading, solving and giving an answer. In case of energy gruzzlers, it is relevant to have a look at the P'-value and the Rir-value of a question, but also to check whether the required time is in accordance with expectations. It is possible that the phrasing of the question is unnecessary complex.
Interaction analysis	
A-value:	❾*Attractiveness of the distractor.* The A-value is the proportion of candidates or the percentage of candidates who chose the distractor in question as an answer to the multiple-choice question.
(Rit) and Rat:	❿*Distinctive capacity.* Rit shows the correlation between the test scores of candidates who chose the correct alternative (Rit) and the scores on the whole test. A similar correlation calculation is made for distractors: Rat-value. These indices turn out higher than the Rir- and Rar-values, because the item score is calculated as part of the total score, so the correlation is flattered.

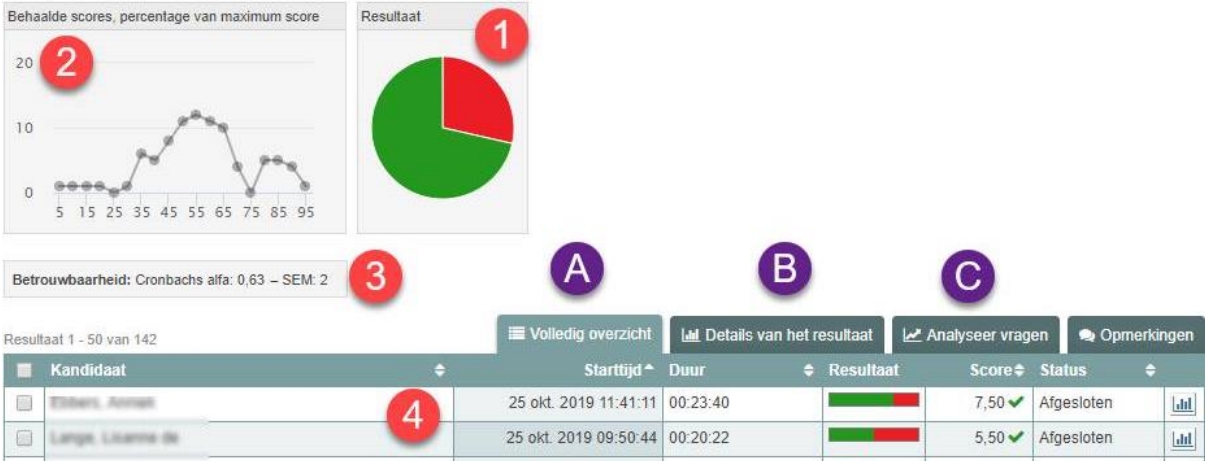
Explanation and interpretation of analysis results in Remindo: *Learning Analytics*

When discussing and interpreting the test analysis (Learning Analytics), we follow three tabs: **Tab A:** *Full overview;* **Tab B:** *Details of the results*, and **Tab C:** Analyze questions, in which you can click **D:** *Interaction analysis* (analysis of the details of a question).

Tab (A). Full overview

The tab "Full overview" exists of four parts, of which the first three concern the exam as a whole: success rates, scores and reliability of the test.

❶Pie chart with fail and success rates
❷ Curve / frequency distribution of scoring rates in relation to the maximum amount of point to be obtained.
❸ Exam’s reliability: Cronbach’s α, and the derived SEM (standard error of measurement), and ❹ a chart with the results per candidate.



Betrouwbaarheid: Cronbachs alfa: 0,63 – SEM: 2

Resultaat 1 - 50 van 142

□	Kandidaat	Starttijd	Duur	Resultaat	Score	Status	
□	[redacted]	25 okt. 2019 11:41:11	00:23:40	🟩🟥	7,50 ✓	Afgesloten	
□	[redacted]	25 okt. 2019 09:50:44	00:20:22	🟩🟥	5,50 ✓	Afgesloten	

Figure 1. Home screen “Full overview” in Remindo

❶ *Pie chart with success and fail rates*

By placing the cursor on the red/green area of the pie chart, the success and fail rates of this exam are shown. In the example (figure 1), 17% has failed and 83% has passed.

❷ *Frequency distribution*

The curve with frequency distribution provides information on how scores are divided. Please note that the scores in the chart are shown as percentages of the maximum score that can be obtained (100%). If you move the cursor over one of the dots, the percentage of participants with the relevant score in relation to the maximum score (in %) is shown. For example, as shown in figure 1, 10% of the participants have scored 65% of the maximum score that can be obtained. Please note: this example also shows the highest score is 95%. A 100% score is obtained by none of the candidates. Much of statistics and interpretation of parameters is based on a ‘normally divided’ score pattern and based on the fact that the knowledge of tested candidates follows normal distribution: extreme scores are rare, and the majority of scores are around average.
To distinguish between high performing and low performing candidates, but also between close passes and close fails, the frequency distribution can show a variety of scores. The probability increases when the group of candidates is larger and more diverse in terms of knowledge of the material.
If the exam is taken by a large group of candidates, it can be expected that the exam exposes mutual knowledge differences. Also, the more questions the exam contains, the more accurate and reliable performances and mutual differences can be determined.
If the highest scoring rate is much lower than the maximum of 100%, and the number of participating candidates is high, there could be something wrong with the level of difficulty of the exam and the manageability of the separate questions.

❸ *Exam’s reliability: Coefficient alpha (α)*

The coefficient alpha is an indicator of the reliability of the exam. The coefficient α can reach a maximum of 1 (fully reliable) and a minimum of 0 (fully unreliable: the scores are realized coincidentally). The more reliable the exam, the more accurate scores can be interpreted. For an exam of which consequences in passing or failing are big (*high stakes* test), an α of 0,8 is aimed to be able to say something substantive about the candidate’s level of achievement. If the exam is the only test the grade is based on, the aim is a 0,8 standard. If the decision (passing or failing the course) is based on results of other tests as well (open questions, interim tests), a lower reliability than the 0,8 standard is justifiable.

An example: there have been two mc-tests in a course, one test halfway through the period, and one at the end of the course. Both tests contained 40 multiple-choice questions, and both have a reliability of 0,60 (Cronbach’s α). To decide whether candidates pass or fail *for the course*, both test results can be ‘combined’: table 1 shows that the expected reliability for the pass-fail decisions (for both tests combined) increased to α = 0,75.

The rule is that the longer the exam is, the higher the reliability and the better the differentiation rate are. The differentiation rate shows how ‘refined’ the test is. For example, does the test differentiate enough between close passes and close fails? Is the score range wide enough, and are both minimum and maximum score rates to be found in the score pattern? The relationship between the length of the test and an estimation of the reliability can be found in table 1.
Example: A test contains 30 questions (K) and the reliability (α) is 0,40. In this case, the test should be extended to containing 90 questions (3K) to reach a reliability of 0,60 (α’’’).

Table 1: Estimation of reliability (α) in case of extending or shortening of a test, with K-questions

K	1.5 K	2 K	3 K
α	α’	α’’	α’’’
0.20	0.27	0.33	0.40
0.40	0.50	0.57	0.60
0.60	0.69	0.75	0.77
0.80	0.86	0.89	0.91

Analysis of a re-sit is a different story. It can be accounted as reasonable that the reliability will be lower, because the mutual differences in the knowledge of the material in the population (spreading) are smaller than in the case of a first take.
The measured reliability also says something about the amount of inconsistent decisions. A test is never 100% reliable, there are always unjust passes or fails. The less reliable the test, the greater the chance of unjust decisions. In table 2, the percentage of non-consistent decisions in relation to the percentage of fails and reliability (α) are shown. In the shown example, around 15% of candidates had failed, with an α of almost 0,70. In this case, in around 14% of the cases, an inconsistent decision was made: that means 7% has passed unjust and 7% has failed unjust. The question is until what point this is still acceptable. The consequence of the decision that 7% of the candidates did pass, but possibly did not meet qualifications, depends on many factors. For example, the place and interest of the course in the study as a whole, or similar goals coming back later in the studies.

Table 2: Percentages of non-consistent decisions in relation to fail rates and test reliability (α). Source: Dousma, Horsten, Brants, Tentamineren (1997).

Fail % (failed)	**Reliability (α)**						
	0,50	**0,60**	**0,70**	**0,80**	**0,90**	**0,95**	**1,00**
5	8	7	6	5	4	3	0
10	14	12	11	9	6	4	0

15	18	17	**14**	12	8	6	0
20	23	20	17	14	10	7	0
25	26	23	20	16	11	8	0
30	29	25	22	18	12	9	0
35	31	27	23	19	13	9	0
40	32	29	24	20	14	10	0
45	33	**29**	**25**	20	14	10	0
50	33	30	25	20	14	10	0

❸ *Standard error of measurement (SEM)*

The standard error of measurement shows the average (un)accuracy of the measured test scores. It is a measure to estimate to what extent candidates would score the same if taking another, imaginary, similar test on the same learning material.
In the standard error of measurement (SEM), the reliability of the test (α) is calculated by using the formula $Sm=Sa\sqrt{1-\alpha}$ (in which Sa is the measured standard deviation of the test scores).
The formula shows that the more reliable the test is, the smaller the measurement error will be, and the more plausible it is that measured scores of the candidates will correspond with 'real scores'.
If the test is unreliable (high degree of coincidence), the measured scores will not have any meaning.
If the standard error of measurement is 2, as in figure 1, candidates with a measured score of 13 will have a 67% probability of having the knowledge that corresponds with a score between 11 and 15 (13 ± 2). With a probability of 95%, the real score will be between the measured score ± 2xSm = 13 ± 4. An assumption for the interpretation of the standard error of measurement is that the test scores are distributed normally.

❹ *Chart with the results per candidate*

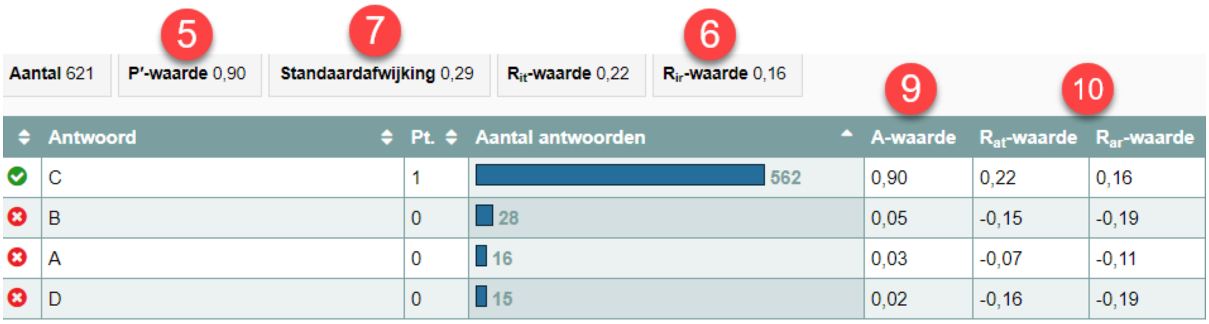
In the chart, the following is shown per candidate. The amount of time he or she spent on the test: *duration, results* in a colored bar (from green to red) and *scores* shown in percentages, the amount of obtained points or the grade.
Not much attention is paid to this table, because the parameters have been made up per candidate and not much meaning can be attributed to them. Duration could be viewed as the average time candidates spend on a specific question. This will be discussed later on, when discussing Tab C.

Tab (B). Details of the results

The tab *Details of the results* is a table in which the following characteristics can be found per question: The number of candidates that has answered the question and:
❺ The P'-value of the question;
❻ The Rit/Rir-value of the question;
❼ STD: standard deviation;
An explanation of the interpretation of these parameters can be found in tab C.

Tab (C). Analyze questions

❺ P'-value
❻ Rir-value
❼ STD (standard deviation)
❽ Average duration



Aantal 621 | **P'-waarde** 0,90 | **Standaardafwijking** 0,29 | **R_{it}-waarde** 0,22 | **R_{ir}-waarde** 0,16

	Antwoord	Pt.	Aantal antwoorden	A-waarde	R_{at}-waarde	R_{ar}-waarde
✓	C	1	562	0,90	0,22	0,16
✗	B	0	28	0,05	-0,15	-0,19
✗	A	0	16	0,03	-0,07	-0,11
✗	D	0	15	0,02	-0,16	-0,19

Figure 2. Screenshot "Analyze questions" in Remindo

❺ *P'-value: level of difficulty (Proportion right)*

P': Max 1; *Min*: 0
Target value: depends on the possible negate guessing

⚠ Remindo puts a flag at the P' of a question if the value is higher than 0,9 (*"possibly the question is too simple?")* or if the P'-value is lower than 0,4 (*"possibly the question is too hard?"*).

The P'-value indicates the level of difficulty of a question. The P'-value in Remindo is a relative measurement. In case of a 4-choice question, a P'-value of 0,5 means that the question was hard, but scores as expected. Question 3 in figure 2 is a 4-choice question as well and has a P'-value of 0,24. This is consistent with the negate guessing (the chance that someone answers a question correctly by guessing).
If P' is lower than the negate guessing, candidates will choose a distractor, and not guessing. It could be possible that something is going on with the question. The matching Rir-value could contain clues for whether it have been especially high performing or lower performing candidates who answered the question correctly.
In case of a question with multiple answers, it is also necessary to take the amount of possibilities into account when interpreting the P'-value. In case of all other question forms, if there is no negate guessing, the P'-value is a good (absolute) indicator of the level of difficulty.
The test is usually composed in a way that there is a good mixture of the levels of difficulty of the questions.
Analysis should confirm this: P'-values provide a varied image. There is a low amount of very difficult questions (low P'-values) in order to separate the 'nines' and 'tens' out of ten, and a low amount of very easy questions (high P'-values) to separate the 'fours' from the 'fives' out of ten. The majority of the questions, however, consists of P'-values that are in the middle (between 0,3 and 0,7) and contribute to the distinction between passing and/or failing for the test.

❻ *Distinctive capacity: Rir-value and/or Rit-value*

Max: 1; *Min*: -1
Target value: positive, higher than 0,10

⚠ Remindo puts a flag at the Rir of a question when the value is negative.

The P'-value indicates the share of candidates who answered the question correctly. Rit (item total score correlation) and Rir (item remaining score correlation) indicate to what extent the question distinguished the high performing candidates from the poor performing candidates. The difference between Rit and Rir-value is that in case of Rit, the correlation between the score for the question and the total score is calculated for candidates who answered the question correctly. Rir is a cleaner standard, because the total score minus the score for the question (the Remaining) is used when calculating the correlation.

If Rir is high (0,3-0,5), the question did its work: high performing candidates answered the question correctly, and poor performing candidates chose a distractor. If Rir is lower, the question may not have been distinguishing. If Rir is negative, something could be wrong. In that case, especially high performing candidates chose for a distractor, while poor performing candidates chose the correct alternative. If Rir is negative, it means per definition that one of the distractors correlates positively with better performing candidates (positive Rar-value). It needs to be checked if something could be said for the distractor, because high performing candidates choose it.

Interpretation of combined P'-values and Rir-values

In order to be able to make a statement about a question's quality, it is of importance to look at both the level of difficulty and the distinctive capacity and the relationship between the measured P'-value and Rir-value. For example, a distinctive capacity (Rir) of almost 0,0 is not alarming if almost all candidates answered the question correctly (high P'-value). And, if all candidates answered the question correctly (P'=1,0), the distinctive capacity (Rir) is 0,0 by definition.

However, if the Rir-value of a question is 0,0 and the P'-value low (<0.3), it is a completely different story.

Table 3 shows the interpretations of several combinations of Rir-values and P'-values. Those are slightly more refined than the ones Remindo works with.

Table 3: Interpretations of possible P'-value and Rir-value combinations

Level of difficulty (P-value)	**Distinctive capacity (Rir-value)**	
	Rir<0,1: Question is not distinctive and with negative values even opposite to expectations: high performing candidates score low on this question.	**Rir> 0,1**: Acceptable to good (Rir>0,3) distinctive question
Hard question **P' < 0,3**	The question has been poorly answered, also by the better performing candidates of the test. *Question: Is the answer key correct? Could more answers be approved? Tip: compare the alternatives to high A-values and positive Rar-values.*	Generally, the question has been answered incorrectly, but it still distinguishes between better and poor performing candidates on the test. *Is the question too hard, too complex? Is it a tricky question? Make sure there are not too many hard questions in the test.*
Practicable question **0,3<P'<0,8**	The question has been answered pretty well but distinguishes unsatisfactory between poor and high performing candidates. *Perhaps other questions can be -partly- approved? Tip: compare the alternatives to high A-values and positive Rar-values.*	The question has been answered fairly good and the distinctive capacity is all right. *In this case there is no need for further action.*
Easy question **P'> 0,8**	The question was easy to answer for the majority of the candidates, regardless of their performance on the test. *Is the question an unintentional giveaway and could it be solved with common sense?*	The question has been answered well and distinguished between high and lower performing candidates. *In this case there is no need for further action.*

❼ *STD: Standard deviation*

Standard deviation (STD) indicates how much question scores vary compared to average scores. Standard deviation is not as much of interest for questions that are either answered correctly (maximum points) or incorrectly (0 points), but they are of interest if more or less points can be obtained gradually. For example, in case of open questions containing more answer elements and closed questions with a 'shared score', in which obtained points and possible subtractions lead to differentiated score patterns.

If the STD is 0 for these types of questions, the conclusion would be that the question has not differentiated.

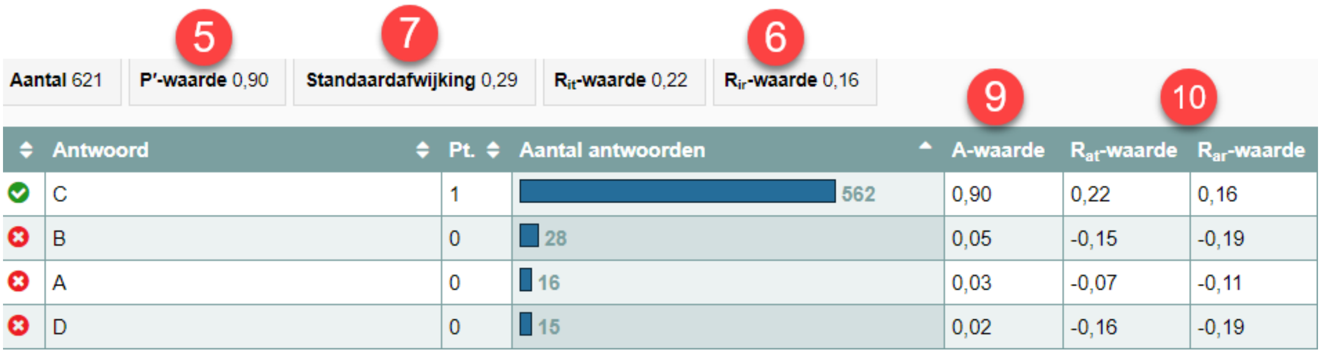
The higher the STD, the greater the spreading in scores. We hope that questions, apart from a reasonable spreading in scores, also differentiate enough between high and poor performing candidates. It is better to have a look at the Rir-value.

❽ *Duration*

The time candidates spend on a question is interesting, and the time for reflection or writing they need possibly says something about the level of difficulty/complexity of the question or the cognitive capacity the question appeals to. As a lecturer, this provides you with substantive feedback on what students regard as hard questions and what they think is easy.
If candidates spend much more time on a question than expected, there could be something wrong with the question or the instruction of the question. The P'-value and Rir-value for this question may give an explanation.
Overall, candidates should get enough time to be able to finish the exam. If the exam contains unintentional 'time-consumers', it could be a reason to avoid such questions the next time around.

Tab (D). Interaction analysis: Analysis details for a specific question

Clicking the image ☷ on the right of the tab 'Analyze questions' guides you to the screen of interaction analysis. The parameters that already have been discussed are shown here, such as the P'-value and Rir-value and standard deviation.
New in this screen are the following parameters:
❾ A-value
❿ Rat-value and Rar-value
Below, the meaning of those parameters is explained on the basis of the screenshot. In Remindo, it is hard that A-values and P-values, and Rir-values and Rar-values (Rit and Rat), are used interchangeably.
In figure 3, the first line (below 5, 6 and 7) shows the parameters for the correctly given answers to the question. Those are the P'-values, Rir/-Rit-value and standard deviation. In the table below (below 9 and 10), only the A-values and Rar- and Rat-values can be found, because of technical reasons. The calculated Rar/Rat of the correct answer is C, so the P'-value, Rir-value and Rit-value of this question.



Aantal 621 | **P'-waarde** 0,90 | **Standaardafwijking** 0,29 | **R_{it}-waarde** 0,22 | **R_{ir}-waarde** 0,16

	Antwoord	Pt.	Aantal antwoorden	A-waarde	R_{at}-waarde	R_{ar}-waarde
✓	C	1	562	0,90	0,22	0,16
✗	B	0	28	0,05	-0,15	-0,19
✗	A	0	16	0,03	-0,07	-0,11
✗	D	0	15	0,02	-0,16	-0,19

Figure 3: Screenshot 'Interaction analysis' in Remindo

❾ *A-value: attractiveness of the distractors (only relevant for closed questions):*

Value: Max 1; *Min* 0

⚠ Target value: the sum of the a-values (proportion wrong) is lower than the P'-value (proportion right) of the question.

A-values provide the proportion (0 until 1,0) of candidates who chose the specific distractor. The A-value is what the P'-value is for the correct answer, but for distractors. Is the A-value similar to or higher than the P'-value, then it means the distractor has been very attractive (could it be a tricky question?). If the A-value is low, the distractor is not very attractive to the candidates, and thus not effective. In that case (for example, none of

the candidates, or only one candidate, chose the distractor), something could be said for deleting the distractor in the future (4-choice question becomes a 3-choice question) or to formulate a better distractor.
In the example of figure 3, the A-value of the correct answer (read: P’-value) is 0,90. This means 90% of the candidates answered this question correctly. The 10% incorrectly given answers are moderately divided amongst the three distractors. It seems plausible that candidates chose a random alternative in case they did not know the answer.

⑩ *Rar-value: discriminating capacity of distractors*

Value range between -1 and 1.
Target value: The sum of all Rar-values is lower than Rir.
The Rar-value has a similar meaning to the Rir-value, but for the distractor/the alternative. The Rar-value provides extra information: which part of the group of candidates chose the correct alternative (Rir) or distractors (Rar). If that were the better performing candidates, Rir will be positive. However, if high performing candidates particularly chose a distractor and poor performing candidates chose the correct answer, something could be wrong with the question.

What to do with analysis information?

If test analyses show that a question falls short, substantive motives should indicate whether or not to make adjustments to the test. Eventually, the goal is for lecturers to explain students why a question is answered correctly or incorrectly and to make that plausible on substantive grounds.
If you want to edit the test and/or results, you can do several things on question level, but also on test level (success and fail rate, and the grade calculation).

1) Nothing changes

The phrasing of the question is correct and the correct alternative as well. The question remains in the test in its original form and nothing is adjusted. Because the question – for indistinct reasons- did not do its job, the lecturer can decide to remove the question from the question file or to take in an improved version of the question.

2) The question will be deleted from the test

This action is only advised if the phrasing of the question is so dubious that it has been impossible to make a choice from the alternatives. An important issue with deleting questions is that candidates who answered the question correctly will feel duped, even if the arguments to do so are substantive.

3) More answers to a question are approved

On the basis of analysis, it is clear which alternatives most candidates (P’-values and A-values) and the best performing candidates chose (Rir- and Rar-values). If the lecturer thinks a distractor (originally an incorrect answer) could also work, it makes sense to approve this alternative. In case of a faulty phrasing of the question, the lecturer can even choose to approve all answers. In comparison to method 2, in which candidates who answered the question correctly will feel duped, the average scores will go up in method 3, so candidates receive the benefit of the doubt.
Please note: In case more alternatives are approved as correct answers, the negate guessing for the question goes up and so does the guessing score for the test. This can be of influence for grades and test results.
In case of a 4-choice question, negate guessing goes up with 0,25 with every extra approved alternative. In case of a 3-choice question that will be 0,33, et cetera. Because negate guessing is often used as a point of reference for the minimum grade (1,0 or 0,0), the adjustment also has

consequences for the grade calculation, unless the lecturer decides to approve more answers, but does not adjust the determination of grades.

4) The success and fail rate (cut score) shifts

There are two reasons to adjust (lower) the standard (cut score) for a pass.

- The test was simply too hard. If there has been poorer performance on the test than in previous years, while there has not been a reason to assume that candidates prepared less, it could be that the test was too hard.
- The test was not reliable enough. If reliability is low, the share of unjust fails is possibly unreasonably high. By lowering the standard, the share of unjust fails can be reduced to acceptable proportions. In this case, you have to take for granted that the number of unjust passes is higher!

5) Method 2, 3 and 4 combined *Editing the cut score*

If the composition of the test has changed afterwards (more alternatives approved, questions that are removed), it is wise to have a look at whether the success and fail rate (cut score) also needs to be adjusted. An exam in which questions have been removed will be shorter and needs an adjustment in the cut score and grade calculation. And if more answers are approved, there could be consequences for the cut score, because the negate guessing has increased and this score is often used as a reference for the minimum grade (1 or 0). A cut score cannot be increased if candidates have already been informed about it. On the other hand, lowering the cut score is always possible (because candidates will definitely not complain about that).

Background/Literature

Dousma, T., Hortsen, A., & Brants, J. (1997). Tentamineren. Groningen: Wolters-Noordhoff.
De Groot, A.D.,&Van Naerssen,R.F. (1969, 1973). Studietoetsen. Den Haag: Mouton.
De Gruijter, D.N.M. (2008). Toetsing en toetsanalyse. Leiden: ICLON.
Milius, J.J. (2007). Schriftelijk tentamineren: een draaiboek voor docenten in het hoger onderwijs. Utrecht: COLUU
Milius, J.J. (2016). Handleiding Remindo toets versie 1.0. Utrecht: Educate-it.

Workshops Toetsanalyse in Remindo, provided by Onderwijsadvies & Training (O&T)

O&T regularly organizes the workshop ‘Toetsanalyse in Remindo’ (1 half-day).
For more information, have a look at O&T’s website.