Memorandum on Advisory Guidelines for Reliability and Item Analysis v. EC PW May 2017

This memorandum on the Advisory Guidelines for Reliability and Item Analysis aims to enable lecturers and examiners to improve the quality of assessment (for example in the case of MC tests and open-question tests) using reliability and item analysis, and to make improvements in preparation for the subsequent academic year. The memorandum discusses the indices that are produced by the reliability and item analysis and seeks to provide guidelines for subsequent improvements based on this quantitative analysis technique.

1. Explanation of the indices used in the reliability and item analysis

- Cronbach's alpha (or KR-20 for multiple-choice tests): the reliability of the test; the extent to which the scores achieved in the test are expected to correspond with the scores achieved in the same test if it were to be repeated (under the same conditions).
- KR-20 (75): the KR-20 standardized to 75 questions, or the expected reliability of a multiple-choice test if it were to consist of 75 questions; indicates the degree of coherence of the test questions (which contributes to the reliability of the test), whereby the effect of the length of the test is corrected[1].
- p-value: proportion of students who answered the relevant item correctly; the difficulty of an item.
- average p-value: the average p-value for all items; the difficulty of the test as a whole.
- p': the p-value corrected for probability; indicates the average level of knowledge of the students in relation to the relevant item. p' = p - (1-p)/(a-1), where a is the number of alternatives in the multiple-choice question. When faced with a multiple-choice question, students always have a chance of answering correctly by guessing. With a 4-choice item, that chance is 25%; a p-value of .25 means that the average knowledge level is 0; a p-value of .7 for a 4-choice question corresponds to a p' of .6.
- rir-value: remaining-item correlation of an item; the extent to which the score for an item corresponds to the scores for the remaining items in the test; the capacity of an item to differentiate between students with a good knowledge of the material and students with less knowledge of the material. Items with a low rir-value have a low capacity to differentiate between students with different levels of knowledge; items with a negative rir-value may indicate an error or a misleading question.
- a-value: proportion of students who chose a particular alternative.

2. Interpretation of the reliability and item analysis

A proper reliability and item analysis assumes that the group of students that participated in the test is representative of the target population, i.e. that the group includes both good students and less good students, leading to a normal distribution of scores, including both higher and lower scores. This means that the first time the test is taken is suited to a reliability analysis; a resit test is not generally suitable for such an analysis. The number of students also affects the quantitative values; in practice, we attach less value to a test analysis involving fewer than 20 students.

[1] N.B. This is not the same as reliability; for the reliability of a test, the length of the test is an influencing factor.

When interpreting the results, we look first at the reliability (Cronbach's alpha / KR-20) of the test. If this is above .80, the reliability is 'high', i.e. the students' scores on the test give an accurate picture of their knowledge, and the test is suitable for summative assessment; if it is below .65, the reliability is 'low', i.e. there is a real risk that students may score higher or lower in the test than their actual level of knowledge; between .65 and .80, the reliability of the test is classed as 'fair'.

The higher the reliability of the test, the more accurate the rir-values will be, and the more relevant these are during the interpretation phase. With lower reliability, rir-values are less meaningful. The p-values are generally less susceptible to higher or lower reliability.

Items with a high rir-value have a high capacity to differentiate between students who know the material well and those who know it less well; this is an indication of the quality of these items. Items with low rir-value[2] have little or no differentiating power, which may indicate that most students (regardless of whether they know the study material well or not) guessed on this question. This may indicate that the students did not have sufficient knowledge of the study material (or that they were not taught this material). A negative R-value for an item indicates that students who knew the study material well (who gave correct answers to the 'rest' of the items) answered this item incorrectly more often than students with less knowledge. This may indicate that an incorrect alternative was indicated as the correct one, or that there is some other problem with the question, such as its formulation, which led students with good knowledge to choose an incorrect answer.

For the p-value, it is less easy to make a quality assessment; there is no such thing as the 'optimal p-value', from a quality perspective. Questions with a very high p-value are answered correctly by many students, and can therefore be answered easily. These are generally items that are more suited to ready-knowledge tests (whereby all students of a particular level are expected to have acquired the relevant knowledge). Items with a very high p-value generally do not have a strong differentiating capacity (rir). For items with a very low p-value, it is sensible to check whether the students have actually covered the study material in question sufficiently. If not, this may be the result of inadequate preparation by the student (e.g. due to lack of time), the way the material was taught or presented (teaching issue) or a failure to appreciate the importance of the subject (e.g. detail question).

We interpret the various values generated from the quantitative analysis of a test in relation to one another. Low rir-values with very high p-values mean very little, for instance, and high rir-values with low reliability do not tell us much either. Looking at the values, we need to make a judgement on how we can improve a particular item or the test as a whole: were the other alternatives also correct or partly correct? Is the phrasing of the question unambiguous or open to interpretation? Was the study material covered adequately and communicated clearly enough? Would the students have been able to prepare properly for the test? Did teaching take place under the right circumstances?

When assessing a question, the content of the question must be considered first, not the quantitative analysis; the latter can only play a signalling role, using the rules of thumb described above. After identifying a potentially problematic item using the quantitative analysis, an analysis of the content of that item will ultimately determine the quality of the question and any possible improvements to be made.

---

[2] In theory, an rir-value of greater than .25 is 'good'; in practice, this does not occur often.

3. Advice on what to do in relation to particular indices

Given that, as mentioned previously, the various values generated by the quantitative analysis must be interpreted in relation to one another and that when assessing the quality of questions, the content of those questions should be the primary consideration, the table below can serve as a guide for further action by lecturers and examiners: how to interpret the quantitative indices, and how to approach an analysis of the content of the test with a view to making quality improvements, either to improve the existing test or to improve the tests used for the next assessment.

Table 1. Step-by-step plan with categorized advice on action following the quantitative analysis

Test as a whole				
Step	**Signal**	**Action**	**Explanation**	
1	resit involving < 20 students	Only perform a qualitative, content-based analysis of the test.	A quantitative analysis is only considered meaningful in relation to the first opportunity to take the test and when a sufficiently large number of students complete the	
2	Cronbach's Alpha or KR-20 > .65	Use the rir and p-values in the analysis; otherwise: only consider the p-values.	If the reliability is high enough, the rir-values are meaningful; p-values remain more robust even with lower reliability.	
For each question				
Step	**Signal**	**Action**	**Explanation**	*****
3	rir < -.10 and p' < .80	Check whether the answer key was correct or whether the question was misleading.	A negative rir-value (and a lower p-value)indicates that better students scored worse than others on a particular question.	B
4	-.05 < rir < .05 and p' < .30 and a' < .20	Check whether the material was covered sufficiently clearly.	If the rir is around 0 and the p and a-values are low, the students were	C
5	rir < .10 and p' < .40 and a' > .30	Check whether another alternative is (also) correct.	If the rir is low, the p is not high, and one a-value is high in relation to the p, another alternative is likely to be (almost) correct.	A
6	rir ≥ .15 and p' < .05	Check whether this item is a detail question.	If p is around equal to the probability of guessing the correct answer, but the better students failed to choose the right answer, it may relate to a detail in the study material.	D
7	rir + p' < .40	Check whether the question was formulated clearly enough.	If p and rir-values are both low, the question does not differentiate between good and less good students and neither did enough students pick out the correct answer from the distractors.	E

* The letters A – E refer to the categories as reported in the multiple-choice examination analysis of the VU Examination Service.

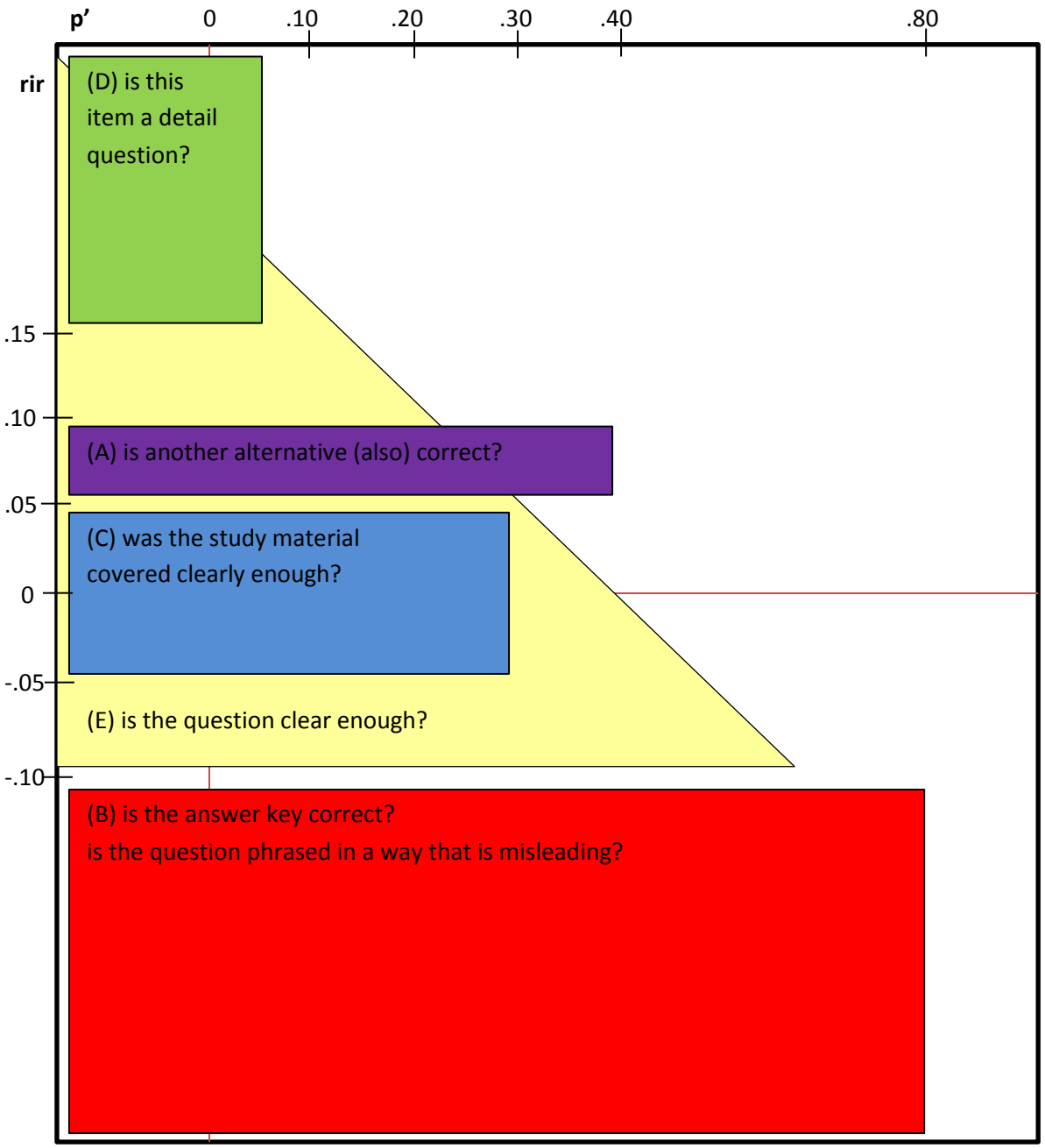

Figure 1. Graphical representation of categorized advice on quantitative analysis, two-dimensional representation of p' and rir.