The reliability of the pass/fail decision for assessments comprised of multiple components

Andreas Möltner[1]
Sevgi Tımbıl[1]
Jana Jünger[1]

1 Ruprecht-Karls-Universität Heidelberg, Kompetenzzentrum Prüfungen in der Medizin Baden-Württemberg, Heidelberg, Deutschland

Abstract

Objective: The decision having the most serious consequences for a student taking an assessment is the one to pass or fail that student. For this reason, the reliability of the pass/fail decision must be determined for high quality assessments, just as the measurement reliability of the point values.

Assessments in a particular subject (graded course credit) are often composed of multiple components that must be passed independently of each other. When "conjunctively" combining separate pass/fail decisions, as with other complex decision rules for passing, adequate methods of analysis are necessary for estimating the accuracy and consistency of these classifications. To date, very few papers have addressed this issue; a generally applicable procedure was published by Douglas and Mislevy in 2010.

Using the example of an assessment comprised of several parts that must be passed separately, this study analyzes the reliability underlying the decision to pass or fail students and discusses the impact of an improved method for identifying those who do not fulfill the minimum requirements.

Method: The accuracy and consistency of the decision to pass or fail an examinee in the subject cluster Internal Medicine/General Medicine/Clinical Chemistry at the University of Heidelberg's Faculty of Medicine was investigated. This cluster requires students to separately pass three components (two written exams and an OSCE), whereby students may reattempt to pass each component twice. Our analysis was carried out using the method described by Douglas and Mislevy.

Results: Frequently, when complex logical connections exist between the individual pass/fail decisions in the case of low failure rates, only a very low reliability for the overall decision to grant graded course credit can be achieved, even if high reliabilities exist for the various components. For the example analyzed here, the classification accuracy and consistency when conjunctively combining the three individual parts is relatively low with κ=0.49 or κ=0.47, despite the good reliability of over 0.75 for each of the three components. The option to repeat each component twice leads to a situation in which only about half of the candidates who do not satisfy the minimum requirements would fail the overall assessment, while the other half is able to continue their studies despite having deficient knowledge and skills.

Conclusion: The method put forth by Douglas and Mislevy allows the analysis of the decision accuracy and consistency for complex combinations of scores from different components. Even in the case of highly reliable components, it is not necessarily so that a reliable pass/fail decision has been reached – for instance in the case of low failure rates. Assessments must be administered with the explicit goal of identifying examinees that do not fulfill the minimum requirements.

Keywords: Assessments, Decision accuracy, Decision consistency, Pass-fail reliability

1. Introduction

Assessments are performance measurements and possess, like all measuring instruments, only a limited accuracy. This must be sufficiently high so that the scores given for assessments reflect the content. Established methods exist for estimating the measurement reliability of the points given on assessments (e.g. Cronbach's α); however, the reliability of the pass/fail decision is hardly taken into consideration in the analysis or evaluation of assessments.

This is remarkable insofar as precisely this aspect clearly has more importance for students in regard to their studies than the measurement reliability of a point value; failing an exam leads to remedial work, lost time, and under circumstances the question of whether to continue or quit medical school. This decision also has importance for the institution administering the assessment: if the examinee possesses the required knowledge and skills to continue the study program, an unjustified failure leads to a greater amount of work. If an examinee is allowed to pass despite not having the qualifications, then not only significant problems in continuing university study are to be expected, but even the endangerment of medical patients in worst case scenarios (see [5]).

Presumably, this topic has also received so little attention in Germany in relation to medical education because the regulations of the medical licensing act (*Ärztliche Approbationsordnung*) have been generally adopted by the academic rules and regulations of most medical schools for multiple-choice testing, the longstanding dominant testing format. With the purely formal definition of a passing score being 60% of all questions asked, this approach does not permit a content-based, criterion-oriented definition of the minimum requirements. To our knowledge, in Germany only the rules and regulations of the Medical Faculty at the University of Heidelberg allow standard setting for multiple-choice tests; this means the ability to deviate from the formal rule of 60% to pass and define a passing score according to content-based criteria and in a standard procedure, similar to the established standard setting for an OSCE [2], [5].

The establishment of new testing formats, with which practical skills, qualifications and necessary competencies for practicing medicine should be assessed in addition to pure subject knowledge, demands definition and, for assessments, the practical setting of minimum requirements. As a result, it is also necessary to pay close attention to the decision accuracy, decision consistency and pass-fail reliability when evaluating tests or testing formats [19]. The decision accuracy indicates the extent to which the examinees that satisfy the minimum requirements pass an actual test and the examinees without sufficient knowledge fail. Decision consistency refers to the agreement of pass/fail between two equivalent tests, meaning two tests that measure *the same knowledge or the same skills equally well*. It needs to be noted here that "same" does not imply that the tests only cover one construct in terms of test theory. An OSCE can contain stations dealing with practical skills and with communicative competencies which are to be regarded as subscales in terms of test statistics. An equivalent test must then have practical and communication stations with the same scope and of the same difficulty.

A series of methods has been developed, particularly since the 1980's, to ascertain the accuracy and consistency in respect to individual tests, even though none of these methods can be viewed as the standard procedure (see [6], [13], [14], [16], [18], [23], [25]). To obtain graded course credit in many medical subjects, multiple individual assessments must be taken, for instance a written exam covering theoretical knowledge and an OSCE to assess practical skills. If these assessment results are combined together into an overall score through weighted averaging or totaling, the entire assessment can be treated as one "single" test.

Often there is another approach, completely justified in terms of content: instead of *compensatory* combination of assessment scores, *all the individual assessments must be passed*. This *conjunctive combination* (logical "and" conjunctions) of the pass/fail decisions has significant effects on the accuracy/consistency of the overall decision, since one single unreliable decision on an individual test can ruin the reliability of the overall decision:

> *...because longer collections of test questions tend to be more reliable than shorter collections of test questions, compensatory scoring tends to be more reliable than conjunctive scoring. In conjunctive scoring, if a student has to pass all of the content areas separately, the least reliable score controls whether a student will pass.* [26]

Practical instances of this include subjects that spread the tested content out over multiple tests to limit the scope of a particular test and subjects in which both theoretical knowledge and practical skills are imparted resulting in a written assessment for the theory and a practical one for the skills. Instead of allowing compensatory scoring in these cases, requiring students to satisfy the minimum on each separate assessment is often justified. Ultimately, a conjunctive combination will also be used for the entire course of study: only those who *have passed in all of the subjects*, will successfully complete the degree program.

Assessment scores can also be combined in other ways. Alongside the conjunctive combinations already mentioned, disjunctive (logical "or" conjunctions) are also possible when only one single component of many must be passed. An example of this would be the repeated assessments. If an assessment can be retaken once, a student has passed if it is passed on the first or second attempt (that a student need not appear for the second administration if he or she has already passed the first attempt is of no interest to logic). In practice at schools and universities even more complex rules apply, such as graded credit must be successfully attained for three of five possible courses.

Only a few studies exist regarding the decision reliability for complex combinations of assessment scores [24], a

generally applicable method of analysis has been proposed by Douglas and Mislevy [7], [8]. Our study applies this method to analyze the assessment for the subject cluster General Medicine/Internal Medicine/Clinical Chemistry that was given at the Faculty of Medicine in Heidelberg during the winter semester 2012-13 and, for the attainment of which, two written exams and one OSCE had to be passed separately. Students had the option of repeating each individual component of the assessment twice.

Graded credit for a cluster of subjects (*fächerübergreifende Leistungsnachweis* or FÜL) is particular to the German medical licensing regulations (*Approbationsordnung*), according to which every medical school must bundle multiple course subjects into one instance of graded course credit. This legal requirement is without significance for the following statistical observations. Douglas and Mislevy's method is directed toward the accuracy and reliability of a complex pass/fail decision that is the result of a combination of individual decisions.

Regardless of the formal legal definitions of a FÜL, the terms "overall test" (for full graded credit) and "individual test" or "component" (for the individual subject assessments) will be used.

The aim of this study is to present a suitable method for the analysis of pass/fail decision reliability using the example of a bundled assessment and establish it as an essential aspect of ensuring the quality of tests.

2. Principles

Decision accuracy and decision consistency

Our starting point is the assumption that the examinees can be classified according to their knowledge or skills into two subgroups, one that fulfills the minimum requirements (master, competent examinee) and one that does not fulfill them (non-master, incompetent examinee). For an assessment in a particular subject, such a definition could be taken from a catalogue of learning objectives, with the definition of a master being someone who – for example – masters 70% of these learning objectives.

For an actual assessment, learning objectives are selected for testing and a passing score is defined. The lowest passing score could then also be set at 70%. A student who has mastered 90% of all learning objectives would with great probability exceed this cut-off, in contrast to someone who has mastered 72% – thus also fulfilling the minimum requirements (master) – but who could possibly be unlucky and fail. The same applies to students who are just under the cut-off for master status, but pass with a bit of luck. A more detailed discussion of the difference between the definition of master (performance standard) and the passing score can be found in [12] (see also [2], [5]).

Depending on the objective of the assessment, the passing score can be varied. If a higher passing score is set, the probability of a non-master passing is reduced, but at the same time the risk of inaccurately classifying a master as a non-master increases. This is analogous to a diagnostic test that compares a gold standard (in this situation the knowledge that a person is a master or non-master) with an actual test score. If one regards the assessment as the diagnosis of non-masters, then this test possesses a certain sensitivity (the probability of failing non-masters) and a specificity (probability that a master passes). Changes to the cut-off point for the test value lead to an increase or decrease in the sensitivity, along with a simultaneous decrease or increase in the specificity.

The degree to which masters and non-masters can be identified using the assessment is referred to as the decision accuracy. The left contingency table in Table 1 presents in full the relative proportions for master/test passed, master/test failed, non-master/ test passed, and non-master/test failed.

If two *equivalent* tests are administered, then the degree of agreement between the two test scores is the *decision consistency* or pass-fail reliability. The corresponding contingency table is shown on the right in Table 1. If the tests are equivalent, then the proportion of students who pass the first test and fail the second must be exactly the same size as the proportion that failed the first and passed the second.

The two values most frequently used in the literature for decision accuracy and decision consistency are the relative number of correct decisions P_a (corresponding to the correct classification rate for diagnostic tests) and agreements P_c [11] and Cohen's κ [4] (for its use in connection with the sensitivity and specificity of diagnostic tests, see [3]). Cohen's κ corrects the number of correct decisions P_a and the agreements P_c for the effects of chance that can be expected in the margin totals of the contingency table. The corresponding values are designated by κ_a and κ_c.

κ assumes the value of 1 in the case of complete agreement. The application of κ as a measure of agreement is criticized in some places (e.g. [10]) and alternatives have been propagated. In our opinion, all the coefficients in this context come with the disadvantage that, with reduction to a single index, important information is lost. Therefore, when analyzing a test, the *entire contingency table* should be drawn upon.

Procedures for estimating the decision accuracy and consistency for individual assessments

In the literature, many methods are presented for determining decision consistency for individual tests. Known are those presented by Livingston-Lewis [16] and Peng-Subkoviak [18]. Overviews and comparisons also exist [6], [13], [14], [23], [25]. In our opinion, it is not currently possible to show a clear preference for any particular one among the various methods.

Table 1: Contingency tables for decision accuracy and decision consistency. The a_i values represent the relative proportions of the scores on a test depending on whether students who fulfill the minimum requirements pass or fail (left). For example the value for a_2 indicates the percentage of students who do not have sufficient knowledge/skills (non-master), but despite this have passed. In the case of two fully equivalent tests, c_i gives the analogous values. As a result of the equivalence of both tests $c_2=c_3$.

Decision accuracy

	Requirements		
Test	met (Master)	unmet (Non-master)	Total
Passed	a_1	a_2	a_{1+2}
Failed	a_3	a_4	a_{3+4}
Total	a_{1+3}	a_{2+4}	1

Decision consistency

	Equivalent test		
	Passed	Failed	Total
Passed	c_1	c_2	c_{1+2}
Failed	c_3	c_4	c_{3+4}
Total	c_{1+3}	c_{2+4}	1

The method of Douglas und Mislevy

Douglas und Mislevy's method [7], [8] serves to determine the decision accuracy and consistency for complex decision rules based on scores from multiple tests. The prerequisite is that the data of the individual tests can be described by a multivariate normal distribution and the reliabilities of the tests are known. In practice, however, scores are not normally distributed, which is why an adequate transformation of the data must be undertaken. For a precise description of the method, reference must be made to the original literature [7], [8].

For the purpose of understanding, let us take a simple, fictional example to determine the decision accuracy with graphic illustration of two individual tests (see Figure 1). Those who passed both individual tests have passed overall (conjunctive combination).

Figure 1a illustrates the distribution of the scores. The examinees whose scores lie within the yellow part of the curve have passed both individual tests and have thus passed overall (in Table 1 this is represented by a_{1+2}). Orange denotes the area of the distribution in which one individual test was passed and one was not. These examinees have not passed overall, just as those who did not pass either of the individual tests (brown area). The proportion of those in the L-shaped section of the curve (orange and brown) – representing those who failed overall – is represented by a_{3+4} in Table 1.

In the method proposed by Douglas and Mislevy, the distribution of the "true values" is determined according to the model of classical test theory and the assumption of normal distribution, meaning the distribution of the values if these had been measured without any error. For this, the reliabilities of the individual tests must be known. The resulting distribution shows a distinctly lower variance. The masters and non-masters are defined on the level of the true values. Figure 1b shows this distribution: the masters are those who satisfy the minimum requirements for both tests (green area; a_{1+3} in Table 1), while non-masters are those who have not satisfied the minimum requirement for one area of both individual tests (red, L-shaped area; a_{2+4} in Table 1).

To determine the decision accuracy, the model now examines how the masters' scores are distributed (see Figure 1c). Due to errors of measurement in the tests, a portion of the masters failed (dark green area). The light green area shows the group of masters who passed overall (a_1 in Table 1); the dark green area indicates the masters who failed overall (a_3 in Table 1).

The corresponding graph for the non-masters is presented in Figure 1d . This is presented from another perspective to make the borderlines more visible. The light red area indicates the portion of non-masters who did not pass overall and dark red those who did pass overall (a_4 and a_2 in Table 1).

If one combines the distributions of the masters and non-masters in Figures 1c and 1d, then the overall distribution of the test scores in Figure 1a is seen again.

3. Method

3.1 Data

The aim of this study is to analyze the scores given for the graded course credit in the bundled subjects of Internal Medicine/General Medicine/Clinical Chemistry at Heidelberg University's Faculty of Medicine during the winter semester 2012-13. The graded assessment consists of the written exam in Internal Medicine/General Medicine, an oral practical assessment (OSCE), and the written exam in Clinical Chemistry. To receive the graded course credit, a case report, a MiniCEX, and Encounter Cards to assess professionalism are also required. Since the pass rate is 100% for each of these, they are of no relevance to this investigation. Only the students who took all three tests were included in the analysis (N=147). The basic data for the tests are presented in Table 2. All in all, seven of the 147 examinees who sat for all three tests failed at least one of the components.

For the written exams in the subjects Clinical Chemistry and Internal Medicine, masters were defined as those who would correctly solve 60% of the questions from the particular question pool for each subject. In terms of the OSCE, the definition of master was those whose mean point totals for the OSCE stations in the subject was at least the number of points set as the standard (performance standard, [5]).

The passing scores for each written exam were defined as 60% of the possible points for the questions actually posed; for the OSCE, the passing score was the mean of

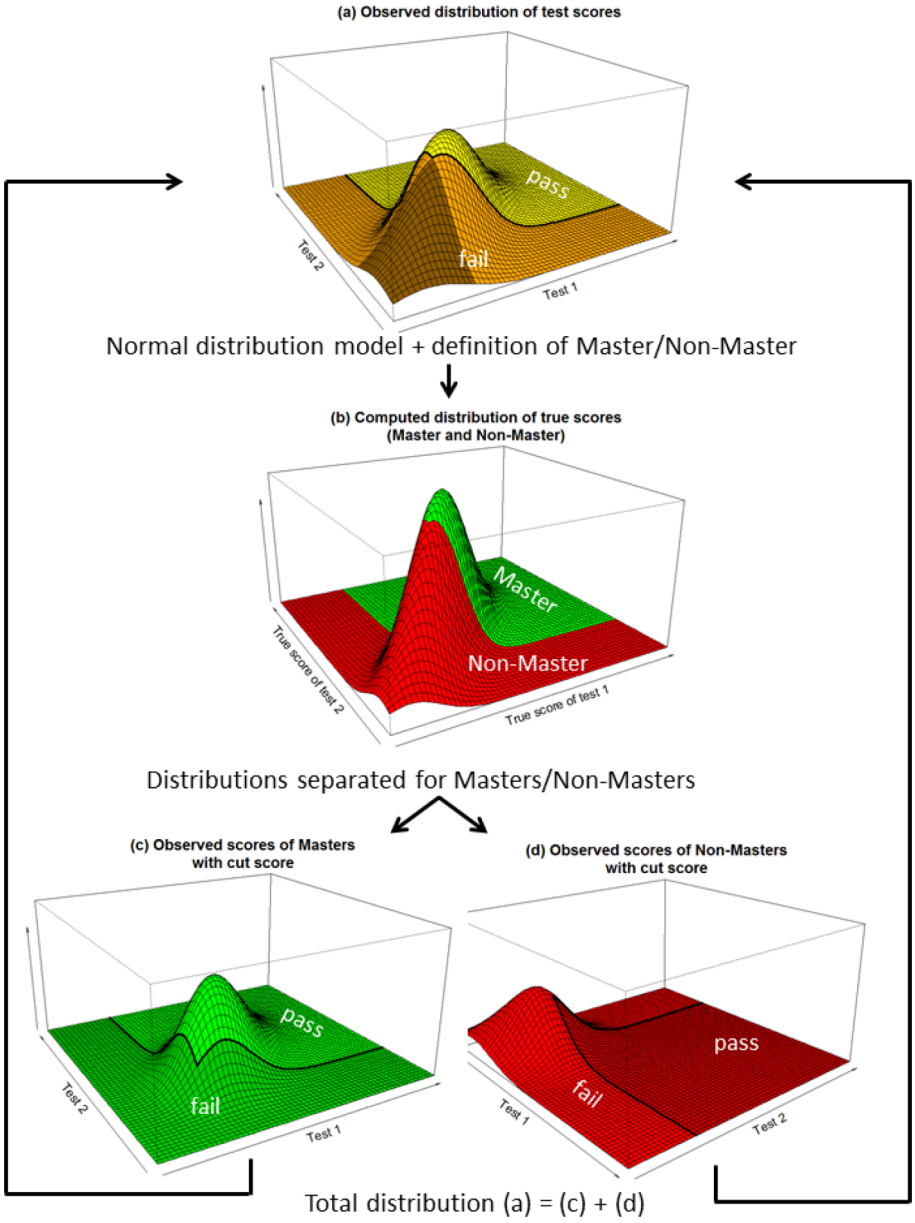


Figure 1: The steps for the method of Douglas and Mislevy: (a) distribution of the test scores for two tests; (b) estimation of the true values and definition of master/non-master according to the model; (c) distribution of the scores achieved by masters; (d) distribution of the scores achieved by non-masters (note: perspective is different). The distribution of the overall results (a) is comprised of the scores achieved by masters and non-masters.

Table 2: Basic data from the tests for graded credit in Internal Medicine/General Medicine/Clinical Chemistry during the winter semester 2012-13 (only examinees who took all three components: *N*=147).

	Internal Medicine exam	Chemistry exam	OSCE
Total points	50.000	52.000	250.000
Mean	39.075	42.537	211.357
Passing score	30.000	31.200	171.500
Number of failures	4	5	1
Guttman's λ_2)*	0.770	0.762	0.752

)* Instead of using Cronbach's α to estimate the reliability, the somewhat more precise estimation using Guttman's λ_2 coefficient was selected [9]. Just like α, λ_2 represents a lower bound for the reliability; the actual reliability can also be higher.

the number of points defined as the standard for the stations used (passing score).

3.2 Statistical analysis

The analysis of the accuracy and consistency of the pass/fail decision was mainly carried out according to the method proposed by Douglas und Mislevy [7], [8].
The method applied by Douglas and Mislevy makes no assumptions about the internal structure of the individual tests in terms of test theory, or about that among the individual tests. In particular, the individual tests are neither required to be homogenous or one-dimensional, nor must a uniform performance dimension be represented by the entirety of the components. However, it is pre-requisite that the data is sufficiently well described by a normal curve of distribution and the measurement reliabilities (reliabilities) of the individual tests are adequately estimated.
Since the point values of the tests each deviate in a highly significant manner from normal distributions (Shapiro-Wilk tests: all $p<0.0008$), the data were subjected to a multivariate Box-Cox transformation [1]. For the transformed data, a test for deviation from trivariate normal distribution using the generalized Shapiro-Wilk test as described by Villasenor-Alva and Gonzalez-Estrada [22] revealed a *p*-value of 0.8467 (MVW=0.9929), so that a sufficiently good adjustment of the data can be assumed.
In contrast to the normalizing rank transformations applied in the study by Douglas and Mislevy, an adjustment to a *multivariate* normal distribution is aimed for with this transformation. To estimate the reliability of the individual tests, Guttman's λ_2 was selected as the coefficient allowing for a slightly better estimation of the minimum reliability than Cronbach's α (=Guttman's λ_3) [9].
The contingency tables for the decision accuracy and consistency of the *individual tests* and their *conjunctive combination* were calculated using numerical integration of the multivariate normal distributions with the algorithm of Miwa, Hayter and Kuriki [17].
Taking the two options to repeat each individual test into account, this analysis is only of a theoretical nature insofar as it is assumed that students, who have not passed a test, concentrate on learning for the repeat attempt. In the analysis undertaken here it is assumed that the students taking these tests sit for the second attempt with the same knowledge they possessed for the first. The algorithm of Miwa et al. [17] is unsuited for the integration of a higher-dimensional normal distribution necessary for calculating the statistical values, so this analysis was done with Monte-Carlo integration as in [8]. All in all, 100,000 simulated data sets were generated to ensure sufficient accuracy of the results.

4. Results

4.1 Individual tests

The contingency tables in Table 3, Table 4 and Table 5 cover the individual tests. The estimated number of failing examinees resulting from the model of normal distribution is calculated as the failure rate of the model ×N=0.0331×147=4.9 for the written exam in Internal Medicine, 3.0 for Clinical Chemistry, and 1.9 for the OSCE. It can be seen that these rates deviate only slightly from the number of examinees who actually failed: 4, 5 and 1 (see Table 2). For all three tests, Cohen's κ coefficients κ_a (decision accuracies) and κ_c (decision consistencies) are low.

4.2 Assessments composed of multiple scores

4.2.1 Conjunctive combination of the individual tests

The decision accuracy and consistency for the conjunctive combination of the three tests are presented in Table 6. According to the model of Douglas and Mislevy, it is to be expected that 7.8 examinees would fail (=failure rate of the model ×N=0.0531×47=7.8). Seven candidates did indeed fail (many of the students did not pass more than one test), demonstrating satisfactory agreement between the model and the actual data. The test logic leads to a clear classification of the students who do not meet the requirements; the proportion of non-masters who pass all three tests is now 0.004 in total (although consideration must be given to the fact that their overall proportion is only 0.0232). The sensitivity to uncover non-masters is 82%, the specificity 97%; however, the positive predictive value is low with 36%.
The decision consistency (three administrations of equivalent tests) does not reach a satisfactory value with $\kappa_c=0.474$. Classification of 94.7% of the examinees would occur right off (P_c), meaning that conflicting information would exist for 5.3% of the examinees about successfully achieving the full graded credit.

4.2.2 Complex conjunctive and disjunctive combination for repeat tests

Each of the three tests in Internal Medicine, Clinical Chemistry and the OSCE can be retaken a total of two times before the student has definitively failed. Logically, this means that a student must pass one of three written exams in Internal Medicine, one of three written exams in Clinical Chemistry, and one of three OSCEs. Within each testing format, the pass/fail decision is then disjunctively combined, and the three component decisions thus conjunctively (see Figure 2). The fact that a student who has passed a test on the first attempt does not appear for another attempt in the same subject is not of importance to the decision logic.

Table 3: Decision accuracy and consistency for the exam in Internal Medicine

	Decision accuracy			Decision consistency		
	Requirements			Equivalent test		
Test	met	unmet	Total	Passed	Failed	Total
Passed	0.9616	0.0053	0.9669	0.9479	0.0190	0.9669
Failed	0.0202	0.0129	0.0331	0.0190	0.0141	0.0331
Total	0.9818	0.0182	1.0000	0.9669	0.0331	1.0000
	κ_a=0.4919, P_a=0.9746			κ_c=0.4060, P_c=0.9620		

Table 4: Decision accuracy and consistency for the exam in Clinical Chemistry

	Decision accuracy			Decision consistency		
	Requirements			Equivalent test		
Test	met	unmet	Total	Passed	Failed	Total
Passed	0.9763	0.0030	0.9793	0.9664	0.0128	0.9793
Failed	0.0139	0.0068	0.0207	0.0128	0.0079	0.0207
Total	0.9902	0.0098	1.0000	0.9793	0.0207	1.0000
	κ_a=0.4393, P_a=0.9831			κ_c=0.3672, P_c=0.9743		

Table 5: Decision accuracy and consistency for the OSCE in Internal Medicine

	Decision accuracy			Decision consistency		
	Requirements			Equivalent test		
Test	met	unmet	Total	Passed	Failed	Total
Passed	0.9852	0.0016	0.9868	0.9781	0.0087	0.9868
Failed	0.0096	0.0036	0.0132	0.0087	0.0044	0.0132
Total	0.9948	0.0052	1.0000	0.9868	0.0131	1.0000
	κ_a=0.3835, P_a=0.9888			κ_c=0.3280, P_c=0.9825		

Table 6: Decision accuracy and consistency for graded credit in the subject cluster Internal Medicine/General Medicine/Clinical Chemistry (conjunctive combination)

	Decision accuracy			Decision consistency		
	Requirements			Equivalent test		
Test	met	unmet	Total	Passed	Failed	Total
Passed	0.9428	0.0040	0.9469	0.9204	0.0265	0.9469
Failed	0.0340	0.0191	0.0531	0.0265	0.0267	0.0531
Total	0.9768	0.0232	1.0000	0.9469	0.0531	1.0000
	κ_a=0.4852, P_a=0.9620			κ_c=0.4740, P_c=0.9471		

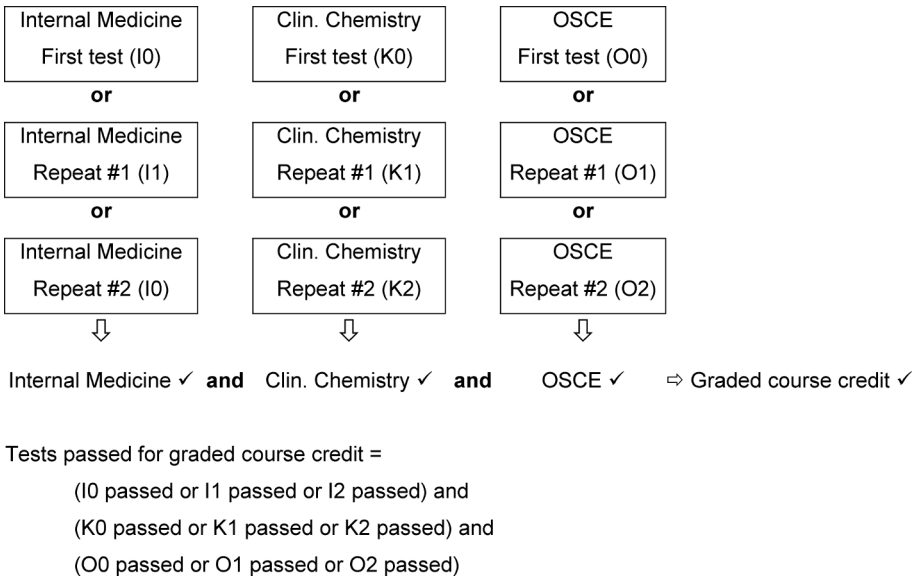


Figure 2: Decision rules for obtaining graded course credit for the subject cluster Internal Medicine/General Medicine/Clinical Chemistry.

Table 7: Decision accuracy and consistency for graded credit in the subject cluster Internal Medicine/General Medicine/Clinical Chemistry with the chance to repeat each test twice (see Figure 2)

	Decision accuracy			Decision consistency		
	Requirements			Equivalent test		
Test	met	unmet	Total	Passed	Failed	Total
Passed	0.9746	0.0124	0.9870	0.9809	0.0062	0.9870
Failed	0.0022	0.0108	0.0130	0.0062	0.0068	0.0130
Total	0.9768	0.0232	1.0000	0.9870	0.0130	1.0000
	κ_a=0.5890, P_a=0.9854			κ_c=0.5181, P_c=0.9877		

Table 7 contains the contingency tables for the decision accuracy and consistency with the assumption that a student takes all tests with the same level of knowledge. Of significance here is primarily that of the 2.32% of the students (a_{2+4}=0.0232) who do not meet the requirements (non-master) more than half (a_2=0.0124) would ultimately receive the graded credit, meaning that, as a result of the possibility to repeat tests, only a portion of the students who do not fulfill the requirements are stopped from continuing the program (note the substantial difference in regard to the results in the section above, in which the corresponding value with a_2=0.0040 in Table 6 is clearly lower than the value of 0.0124 in Table 7).

5. Discussion

Individual tests

Decision accuracy: all three of the individual tests demonstrate an overall satisfactory reliability (see Table 2). Of the non-masters, who altogether represent only 0.5 – 1.8% of the examinees (see Tables 3 to 5: a_{2+4}), about one-third pass each of the tests (a_2). The relevant percent of the masters who do not pass the test is low in all cases (a_3); however, in absolute numbers this is distinctly more than there are non-masters taking the test, so that for all three tests more than double the number of assumed non-masters in the group fail.

Decision consistency: The reliability of the decision to fail an examinee must be assessed as unsatisfactory. Of those who fail, about 60–65% would pass an equivalent repeat test. The poor decision consistency is also seen in the low κ_c values of 0.33–0.41.

Conjunctive and complex combinations of the test scores

Decision accuracy: The contingency tables regarding the decision accuracy for the conjunctive combination of the three tests (see Table 6) show that of the 2.3% non-masters (the students who do not meet the minimum requirements in at least one of the three subjects) only 17% pass (a_2/a_{2+4}=0.040/0.232=0.0172). The relative percent of masters who fail though increases to 3.5% (a_3/a_{1+3}=0.0348); for the individual tests this percent was at the highest 2%. In this case also, distinctly more examinees fail (5.3%, a_{3+4}) than there are non-masters among the candidates. Cohen's κ_a is with 0.49 almost just as high as the value of the best κ_a for the individual tests (written exam in Internal Medicine); the percentage of correct classifications is lower with P_a=0.96. According to this, the assertion that in conjunctive combinations the test with the poorest decision accuracy dominates must be evaluated with more precision.

If the fact that each student has two opportunities to repeat a test is taken into consideration (see Table 7), then assuming that the students attend equivalent repeat tests with the same level of knowledge or skills, only 47% of the non-masters do not in the end receive the graded credit (a_4/a_{2+4}=0.0108/0.0232=0.4655). For the masters, this is negligibly small with 2.3‰ (a_3/a_{1+3}=0.0022/0.9768=0.0023). Thus, the testing structure with the two options to retake each individual test is obviously poorly suited for reliably recognizing the non-masters.

Decision consistency: when conjunctively combining the three tests, the stability of the decision "fail" is also not satisfactory, but somewhat better than for the individual tests. In the case of an equivalent test complex consisting of the tests in the three subjects, a little more than half the examinees would pass the test. If κ_c is used as the consistency index, then this is higher than for each individual test with a value of 0.47.

When taking the possibility to repeat tests into account, a similar situation emerges: only somewhat more than half of the students who ultimately fail would be forced to end their studies again if they started over from the beginning.

Summary

In conclusion, it is clear that the pass/fail decision for the tests administered here needs improvement not only in terms of its accuracy, but also its consistency. "Sifting out" the non-masters is not possible in a reliable manner because tests may be repeated. On the other hand, there is hardly any danger that someone who meets the requirements will have to discontinue university study due to one or more instances of bad luck on tests.

To start with, the reason for this result could be seen in the model of normal distribution. To achieve an acceptable decision accuracy and consistency in the case of low failure rates, an extremely high reliability is necessary (a corresponding table for the κ_c coefficients is presented in [21]). This characteristic however is not specific for the normal distribution; not presented here are analyses for other assumed distributions that lead to similar results. Making the usual assumptions about the distribution

form of the point totals on tests, most non-masters will fall close to the passing score if there is a *low failure rate and no excessively high reliabilities*. This does not depend on whether a formal (e.g. required by law), norm-oriented, or criterion-oriented cut-off is involved. This is why there is a relatively high probability that non-masters pass with a bit of luck, so that high levels of accuracy or consistency cannot be expected in these cases.

Limitations of Douglas und Mislevy's method

The major limitation of the method proposed by Douglas und Mislevy is its assumption of a multivariate normal distribution. For the tests analyzed here, an acceptable normalization of the data was possible through a multivariate Box-Cox transformation, something that would not work in every case for data from other tests. Furthermore, the assumption of a multivariate normal distribution for the true values and measurement errors implies a constant error of measurement. However, the error of measurement can be distinctly higher at the cut-off point and lead to an overestimation of the decision accuracy. On the other hand, the distributions of the observed point values are clearly skewed to the left. As a result of the normalizing transformation, the values for very poor students have been moved "closer" to the passing score, so that in the analysis they belong more to the group for which, due to the error of measurement, inaccurate or inconsistent decisions are to be expected, although on the original scale they are reliably identified as non-masters.

Low decision accuracy and consistency: consequences for testing

With low failure rates, as for the tests analyzed here, a highly reliable test would be necessary to achieve a sufficient reliability for the decision to pass or fail. This is not surprising to the extent that on a test aiming for the usual measurement reliabilities, a large portion of the questions display good discriminatory properties for the majority of the examinees, but give little information regarding the separation of the sparsely populated extreme groups. One approach – albeit difficult to implement at universities – would be the administration of two tests: the first serving the usual assessment of student performance and the second specifically for identifying masters and non-masters with questions specifically selected for this purpose (Kane [12], p. 430 has already suggested the latter). For the first test, a relatively high passing score is set, with which the probability of a non-master passing remains very low. The remaining group then consists of (poor) masters and non-masters who can be separated as well as possible by the second test. Methods for optimal question selection can be found in the literature using "(computerized) classification tests" (CCT) (e.g. [20], [15]).

6. Conclusion and outlook

The method of Douglas and Mislevy is suitable for analyzing the decision accuracy and consistency of overall decisions concerning assessments composed of multiple parts and for which the overall pass/fail decision is the result of a complex combination of individual scores. Above all, the conjunctive combinations (each individual test must be passed) and disjunctive combinations (only one of multiple tests must be passed; this applies for repeated tests) are of practical importance.

The graded course credit for a cluster of subjects (*fächerübergreifender Leistungsnachweis*) was selected as being exemplary of German medical education at present. In this testing situation, theoretical and practical assessments in different subjects are combined and, in order to pass overall, all of the components must be passed. Students have the possibility to repeat each individual test twice.

Using the method of Douglas and Mislevy, the decision accuracy and consistency for giving the graded course credit could be successfully analyzed; there was a high degree of congruence between the model and the data. The analysis also revealed a significant issue concerning tests and low failure rates: these can only be reliably identified with difficulty for tests that comply with the usual demands for a sufficient reliability. Identifying masters and non-masters would require targeted classification tests with an appropriate selection of questions. An analysis of the decision accuracy and consistency should generally be carried out on the relevant tests. The limitation of using the normal distribution model still needs to be viewed as a substantially limiting factor; it is to be hoped that suitable methods with weaker distribution assumptions (e.g. multivariate beta-binomial distributions) or distribution-free methods are developed.

Competing interests

The authors declare that they have no competing interests.

References

1. Andrews DF, Gnanadesikan R, Warner JL. Transformations of multivariatedata. Biometrics. 1971;27:825–840. DOI: 10.2307/2528821
2. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Med Teach. 2008;30(9-10):836–845. DOI: 10.1080/01421590802402247
3. Brenner H, Gefeller O. Chance-corrected measures of the valdity of a binary diagnostic test. J Clin Epidemiol. 1993;47(6):627–633. DOI: 10.1016/0895-4356(94)90210-0
4. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measure. 1960;20:37–46. DOI: 10.1177/001316446002000104

5. Cusimano MD. Standard setting in medical education. Acad Med. 1996;71:112–120. DOI: 10.1097/00001888-199610000-00062
6. Deng N. Evaluating IRT-and CTT-based Methods of Estimating Classification Consistency and Accuracy Indices from Single Administrations. Massachusetts: University of Massachusetts; 2011. Open Access Dissertations. Paper 452. Zugänglich unter/available from: http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1451&context=open_access_dissertations
7. Douglas KM. A general method for estimating the classification reliability of complex decisions based on configural combinations of multiple assessment scores.Unpublished dissertation. Maryland: University of Maryland; 2007.
8. Douglas KM, Mislevy RJ. Estimating classification accuracy for complex decision rules based on multiple scores. J Educ Behav Stat. 2010;35:280–306. DOI: 10.3102/1076998609346969
9. Guttman LA. A basis for analyzing test-retest reliability. Psychomet. 1945;10:255–282. DOI: 10.1007/BF02288892
10. Gwet KL. Handbook of inter-rater reliability (2nd ed.). Gaithersburg: Advanced Analytics, LLC; 2010.
11. Hambleton RK, Novick MR. Toward an integration of theory and method for criterion-referenced tests. J Educ Meas. 1973;10:159–96. DOI: 10.1111/j.1745-3984.1973.tb00793.x
12. Kane M. Validating the performance standards associated with passing scores. Rev Educ Res. 1994;64:425–461. DOI: 10.3102/00346543064003425
13. Kim DI, Choi SW, Um KR. A comparison of methods for estimating classification consistency. Paper presented at the 2006 Annual Meeting of the National Council on Education in Measurement. San Francisco, CA: National Council of Edudation in Measurement; 2006.
14. Lee WC. Classifcation consistency and accuracy for complex assessments using item response theory. CASMA Research Report No. 27. Iowa City, IA: University of Iowa; 2007.
15. Lin CJ. Item selection criteria with practical constraints for computerized classification testing. Educ Psychol Meas. 2011;71:20-36. DOI: 10.1177/0013164410387336
16. Livingston SA, Lewis C. Estimating the consistency and accuracy of classifications based on test scores. J Educ Meas. 1995;32:179–197. DOI: 10.1111/j.1745-3984.1995.tb00462.x
17. Miwa A, Hayter J, Kuriki S. The evaluation of general non-centred orthant probabilities. J Royal Stat Soc. 2003;65:223-U234. DOI: 10.1111/1467-9868.00382
18. Peng CJ, Subkoviak MJ. A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. J Educ Meas. 1980;17:359–368. DOI: 10.1111/j.1745-3984.1980.tb00837.x
19. Schuwirth L, Colliver J, Gruppen L, Kreiter C, Mennin S, Onishi H, Pangaro L, Ringsted C, Swanson D, van der Vleuten C, Wagner-Menghin M. Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):224–233. DOI: 10.3109/0142159X.2011.551558
20. Spray JA, Reckase MD. Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. J Educ Behav Stat. 1996;21:405–414. DOI: 10.3102/10769986021004405
21. Subkoviak MJ. A practitioner's guide to computation and interpretation of reliability indices for mastery tests. J Edc Meas. 1988;25:47–55. DOI: 10.1111/j.1745-3984.1988.tb00290.x
22. Villasenor-Alva JA, Gonzalez-Estrada E. A generalization of Shapiro-Wilk's test for multivariate normality. Communication Stat Theo Method. 2009;38:1870–1883. DOI: 10.1080/03610920802474465
23. Wan L,Brennan RL, Lee W. Estimating classification consistency for complex assessments. CASMA Research Report No. 22. Iowa City, IA: University of Iowa; 2007.
24. Wheadon C, Stockford I. Estimation of composite score classification accuracy using compound probability distributions. Psychol Test Assess Mod. 2013;55:162–180.
25. Zhang B. Assessing the accuracy and consistency of language proficiency classification under competing measurement models. Lang Test. 2010;27:119–140. DOI: 10.1177/0265532209347363
26. Zieky M, Perie M. A Primer on Setting Cut Scores on Tests of Educational Achievement. Washington/DC: Educational Testing Service; 2006. Zugänglich unter/available from: http://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf

Corresponding author:
Dr. phil. Andreas Möltner
Ruprecht-Karls-Universität Heidelberg, Kompetenzzentrum Prüfungen in der Medizin Baden-Württemberg, Im Neuenheimer Feld 346, 69120 Heidelberg, Deutschland, Tel.: +49 (0)6221/56-8249, Fax: +49 (0)6221/56-7175
andreas.moeltner@med.uni-heidelberg.de

Please cite as
Möltner A, Tımbıl S, Jünger J. The reliability of the pass/fail decision for assessments comprised of multiple components. GMS Z Med Ausbild. 2015;32(4):Doc42.
DOI: 10.3205/zma000984, URN: urn:nbn:de:0183-zma0009843

This article is freely available from
http://www.egms.de/en/journals/zma/2015-32/zma000984.shtml

Received: *2013-12-20*
Revised: *2014-03-12*
Accepted: *2014-05-26*
Published: *2015-10-15*

OPEN ACCESS

This is the translated (German) version.
The original (English) version starts at p. 1.

Die Zuverlässigkeit der Entscheidung „bestanden/durchgefallen“ bei zusammengesetzten Prüfungen

Andreas Möltner[1]
Sevgi Tımbıl[1]
Jana Jünger[1]

1 Ruprecht-Karls-Universität Heidelberg, Kompetenzzentrum Prüfungen in der Medizin Baden-Württemberg, Heidelberg, Deutschland

Zusammenfassung

Zielsetzung: Die gravierendsten Konsequenzen für einen Studierenden bei einer Prüfung besitzt die Entscheidung über „bestanden“ oder „durchgefallen“. Wie die Messzuverlässigkeit der Punktwerte muss bei qualitativ hochwertigen Prüfungen deshalb auch die Zuverlässigkeit der Entscheidung „bestanden“ oder „durchgefallen“ bestimmt werden. Oft setzen sich Prüfungen eines Fachs (Leistungsnachweise) aus mehreren Teilprüfungen zusammen, die z. B. unabhängig voneinander bestanden werden müssen. In diesem Fall einer „konjunktiven“ Verknüpfung der Einzelentscheidungen „bestanden/durchgefallen“ wie auch bei anderen komplexen Bestehensregelungen sind zur Abschätzung der Genauigkeit und Konsistenz der Entscheidung „bestanden/durchgefallen“ adäquate Auswertungsverfahren erforderlich. Bislang liegen zu dieser Problemstellung nur wenige Arbeiten vor, ein allgemein verwendbares Verfahren wurde 2010 von Douglas und Mislevy publiziert. In der Studie soll am exemplarischen Beispiel einer zusammengesetzten Prüfung, bei der mehrere Teilprüfungen unabhängig voneinander bestanden werden müssen, eine Analyse der Zuverlässigkeit der Entscheidung „bestanden/durchgefallen“ durchgeführt und Konsequenzen für eine verbesserte Methodik zur Identifikation von Studierenden, die die gestellten Mindestanforderungen nicht erfüllen, diskutiert werden.
Methodik: Untersucht wird die Entscheidungsgenauigkeit und -konsistenz von „bestanden/durchgefallen“ des Leistungsnachweises Innere Medizin/Allgemeinmedizin/Klinische Chemie der medizinischen Fakultät Heidelberg. Für diesen müssen drei Teilprüfungen (zwei Klausuren und ein OSCE) unabhängig voneinander bestanden werden, wobei jede Teilprüfung für sich zweimal wiederholt werden kann. Die Analyse erfolgt mit dem Verfahren von Douglas und Mislevy.
Ergebnisse: Auch bei hohen Reliabilitäten von Teilprüfungen lässt sich bei komplexen logischen Verknüpfungen der Einzelentscheidungen „bestanden/durchgefallen“ im Fall niedriger Nichtbestehensquoten häufig nur eine geringe Zuverlässigkeit der Gesamtentscheidung erreichen. So ist im hier untersuchten Beispiel trotz der bei allen drei Teilprüfungen guten Reliabilitäten von über 0,75 die Entscheidungsgenauigkeit und -konsistenz bei konjunktiver Verknüpfung der drei Prüfungsteile mit κ=0,49 bzw. κ=0,47 relativ niedrig. Die Möglichkeit, die Teilprüfungen jeweils zweimal zu wiederholen, führt dazu, dass von den Studierenden, die den Mindestanforderungen nicht genügen, nur etwa die Hälfte endgültig die Gesamtprüfung nicht bestehen würde, die andere Hälfte jedoch trotz mangelhafter Kenntnisse/Fertigkeiten ihr Studium fortsetzen kann.
Schlussfolgerung: Das Verfahren von Douglas und Mislevy erlaubt, Entscheidungsgenauigkeit und -konsistenz komplexer Verknüpfungen von Teilprüfungen zu analysieren. Auch bei hochreliablen Teilprüfungen wird – etwa im Fall niedriger Nichtbestehensquoten – nicht notwendigerweise eine zuverlässige Entscheidung über „bestanden“ oder „durchgefallen“ erreicht. Hierzu müssten Prüfungen mit dem expliziten

Ziel der Identifizierung von Studierenden, die den Mindestanforderungen nicht genügen, durchgeführt werden.

Schlüsselwörter: Prüfungen, Entscheidungsgenauigkeit, Entscheidungskonsistenz, pass-fail-reliability

1. Einleitung

Prüfungen sind Messinstrumente für die Leistungsfähigkeit und besitzen wie alle Messinstrumente nur eine begrenzte Genauigkeit. Diese muss ausreichend hoch sein, damit Prüfungsergebnisse auch eine inhaltliche Aussagekraft aufweisen können. Für die Abschätzung der Messzuverlässigkeit der vergebenen Punktwerte in Prüfungen stehen etablierte Verfahren zur Verfügung (z. B. Cronbachs α), die Zuverlässigkeit der Entscheidung „bestanden/durchgefallen“ findet jedoch bei der Analyse und Bewertung von Prüfungen kaum Beachtung.

Dies ist insofern bemerkenswert, als gerade diese für den Studierenden eine deutlich höhere Bedeutung für den Studienverlauf aufweist als die Messzuverlässigkeit eines Punktwerts, ein „nicht bestanden“ führt zu Nacharbeitungsaufwand, Zeitverlust und u. U. zur Frage nach Fortsetzung oder Beendigung des Studiums. Auch auf Seiten der prüfenden Institution ist die Entscheidung von Bedeutung: Besitzt der Studierende die für die Fortsetzung des Studiums erforderlichen Kenntnisse und Fertigkeiten, führt ein ungerechtfertigtes „durchgefallen“ ebenfalls zu einem höheren Arbeitsaufwand. Lässt man andererseits einen Studierenden trotz fehlender Qualifikation bestehen, so sind nicht nur erhebliche Probleme bei der Fortführung des Studiums zu erwarten, sondern im medizinischen Bereich schlimmstenfalls Gefährdungen von Patienten nicht auszuschließen (vgl. [5]).

Das Thema hat in Deutschland im Bereich der medizinischen Ausbildung vermutlich auch deshalb bislang wenig Beachtung gefunden, als in den Studien- oder Prüfungsordnungen der meisten Fakultäten für das nach wie vor dominierende Prüfungsformat der Multiple-Choice-Prüfungen die Regularien der Ärztlichen Approbationsordnung im Wesentlichen übernommen wurden. Mit der dort rein formal festgelegten Bestehensgrenze von 60% der gestellten Aufgaben ist eine inhaltlich begründete, kriteriumsorientierte Festlegung der Mindestanforderungen nicht möglich. Unseres Wissens lässt es in Deutschland lediglich die Studienordnung der medizinischen Fakultät Heidelberg zu, bei Multiple-Choice-Klausuren ein Standard-Setting durchzuführen, d. h. von der formalen Regel einer 60%-Grenze abzuweichen und – ähnlich zum etablierten Standard-Setting bei einem OSCE – mit einem Standardvorgehen eine an inhaltlichen Kriterien orientierte Bestehensgrenze zu definieren [2], [5].

Die Etablierung neuer Prüfungsformate, mit denen neben reinem Fachwissen auch praktische Fertigkeiten, Qualifikationen und für die Ausübung des Arztberufs erforderliche Kompetenzen geprüft werden sollen, erfordert jedoch die Definition und bei Prüfungen die praktische Festlegung von Mindestanforderungen. Damit wird es auch erforderlich, bei der Beurteilung von Prüfungen oder Prüfungsformen der *Entscheidungsgenauigkeit* („decision accuracy“) und der *Entscheidungskonsistenz* („decision consistency“,„pass-fail-reliability“) eine hohe Aufmerksamkeit zu widmen [19]. Dabei bezeichnet die Entscheidungsgenauigkeit das Ausmaß, in dem Studierende, die den Mindestanforderungen genügen, in einer konkreten Prüfung bestehen und Studierende ohne hinreichende Kenntnisse durchfallen. Die Entscheidungskonsistenz ist die Übereinstimmung von „bestanden/durchgefallen“ in zwei äquivalenten Prüfungen, d. h. in zwei Prüfungen, die *das selbe Wissen/die selben Fertigkeiten gleich gut* messen. Man beachte hier, dass das „Selbe“ hier nicht impliziert, das die Prüfungen im testtheoretischen Sinn nur ein Konstrukt abfragen. Ein OSCE kann Stationen zu praktischen Fertigkeiten („Skills“) und zu kommunikativen Kompetenzen enthalten, die teststatistisch wie Unterskalen aufzufassen sind. Eine hierzu äquivalente Prüfung müsste dann auch in gleichem Umfang und Schwierigkeit praktische und Kommunikationsstationen enthalten.

Für den Fall einzelner Prüfungen sind – insbesondere seit den 1980er Jahren – eine Reihe von Verfahren zur Bestimmung von Genauigkeit und Konsistenz entwickelt worden, wenngleich noch keine dieser Methoden als „Standardprozedur“ angesehen werden kann (vgl. [6], [13], [14], [16], [18], [23], [25]). Zur Erlangung von Leistungsnachweisen in vielen medizinischen Fächern sind jedoch mehrere einzelne Prüfungen abzulegen, etwa eine schriftliche Prüfung für das theoretische Wissen und ein OSCE zur Prüfung der praktischen Fertigkeiten. Werden diese Prüfungsleistungen durch gewichtete Mittelungen oder Summierungen zu einem Gesamtscore verrechnet, kann die gesamte Prüfung wie eine „einzige“ behandelt werden.

Oft findet sich aber eine andere, inhaltlich durchaus begründete, Praxis: Statt die Prüfungsleistungen *kompensatorisch* zu verrechnen, müssen *sämtliche Einzelprüfungen bestanden* werden. Diese *konjunktive Kombination* (logische „und“-Verknüpfung) der Entscheidungen „bestanden“/durchgefallen“ hat erhebliche Auswirkungen auf die Genauigkeit/Konsistenz der Gesamtentscheidung, da eine einzige unzuverlässige Entscheidung bei einer Teilprüfung die Zuverlässigkeit der Gesamtentscheidung zunichte machen kann:

> *...Because longer collections of test questions tend to be more reliable than shorter collections of test questions, compensatory scoring tends to be more reliable than conjunctive scoring. In conjunctive scoring, if a student has to pass all of the content areas separately, the least reliable score controls whether a student will pass*. [26]

Praktische Anwendungsfälle sind hier z. B. Fächer, die die zu prüfenden Lehrinhalte auf mehrere Teilprüfungen aufteilen, um den Umfang einer einzelnen Prüfung zu begrenzen oder Fächer, in denen sowohl theoretisches Wissen wie auch praktische Fertigkeiten vermittelt werden und die deshalb eine schriftliche Prüfung für die Theorie und eine praktische für die Fertigkeiten durchführen. In diesen Fällen ist es häufig gerechtfertigt, das Erreichen von Mindeststandards in jeder Einzelprüfung zu fordern, statt eine Kompensation zu ermöglichen. Schließlich wird auch für das gesamte Studium eine konjunktive Kombination angewandt: Nur wer in *allen Fächern bestanden hat*, hat das Studium erfolgreich beendet.

Prüfungsleistungen können auch noch auf andere Weisen kombiniert werden. So sind neben den bereits erwähnten konjunktiven Verknüpfungen auch disjunktive (logische „oder“-Verknüpfungen) möglich, bei denen von mehreren Prüfungsteilen nur eine einzige bestanden werden muss. Ein Beispiel hierfür wären Wiederholungsprüfungen. Kann eine Prüfung einmal wiederholt werden, hat man insgesamt bestanden, wenn man die erste Prüfung besteht oder die zweite (dass ein Studierender zur zweiten Prüfung nicht antreten muss, wenn er bereits die erste bestanden hat, ist für die Logik ohne Belang). In der schulischen und universitären Praxis sind auch noch komplexere Regularien anzutreffen, wie z. B., dass drei von fünf möglichen Leistungsscheinen erworben werden müssen. Zur Entscheidungszuverlässigkeit bei komplexen Kombinationen von Prüfungsleistungen liegen nur wenige Arbeiten vor [24], ein allgemein einsetzbares Analyseverfahren wurde von Douglas und Mislevy vorgeschlagen [7], [8]. Mit diesem soll in der vorliegenden Studie exemplarisch der fächerübergreifende Leistungsnachweis Allgemeinmedizin/Innere Medizin/Klinische Chemie der medizinischen Fakultät Heidelberg des Wintersemesters 2012/13 untersucht werden, für dessen Erwerb zwei Klausuren und ein OSCE unabhängig voneinander bestanden werden müssen. Dabei steht jedem Studierenden für jede Einzelprüfung die Möglichkeit zweier Prüfungswiederholungen offen.

Der „fächerübergreifende Leistungsnachweis“ (FÜL) ist eine Besonderheit der deutschen Approbationsordnung, nach der im Medizinstudium jede Fakultät mehrere Fächer zu einem gemeinsamen Leistungsnachweis bündeln muss. Diese juristische Vorgabe ist für die folgenden statistischen Betrachtungen jedoch ohne Bedeutung, das Verfahren von Douglas und Mislevy zielt auf die Genauigkeit und Zuverlässigkeit einer komplexen Entscheidung über „bestanden“/„nicht bestanden“ ab, die durch eine Kombination von Einzelentscheidungen gewonnen wird. Ungeachtet der formaljuristischen Begrifflichkeit bei einem FÜL sollen auch hier die Bezeichnungen „Gesamtprüfung“ (für den gesamten Leistungsnachweis) und „Einzel-“ oder „Teilprüfungen“ (für die einzelnen Fachprüfungen) Verwendung finden.

Intention der Arbeit ist, ein für die Analyse der Entscheidungszuverlässigkeit von „bestanden/durchgefallen“ geeignetes Verfahren am Beispiel einer zusammengesetzten Prüfung darzustellen und damit als wesentlichen Bestandteil der Qualitätssicherung von Prüfungen zu etablieren.

2. Grundlagen

Entscheidungsgenauigkeit und Entscheidungskonsistenz

Ausgangspunkt ist die Annahme, dass die zu prüfenden Studierenden bezüglich ihrer Kenntnisse/Fertigkeiten unterteilt werden können in solche, welche die Mindestanforderungen erfüllen („Master“, „competent examinee“) und solche, die ihnen nicht genügen („Non-Master“, „incompetent examinee“). Bei einer Prüfung in einem Fach könnte eine solche Definition z. B. darin bestehen, dass ein Lernzielkatalog existiert und als „Master“, definiert wird, welcher z. B. 70% dieser Lernziele beherrscht.

In einer konkreten Prüfung wird dann eine Auswahl von Lernzielen getroffen, die geprüft werden und eine Bestehensgrenze festgelegt. Diese Bestehensgrenze könnte dann z. B. ebenfalls mit 70% angesetzt werden. So würde z. B. ein Studierender, der 90% aller Lernziele beherrscht, mit großer Wahrscheinlichkeit diese Grenze überschreiten, hingegen jemand, der 72% beherrscht und demzufolge ebenfalls die Mindestanforderungen erfüllt („Master“), wird aber möglicherweise Pech haben und durchfallen. Analoges gilt für Studierende knapp unterhalb der Grenze zum Master, die mit etwas Glück bestehen. Eine eingehendere Diskussion des Unterschieds zwischen der Definition eines Master („performance standard“) und der Bestehensgrenze („passing score“) findet sich etwa in [12] (s. auch [2], [5]).

Abhängig vom Ziel der Prüfung kann die Bestehensgrenze variiert werden, prüft man strenger (höhere Bestehensgrenze) vermindert man die Wahrscheinlichkeit, einen Non-Master bestehen zu lassen, erhöht aber gleichzeitig das Risiko, einen Master fehlzuklassifizieren, indem er durchfällt. Dies ist völlig analog zu diagnostischen Test, die einem „Goldstandard“ (das entspräche dem Wissen, ob jemand Master oder Non-Master ist) ein tatsächliches Testergebnis gegenüberstellen. Fasst man die Prüfung als Test zur „Diagnose“ der Non-Master auf, so besitzt dieser eine gewisse Sensitivität (die Wahrscheinlichkeit, Non-Master durchfallen zu lassen) und eine Spezifität (Wahrscheinlichkeit, dass ein Master besteht). Änderungen des „Cut-Off“-Punkts des Testwerts führen zu einer Erhöhung oder Verringerung der Sensitivität bei gleichzeitiger Verringerung bzw. Erhöhung der Spezifität.

Das Ausmaß, mit dem man durch die Prüfung Master und Non-Master erkennen kann, wird als „Entscheidungsgenauigkeit“ bezeichnet. Der vollständigen Darstellung dient die linke Vierfeldertafel in Tabelle 1, die die Relativanteile für Master/Prüfung bestanden, Master/Prüfung nicht bestanden, Non-Master/bestanden und Non-Master/nicht bestanden aufführt.

Werden zwei *äquivalente* Prüfungen durchgeführt, so ist der Grad der Übereinstimmung der beiden Prüfungsergebnisse die *Entscheidungskonsistenz* („decision consis-

Tabelle 1: Vierfeldertafeln der Entscheidungsgenauigkeit und Entscheidungskonsistenz Die a_i repräsentieren die Relativanteile der Ergebnisse einer Prüfung in Bezug darauf, ob Studierende, die den Minimalanforderungen genügen, bestehen oder nicht (links). So gibt etwa a_2 den Anteil der Studierenden an, die ungenügende Kenntnisse/Fertigkeiten aufweisen (Non-Master), aber dennoch bestanden haben. Für den Fall zweier völlig äquivalenter Prüfungen geben die c_i die analogen Werte an. Aus der Äquivalenz der beiden Prüfungen folgt $c_2 = c_3$.

Entscheidungsgenauigkeit

	Anforderungen		
	erfüllt	nicht erfüllt	
Prüfung	(Master)	(Non-Master)	Summe
bestanden	a_1	a_2	a_{1+2}
durchgefallen	a_3	a_4	a_{3+4}
Summe	a_{1+3}	a_{2+4}	1

Entscheidungskonsistenz

	Äquivalente Prüfung		
	bestanden	durchgefallen	Summe
bestanden	c_1	c_2	c_{1+2}
durchgefallen	c_3	c_4	c_{3+4}
Summe	c_{1+3}	c_{2+4}	1

tency“, „pass-fail reliability“). Die analoge Vierfeldertafel zeigt Tabelle 1 rechts. Bei Äquivalenz der Prüfungen muss der Anteil von Studierenden, der in der ersten Prüfung besteht und in der zweiten nicht, genau so groß sein, wie der, die in der ersten durchfallen und in der zweiten bestehen.

Die beiden in der Literatur am häufigsten verwendeten Kennmaße für die Entscheidungsgenauigkeit und die -konsistenz sind die relative Zahl der Korrektentscheidungen P_a (entspricht der „Korrektklassifikationsrate“ in diagnostischen Tests) bzw.Übereinstimmungen P_c [11] und Cohens κ [4] (für seine Verwendung im Zusammenhang mit Sensitivität und Spezifität diagnostischer Tests s. [3]). Cohens κ korrigiert die Zahl der Korrektentscheidungen P_a und der Übereinstimmungen P_c um den Effekt, der unter Zufall aus den Randsummen der Vierfeldertafel zu erwarten wäre. Die entsprechenden Werte seine durch κ_a bzw. κ_c bezeichnet.

κ nimmt bei vollständiger Übereinstimmung maximal den Wert 1 an. Die Verwendung von κ als Maß der Übereinstimmung wird mancherorts kritisiert (z. B. [10]) und andere Alternativen propagiert. Unseres Erachtens besitzen in diesem Zusammenhang alle Koeffizienten jedoch den Nachteil, dass bei Reduktion auf einen einzigen Index wesentliche Informationen verlorengehen. Es sollte deshalb zur Beurteilung einer Prüfung immer die *gesamte Vierfeldertafel* herangezogen werden.

Verfahren zur Abschätzung der Entscheidungsgenauigkeit und –konsistenz bei einzelnen Prüfungen

In der Literatur wird eine Vielzahl von Verfahren zur Bestimmung der Entscheidungskonsistenz von einzelnen Prüfungen dargestellt, bekannt sind etwa das Verfahren von Livingston-Lewis [16] oder das von Peng-Subkoviak [18]. Übersichten und Vergleiche finden sich etwa bei [6], [13], [14], [23], [25]. Unseres Erachtens kann zum gegenwärtigen Zeitpunkt keine eindeutige Präferierung unter den verschiedenen Methoden vorgenommen werden.

Das Verfahren von Douglas und Mislevy

Das Verfahren von Douglas und Mislevy [7], [8] dient zur Bestimmung der Entscheidungsgenauigkeit und konsistenz bei komplexen Entscheidungsregeln aus den Einzelprüfungen. Voraussetzung ist, dass die Daten der Einzelprüfungen durch eine multivariate Normalverteilung beschrieben werden können und die Reliabilitäten der Prüfungen bekannt sind. In der Praxis sind die Verteilungen von Prüfungsergebnissen jedoch nicht normalverteilt, weshalb eine adäquate Transformation der Daten vorgenommen werden muss. Für die genaue Beschreibung des Vorgehens muss hier auf die Originalliteratur [7], [8] verwiesen werden.

Zum Verständnis sei ein einfaches fiktives Beispiel für die Bestimmung der Entscheidungsgenauigkeit mit zwei Einzelprüfungen graphisch dargestellt (siehe Abbildung 1). Insgesamt hat bestanden, wer beide Einzelprüfungen bestanden hat (konjunktive Verknüpfung).

Abbildung 1a stellt die Verteilung der Prüfungsergebnisse dar. Die Teilnehmer, deren Ergebnisse im gelben Teil der Verteilung liegen, haben beide Einzelprüfungen und somit auch insgesamt bestanden (in der Tabelle 1 ist das der Anteil a_{1+2}). Orange unterlegt ist der Teil der Verteilung, bei dem eine Einzelprüfung bestanden und eine nicht bestanden wurde. Insgesamt haben diese Personen damit nicht bestanden, ebenso natürlich wie diejenigen, die keine der beiden Einzelprüfungen bestanden haben (braun unterlegt). Der Anteil des im Grundriss "L-förmige" Bereichs (orange und braun) derjenigen, die insgesamt nicht bestehen ist in Tabelle 1 mit a_{3+4} bezeichnet.

Im Verfahren von Douglas und Mislevy wird nach dem Modell der klassischen Testtheorie und der Normalverteilungsannahme die Verteilung der „wahren Werte“ bestimmt, also die Verteilung der Werte, wenn diese messfehlerfrei gemessen worden wären. Hierzu müssen die Reliabiltäten der Einzelprüfungen bekannt sein. Die resultierende Verteilung besitzt eine deutlich geringere Varianz. Auf der Ebene der wahren Werte sind „Master“ und „Non-Master“ definiert. Abbildung 1b zeigt diese Verteilung, die Master sind diejenigen, die in beiden abgeprüften Inhalten die Mindestanforderungen erfüllen (grüner Bereich, in Tabelle 1 a_{1+3}), Non-Master die, die mindestens bei einem Gebiet der beiden Einzelprüfungen die Mindest-

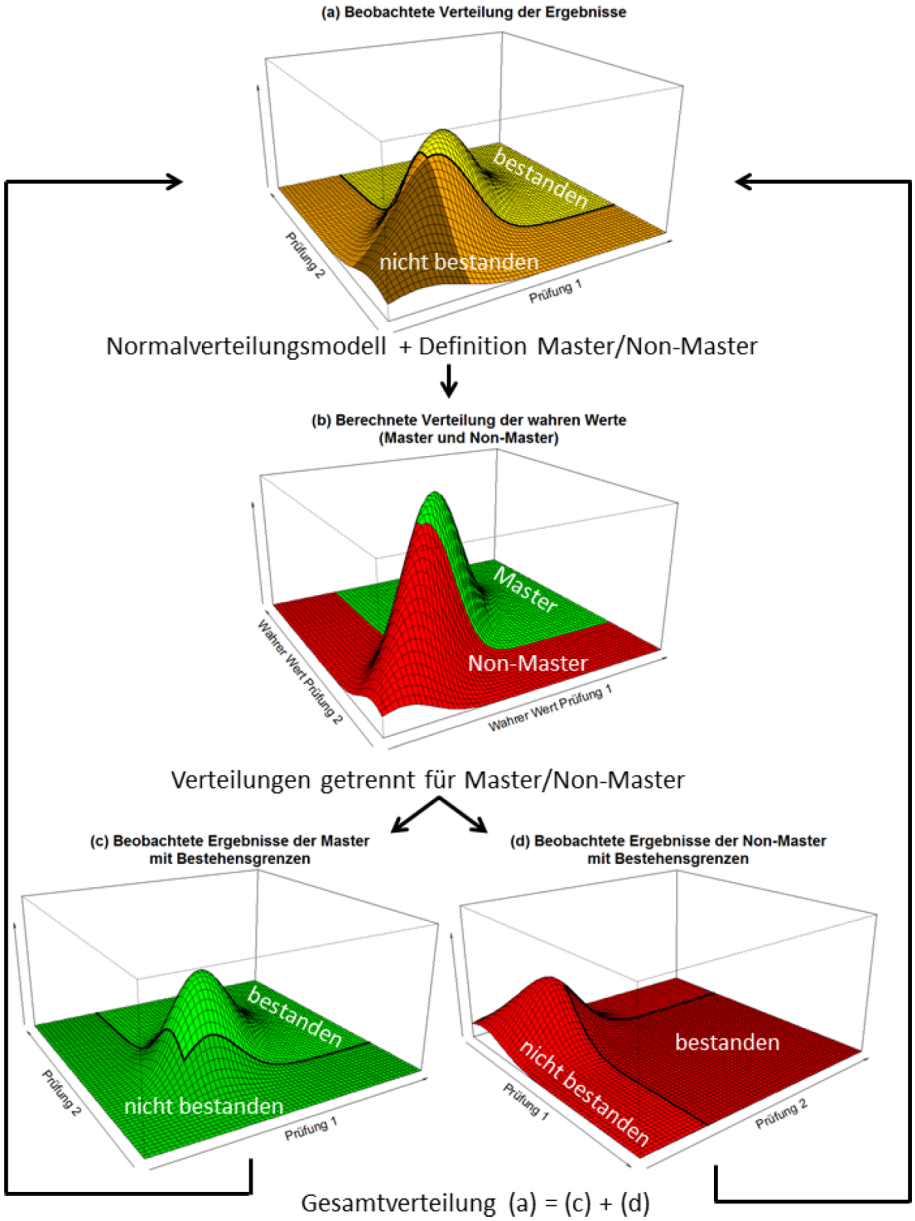


Abbildung 1: Die Schritte des Verfahrens von Douglas und Mislevy: (a) Verteilung der Prüfungsergebnisse zweier Prüfungen. (b) Schätzung des Modells der wahren Werte und Definition von Master/Non-Master. (c) Verteilung der Ergebnisse der Master. (d) Verteilung der Ergebnisse der Non-Master (Beachte: Ansicht gedreht!). Die Verteilung der Gesamtergebnisse (a) setzt sich zusammen aus den Ergebnissen der Master und Non-Master.

anforderung nicht erfüllen (im Grundriss "L förmiger" roter Bereich, a_{2+4} in Tabelle 1).
Zur Bestimmung der Entscheidungsgenauigkeit wird nun im Modell betrachtet, wie die Ergebnisse der Master verteilt sind (siehe Abbildung 1c). Aufgrund der Messfehler der Prüfungen fällt ein Teil der Master durch (dunkelgrüner Bereich). Der hellgrüne Bereich stellt also den Anteil der Master dar, die insgesamt bestehen (in der Tabelle 1 das a_1), der dunkelgrüne den der Master, die insgesamt durchfallen (siehe Tabelle 1 a_3).
Die entsprechende Abbildung für die Non-Master ist Abbildung 1d. Diese ist zur besseren Sichtbarkeit der Grenzlinien aus einer anderen Perspektive dargestellt. Hellrot ist der Anteil der Non-Master, die insgesamt nicht bestehen, dunkelrot derer, die insgesamt bestehen (in Tabelle 1 a_4 bzw. a_2).

Fasst man die beiden Verteilungen der Master und Non-Master in Abbildung 1c und 1d zusammen, so ergibt sich wieder die Gesamtverteilung der Prüfungsergebnisse der Abbildung 1a .

3. Methodik

3.1 Daten

Ziel der Studie ist eine Analyse der Ergebnisse der Prüfungen für den fächerübergreifenden Leistungsnachweis Innere Medizin/Allgemeinmedizin/Klinische Chemie an der medizinischen Fakultät Heidelberg des Wintersemesters 2012/2013. Der Leistungsnachweis besteht aus der schriftlichen Klausur Innere Medizin/Allgemeinmedizin,

Tabelle 2: Basisdaten der Prüfungen des Leistungsnachweises Innere Medizin/Allgemeinmedizin/Klin. Chemie im Wintersemester 2012/2013 (nur Teilnehmer, die an allen drei Prüfungen teilgenommen haben: *N*=147).

	Klausur Inn. Med.	Klausur Chemie	OSCE
max. Punktzahl	50.000	52.000	250.000
Mittel	39.075	42.537	211.357
Bestehensgrenze	30.000	31.200	171.500
Anzahl „nicht bestanden“	4	5	1
Guttmans λ_2)*	0.770	0.762	0.752

)* Zur Abschätzung der Reliabilität wurde statt Cronbachs α die etwas genauere Abschätzung durch den Koeffizienten λ_2 von Guttman gewählt [9].

einer praktisch-mündlichen Prüfung (OSCE) und der Klausur Klinische Chemie. Zusätzlich sind zur Erlangung des Leistungsnachweises noch ein Patientenbericht zu erstellen, ein MiniCEX abzulegen und zur Prüfung professionellen Verhaltens Encounter Cards einzuholen. Da bei diesen die Bestehensrate jeweils 100% beträgt, besitzen sie hier keine Relevanz. Für die Auswertung wurden nur die Studierenden berücksichtigt, die an allen drei Prüfungen teilgenommen haben (*N*=147). Die Basisdaten der Prüfungen sind in Tabelle 2 aufgeführt. Insgesamt sind 7 der 147 Teilnehmer an allen drei Prüfungen bei wenigstens einer Teilprüfung durchgefallen.

Für die Klausuren in den Fächern Klinische Chemie und Innere Medizin wurde als Master definiert, wer 60% der Aufgaben im zugrundeliegenden Aufgabenpool der jeweiligen Fächer zutreffend löst. Für den OSCE ist als Master definiert, wessen durchschnittlich erreichte Punktzahl in OSCE-Stationen des Faches die durch das Standard-Setting festgelegte Punktzahl erreicht („performance standard“, [5]).

Als Bestehensgrenzen für die Klausuren wurden jeweils 60% der erreichbaren Punktzahlen bei den tatsächlich gestellten Aufgaben gewählt, beim OSCE war Bestehensgrenze das Mittel der im Standard-Setting festgelegten Punktzahlen der verwendeten Stationen („passing score“).

3.2 Statistische Analyse

Die Analyse der Entscheidungsgenauigkeit und der -konsistenz von „bestanden“/„durchgefallen“ erfolgt im Wesentlichen mit dem von Douglas und Mislevy vorgeschlagenen Verfahren [7], [8].

Das Verfahren von Douglas und Mislevy macht keine Annahmen über die interne testtheoretische Struktur der Einzelprüfungen noch über die zwischen den einzelnen Prüfungen. Insbesondere müssen die Einzelprüfungen nicht homogen oder eindimensional sein, noch muss durch das Gesamt der Prüfungen eine „einheitliche“ Leistungsdimension abgebildet werden. Voraussetzung ist jedoch, dass die Daten hinreichend gut durch eine Normalverteilung beschrieben werden und die Messzuverlässigkeiten (Reliabilitäten) der Einzelprüfungen adäquat abgeschätzt werden.

Da die Punktwerte der Prüfungen jeweils hochsignifikant von Normalverteilungen abweichen (Shapiro-WilksTests: alle p<0,0008), wurden die Daten einer multivariaten Box-Cox-Transformation unterworfen [1]. Für die so transformierten Daten ergab ein Test auf Abweichung von einer trivariaten Normalverteilung mittels des verallgemeinerten Shapiro-Wilks-Tests von Villasenor-Alva und Gonzalez-Estrada [22] ein p=0,8467 (MVW=0,9929), so dass von einer hinreichend guten Anpassung der Daten ausgegangen werden kann. Im Unterschied zu der in der Arbeit von Douglas und Mislevy verwendeten normalisierenden Rangtransformation, wird mit dieser Transformation eine Anpassung an eine *multivariate* Normalverteilung angestrebt. Zur Abschätzung der Reliabilität der Einzelprüfungen wurde der Koeffizient λ_2 von Guttman gewählt, der eine leicht bessere Schätzung für die Mindestreliabilität als Cronbachs α (=Guttmans λ_3) erlaubt [9].

Die Vierfeldertafeln von Entscheidungsgenauigkeit und -konsistenz für die *Einzelprüfungen* und ihrer *konjunktiven Verknüpfung* wurden durch numerische Integration der multivariaten Normalverteilungen mit dem Algorithmus von Miwa, Hayter und Kuriki [17] bestimmt.

Die Analyse unter *Berücksichtigung zweier Wiederholungsmöglichkeiten* für jede Einzelprüfung ist insofern eher theoretischer Natur, als anzunehmen ist, dass Studierende, die eine Prüfung nicht bestanden haben, auf die Wiederholungsprüfung konzentriert lernen. In der hier vorgenommenen Analyse wird angenommen, dass die Studierenden diese Prüfungen mit demselben Wissenstand ablegen wie die erste. Für die zur Bestimmung der Kennwerte erforderliche Integration einer höherdimensionalen Normalverteilung ist der Algorithmus von Miwa et al. [17] ungeeignet, so dass diese Analyse wie in [8] durch Monte-Carlo-Integration erfolgte. Insgesamt wurden hierzu 100.000 simulierte Datensätze erzeugt um eine ausreichende Genauigkeit der Ergebnisse zu gewährleisten.

4. Ergebnisse

4.1 Einzelprüfungen

Für die einzelnen Prüfungen ergeben sich die Vierfeldertafeln in Tabelle 3, Tabelle 4 umd Tabelle 5. Die aus dem Normalverteilungsmodell resultierende geschätzten Anzahl von Prüfungsteilnehmern, die die Prüfung nicht bestehen, sind: Durchfallquote des Modells ×N=0,0331×147=4,9 für die Klausur Innere Medizin, 3,0 für Klinische Chemie und 1,9 für den OSCE. Damit weichen diese Raten nur wenig von der Zahl der tatsächlich durchgefallenen Studierenden 4, 5 bzw. 1 ab (siehe

Tabelle 3: Entscheidungsgenauigkeit und -konsistenz der Klausur Innere Medizin

	Entscheidungsgenauigkeit			Entscheidungskonsistenz		
	Anforderungen			Äquivalente Prüfung		
Prüfung	erfüllt	nicht erfüllt	Summe	bestanden	durchgefallen	Summe
bestanden	0.9616	0.0053	0.9669	0.9479	0.0190	0.9669
durchgefallen	0.0202	0.0129	0.0331	0.0190	0.0141	0.0331
Summe	0.9818	0.0182	1.0000	0.9669	0.0331	1.0000
	κ_a=0.4919, P_a=0.9746			κ_c=0.4060, P_c=0.9620		

Tabelle 4: Entscheidungsgenauigkeit und -konsistenz der Klausur Klinische Chemie

	Entscheidungsgenauigkeit			Entscheidungskonsistenz		
	Anforderungen			Äquivalente Prüfung		
Prüfung	erfüllt	nicht erfüllt	Summe	bestanden	durchgefallen	Summe
bestanden	0.9763	0.0030	0.9793	9.9664	0.0128	0.9793
durchgefallen	0.0139	0.0068	0.0207	0.0128	0.0079	0.0207
Summe	0.9902	0.0098	1.0000	0.9793	0.0207	1.0000
	κ_a=0.4393, P_a=0.9831			κ_c=0.3672, P_c=0.9743		

Tabelle 5: Entscheidungsgenauigkeit und -konsistenz des OSCE Innere Medizin

	Entscheidungsgenauigkeit			Entscheidungskonsistenz		
	Anforderungen			Äquivalente Prüfung		
Prüfung	erfüllt	nicht erfüllt	Summe	bestanden	durchgefallen	Summe
bestanden	0.9852	0.0016	0.9868	0.9781	0.0087	0.9868
durchgefallen	0.0096	0.0036	0.0132	0.0087	0.0044	0.0132
Summe	0.9948	0.0052	1.0000	0.9868	0.0131	1.0000
	κ_a=0.3835, P_a=0.9888			κ_c=0.3280, P_c=0.9825		

Tabelle 2). Bei allen drei Prüfungen sind Cohen's κ-Koeffizienten κ_a (Entscheidungsgenauigkeiten) und κ_c (Entscheidungskonsistenzen) niedrig.

4.2 Zusammengesetzte Prüfungen

4.2.1 Konjunktive Verknüpfung der Einzelprüfungen

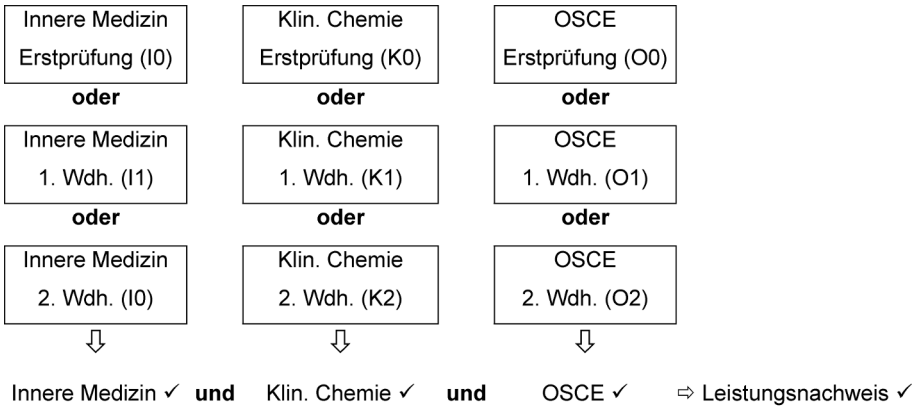
Für die konjunktive Kombination der drei Prüfungen sind Entscheidungsgenauigkeit und -konsistenz in Tabelle 6 aufgeführt. Gemäß dem Modell von Douglas und Mislevy wäre zu erwarten, dass 7,8 Teilnehmer (= Durchfallquote des Modells ×N=0,0531×147=7,8) nicht bestehen, tatsächlich sind 7 Teilnehmer durchgefallen (mehrere der Studierende haben mehr als eine Prüfung nicht bestanden), so dass auch hier eine zufriedenstellende Übereinstimmung des Modells und der tatsächlichen Daten vorliegt. Die Prüfungslogik führt zu einer klaren Aussortierung der Studierenden, die den Anforderungen nicht genügen, der Anteil von Non-Mastern, die bei allen drei Prüfungen bestehen, beträgt insgesamt nur 0,004 (wobei jedoch berücksichtigt werden muss, dass deren Gesamtanteil lediglich bei 0,0232 liegt). Die „Sensitivität“ zur Entdeckung von Non-Mastern beträgt 82%, die „Spezifität“ liegt bei 97%, der positive Vorhersagewert ist mit 36% jedoch gering.
Die Entscheidungskonsistenz (Wiederholung mit drei jeweils äquivalenten Prüfungen) erreicht mit κ_c=0,474 keinen befriedigenden Wert. 94,7% der Prüfungsteilnehmer würden gleich klassifiziert werden (P_c) d. h. bei 5,3% der Teilnehmer erhielte man unterschiedliche Aussagen zum Bestehen des gesamten Leistungsnachweises.

4.2.2 Komplexe konjunktive und disjunktive Verknüpfung bei Prüfungswiederholungen

Insgesamt kann jede der drei Prüfungen Innere Medizin, Klinische Chemie und der OSCE zweimal wiederholt werden, bevor der Studierende endgültig nicht bestanden hat. Logisch bedeutet das, dass ein Studierender eine von drei Klausuren Innere Medizin, eine von drei Prüfungen in Klinischer Chemie und einen von drei OSCEs bestanden haben muss. Innerhalb eines Prüfungsformats wird die Entscheidung bestanden/nicht bestanden also disjunktiv verknüpft, diese drei Teilentscheidungen sodann konjunktiv (siehe Abbildung 2). Die Tatsache, dass ein Studierender, der eine erste Prüfung bestanden hat, gar nicht zu einer weiteren antritt, ist für die Entscheidungslogik ohne Belang.
In Tabelle 7 sind die Vierfeldertafeln für die Entscheidungsgenauigkeit und -konsistenz unter der Annahme dargestellt, dass ein Studierender in allen Prüfungen mit dem selben Wissensstand antritt.
Bedeutend ist hier vor allem, dass von den 2,32% der Studierenden (a_{2+4}=0,0232), die die Anforderungen nicht erfüllen (Non-Master), mehr als die Hälfte (a_2=0,0124) den Leistungsnachweis schlussendlich doch erhalten würde, d. h. durch die Möglichkeit der Wiederholungen wird nur noch ein Teil der Studierenden, die den Anforderungen nicht genügen, vom Weiterstudium ausgeschlossen (man beachte den substantiellen Unterschied zu den

Tabelle 6: Entscheidungsgenauigkeit und -konsistenz des Leistungsnachweises Innere Medizin/Allgemeinmedizin/Klinische Chemie (konjunktive Kombination)

	Entscheidungsgenauigkeit			Entscheidungskonsistenz		
	Anforderungen			Äquivalente Prüfung		
Prüfung	erfüllt	nicht erfüllt	Summe	bestanden	durchgefallen	Summe
bestanden	0.9428	0.0040	0.9469	0.9204	0.0265	0.9469
durchgefallen	0.0340	0.0191	0.0531	0.0265	0.0267	0.0531
Summe	0.9768	0.0232	1.0000	0.9469	0.0531	1.0000
	κ_a=0.4852, P_a=0.9620			κ_c=0.4740, P_c=0.9471		



Prüfungen zu Leistungsnachweis bestanden =

(I0 bestanden oder I1 bestanden oder I2 bestanden) und

(K0 bestanden oder K1 bestanden oder K2 bestanden) und

(O0 bestanden oder O1 bestanden oder O2 bestanden)

Abbildung 2: Entscheidungslogik für das Erreichen des fachübergreifenden Leistungsnachweises Innere Medizin/Allgemeinmedizin/Klinische Chemie.

Tabelle 7: Entscheidungsgenauigkeit und -konsistenz des Leistungsnachweises Innere Medizin/Allgemeinmedizin/Klin. Chemie mit zwei Wiederholungsmöglichkeiten (siehe Abbildung 2)

	Entscheidungsgenauigkeit			Entscheidungskonsistenz		
	Anforderungen			Äquivalente Prüfung		
Prüfung	erfüllt	nicht erfüllt	Summe	bestanden	durchgefallen	Summe
bestanden	0.9746	0.0124	0.9870	0.9809	0.0062	0.9870
durchgefallen	0.0022	0.0108	0.0130	0.0062	0.0068	0.0130
Summe	0.9768	0.0232	1.0000	0.9870	0.0130	1.0000
	κ_a=0.5890, P_a=0.9854			κ_c=0.5181, P_c=0.9877		

Ergebnissen des vorigen Abschnitts, bei dem der entsprechende Wert mit a_2=0,0040 in Tabelle 6 gegenüber 0,0124 in Tabelle 7 deutlich niedriger ist).

5. Diskussion

Einzelprüfungen

Entscheidungsgenauigkeit: Alle drei Einzelprüfungen weisen eine insgesamt zufriedenstellende Reliabilität auf (siehe Tabelle 2). Von den „Non-Mastern“, die insgesamt nur einen Anteil von 0,5 – 1,8% (siehe Tabellen 3 bis 5, a_{2+4}) der Prüfungsteilnehmer ausmachen, besteht aber jeweils ein knappes Drittel die Prüfungen (a_2). Der relative Anteil der Master, die die Prüfung nicht bestehen ist in allen Fällen gering (a_3), in absoluten Zahlen sind dies aber jeweils deutlich mehr als Non-Master an der Prüfung teilnehmen, so dass bei allen drei Prüfungen mehr als doppelt so viel Kandidaten durchfallen als Non-Master in der Gruppe anzunehmen sind.

Entscheidungskonsistenz: Die Zuverlässigkeit der Entscheidung „durchgefallen“ muss als unzufriedenstellend eingestuft werden. Von denjenigen, die durchfallen, würden etwa 60–65% bei einer äquivalenten Wiederholungsprüfung bestehen. Die geringe Entscheidungskonsistenz zeigt sich auch in den niedrigen κ_c-Werten von 0,33–0,41

Konjunktive und komplexe Verknüpfungen der Prüfungsergebnisse

Entscheidungsgenauigkeit: Die Vierfeldertafel der Entscheidungsgenauigkeit für die konjunktive Verknüpfung der drei Prüfungen (siehe Tabelle 6) zeigt, dass von den

2,3% Non-Mastern (also Studierende, die in mindestens einem der drei Fächer den Mindestansprüchen nicht genügen), lediglich 17% bestehen (a_2/a_{2+4}=0,040/0,232=0,0172). Es erhöht sich jedoch der relative Anteil an Mastern, die durchfallen auf 3,5% (a_3/a_{1+3}=0,0348), bei den Einzelprüfungen lag dieser Anteil bei höchstens 2%. Auch hier fallen deutlich mehr Prüflinge durch (5,3%, a_{3+4}) als Non-Master teilnehmen. Cohens κ_a ist mit 0,49 fast genauso hoch wie das des besten κ_a bei den Einzelprüfungen (Klausur Innere Medizin), geringer ist der Anteil an Korrektklassifikationen mir P_a=0,96. Die Aussage, dass bei konjunktiven Verknüpfungen die Prüfung mit der schlechtesten Entscheidungsgenauigkeit dominiert ist demzufolge etwas differenzierter zu beurteilen.

Berücksichtigt man die Tatsache, dass jedem Studierenden zwei Wiederholungsmöglichkeiten zur Verfügung stehen (siehe Tabelle 7), so müssen unter der Annahme, dass die Studierenden mit dem selben Wissen oder Können in äquivalente Wiederholungsprüfungen gehen, nur 47% der Non-Master den Leistungsnachweis endgültig nicht erhalten (a_4/a_{2+4}=0,0108/0,0232=0,4655). Bei den Mastern ist der Anteil mit 2.3‰ verschwindend gering (a_3/a_{1+3}=0,0022/0,9768=0,0023). Damit ist die Prüfungsstruktur mit den beiden Wiederholungsmöglichkeiten für jede Einzelprüfung offensichtlich nur wenig geeignet, die Non-Master zuverlässig zu erkennen.

Entscheidungskonsistenz: Bei der konjunktiven Verknüpfung der drei Prüfungen ist die Stabilität der Entscheidung „durchgefallen“ ebenfalls nicht zufriedenstellend, aber etwas besser als in den Einzelprüfungen. Bei einem äquivalenten Prüfungskomplex, bestehend aus den Prüfungen in den drei Fächern, würde etwas mehr als die Hälfte die Prüfung bestehen. Wird als Index für die Konsistenz κ_c verwendet, so ist dieser mit 0,47 höher als bei jeder Einzelprüfung.

Bei Berücksichtigung der Wiederholungsmöglichkeiten zeigt sich ein ähnliches Bild, nur etwas mehr als die Hälfte der Studierenden, die letztendlich durchfallen, würden bei einem „Neustart von Anfang an“ erneut ihr Studium abbrechen müssen.

Resümee

Zusammenfassend muss festgestellt werden, dass die Entscheidung „bestanden“/„durchgefallen“ sowohl hinsichtlich ihrer Genauigkeit als auch ihrer Konsistenz mit den durchgeführten Prüfungen einer Verbesserung bedarf. Durch die Wiederholungsmöglichkeiten ist auch ein „Aussieben“ der Non-Master nicht zuverlässig möglich, andererseits besteht kaum Gefahr, dass jemand, der den Anforderungen genügt, auf Grund ein- oder mehrfachen Pechs bei Prüfungen sein Studium beenden muss.

Als Grund hierfür könnte zunächst das Normalverteilungsmodell für das Ergebnis verantwortlich gemacht werden. Um bei niedrigen Nichtbestehensquoten eine akzeptable Entscheidungsgenauigkeit und -konsistenz zu erreichen, benötigt man eine äußerst hohe Reliabilität (für den κ_c-Koeffizienten ist eine entsprechende Tabelle in [21] angegeben). Diese Eigenschaft ist jedoch nicht spezifisch für die Normalverteilung, hier nicht dargestellte Analysen für andere Verteilungsannahmen führen zu ähnlichen Resultaten. Bei den üblichen Annahmen für die Verteilungsform der Punktzahlen in Prüfungen liegen bei *niedrigen Nichtbestehensquoten und nicht exzessiv hohen Reliabilitäten* die meisten Non-Master in der Nähe der Bestehensgrenze. Dies ist unabhängig davon, ob es sich um eine (z. B. gesetzlich vorgegebene) formale, norm- oder kriteriumsorientierte Grenze handelt. Deshalb besteht eine relativ hohe Wahrscheinlichkeit, dass die Non-Master „mit etwas Glück“ bestehen, so dass bei diesen weder eine hohe Genauigkeit noch Konsistenz zu erwarten ist.

Einschränkungen des Verfahrens von Douglas und Mislevy

Die wesentliche Beschränkung des Verfahrens von Douglas und Mislevy ist die Annahme einer multivariaten Normalverteilung. Bei den hier analysierten Prüfungen war durch eine multivariate Box-Cox-Transformation eine akzeptable Normalisierung der Daten möglich, was für die Daten anderer Prüfungen nicht in jedem Fall gelingen wird. Weiter impliziert die Annahme einer multivariaten Normalverteilung für die wahren Werte und Messfehler einen konstanten Messfehler. An der Bestehensgrenze kann der Messfehler jedoch deutlich höher sein und zu einer Überschätzung der Entscheidungsgenauigkeit führen. Andererseits sind die Verteilungen der beobachteten Punktwerte deutlich linksschief, durch die normalisierende Transformation werden die Werte sehr schlechter Studierender „näher’ an die Bestehensgrenze gerückt, womit sie in der Analyse eher zur Gruppe derjenigen zählen, für die aufgrund der Messungenauigkeit fehlerhafte oder inkonsistente Entscheidungen zu erwarten sind, obwohl sie in der Originalskala zuverlässig als Non-Master erkannt werden.

Niedrige Entscheidungsgenauigkeit und -konsistenz: Konsequenzen für Prüfungen

Bei niedrigen Nichtbestehensraten, wie sie in der analysierten Prüfung auftreten, wäre zum Erreichen einer ausreichenden Zuverlässigkeit der Entscheidung „bestanden“/„durchgefallen“ eine hochreliable Prüfung erforderlich. Dies ist insofern nicht überraschend, als bei einer Prüfung mit den üblicherweise angestrebten Messzuverlässigkeiten ein Großteil der Aufgaben gute Trenneigenschaften für den Großteil der Probanden aufweist, für die Separation der dünn besetzten Extremgruppen aber wenig Informationen liefert. Ein – in der üblichen Prüfungspraxis der Universitäten zwar schwer etablierbares – Vorgehen wäre die Durchführung von zwei Prüfungen: Die erste dient der üblichen Bewertung der studentischen Leistungen, die zweite wird speziell zur Identifikation von Mastern und Non-Mastern mit spezifisch für diesen Zweck selegierten Aufgaben durchgeführt (auf letzteres hat bereits Kane [12], p. 430) hingewiesen. In der ersten Prüfung

wird eine relativ hohe Bestehensgrenze eingesetzt, mit der die Wahrscheinlichkeit, dass ein Non-Master besteht, sehr gering bleibt. Die verbleibende Gruppe besteht dann aus (schlechten) Mastern und Non-Mastern, die im zweiten Test möglichst gut zu separieren ist. Methoden der optimalen Aufgabenwahl finden sich in der Literatur zu „(computerized) classification tests“ (CCT) (z. B. [20], [15]).

6. Zusammenfassung und Ausblick

Das Verfahren von Douglas und Mislevy ist dazu geeignet, Prüfungen, die sich aus mehreren Teilprüfungen zusammensetzen und bei denen die Gesamtentscheidung über „bestanden/durchgefallen“ das Resultat einer komplexen Verknüpfung der Einzelergebnisse ist, hinsichtlich der Entscheidungsgenauigkeit und -konsistenz dieser Gesamtentscheidung zu analysieren. Praktisch bedeutsam sind vor allem konjunktive Verknüpfungen (jede einzelne Prüfung muss bestanden werden) oder disjunktive Verknüpfungen (von mehreren Prüfungen muss nur eine bestanden werden, dies gilt etwa für Wiederholungsprüfungen).

Als Beispiel wurde der für die gegenwärtige deutsche Medizinerausbildung bedeutsame Fall eines „fächerübergreifenden Leistungsnachweises“ gewählt. In diesem Beispiel werden theoretische und praktische Prüfungen verschiedener Fächer kombiniert, zum Bestehen ist das Bestehen jeder einzelnen Prüfung erforderlich. Für jede Einzelprüfung stehen dem Studierenden zwei Wiederholungsmöglichkeiten zur Verfügung.

Mit dem Verfahren von Douglas und Mislevy konnten Entscheidungsgenauigkeit und konsistenz des Leistungsnachweises erfolgreich analysiert werden, es zeigte sich eine hohe Übereinstimmung des Modells mit den Daten. Die Analyse zeigte auch eine wesentliche Problematik von Prüfungen bei niedrigen Durchfallquoten auf: Mit Prüfungen, die den gewöhnlichen Ansprüchen an eine hinreichende Reliabilität entsprechen, sind diese nur schwer zuverlässig zu identifizieren. Erforderlich wären zielgerichtete Klassifikationstests mit entsprechender Aufgabenwahl zur Identifikation der „Master“ und „Non-Master“.

Eine Analyse von Entscheidungsgenauigkeit und -konsistenz sollte bei relevanten Prüfungen allgemein durchgeführt werden. Die Beschränkung auf das Normalverteilungsmodell muss noch als erheblicher limitierender Faktor betrachtet werden, es ist zu hoffen, dass geeignete Verfahren mit schwächeren Verteilungsannahmen (z. B. multivariate Betabinomialverteilungen) oder verteilungsfreie Methoden entwickelt werden.

Interessenkonflikt

Die Autoren erklären, dass sie keine Interessenkonflikte im Zusammenhang mit diesem Artikel haben.

Literatur

1. Andrews DF, Gnanadesikan R, Warner JL. Transformations of multivariatedata. Biometrics. 1971;27:825–840. DOI: 10.2307/2528821
2. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Med Teach. 2008;30(9-10):836–845. DOI: 10.1080/01421590802402247
3. Brenner H, Gefeller O. Chance-corrected measures of the valdity of a binary diagnostic test. J Clin Epidemiol. 1993;47(6):627–633. DOI: 10.1016/0895-4356(94)90210-0
4. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measure. 1960;20:37–46. DOI: 10.1177/001316446002000104
5. Cusimano MD. Standard setting in medical education. Acad Med. 1996;71:112–120. DOI: 10.1097/00001888-199610000-00062
6. Deng N. Evaluating IRT-and CTT-based Methods of Estimating Classification Consistency and Accuracy Indices from Single Administrations. Massachusetts: University of Massachusetts; 2011. Open Access Dissertations. Paper 452. Zugänglich unter/available from: http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1451&context=open_access_dissertations
7. Douglas KM. A general method for estimating the classification reliability of complex decisions based on configural combinations of multiple assessment scores.Unpublished dissertation. Maryland: University of Maryland; 2007.
8. Douglas KM, Mislevy RJ. Estimating classification accuracy for complex decision rules based on multiple scores. J Educ Behav Stat. 2010;35:280–306. DOI: 10.3102/1076998609346969
9. Guttman LA. A basis for analyzing test-retest reliability. Psychomet. 1945;10:255–282. DOI: 10.1007/BF02288892
10. Gwet KL. Handbook of inter-rater reliability (2nd ed.). Gaithersburg: Advanced Analytics, LLC; 2010.
11. Hambleton RK, Novick MR. Toward an integration of theory and method for criterion-referenced tests. J Educ Meas. 1973;10:159–96. DOI: 10.1111/j.1745-3984.1973.tb00793.x
12. Kane M. Validating the performance standards associated with passing scores. Rev Educ Res. 1994;64:425–461. DOI: 10.3102/00346543064003425
13. Kim DI, Choi SW, Um KR. A comparison of methods for estimating classification consistency. Paper presented at the 2006 Annual Meeting of the National Council on Education in Measurement. San Francisco, CA: National Council of Edudation in Measurement; 2006.
14. Lee WC. Classifcation consistency and accuracy for complex assessments using item response theory. CASMA Research Report No. 27. Iowa City, IA: University of Iowa; 2007.
15. Lin CJ. Item selection criteria with practical constraints for computerized classification testing. Educ Psychol Meas. 2011;71:20-36. DOI: 10.1177/0013164410387336
16. Livingston SA, Lewis C. Estimating the consistency and accuracy of classifications based on test scores. J Educ Meas. 1995;32:179–197. DOI: 10.1111/j.1745-3984.1995.tb00462.x
17. Miwa A, Hayter J, Kuriki S. The evaluation of general non-centred orthant probabilities. J Royal Stat Soc. 2003;65:223-U234. DOI: 10.1111/1467-9868.00382
18. Peng CJ, Subkoviak MJ. A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. J Educ Meas. 1980;17:359–368. DOI: 10.1111/j.1745-3984.1980.tb00837.x

19. Schuwirth L, Colliver J, Gruppen L, Kreiter C, Mennin S, Onishi H, Pangaro L, Ringsted C, Swanson D, van der Vleuten C, Wagner-Menghin M. Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):224–233. DOI: 10.3109/0142159X.2011.551558

20. Spray JA, Reckase MD. Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. J Educ Behav Stat. 1996;21:405–414. DOI: 10.3102/10769986021004405

21. Subkoviak MJ. A practitioner's guide to computation and interpretation of reliability indices for mastery tests. J Edc Meas. 1988;25:47–55. DOI: 10.1111/j.1745-3984.1988.tb00290.x

22. Villasenor-Alva JA, Gonzalez-Estrada E. A generalization of Shapiro-Wilk's test for multivariate normality. Communication Stat Theo Method. 2009;38:1870–1883. DOI: 10.1080/03610920802474465

23. Wan L,Brennan RL, Lee W. Estimating classification consistency for complex assessments. CASMA Research Report No. 22. Iowa City, IA: University of Iowa; 2007.

24. Wheadon C, Stockford I. Estimation of composite score classification accuracy using compound probability distributions. Psychol Test Assess Mod. 2013;55:162–180.

25. Zhang B. Assessing the accuracy and consistency of language proficiency classification under competing measurement models. Lang Test. 2010;27:119–140. DOI: 10.1177/0265532209347363

26. Zieky M, Perie M. A Primer on Setting Cut Scores on Tests of Educational Achievement. Washington/DC: Educational Testing Service; 2006. Zugänglich unter/available from: http://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf

Korrespondenzadresse:
Dr. phil. Andreas Möltner
Ruprecht-Karls-Universität Heidelberg,
Kompetenzzentrum Prüfungen in der Medizin
Baden-Württemberg, Im Neuenheimer Feld 346, 69120 Heidelberg, Deutschland, Tel.: +49 (0)6221/56-8249,
Fax: +49 (0)6221/56-7175
andreas.moeltner@med.uni-heidelberg.de

Bitte zitieren als
Möltner A, Tımbıl S, Jünger J. The reliability of the pass/fail decision for assessments comprised of multiple components. GMS Z Med Ausbild. 2015;32(4):Doc42.
DOI: 10.3205/zma000984, URN: urn:nbn:de:0183-zma0009843

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2015-32/zma000984.shtml

Eingereicht: *20.12.2013*
Überarbeitet: *12.03.2014*
Angenommen: *26.05.2014*
Veröffentlicht: *15.10.2015*