Metodološki zvezki, Vol. 7, No. 2, 2010, 151-166

Allowing Examinee Choice in Educational Testing

Gašper Cankar[1]

Abstract

Achievement tests sometimes entertain examinee choice – a situation where an examinee is presented with a set of items among which s/he has to choose one (or few) to answer that will be scored in her/his total test score. Basic assumption is that choice items are equivalent regarding both content and psychometric characteristics and therefore it doesn't matter which particular item examinee selects. Choice items often also share same maximum number of points and examinee score on a test is usually achieved by summing scores from all items taken by the examinee regardless of their combination. Choice items present conceptual problems like why enabling choice when items should be equivalent in the first place and methodological ones like how item scores from different combinations of items should contribute to comparable total score on test. Author used Rasch's model within Item Response Theory framework to test the assumption of equivalence of choice items by scaling all item difficulties on same scale. Physics 2008 and 2009 tests from Slovenian General Matura[2] examination are analyzed as an example to explore equivalence of choice items. Differences in difficulty of choice items in those tests are presented and discussed. It seems that examinee choice still doesn't work in educational testing and should be avoided when possible.

1 Introduction

Present paper is focused on examinee choice in educational testing – a situation when an examinee is presented with broader set of items from which s/he has to choose one or few to contribute to her/his total score on test. This kind of choice differs substantially from a more common situation in computer adaptive testing (CAT) where examinee is presented with a set of items, iteratively selected from

[1] National Examinations Centre, Ob železnici 16, 1000 Ljubljana; gasper.cankar@guest.arnes.si

[2] Data from Slovenian General Matura were obtained through National Examinations Centre.

large pool of pretested items according to her/his previous answers. In later case examinee can not consciously select items – they are selected adaptively by a computer algorithm to ensure reliable score and optimal item use. In first case, however, choice of items to answer is a conscious act on the part of the examinee who first sees the test with all items and then chooses a subset according to instructions in a test. The choice items can be used only in a portion of a test or in a whole test. We are interested in problem of item choice in educational testing scientifically, since available research on choice items in testing is rare, but also practically, since Slovenian General Matura examinations are main high stakes examinations in Slovenia and they provide the certificate of secondary education as well as enable the student the admission to the university with Matura results being used in selection procedures.

2 Examinee choice

Examinee choice is not often used in educational testing, especially in the field of external public examinations[3] and certifications where objective measurement of proficiency is paramount. This coincides with published research on examinee choice. Bridgeman, Morgan and Wang (1996) assert that the choice of essay topic should be left to the examinee only when the objective of testing is the proficiency to organize facts, shape solid arguments, etc. about a topic they are familiar with. When the objective is measurement of specific knowledge, selection of questions and items should be left to test experts and leaving the item choice to the examinee is inappropriate.

Burton (1993) similarly proposes that the choice should be allowed to the examinee when we wish to measure the ability of choosing and not when we wish to measure knowledge. For example, when we wish to measure the ability to read less demanding texts in foreign language, we can leave the selection of the text to the examinee. If s/he chooses text s/he is not familiar with, s/he will have to show greater ability. Choice is welcome when it doesn't interfere with measurement objective (or when it IS the measurement objective). Advocates encouraging item choice by an examinee can be found on relatively distinct area of testing proficiency to write an essay or to read a longer passage. Even with tests of divergent thinking where one might expect item choice would benefit the domain authors don't agree on positive aspects of item choice (Powers & Bennett, 1998)

White (1994) discards widespread opinion that prescribed topic of an essay limits examinees imagination and creativity and that choice should be a part of the task. White argues that such freedom is only imaginary – examinee still has to guess what rater wants.

[3] External refers to examinations, prepared outside schools, usually on a national or regional level of school authority that are same for all students in all schools.

Choice in educational testing is encouraged by advocates of performance assessment where students are presented with a complex performance task and are assessed on process and end result of their performance. Choice is supposed to increase authenticity of testing (Fitzpatrick in Yen, 1995) to the real life tasks, although authors don't explain this premise, since in real world we usually can not choose a problem situation we want. Gordon (1992) without further arguments states that choice is a necessary condition for fair testing. Wiggins (1993) argues that choice increases examinee's motivation since it enables the student to show her/his strengths. Choice therefore increases internal motivation which is always welcome in learning process.

To sum very different views on the subject – leaving choice to the examinee should be introduced in testing situation only in areas or in ways that can't influence achievement and scoring. One such example might be figure skating, where skaters can select their own music or most aspects of their clothing which greatly influences visual impression, but those aspects of their performance are not judged by the judges. In this case their choice may have profound motivational and aesthetic effects but in principle doesn't influence the result of the judging process.

Choice of items leads to different subgroups of students taking different combinations of test items. Measurement characteristics for whole test are hard to assess since no one answered all items in the test. Since different items are chosen by different subgroups we can not simply obtain comparable estimates of difficulty, discriminativity and other characteristics of the items via classical test theory without great complications.

Imagine a situation where we took pretested items and constructed a test of only two items. One of the items is more difficult, other easier. We do know the differences in difficulty but examinees don't and their choice is not afflicted by it. Since different choice items would be selected by different students it could happen that more able students would select hard item and solve it successfully while less able students would select an easy one and struggle with it. If we would use the data from both groups to calculate percent correct (usual difficulty index in classical test theory) results would show easy item as more difficult one and hard item as an easy one!? Such results would be clearly misleading.

It could also happen that more able students would choose harder item since it offers greater challenge to them and end up with lower score than if they took the easy one. They would end up penalized for their choice. That would make a disincentive to select items that require greater effort and knowledge and that is not good for any educational test.

Choice can be problematic from the perspective of content coverage (curriculum). When a student can choose from given set of items and items by design cover different areas of content s/he can skip some areas of knowledge while learning since s/he can skip the choice item that covers the same content in the test. Her/his score on test doesn't reflect performance on whole domain but only on content covered in selected items instead. Since the intent of test is to generalize results on broader domain of knowledge such implications thwart the very essence of the test.

Motivation of students can increase when given opportunity to interact with test and choose items (Wiggins, 1993). Her/his score is not determined completely in advance but s/he can select areas where s/he can show her/his strengths. We implicitly assume s/he will be able to choose wisely – to maximize her/his score. If s/he has enough time s/he could solve all items and then decide which would count to her/his total score. Usually time limit prevents this and choice then necessary also means frustration. Examinee must make a decision in a moment when s/he is already under stress and s/he should focus all her/his efforts on the tasks at hand. Choosing an item takes time, which is deducted from time for solving it.

Important aspect of item choice is that we lose unified scale to measure performance of students. Unless items are pretested we have no argument that sums of scores from different combinations of items represents fair measurement for all students. Items can differ in their difficulty and content and students can differ in their ability to choose wisely.

Most comprehensive synthesis of research on item choice is given by Wainer and Thissen (1994) who note that students indeed are not equal in their ability to choose items well and that this ability varies among different subgroups (by gender, ethnicity, etc.) in population. Their research showed that students do make suboptimal choices – sometimes select items where they don't perform best. To further complicate the issue this ability to choose well can be negatively associated to the ability the test is measuring – more able students tend to select more difficult items. Same conclusion is echoed by Powers et al. (1992) who note that more able students see more difficult items as bigger challenge and choose them even if their score would be higher on some other item.

3 Examinee choice in Slovenian General Matura

Slovenian General Matura is a modern external examination that enables student to complete upper secondary education and gain admission to university. It is prepared by committees of subject experts and tests in one session are same for all students in all schools. It consists of examinations in five school subjects (Slovene, mathematics and first foreign language are obligatory, last two are selected by student from wide array of subjects). When choosing among different school subjects for the last two exams examinee choice is welcome and highly regarded. It is understood that choice of a school subject gives to a student a chance to display strong areas of her/his knowledge and therefore increases motivation. Since each candidate can choose only five subjects and number of school subjects is much larger the choice of subjects also allows all main school subjects to be assessed.

Each subject examination typically includes external (mostly written) part, prepared by subject experts on national level and second part that is conducted internally on schools by teachers following nationally prescribed instructions. In final score internal parts have smaller weight (usually 20-25%). Written tests are constructed from wide range of item types that include multiple choice items, open ended tasks, essays and problem solving tasks. Total score on examination is transformed into grades (1-8 for mother tongue and higher level examinations and 1-5 for basic level examinations) and sum of grades for all five subjects is student's result on General Matura that is used in admission to university when there's a limitation of admittance ('numerus clausus'). The choice of subjects and equivalence of grades awarded is not the focus of this paper although it is an important aspect of examinee choice in itself. We are concerned here only with choice of items within single domain of knowledge usually represented by a single school subject examination.

Item choice can be traced in seven subjects: Physics, Biology, Biotechnology, Sociology, Psychology, History of art and Economy. Table 1 shows relevant data on those subjects.

Table 1: School subjects with choice items in Slovenian General Matura.

Subject	Prop. in test	N of items	Needed items	Possible comb.	Minimal overlap	N(2008)	N(2009)	Comb. in 2008	Comb. in 2009
Physics	50%	5	4	5	87,5%	1578	1498	5	5
Biology	50%	9	5	126	60%	1206	1253	123	114
Biotechnology	50%	6, 3*	4, 2*	45	60%	110	62	26	21
Sociology	100%	5, 4*	2, 2*	60	0%	1228	1185	36	20
Psychology	84%	3, 4*	2, 3*	12	65,3%	1617	1711	12	12
History of art	100%	9	7	36	71,4%	399	391	25	26
Economy	16,7%	3, 3*	1, 1*	9	83,3%	719	652	9	9

*some tests have more than one set of choice items.
Legend:
Prop. in test - Proportion of choice items in written test
N of items - Number of all choice items
Needed items - Number of needed choice items
Possible comb. – Number of all possible combinations of items
Minimal overlap – Smallest overlap between different combinations of items – this is the overlap of items in two most different combinations of items (in % of score points)
N(2008) – Number of students taking examination in spring 2008
N(2009) – Number of students taking examination in spring 2009
Comb. in 2008 – Number of combinations of items observed in 2008
Comb. in 2009 – Number of combinations of items observed in 2009

Number of combinations can quickly become very large and that alone poses great challenge in equating test results over all combinations. As long as all examinees take whole test it doesn't matter how easy or hard are specific items – everyone was tested under same conditions and as long as test is internally consistent scores can be summed into total score and will be at least ordinally comparable across students. If students didn't take same items, such comparison isn't valid since same scores don't represent same proficiency. And since items in General Matura examinations are not pretested we can not in advance establish sets of items with equivalent difficulties that could be reasonably used to choose from. Subject experts constructing items are aware that choice items should be equivalent but can not verify equivalence when they construct a test. It should be noted that items in General Matura examinations differ in content and form. Subject experts could easily construct basically equivalent items with only values changed between different items but that would render choice meaningless since students would be facing essentialy same item regardless of their choice.

Given the psychometric complications of the item choice one might wonder about specific reasons for the use of this technique by seven listed subjects. We explored historical data like reports from meetings, public documents and statements from the beginnings of modern General Matura in search for clues about such reasons. Before General Matura was launched in 1995 for whole population of students there was a pretest in 1994, where some schools conducted Matura-like examinations instead of usual school leaving (internal) exams in order to check various aspects of Matura implementation. All subjects, listed in Table 1 (except Biotechnology that was first introduced in Matura 2006) had choice items already in the 1994 pretests.

The National Matura Board (Republiška maturitetna komisija - RMK) proposed to the subject experts use of item choice in November 1994 – before the first Matura examinations in 1995 on the grounds that it allows students to choose only items that completely cover the curriculum, taught to them (RMK, 1994). There was uncertainty if all students taking Matura in 1995 were taught exactly same curriculum (same content and same depth) and item choice was seen as the way to address this issue. From the proposition it could be read that this was a temporary measure recommended only for the 1995 Matura examinations. The National Matura Board was being prudent and proposed item choice to address unknown diversity of curriculum taught and to win public vote for implementation of General Matura. Since General Matura was successfully implemented the goal was accomplished, but item choice remained.

4 Test of item equivalence

The assumption of item equivalence, condition '*sine qua non*' for the use of choice items, is hard to test under classical test theory framework because item difficulties depend on the proficiency of students that took them. Indeed, Gulliksen (1950) already noted that in the presence of choice items final scores can not be adjusted without inordinate amount of effort and choice items should therefore be avoided.

Item difficulty under classical test theory is for item *i* given as M_i/k, where M_i is the arithmetic mean of the scores on item *i* and *k* the maximum possible number of points on item *i*. In case of dichotomous items (scores 0/1) the item difficulty equals to percentage of examinees who answered correctly.

Since difficulty of items under classical test theory depends on proficiency of examinees who took them we can not directly compare item difficulties, calculated on different groups of examinees. It could be for example possible, that items would be equal in their difficulty but taken by subgroups of different proficiency. Difficulties as calculated within classical test theory would then differ. It could also happen that subgroups would be similar, but item difficulties would in fact differ. We can expect a combination of both situations in practice. To compare items we have to use an approach that can estimate item difficulties simultaneously and account for the fact that all of the items were not taken by every examinee. This is possible through Item Response Theory (IRT) that models examinee's response to an item and when model fits the data enables us to estimate comparable item difficulties by placing them on same scale.

There are many IRT models with different assumptions and characteristics. To test equivalency of item difficulties, Simple Rasch Model (one dimensional IRT model) was used since it provides robust estimates of item difficulties. Rasch’s model models the probability of a correct response given the proficiency of the examinee and the difficulty of the item. The probability of a correct answer of a person *j* with proficiency β_j to an item *i* with difficulty δ_i is given by the Rasch model as:

$$P\left\{x_{ji} = 1 \mid b_j, d_i\right\} = \frac{e^{\left(b_j - d_i\right)}}{1 + e^{\left(b_j - d_i\right)}}$$

Probability of a correct answer is dependant only on the difficulty of the item and proficiency of the examinee, both measured on the same scale and when difficulty of the item equals proficiency of the candidate, probability is 0,5.

When test includes choice items we don’t have a response from every examinee to every item. Items without response present missing values which complicate the analysis. IRT can deal with missing values since estimation methods do not require any imputations of missing data or case-wise or pair-wise omission. When data fits the model, the missing data decreases the precision of estimates but does not produce biased estimates of difficulties. Since there is no straightforward association between choice of items and proficiency of examinees, we can assume the scores are missing at random. For stable estimates and fit to the model, each item must be applied to reasonable sample and there must be some overlap between items/persons either via common items (items taken by more than one group) or common people (group that has taken more than one set of the items). In the case of tests in this article, with large overlap of common items between different combinations, we have enough ‘anchor items’ to achieve stable estimates. Results of IRT analysis are item difficulties and person proficiency scores reported on the same scale.

5 Method

To explore examinee choice in Slovenian General Matura examinations we have chosen two written parts of tests on Physics from Spring sessions of 2008 and 2009. Physics was chosen since the total number of combinations is low (5) and overlap between different combinations allows valid scaling of item difficulties to common scale. Matura is conducted in two sessions – Spring session right after finishing the upper secondary school and Autumn session just before next school year. Most examinees take examinations in Spring session and that sample is most representative of the population of examinees, since in Autumn session most examinees come to retake a failed exam or improve their result from Spring session. Since we are not interested in specific subgroup, results from all examinees from Spring sessions of 2008 and 2009 were used in two separate analysis (1578 and 1498 examinees in 2008 and 2009 respectively). Two subsequent years were used to show oscillations in results.

Test structure didn't change over time, it consisted of 40 multiple choice items worth 1 point each, followed by 5 structured items that score 10 points each[4]. Examinee has to answer all 40 multiple choice items and choose four out of five structured items. Maximum number of points on written part of the test is 80.

Items were analyzed using Rasch’s model[5] and since tests have quite large overlap in common items with at least 87.5% of common items for two random candidates, difficulty estimates of items are stable.

6 Results and discussion

Both tests demonstrated high reliability. Since choice items imply missing data in dataset, usual estimates of reliability (ie. Guttman-Cronbach's alpha) didn't apply. We could sum set of choice items together into one item and then treat them with Guttman-Cronbach's alpha or similar approach but in case of Physics test in either year choice items represent half of all possible score points on a written test (40). Tests in other subjects, listed in Table 1 have even higher proportion of points, achieved with choice items. We used Person Separation Index (Andrich, 1982) which is based on Item Response Theory and gives values close to Guttman-Cronbach's alpha reliability index when there's no missing data in dataset. Person Separation Index for Physics 2008 and 2009 written tests were 0.89 and 0.90 respectively which is an indication of high reliability and hence good person

[4] Structured items are scored to a points precision but since their inner structure differs, they have to be used in analysis as whole items.

[5] Specific software Rumm2020 and Winsteps 3.68.0. were used for analysis

Table 2: Item difficulties and standard errors for Physics 2008 and 2009 tests.

Physics 2008					Physics 2009				
Item	Difficulty	Points	S.E.	Type	Item	Difficulty	Points	S.E.	Type
I2.1	0,29	10	0,02	choice	It2.1	-0,33	10	0,02	choice
I2.2	0,78	10	0,016	choice	It2.2	0,51	10	0,02	choice
I2.3	0,29	10	0,02	choice	It2.3	0,95	10	0,02	choice
I2.4	1,24	10	0,02	choice	It2.4	0,69	10	0,02	choice
I2.5	0,97	10	0,02	choice	It2.5	0,55	10	0,02	choice
I1.1	-0,33	1	0,06	regular	It1.1	-3,60	1	0,24	regular
I1.2	-3,33	1	0,19	regular	It1.2	-1,32	1	0,09	regular
I1.3	-0,52	1	0,06	regular	It1.3	0,01	1	0,06	regular
I1.4	-2,50	1	0,132	regular	It1.4	0,46	1	0,06	regular
I1.5	-3,08	1	0,172	regular	It1.5	-0,62	1	0,07	regular
I1.6	1,08	1	0,054	regular	It1.6	1,36	1	0,06	regular
I1.7	-0,91	1	0,07	regular	It1.7	-1,92	1	0,11	regular
I1.8	-0,33	1	0,06	regular	It1.8	-0,73	1	0,07	regular
I1.9	0,16	1	0,06	regular	It1.9	0,30	1	0,06	regular
I1.10	0,14	1	0,06	regular	It1.10	-0,85	1	0,07	regular
I1.11	0,71	1	0,05	regular	It1.11	-0,13	1	0,06	regular
I1.12	-1,61	1	0,09	regular	It1.12	-0,18	1	0,06	regular
I1.13	0,27	1	0,06	regular	It1.13	0,01	1	0,06	regular
I1.14	1,38	1	0,05	regular	It1.14	-3,25	1	0,19	regular
I1.15	-1,75	1	0,09	regular	It1.15	1,74	1	0,06	regular
I1.16	2,22	1	0,06	regular	It1.16	0,23	1	0,06	regular
I1.17	-0,41	1	0,06	regular	It1.17	0,50	1	0,06	regular
I1.18	0,94	1	0,05	regular	It1.18	-1,53	1	0,09	regular
I1.19	-0,34	1	0,06	regular	It1.19	-1,42	1	0,09	regular
I1.20	-0,90	1	0,07	regular	It1.20	2,22	1	0,06	regular
I1.21	-0,90	1	0,07	regular	It1.21	1,05	1	0,06	regular
I1.22	1,85	1	0,06	regular	It1.22	1,69	1	0,06	regular
I1.23	-0,04	1	0,06	regular	It1.23	-0,71	1	0,07	regular
I1.24	1,66	1	0,06	regular	It1.24	1,50	1	0,06	regular
I1.25	0,80	1	0,05	regular	It1.25	-1,03	1	0,08	regular
I1.26	1,34	1	0,05	regular	It1.26	2,59	1	0,07	regular
I1.27	-0,56	1	0,07	regular	It1.27	-0,45	1	0,07	regular
I1.28	-1,22	1	0,08	regular	It1.28	0,12	1	0,06	regular
I1.29	0,85	1	0,05	regular	It1.29	-0,08	1	0,06	regular
I1.30	0,76	1	0,05	regular	It1.30	-1,03	1	0,08	regular
I1.31	0,03	1	0,06	regular	It1.31	1,39	1	0,06	regular
I1.32	0,59	1	0,05	regular	It1.32	-0,06	1	0,06	regular
I1.33	0,77	1	0,05	regular	It1.33	0,44	1	0,06	regular
I1.34	-0,36	1	0,06	regular	It1.34	-0,07	1	0,06	regular
I1.35	1,26	1	0,05	regular	It1.35	-0,49	1	0,07	regular
I1.36	1,34	1	0,06	regular	It1.36	-0,77	1	0,07	regular
I1.37	-1,17	1	0,08	regular	It1.37	0,38	1	0,06	regular
I1.38	0,18	1	0,06	regular	It1.38	0,91	1	0,06	regular
I1.39	-1,90	1	0,1	regular	It1.39	-0,22	1	0,06	regular
I1.40	0,29	1	0,06	regular	It1.40	1,19	1	0,06	regular

separation. Table 2 presents essential data on both tests including item difficulty, raw maximum points for item, standard error of difficulty estimate and indication whether it's a choice item. Item difficulties are in logit units as given by Rasch model. Value of 0 is set at the average item difficulty of the test. A person with average proficiency ($\beta_j=0$) would solve item with a difficulty of 0 with 50% probability and an item with difficulty of 1 (more difficult item) with 26,9% probability.

The probability is dependant only on a person's proficiency and difficulty of the item, both measured on a same scale. The common scale to which difficulties of all items are scaled at the same time represents the scale for measuring proficiency of persons.

6.1 Fit to the Rasch model and difficulties of the choice items for Physics 2008

Differences in item difficulties for all items together can be compared in bubble chart (Figures 1 and 3) and specifically only for choice items in ICC[6] chart (Figures 2 and 4) for Physics 2008 and Physics 2009 respectively.

One big advantage of IRT is that it measures person's proficiency on same scale as item difficulties. One logit on a scale can then be interpreted as follows: An examinee with certain proficiency (i.e. 1) has by definition 50% probability of correct answer on items of difficulty equal to her/his proficiency. S/he would solve items with difficulties one logit above her/his proficiency with only 27% probability and items one logit under his proficiency with 73% probability of correct answer.

Bubble chart plots on x axis the measure of fit to the Rasch model (given as information weighted mean square statistic and called 'Infit Mean Square') and difficulty of the items on the y axis. Size of points is relative to the standard error of the estimate. Fit estimates are well between 0.7 and 1.3 which is common rule of thumb (Bond & Fox, 2007) for reasonable fit of items to Rasch model. When fit estimates are low, the item deviates from the model in a random fashion. That could be indication of an item measuring something else than other items or it could mean that an item isn't suitable for measurement. Too high infit values mean that the item discriminates the proficiency much sharper than other items. Often this is an indication of an item that besides the construct under attention measures also some other covariating construct. Items in mathematics with very long text introductions tend to measure both knowledge of mathematic and reading ability. From Figure 1 can be concluded that dataset for Physics 2008 test fits the

[6] Item characteristic curves (ICC) are charts of modeled response for participants on particular item. They show probability of a correct answer given the proficiency of the person.

model well and results can be interpreted meaningfully. Present analysis is interested in vertical spread of five choice items (I2.1-I2.5) and we can observe that their difficulties vary between 0.3 and 1.2 on a common scale.

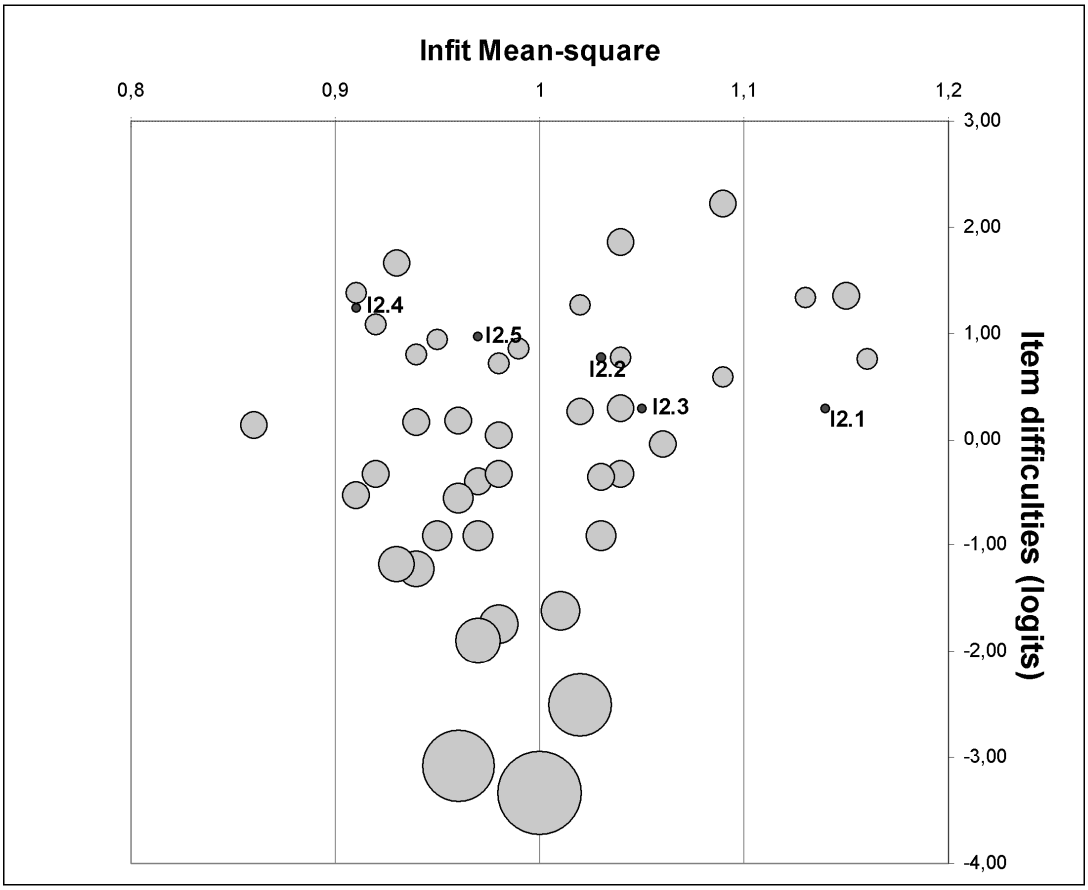

Figure 1: Bubble chart of item fit for Physics 2008 test. Choice items are marked.

When we explore item characteristic curves (ICC's) of five choice items we can readily observe differences already noted in difficulty estimates. Here we can also observe considerable differences in slopes. In Rasch model slopes are held constant for dichotomous items and thresholds of the polytomous items. Each of the five choice items has 10 points and ICC for whole item can differ in slope from ICC’s of the dichotomous items. ICC for a 10 point item can be seen as a characteristic curve of short 10 point test. Differences in item difficulties indicate that items are not equivalent in their overall difficulty. Differences in slopes of the items indicate that items also differ in the relative difficulties of their thresholds (single points within a 10 point item).

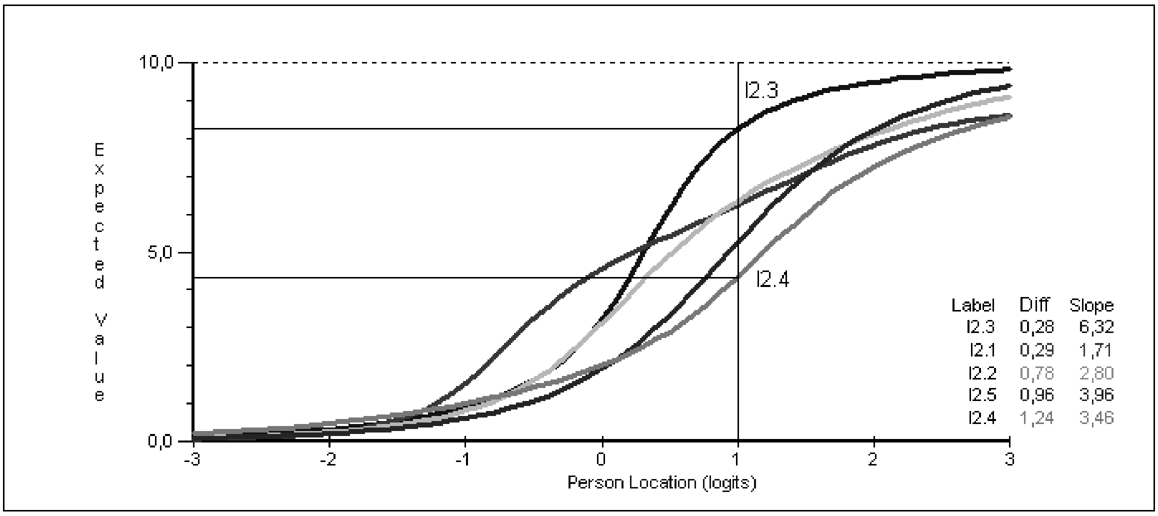

Figure 2: Item characteristic curves (ICC) for five choice items in physics 2008 test. Diff = Item difficulty, Expected value = The most likely score on the item, Person Location = Proficiency of persons on same scale as difficulty of items (logits).

6.2 Fit to the Rasch model and difficulties of the choice items for Physics 2009

Physics 2009 test shows pretty similar characteristics to a 2008 test with similar spread of all items, relatively small standard errors for choice items due to large number of raw points and position of choice items' difficulties in the middle of all difficulties. Fit estimates for 2009 test are between 0.8 and 1.3 which is within reasonable limits and shows that data conforms to the model used. Choice items (It2.1-It2.5) vary in their difficulties between -0.3 and 0.9.

Differences in item characteristic curves for five choice items in 2009 are smaller than year before with much more similar slopes, but they still demonstrate unequal difficulties.

As with any real data it wouldn't be plausible to expect exactly same difficulty estimates for all five items within one year. Since the standard errors of estimates for all five items in both years are very small, largest differences in difficulty between five items are statistically significant. The question therefore isn't ‘are there any differences’, but are there (practically) significant differences in observed item difficulties.

What does a range of 0.96 (1.24-0.28) or 1.28 (0.95-(-0.33)) logits mean? To answer this question one must ponder the nature of the logit scale on which difficulties are estimated. As explained earlier the difference of 1 logit can mean a differences in probabilities of correct response from 27% to 50% or from 50% to 73%.

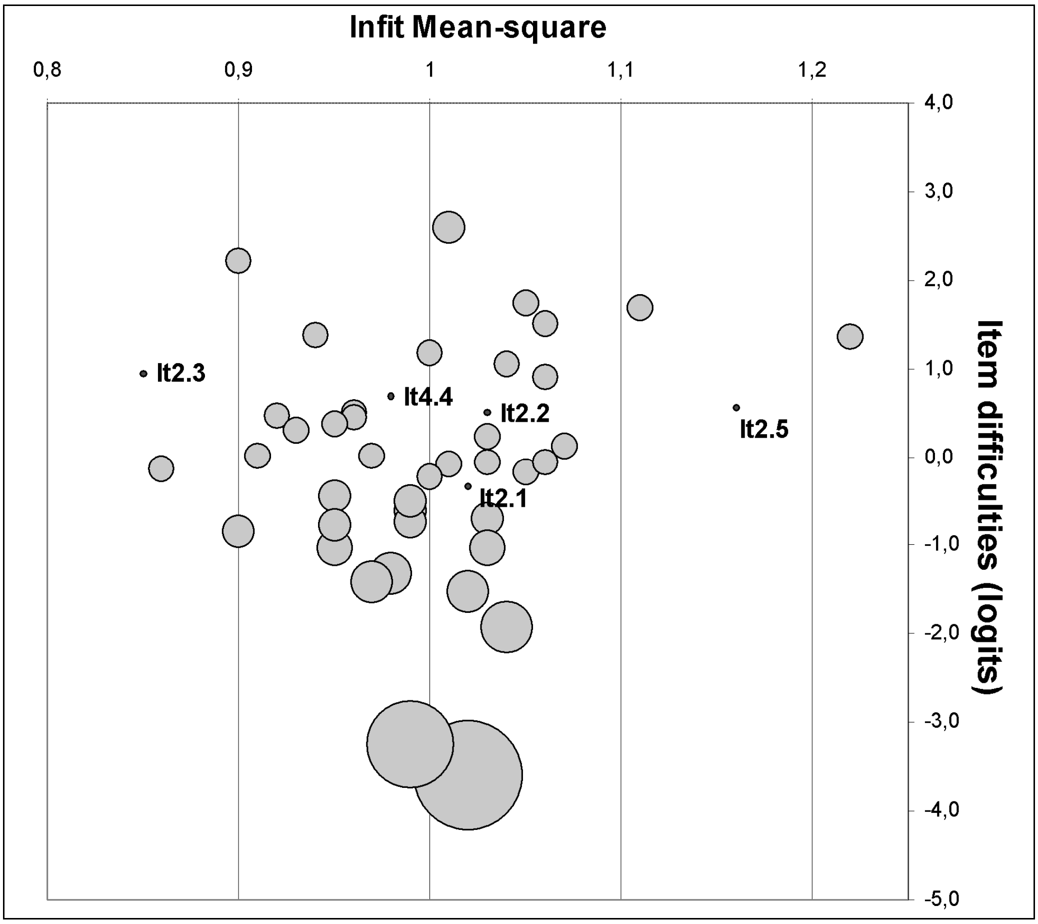

Figure 3: Bubble chart of item fit for Physics 2009 test. Choice items are marked.

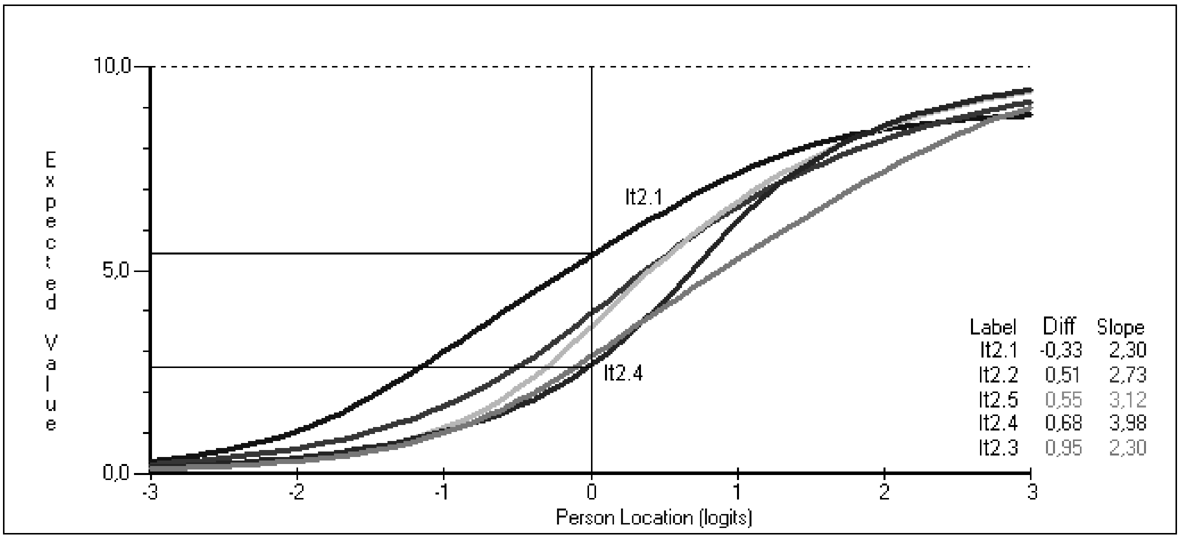

Figure 4: Item characteristic curves for five choice items in physics 2009 test. Diff = Item difficulty, Expected value = The most likely score on the item, Person Location = Proficiency of persons on same scale as difficulty of items (logits).

We can also make other type of comparisons. From Figure 2 we can observe that in Physics 2008 test an examinee with estimated proficiency of 1 who chose Item I2.3 had much higher expected score than an examinee with otherwise equal estimated proficiency that chose Item I2.4. A difference of about 4 points is certainly not trivial on a 80 point test. In Physics 2009 test (Figure 4) there's less difference in slopes but we can still observe about 3 points of advantage of an examinee with proficiency 0 who chose Item It2.1 against an examinee with same proficiency that chose Item It2.4. Differences can also be stated in terms of final grades received. Candidates receive five grades with 1 being the negative grade and 2-5 positive. Difference of one grade on Physics test corresponds to 13 points on a test. Standard deviations of points on a test were 12.3 and 12.2 in 2008 and 2009 respectively. Effect sizes would depend on the proficiency of the candidate and combination of items selected but 4 points in 2008 correspond to an effect of 0.33 and 3 points in 2009 to an effect of 0.24.

We can conclude that both tests demonstrated statistically and practically significant differences in difficulties of five choice items. This contradicts the underlying assumption for their use.

7 Conclusion

We tested assumption of item equivalence on two tests from Slovenian General Matura and through practical application demonstrated problems of item choice by an examinee in educational testing.

Assumption of equivalence of item difficulties can be fairly well tested in IRT framework only in cases when there's enough overlap between different combinations of choice items to allow consistent estimates. In Physics 2008 and 2009 written parts of tests at least 87.5% of common items were answered between two randomly selected examinees, but tests for other subjects listed in Table 1 don't have similar overlap. Problem of choice items can be detected and analyzed only under conditions of large sample of persons and great overlap between combinations of items. In other cases this may not be always possible.

Analysis of Physics 2008 and 2009 tests demonstrate significant practically important differences in item difficulties. Items are not equal and that violates the basic assumption underlying their use. That also echoes whole range of questions and issues brought up by other researchers already noted in introduction. How unequal are results of different examinees obtained with different combinations of items? Their scores are currently summed together as equivalent and total score used interchangeably regardless of actual items chosen. Of course different combinations of items could in theory be equated to produce comparable results regardless of the combination selected to ameliorate the problem. This is however operationally implausible for large number of possible combinations or some combinations that were rarely selected. In cases of small overlap between combinations (few anchor items) problem could only be solved if examinees would also solve the items not selected (with equal motivation) which questions the point of choice in the first place. It also conflicts with the idea that an examinee should in advance know how much will each item contribute to her/his final score. Equating could mean that fewer raw points on a harder item would be worth more than more raw points on an easy item. We may speculate that an examinee would have chosen differently if s/he knew that in advance.

Results of present analysis are congruent with reports from other researchers, cited in introduction – choice items makes sense when act of choosing itself is being evaluated or choice is irrelevant for object of measurement (Wainer and Thissen, 1994). In all other cases choice introduces error in measurement and consequently threatens reliability and validity.

Possible positive effects of choice from increased motivation are confronted by suboptimal choice of items. Problem can be summarized in following paradox, asserted by Wainer and Thissen (1994) – examinee choice is based on assumptions that make it unnecessary. When items are indeed equivalent in difficulty and content, when they need equal proficiency to solve them, then choice is only additional burden to the examinee – from assumptions it follows that s/he would get equal result regardless of item chosen. When items are not that similar, we can reasonably question equivalence of choice items in a situation of objective measurement. If use of choice items was justified when implementing General Matura in 1995 on grounds of differences in curriculum taught to the candidates, it should be reconsidered after 15 years of already implemented examination that consolidated the curriculum taught. In view of differences in item difficulties, presented in this article, use of choice items should be reconsidered and avoided wherever possible. If choice items are essential for certain subject and can not be avoided the problem could be addressed by pretesting items and selecting only those that are demonstrably equally difficult.

Acknowledgement

Author wishes to express acknowledgement to National Examinations Centre for datasets and other data provided. Opinions or interpretations, presented in this paper pertain to the author and are not affiliated with National Examination Centre or any other institution.

References

[1] Andrich, D. (1982): An index of person separation in latent trait theory, the traditional KR.20 index and the Guttman Scale response pattern. *Educational Research and Perspectives*, **9**, 95-104.

[2] Bond, T.G. and Fox, C.M. (2007): *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, Second Edition (2. ed., p. 352). Lawrence Erlbaum.

[3] Fitzpatrick, A.R. and Yen, W.M. (1995): The psychometric characteristics of choice items. *Journal of Educational Measurement*, **32**, 243-259.

[4] Gabrielson, S., Gordon, B., and Engelhard, G. (1995): The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education*. **8**, 273-290.

[5] Gordon, E.W. (1992): *Implications of diversity in human characteristics for authentic assessment*. (CSE Technical Report 341). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

[6] Gulliksen, H.O. (1950): *A Theory of Mental Tests*. New York: Wiley.

[7] Powers, D.E. and Bennett, R.E. (1998): *Effects of Examinee Choice on a Test of Divergent Thinking*. GRE Board Final Report No. 93-20P, Princeton, NJ: ETS.

[8] Powers, D.E., Fowles, M.E., Farnum, M., and Gerritz, K. (1992): *Giving a choice of topics on a test of basic writing skills: Does it make any difference?* (Research Report No. RR-92-19). Princeton, NJ: Educational Testing Service.

[9] RMK (1994): Priprava izpitnih pol za maturo 1995 – priporočilo RMK [Recommendations of RMK for the preparation of Matura 1995 examinations – signed letter to the subject comittes].

[10] Bridgeman, B., Morgan, R., and Wang, M. (1996): *Choice among Essay Tonics: Impact on Performance and Validity* (ETS Research Report 96-4). Princeton, NJ: Educational Testing Service.

[11] Shavelson, R.J., Baxter, G.P., and Pine, J. (1992): Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, **2**, 22-27.

[12] Wainer, H., and Thissen, D. (1994): *On Examinee Choice in Educational Testing* (GRE Report No. 91-17, ETS Research Report RR-94-3 1). Princeton, NJ: Educational Testing Service.

[13] Wang, X.B., Wainer, H., and Thissen, D. (1993): *On the Viability of some Untestable Assumptions in Equating Exams that Allow Examinee Choice* (ETS Technical Report RR-93-21). Princeton, NJ: Educational Testing Service.

[14] White, E.M. (1994): *Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating, and Improving Student Performance* (2nd ed.). San Francisco: Jossey-Bass.