

Introduction

The quality of tests has always been of great concern to test developers. In the past quality was expressed as a reliability coefficient, and later, within the IRT framework, by the Fisher information function. More recently there has also been some attention to classification reliability of tests. For a survey of the literature see Douglas (2007). Verstralen and Bechger (2008) wrote a report on the classification accuracy of educational tests as well. This report used a later version of program developed by Verstralen in 1997 to be used in Sluijter (1998). However, usually the results of a test are not used in isolation. Often tests are part of a battery of tests, and their results are combined to decide on a classification of the testee. For instance with exams, the main decision for the examinee is whether or not she passes. As also noted in Douglas (o.c.) this subject has had little attention in the literature.

In a previous report Verstralen (July, 2009) the accuracy of exams was studied within a Classical Test Theory (CTT) framework. Douglas, like Verstralen, did her study within the framework of CTT using simulation. However, she assumed the standard error for all true scores a constant, whereas Verstralen, following Feldt (1985), used a conditional variance of observed scores given the true score.

It is common practice in the literature on reliability of decisions to distinguish between accuracy and consistency. Accuracy refers to the similarity of true classifications and observed classifications, whereas consistency concerns the similarity of classifications of observed classifications based on two parallel tests or exams.

The main conclusion in Verstralen (2009) was that a Resit had a very positive effect on the quality of the decision procedure. Actually, the emphasis was on a large improvement on Undue Failures. Now these are an essential part of accuracy, but do not cover accuracy sufficiently. In the present report this finding is further scrutinized by also using an Item Response Theory (IRT) framework. Further, the pass/fail decision rules have been more polished, by systematically varying the main criteria that are used in practical exams.

Research questions

There are so many variables that can be varied in an investigation like this, that the results tend to become incomprehensible. Therefore, a selection on which variables to focus is necessary. First the variables not investigated are discussed.

Variables held constant (no research question)

As already mentioned in the introduction the main focus here is on scrutinizing the favorable influence of a Resit on the classification quality of exams. This means that in some respects this study is more elaborate than the first, but also that some of the questions of the first investigation are neglected here. Most prominently in this investigation the number of tests of an exam will not be varied. The number of tests in the exam is fixed at seven. Also the covariance between tests is kept constant. An investigation by Van Rijn a.o. (2009) revealed that the average of true correlations between grades of tests in the Dutch exams for pre university education were about 0.50. Because grades are for the most part linear in the test scores, and abilities are more or less linear in the test scores, this value is adopted here. So the correlation between true scores on all pairs of tests is assumed to be 0.50, as are the correlations between the true abilities on all pairs of tests. All items are assumed binary scored. As with the previous investigation the method used is simulation. Per simulation 10.000 students are drawn from the multivariate distribution.

IRT

Also not varied are the parameters of the items. For every test they are equally spaced between -0.25, and 0.45. The discrimination parameters are all equal to 3. So, actually, the Rasch model has been used. The mean and standard deviation of the ability θ on all tests is set at (0.35,0.33). Five cutoff points for each test have to be defined to be able to use all passing rules. They were set for every test at -0.4, -0.2, 0.1, 0.4, and 0.6 on the latent scale, with the third (0.1) the Fail/Pass point. This setup results in pass percentages around 51%, although still with a large variation depending on passing procedure. Keeping the item parameters fixed, instead of sampling them e.g. uniformly, reduces the variability of results by irrelevant sources.

CTT

As test length is only instrumental in obtaining scores test length is invariably set at 40, as in the previous study. The measurement accuracy of the test is varied using the KR20 (see below). Means and standard deviations of the true scores on the tests are set at (28.7, 6). The five cutoff points on the score scale are 15, 20, 25, 30, 35. As in our previous CTT study the local score variation given the true score τ is governed by

$$S_E(\tau) = \sqrt{\frac{\tau(k-\tau)(1-r_{x'x})}{(k-1)(1-r_{21})}}$$

where $r_{x'x}$ is the KR20 and r_{21} the KR21

$$r_{21} = \frac{ks_x^2 - \mu \langle k - \mu \rangle}{(k - 1)s_x^2} = 0.795$$

(Feldt a.o. 1985). The value for r_{21} is obtained with the chosen values for $k = 40$, $(\mu, s_x^2) = (28.7, 36)$. At $\tau = 25$ the following values are obtained for $S_E(\tau)$ at the values used for the KR20 (see below):

Table 1: Values of $S_E(\tau)$ (CTT)

KR20	$S_E(15)$	$S_E(20)$	$S_E(25)$	$S_E(30)$	$S_E(35)$
0.64	4.106	4.240	4.106	3.672	2.805
0.78	3.210	3.315	3.210	2.871	2.193
0.88	2.370	2.448	2.370	2.120	1.619
0.94	1.676	1.731	1.676	1.499	1.145

At this point some scrutiny on this CTT approach to local score variance is in order. Table 1 clearly shows that the local score variance of the test strongly depends on the reliability. However, from the general concept of reliability as the ratio of true and observed score variance

$$r = \frac{\sigma_\tau^2}{\sigma_x^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_E^2}$$

one sees that keeping constant the average local variance σ_E^2 , and increasing the true score variance σ_τ^2 increases the reliability. So an increase in reliability does not necessarily imply a decrease of the local score variance as is the suggestion of Table 1. However, because here the population and so the true score variance σ_τ^2 is kept constant an increase in KR20 necessarily involves a decrease in the mean local variance σ_E^2 .

A complete comparison with the IRT setup is not possible since there reliability is manipulated via the test length. However, we can compare the IRT 40 item test, which has a reliability of 0.88 with the third line in Table 1. As is well known the local variance of the sufficient statistic of an exponential model is given by the Fisher information function. Because all items have discrimination index equal to 3 that information has to be divided by 9 to get the variance of the raw score. Taking the square root of the result gives the entries in Table 1b.

Table 1b: Values of $S_E(\vartheta)$

KR20	$S_E(-0.4)$	$S_E(-0.2)$	$S_E(0.1)$	$S_E(0.4)$	$S_E(0.6)$
0.88	2.43	3.00	3.08	2.67	2.29

Comparison of Table 1b with the third line in Table 1 makes one conclude that given the choices made here respectively within the IRT and CTT frameworks the IRT framework shows the higher local score variances.

Variables varied (research questions)

From our previous research it was found that reliability (KR20) of the tests had the clearest positive influence on the classification accuracy of the exam. In the present application we adopt two measurement models, CTT and as an IRT model OPLM for dichotomous questions. For CTT the KR20 can be varied again. However, in OPLM the length of the test and the discrimination of the items in the test are the main determinants of the reliability of the test. Because the asymptotic measurement accuracy is governed by the information function, one would expect that it is immaterial whether a ten item test will be lengthened to 40, or its discrimination indices are multiplied by two. Both measures result in a fourfold increase in the information function, and also result in the same increased KR20. Here only test length is used with values 10, 20, 40, 80. In combination with the item and population parameters this results in a KR20 of respectively 0.64, 0.78, 0.88, and 0.94. These KR20 values will also be used in the CTT study.

Compared to the previous research the pass/fail procedures have been better organized. Whereas they were more or less inspired by common practices in actual exams, the essential elements have been extracted and systematically varies in the present investigation. These are

1. Number of Resits: (values 0,1,2). Whereas in the previous study the number of Resits was maximally one, in the

present study maximally 2 Resits are allowed. If one fails the exam, the test or tests with the worst results are redone using the same true score or ability. Especially with compensatory pass procedures, more complicated strategies are imaginable. However, here invariably the tests with the worst results are redone. Moreover, although in practice a student can try to increase her ability between the first test administration and the Resit, here it is assumed that the ability, or true score at the Resit remains unchanged.

2. Levels of compensation with the compensatory passing rule. An example can clarify this. In the Dutch educational system grades range from 1 (worst) to 10 (excellent). Just satisfactory is associated with a grade 6 (≥ 5.5). At one level of compensation one still passes if one obtained a grade 5 for a test if for another test one obtains a 7. At two levels of compensation one can also compensate a 4 by an 8. We use 1, and 2 levels of compensation. Were the number of levels equal to 0 the procedure would be simply conjunctive: one has to pass for all tests. This situation is achieved using the next variable.
3. Maximal number of compensations. We use the values 0,1,2,3. At 0, like with 0 levels, the compensation procedure does not compensate at all and is equal to a conjunctive procedure. With maximally 3 compensations one may compensate three bad results by three good results. E.g. combined with two levels three 4's by three 8's.
4. Size of the group of tests that has to be passed. For some exams there are certain subjects that you may not fail on to pass the exam, because they are considered essential in general, or for the specific type of education. Values range from 0 to 3.
5. Passing rule. Besides a compensatory passing rule, also a leniency rule has been used. This rule operates with one exception similar to the compensatory rule except that no compensations are required. So at one level of leniency and maximal number of compensations equal to one you can pass with one grade 5. With two levels, however, the lower level is weighed by two. So if, with two levels, the maximum number of compensations equals 3, you may pass by either having three 5's or by having one 4 and one 5.

Summarizing the passing rules. Two types are distinguished

- Compensatory
- Leniency

Both types are fine tuned by number of levels and Max number of Comp (compensations or leniencies). With the Compensatory rules one can compensate a bad result by a good result, with the Leniency rule some bad results are just allowed. How many bad results can be compensated is governed by MaxComp, and the allowed badness of test results that can be compensated/allowed is governed by Level of compensation/leniency. In the sequel the PassTypes are numbered 1 through 4 as follows

1. Compensatory one level of compensation
2. Compensatory two levels of compensation
3. Leniency one level
4. Leniency two levels

IRT and CTT are compared only concerning the effect of Resits, as a continuation of the previous report (Verstralen, 2009). The other variables are studied only in the context of IRT, because this theory offers a better framework for local score variability than CTT.

Classifying candidates

Each of the pass/fail decision rules is based on the classification of a candidate with respect to the cutoff points on each of the tests in the exam. For each student and each test we obtain a true classification and an observed classification. The true pass fail decision is based on the true classifications, and the observed pass/fail decision on the observed classifications. The true classification on a test is based on his ability or true score on the test. In the CTT case the cutoff points are defined on the score scale, so given the true score on a test the true classification is determined. The same holds for the observed classification which is determined by the observed score. In the IRT case it is a little more involved. The cutoff points of a test are defined on the ability scale of the test. So given the (true) ability of a candidate his true classification with respect to the cutoff points is known. After doing the test we obtain a vector of item scores. This score vector is transformed to a weighted score, the sufficient statistic in the OPLM model. For each cutoff point on the latent scale of the test the expected weighted score is determined. The classification of the observed weighted score is determined by its position with respect to the expected weighted scores of the cutoff points. This amounts to using the ML ability estimator for obtaining an observed classification of a candidate with respect to the original cutoff points on the latent scale.

Measures of decision accuracy

Given a pass/fail procedure, for each candidate a true and an observed pass/fail decision is obtained. By simulating this process for 10.000 candidates each independently randomly drawn from the specified multivariate distribution one obtains for each combination of true and observed passes or fails a count. E.g. the number of times a candidate ought to pass according to his true values but failed according to his observed results on the tests, the so-called undue failures. The basic data on the accuracy of a decision process is, therefore, a 2×2 table

	Observed	Pass	Fail
True			
Pass		a	b
Fail		c	d

with e.g. b the number of candidates out of 10.000 who unduly failed. Clearly the larger a and d the better the procedure succeeds in correct classification. To investigate the effect of the investigated variables much more clarity will be gained if accuracy can be summarised with one number instead of the four numbers in the 2×2 table. Initially four measures to summarize the accuracy inherent in the 2×2 table were tried.

- Tetrachoric correlation. In the previous research the cosine pi approximation was used. But this approximation appeared to be really very approximate. Its error can exceed 0.5 (Divgi, 1979). Therefore, in this study an alternative approximation developed by Divgi has been tried.
- Pearson's χ^2 with Yates's correction for column and row totals (Guilford and Fruchter, 1978, pg 205)

$$\chi^2 = \frac{n \{ |ad - bc| - n/2 \}^2}{(a+b)(a+c)(b+d)(c+d)}$$

with $n = a + b + c + d$, the total number of observations.

- Coefficient ϕ

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Obviously there is a close relationship between χ^2 and ϕ : neglecting the correction term $n/2$ one has $\chi^2 = n\phi^2$

- Cohen's κ

$$\kappa = \frac{p_a - p_r}{1 - p_r}$$

where

$$p_a = (a + d)/n,$$

the observed proportion of agreement, and

$$p_r = ((a + b) * (a + c) + (b + d) * (c + d))/n^2$$

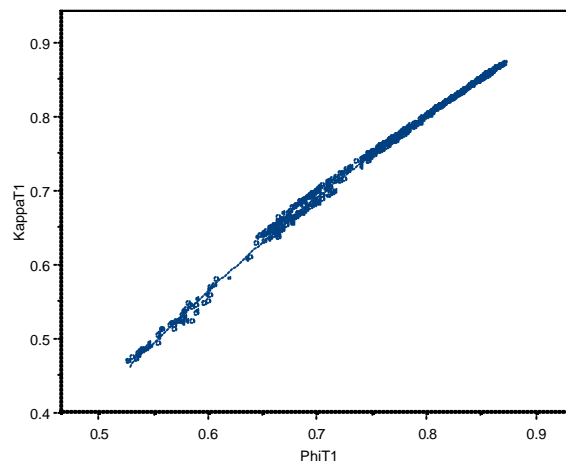
the estimated conditional probability of random agreement given only the marginal totals.

And further the following more direct (in)accuracy measures are at our disposal.

- Proportion correct $(a + d)/n$, also called %Correct, actually the most simple and direct accuracy indicator
- Proportion of undue failures b/n , if one is especially concerned with undue failures, as in general education.
- Proportion of undue passes c/n , if one is especially concerned with undue passes, as in professional education.

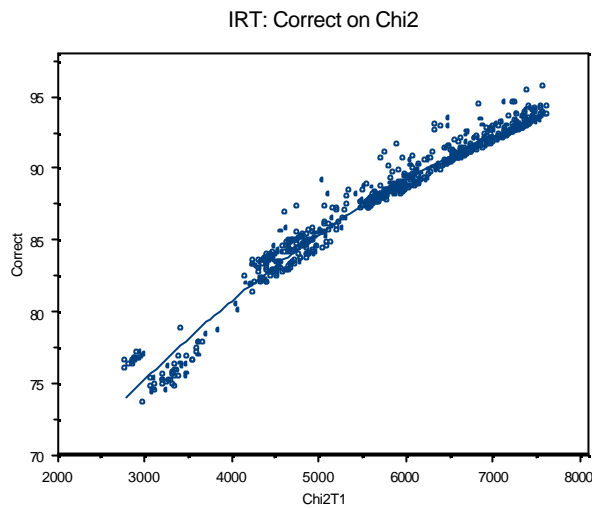
As expected χ^2 and ϕ give about the same picture. But also κ and ϕ are very close, as shown in Figure .

IRT: Kappa on Phi



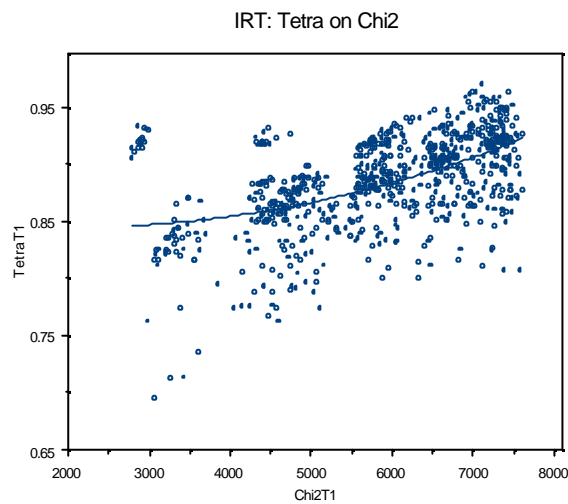
IRT: Regression of κ on ϕ

Although less tight the relationship between χ^2 and %Correct is also very strong (Figure)



IRT: Regression of %Correct on χ^2

The only exception is the tetrachoric coefficient that shows a very loose relationship with e.g. χ^2 (Figure).



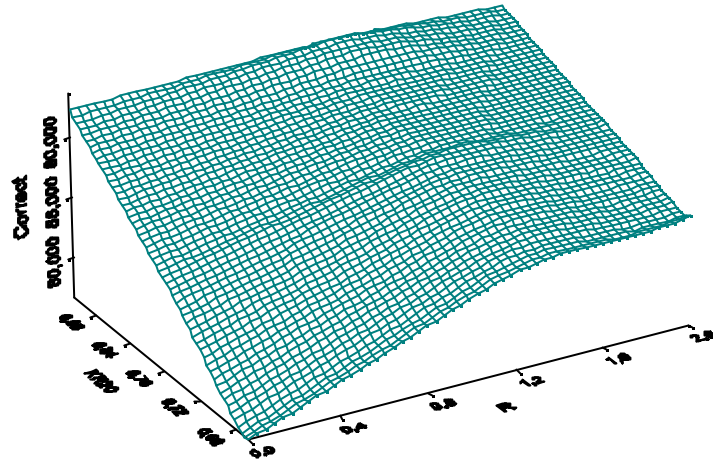
However, the tetrachoric coefficient showed a less clear relationship with the investigated variables. Therefore, the effect of the variables of interest is evaluated using %Correct or one of the two undue decisions. Because these are the most simple to interpret, and, obviously χ^2 , ϕ , and κ show about the same pictures as %Correct.

Results

IRT results

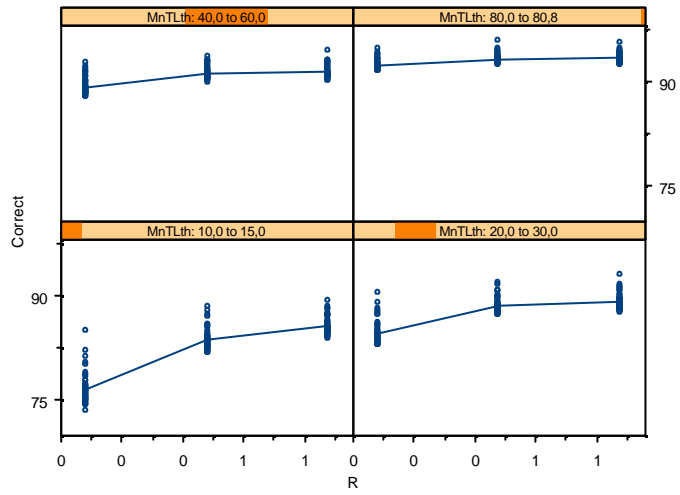
As shown in Figures , , and test length (and so KR20) has the highest influence on accuracy as measured by %Correct. After the KR20 the number of Resits has the largest influence, especially for the shorter tests. The opportunity of a second Resit hardly gives an improvement compared to maximally one Resit. It is also clear that with tests of length 10 one Resit offers even better results than using tests of twice this length without a Resit.

IRT: Correct on Resits and KR20

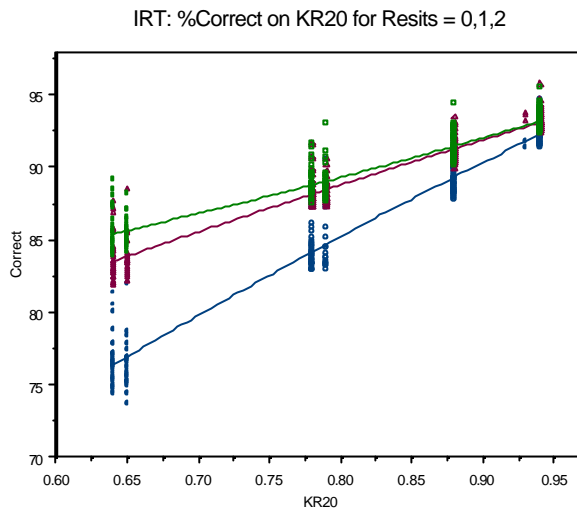


IRT: Cubic spline of %Correct on Resits and Test Length

IRT: Correct on Resits conditional on Test length (KR20)



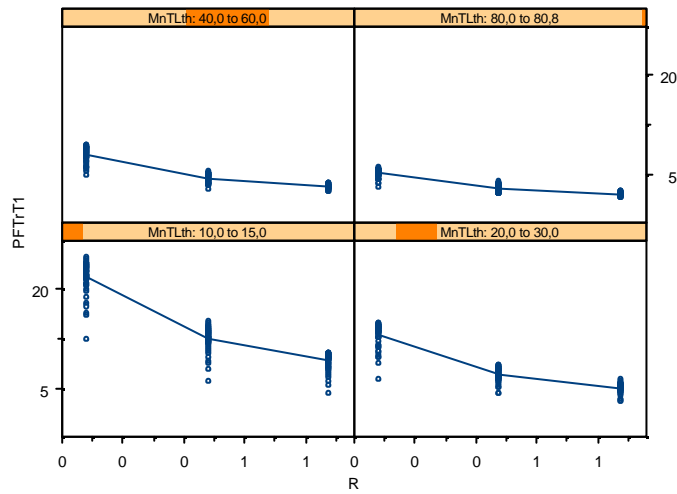
IRT: %Correct on Resits conditional on Test length (KR20).



IRT: %Correct on KR20 (Test length) for Resits = 0,1,2. The lowest

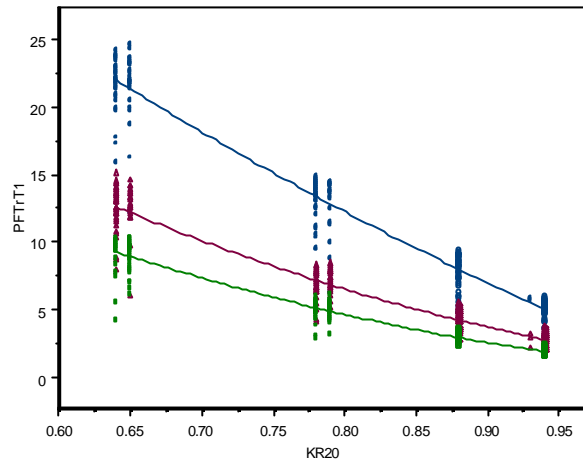
Figure shows more clearly than Figure the impact of Resits on %Correct, and also that the main gain is about exhausted after one Resit. For the shorter tests (10 and 20 items) the second Resit still results in a small gain. For the longer tests, however, a second Resit has no additional benefit. Figures (,) and (,) show respectively the regression of undue Failures and undue Passes on Resits conditional again on Test length. Clearly, the positive effect of Resits on undue Failures as found in the previous report is reproduced here within the IRT framework, as was to be expected.

IRT: Undue Failures on Resits conditional on Test length



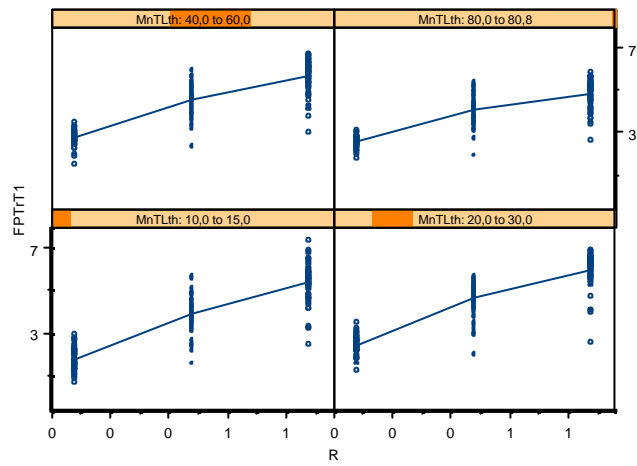
IRT: Undue Failures on Resits conditional on Test length

IRT: Undue Failures on KR20 (Test length) for Resits = 0,1,2



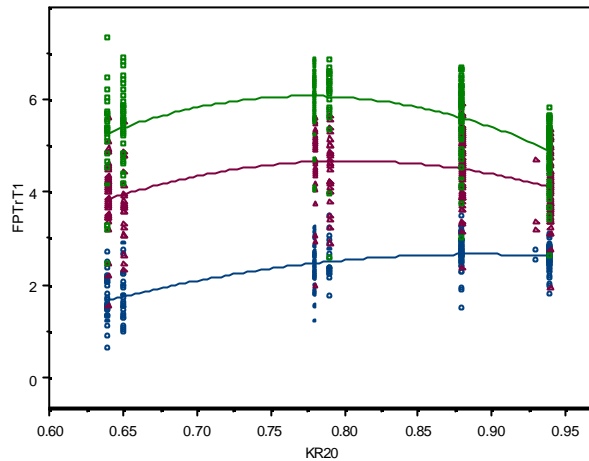
IRT: Regression of Undue Failures on KR20 (Test length) for

IRT: Undue Passes on Resits conditional on Test length



IRT: Undue passes on Resits conditional on Test length

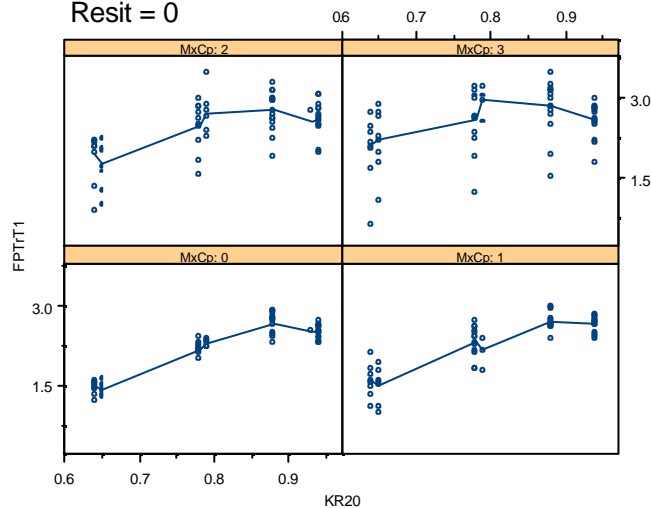
IRT: Undue Passes on KR20 (Test length) for Resits = 0,1,2



IRT: Regression of Undue Passes on KR20 (Test length) for Resits =

Figure shows the regression of undue Passes on KR20. The lowest line represents the procedures without a Resit. This line shows strongest the very peculiar phenomenon that the percentage of undue Passes increases with KR20 or test length, whereas one would expect them to be mitigated with increasing KR20. Only for the largest test length the undue Passes decline a little, but still are larger than for exams with the smallest tests. The picture becomes even clearer if this regression is depicted conditionally upon maxComp, as shown in Figure . The picture with no Resits is clearest with MaxComp=0, and tends to become more blurred with increasing MaxComp. Such phenomena that are impossible in a frame of mind where abilities, test scores etc. are continuous, are typically caused by the discrete discontinuous nature of test scores. The expected scores associated with the pass/fail ϑ -value are just a little over half of the maximum score of the test. e.g. for the ten item tests the expected weighted score for $\vartheta = 0.1$ equals 15.05, for the 80 item test 120.36. Remember that all discrimination parameters are equal to 3. For an ability just below 0.1, having one more item correct than 5 on a ten item test has a lower probability, than having one more item correct than 40 on an eighty item test.

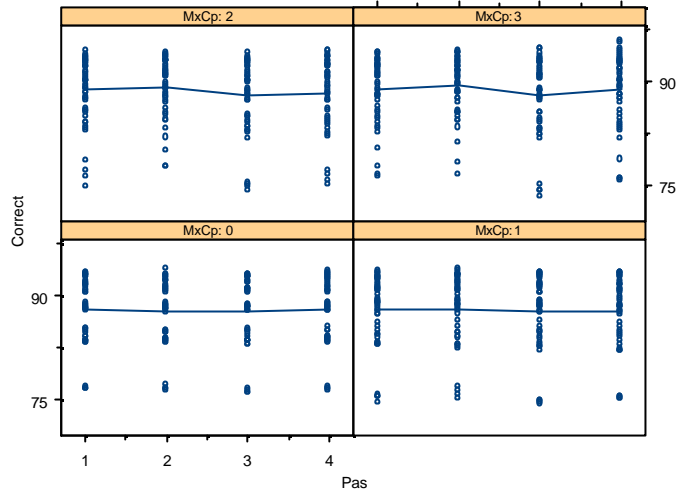
IRT: Undue Passes on KR20 (Test length) conditional on MaxComp
Resit = 0



IRT: Resits = 0. Regression of undue Passes on KR20 (test length)

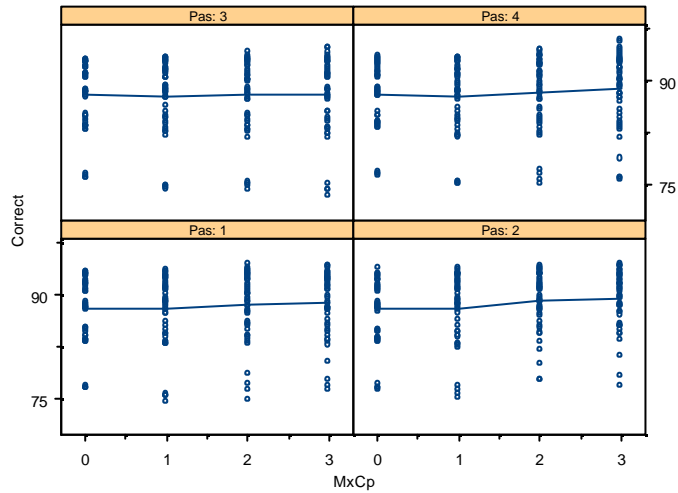
There was also found a little dependency of Percentage Correct on PassType and MaxComp. If MaxComp equals 0 then all PassTypes are equivalent to the conjunctive rule: one has to pass all tests to pass the exam. So in the lower left picture of Figure shows about four equivalent vertical dot patterns and a horizontal line. The largest differences between the different PassTypes occurs with the maximum freedom with MaxComp = 3. With MaxComp = 3 the two level compensatory rule (PassType=2) gives the highest accuracy, and the two level leniency rule (PassType=4) comes close (respective means 89.40, and 88.78). Figure shows that accuracy increases slightly by allowing more compensations or by being more lenient.

IRT: Correct on PassType conditional on MaxComp



IRT: %Correct on PassType conditional on MaxComp

IRT: Correct on MaxComp conditional on PassType



IRT: %Correct on PassType conditional on MaxComp

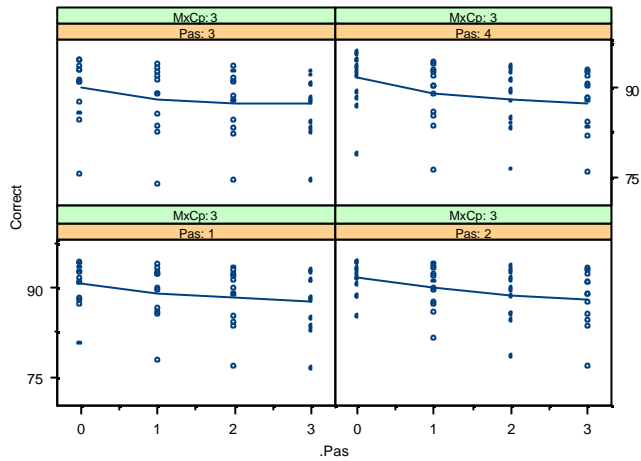
The means of the %Correct for the sixteen combinations of PassType and MaxComp are shown in Table 2. Clearly the two Compensation PassTypes perform better than the two Leniency PassTypes and two level Compensation with highest MaxComp performs best.

Table 2: Mean %Correct on PassType and MaxComp

PassType	MaxComp			
	0	1	2	3
1	87.78	87.76	88.50	88.76
2	87.74	87.94	89.12	89.39
3	87.65	87.66	87.92	87.94
4	87.82	87.59	88.24	88.79

The last question that was to be addressed concerns the number of tests that one has to pass, that is the number of tests that are exempted from complementary or leniency measures. Clearly with MaxComp=0 where one has to pass for all tests this variable has no influence, because there are no compensations or leniencies, one simply has to pass all tests not just the exempted tests. Therefore, the influence of the number of tests that one has to pass increases with and shows most clearly with the maximum value of MaxComp (3), which is shown in Figure . These figures show a serious decrease of accuracy by the introduction of a single 'have to pass test ', and after that a steady decline of accuracy with the introduction of more test that have to be passed.

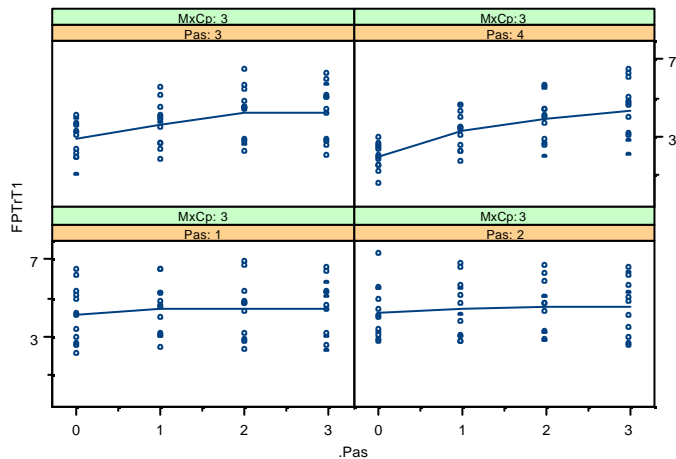
IRT: %Correct on #Pass for PassType = 1,2,3,4, and MaxComp=3



IRT: %Correct on Number of tests to pass, for maxComp=3 and

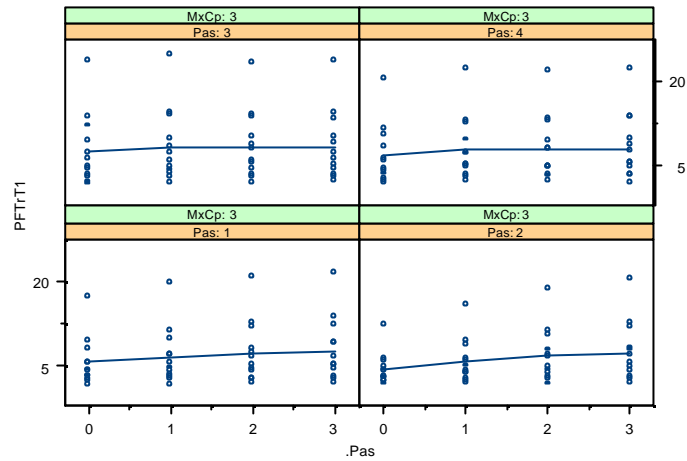
It is quite interesting to see how the decline of %Correct is divided over undue Passes (Figure) and undue Failures (Figure). Particularly the rise of undue Passes with Number to Pass for Leniency level 2 (PassType=4) is surprising, as is the increase of undue Failures for Complementary level 2 (PassType=2), especially for the short tests (the highest row of dots pertains to Test length = 10).

IRT: Undue Passes on #Pass for PassType=1,2,3,4, and MaxComp = 3



IRT: Undue Passes on Number to pass conditional on PassType and

IRT: Undue Failures on #Pass for PassType=1,2,3,4, and MaxComp = 3

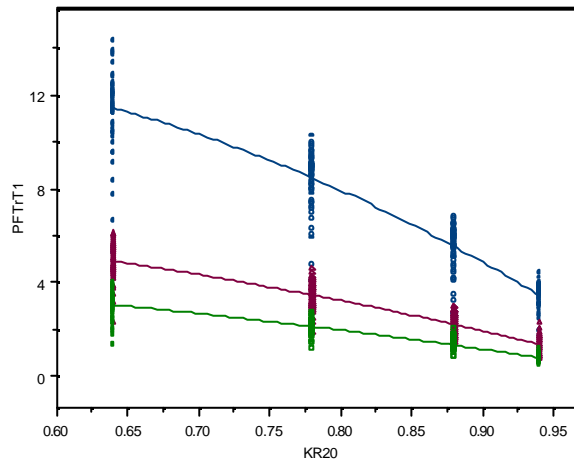


IRT: Undue Failures on Number to pass conditional on PassType and

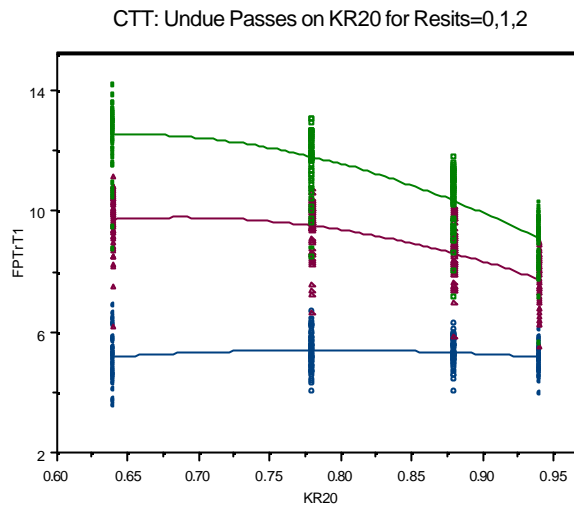
CTT Results

The clear picture about the effect of Resits on accuracy that emerged from the previous subsection using IRT did this time not arise in the CTT framework. Like in our previous Report (Verstralen, 2009) a clear positive effect of about the same size on Undue Failures is shown in Figure . However, the Undue Passes in the present example are appreciably higher (Figure) than was found in Verstralen (2009)

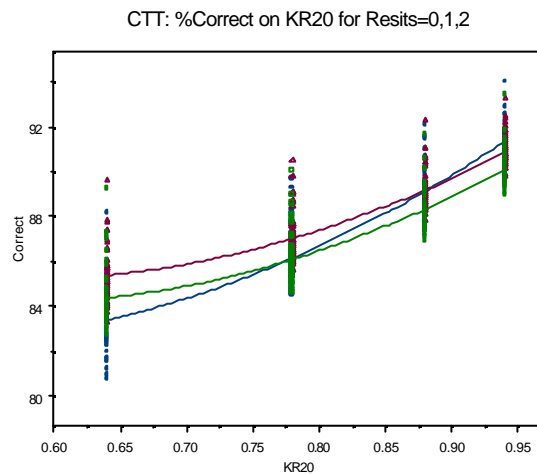
CTT: Undue Failures on KR20 for Resits=0,1,2



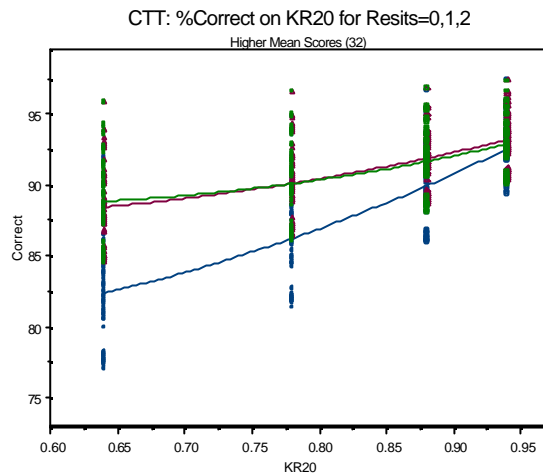
CTT: Undue Failures on KR20 for Resits=0,1,2. The highest line



CTT: Undue Passes on KR20 for Resits=0,1,2. The highest line
 Consequently, the effect of a Resit on accuracy gives an unclear picture as shown in Figure . For the highest Test length no Resit even performs slightly better than one or two Resits.



CTT: %Correct on KR20 for Resits=0,1,2. At KR20=0.64 the
 Probably the main difference with the previous study (Verstralen, 2009) is the severity of the exam. Whereas the mean percentage passing the exam there with seven tests was about 70%, ranging from 48% to 93%, in this study the CTT mean equals 53% and it ranges from 27% to 80%. This decrease in successes causes the increase in false positives from 1.51% with no Resit and 2.30% with one Resit, to in the present CTT study 5.27%, 8.90%, and 10.95% for respectively 0,1, and 2 Resits. Therefore, the CTT case was repeated with a higher mean score for the tests, it was raised from 27.8 to 32. This results in almost 80% passes for the exam, ranging from 52% to 96%, so even higher than in the previous study. Undue pass percentages for respectively 0, 1, and 2 Resits are 3.49%, 6.01%, and 7.21%. So these are also higher than in the previous study although less than with the lower mean of 27.8. Nevertheless in this case one Resit has a large impact on accuracy as shown in Figure . At the lower KR20 = 0.64 one Resit gives a substantial improvement from 82.5% correct to 88%. A second Resit, however, shows no added value at all.

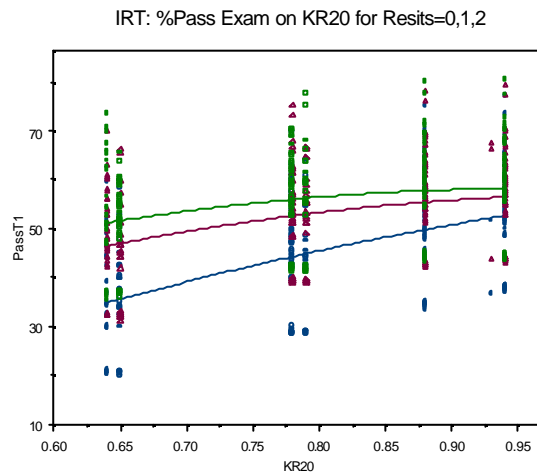


CTT: Higher Mean (32) %Correct on KR20 for Resits = 0,1,2.

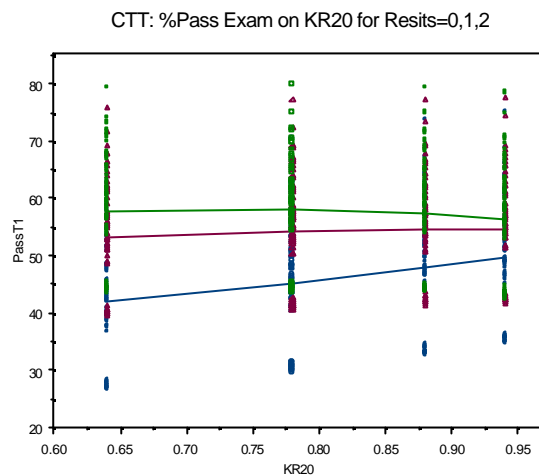
So whether a Resit improves the accuracy or not as evaluated within the CTT framework depends on the %Passing the exam.

Discussion

The effect of Resits on %Correct was different for IRT and CTT. Another difference between IRT and CTT is the number of Passes for the exam. With IRT for each value of Resits the number of Passes increases with KR20. For CTT this is only the case for no Resits, and less pronounced than for IRT, as shown in Figures and . It also strikes that for each combination of values of KR20 and Resits there still is an enormous variability in the percentage passing the exam, due to changes in procedure.



IRT: %Passes for the exam on KR20 for Resits=0,1,2. The



CTT: %Passes for the exam on KR20 for Resits=0,1,2. The

In the IRT study, although the situation with respect to %Passing the exam is even worse than in the CTT case, mean 51% range from 20% to 80%, the amount of undue Passes is much more favorable, the means for 0, 1, and 2 Resits being 2.36%, 4.28%, and 5.47%. In view of the results shown in Tables 1 and 1b above, where at least for the forty item test a larger local score variance was found for the IRT case than for the CTT case, this finding was unexpected. Most probably the discontinuous nature of scores and the relative positions of the edges with respect to the scores are again to blame here. These lower undue passes keep the positive effect on accuracy of Resits upright. I think the final conclusion of this study must be that the positive effect of Resits on accuracy is probably sensitive to

1. Percentage passing
2. The exact position of the edges with respect to surrounding discontinuous scores.

However, because these variables were not part of the present investigation this conclusion can only be provisional, and further research focused on these variables is needed to obtain a clearer picture.

References

- Douglas, K.M. (2007). *A general method for estimating the classification reliability of complex decisions based on configural combinations of multiple assessment scores*. Unpublished doctoral dissertation.
- Feldt, L.S., Steffen, M., & Gupta, N.C. (1985). Comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, 9,4,351-361.
- Rijn, van P.W., Verstralen, H.H.F.M., & Béguin, A.A. (2009). *Classification accuracy of multiple test-based decisions using item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Sluijter, C. (1998). *Toetsing bij doorstroombeslissingen in het voorgezet onderwijs*. Doctoral dissertation, Amsterdam University.
- Verstralen H. (2009). Quality of certification decisions. Unpublished report, POK, Cito, Arnhem, the Netherlands.
- Verstralen H. & Bechger, T. (2008). *Classification accuracy of educational tests*. POC, Cito, Arnhem, The Netherlands.