Edukacja 2015, 3(134), 48–62
ISSN 0239-6858

Sex differences in guessing and item omission

Karolina Świst, Paulina Skórska, Maciej Koniewski,
Aleksandra Jasińska-Maciążek
Educational Research Institute*

Guessing and item omission may be regarded as risk-taking or risk-avoidance strategies – sex specific adaptations to testing situations. In this article, these phenomena were analysed by (a) percentage of omissions by sex, (b) negative binomial regression to asses sex differences in the number of omissions, (c) *c*-DIF analysis using IRT-LR test and (d) linear regression using item attributes, to assess whether the *c*-parameter is sex differentiated by the percentage of omits (controlling item difficulty). The data set analysed were from the 2012–2014 Polish lower-secondary schools final exams, comprising tests in maths, language, science and humanities. Contrary to the vast body of literature, boys omitted items slightly more frequently than girls. Possible explanations of this finding – specific to the Polish examination system – were provided. The hypothesis of a higher *c*-parameter for boys did not find strong support from this study. It was shown that the *c*-parameter should not only be interpreted as resulting from item non-omission. This supports the modern concept of the *c*-parameter as a consequence not only of random guessing, but also problem solving, creative guessing or cheating.

Keywords: sex differences, guessing, item omission, *c*-DIF.

Risk taking can be perceived as environmental adaptation. People successfully adapt by systematically taking certain risks whilst avoiding others (Byrnes, 1998). From such a perspective, test-taking strategy can be conceptualised as an adaptation to a testing situation. In the test situation, student knowledge and problem solving skills are pitted against scoring rules, which may either penalise them or not for guessing and item omission. For example, if the scoring rules penalise an incorrect answer, a student might be less likely to take the risk of providing an uncertain answer. However, if the student can make an educated guess, the risk of providing an incorrect answer is reduced and might be considered as worth taking. Omitting test items may be considered as an indicator of the tendency to avoid this risk. When a student has insufficient knowledge and skills, and nevertheless answers an item, while this answer turns

The article is an extended version of the presentation "Gender differences in pseudo(guessing) and item omission", presented by the authors at the EduMetric 2014 International Seminar on Educational Research and Measurement, Cracow, December 6–8. The seminar was organised by the Jagiellonian University, as part of the system level project "Quality and effectiveness of education – strengthening institutional research capabilities" carried out by the Educational Research Institute and co-financed by the European Social Fund (Human Capital Operational Programme 2007–2013, Priority III High quality of the education system).

* Address: ul. Górczewska 8, 01-180 Warszawa, Poland. E-mail: k.swist @ibe.edu.pl

out to be incorrect, it places the student at risk of social disapproval, lowered senses of self-esteem and self-efficacy. Both guessing and item omission, considered as risk-taking and risk-avoiding strategies, may differ in scale between boys and girls. Sex differences in willingness to risk providing an uncertain answer and the penalty for a mistake might be conceptualised as sex-specific adaptation to the test situation.

Theoretical framework for risk behaviour

Lopes (1987) indicated two strands of development of the psychology of risk. Personality theories use an idiographic approach to explain risky behaviour. Personologists are concerned with the things that make people differ in their perception of, and reaction to, a situation involving risk. Individual differences influence how individuals adapt to the environment, which includes risk selection and management. Byrnes, Miller and Schafer (1999) referred to Zuckerman's (1991) account of the sensation-seeking personality and the "risk as value" hypothesis described by Kelling, Zirkes and Myerowitz (1976) as two examples of such theories. In general, such theories suggest that sex differences would be constant between contexts. Men would always take more risks than women and the gap would remain relatively similar regardless of context and type of task. Sex is viewed as a characteristic explicitly indicating risk-seeking or risk-averse personality.

The second strand of development of the psychology of risk is from the experimental perspective, typically taking the nomothetic approach, aimed at generalising research findings to the population level and understanding the ways in which people are alike. A good example would be Kahneman and Tversky's (1979) prospect theory or Coombs's (1975) "portfolio theory". In this approach, the researcher manipulates the situation to observe a response, assuming no differences between individual characteristics, as the possible differences are randomly distributed between experimental conditions. Such theories attempt to explain differences between situations that promote risk taking or aversion. The effect of sex in predicting risk-seeking or risk-averse behaviour is by definition zero.

While personologists focus largely on variables describing people, experimentalists deal with situational variables. A two-factor theory for risky choice was built on the interplay between the person-centred (situation held constant) and situation-centred (individual differences held constant) theories. These two factors are the security versus potential factor (dispositional factor) and the aspiration level factor (situational factor).

> The dispositional factor describes the underlying motives that dispose people to be generally oriented to achieving security (i.e. risk adverse in conventional terminology) or to exploit potential (i.e. risk seeking in conventional terminology). The situational factor describes people's responses to immediate needs and opportunities (Lopes, 1987, p. 275).

According to such models, sex differences would vary by situation context. Some contexts may promote male risk-seeking behaviour, while others may promote female risk-seeking behaviour.

Byrnes et al. (1999) referred to two other models consistent with the two-factor theory for risky choice. Arnett's (1992) theory of broad and narrow socialisation (BNS) suggests that risk behaviours are a function of both personal attitude and cultural driven expectancies. This theory suggests that in some cultures, women's risk-seeking tendency can be dampened by cultural restrictions causing underestimation of the effect of sex differences in risk-seeking tendencies. Wilson and Daly's (1985) model suggests that men would only be more likely than women to take risks when the context involves both competition and a large spread in rewards between winners and losers.

In line with the two-factor theory for risky choice, it was assumed that sex-specific adaptation to a test situation might be caused by factors associated with sex, as well as factors associated with the test situation (e.g. characteristics of the particular measurement tool).

Research on sex differences in risk taking

Several studies have indicated sex differences in risk-taking. Byrnes et al. (1999) reviewed 150 empirical studies from 1967 to 1997 in which the researchers examined differences in risk-taking and reported a direct comparison between men and women on some risk-taking measures. The results (i.e. 60% of 322 effects) clearly supported the idea that men were more likely to take risks than women. Nearly half (48%) of the effects were greater than 0.20 Cohen's *d* (Cohen, 1992). However, a sizeable minority (i.e. 40%) were either negative or close to zero. For all 322 effects, the weighted mean was found to be 0.13 Cohen's *d* (95% CI of 0.12 to 0.14).

Of particular interest in the context of the research problem posed in this article are the categories of informed guessing and intellectual risk taking. Informed guessing included tasks in which participants earn points or money for correct guesses, but also lose points or money for incorrect guesses (e.g. standardised achievement tests with penalties for incorrect guesses). The category of intellectual risk-taking involved tasks which required mathematical or spatial reasoning skills. Participants were presented with items of various levels of difficulty and asked to indicate their preferred choice. Unlike the tasks in the informed guessing category, points were not subtracted for incorrect answers on intellectual tasks. Thus, participants were mainly concerned about getting stuck on items or exposing their lack of skill if they failed (Byrnes et al., 1999).

The mean effect of sex differences for informed guessing was 0.18 (*n* effects = 11; 95% CI: 0.13 to 0.23), while for intellectual risk taking 0.40 (*n* effects = 7; 95% CI: 0.25 to 0.55) both significant ($p < 0.05$). The effects, analysed according to separate age categories, emerged only to be significant for participants aged 10–13. In the 10–13 age group, the mean effect on informed guess was 0.31, while the effect on intellectual risk taking was 0.68. The strong effect of intellectual risk-taking revealed that girls in the age of 10–13 seemed to be disinclined to take risks even in fairly innocuous situations or when it was a good idea to take the risk (e.g. intellectual risk-taking on practice SATs).

Defining and measuring omissions and guessing

Although Lord (1980, p. 226) claimed that: "if […] proper test directions are given, the examinees understand the directions, and they act in their own self-interest, then there will be no omitted responses", examinees do leave some items unanswered.

There are three types of unanswered items: unreached (due to lack of time), intentionally omitted (answer deliberately not provided), and unintentionally omitted (not seen). Unreached items are usually identified at the end of the response pattern. Items with no answer provided which precede the last answered item in a response string are considered as intentionally or unintentionally omitted (Lord, 1980).

Ben-Shakhar and Sinai (1991) presented several omission measures. The first is simply the total number of items omitted by an examinee. The second is the number of items omitted by an examinee up to the last item answered in a response string, that is to the point prior to the unreached items. The first measure confounds unreached items with omitted items.

Other omission measures (Angoff and Schrader, 1981; Ziller, 1957a; 1957b) are based on the ratio of the estimated number of items guessed to the estimated number of items answered incorrectly. It should be noted that these indicators assume that all errors are caused by pure guessing. What is more, these measures cannot be defined for examinees with maximum scores (Ben-Shakhar and Sinai, 1991).

Guessing may usually occur when an examinee is attempting a multiple-choice question (MCQ)[1]. The examinee may score points by randomly selecting an answer. Though some test developers try to prevent guessing on MCQs, e.g. by penalising incorrect answers (loosing points) or awarding partial points for omitted ones (Han, 2012), creating a "guess-free" MCQ is nearly impossible.

Researchers usually deal with the phenomenon of guessing by estimating the *c*-parameter in a three-parameter logistic model (3PLM) using item response theory (IRT). Traditionally, the *c*-parameter was considered as the guessing parameter. Some researchers however, have suggested that the *c*-parameter estimated in 3PL model should not be interpreted as pure random guessing indicator (Han, 2012; San Martin, Pino and De Boeck 2006). If students do not know the correct answer for a particular item, they often use partial knowledge to eliminate the less likely answers, and rarely guess entirely according to chance. A more appropriate term is therefore the pseudo-guessing parameter (Hambleton, Swaminathan and Rogers, 1991). Recently, the *c*-parameter has been regarded an even more complex phenomenon. Han (2012) proposed conceptualisation of the *c*-parameter as the product of random guessing, logical guessing and problem solving. Meijer (1996a; 1996b) proposed cheating, careless responding, lucky guessing, creative responding, and random responding.

Sex differences in guessing tendency have been examined in empirical studies over several decades. Swineford (1941) found a greater guessing tendency among boys in the ninth grade (though the sample was clearly limited, as only 25% of students answered all items). Slakter and colleagues (1971) showed greater guessing tendency among boys (in primary and high schools). Ben-Shakhar and Sinai (1991) provided evidence that males are less prone to omitting items than females. Differences were visible even for tests in which women traditionally outperformed men, or when there was no statistically significant difference in performance between males and females. Pekkarinen (2014) analysed the omission tendency for men and women in university admission tests. Recruitment depended on the total starting points based on high school results and the entrance exam. When the starting points were controlled, women scored worse than men on the entry exam, as they tended to omit more items. As a result, they lost the advantage of their starting points and were less likely to be accepted by the university. Von Schrader and Ansley (2006) indicated that girls tended to omit more items in mathematics tests (on which they underperform) and boys tended to omit more on language and reading tests (on which girls outperform boys). This might suggest that the omission tendency is associated with ability level – and indeed the higher the ability, the lower the tendency to omit.

Concluding, the majority of research showed that guessing is a phenomenon attributed to boys, while item omission is a phenomenon attributed to girls. Some studies however, did not support these results. For instance, Slakter (1967; 1968a;

[1] An examinee may guess correct answer on all types of selected-response items and even in some constructed-responses (e.g. short answer items), but research on guessing in such cases is not extensive within the IRT framework.

1968b) did not obtain clear sex differences for guessing behaviour of children and college students. More recent research on American College Testing (ACT) also indicated that there was no significant sex difference in the tendency to omit items (Zhu and Thompson, 1995).

Differential item functioning – a comprehensive approach to assess sex differences in performance on test items

Differential item functioning (DIF) analysis is an useful approach to assess sex differences in performance on a particular item. DIF occurs when the probability of the correct answer on a given item depends on factor(s) other than the test-taker's ability level. DIF can be associated with examinee group membership (e.g. boys) as well as items or test attributes (e.g. different booklet versions). Traditionally, DIF is divided into two types, focusing on conditional group differences in item difficulty (uniform DIF) or item difficulty and discrimination (non-uniform DIF) parameters. Little attention has been paid to group differences in item *c*-parameters. Teresi et al. (2007) investigated race and age group differences in *c*-parameter on physical functioning ability and general distress measures, to discover that several items showed DIF with respect to age. Finch and French (2014) investigated group differences in *c*-parameter on inflation of indices for uniform and non-uniform DIF detection – when groups differed on the *c*-parameter, Type I error rates for both uniform and non--uniform DIF increased.

The *c*-DIF is assessed by a direct test of the equality of the c-parameter between groups. Two methods of *c*-DIF detection are widely known: Lord's chi-square test (Lord, 1980), IRT-D2 and the item response theory log-likelihood ratio test (Thissen, Steinberg and Wainer, 1988), IRT-LR. The logic of the *c*-DIF detection procedure was described by Finch and French (2014, p. 29) as follows:

> In particular, to test for *c*-DIF, two models are fit to the item response data. In the first model, the pseudo-guessing parameter for the target item is held equal between the groups, while in the second model the pseudo-guessing parameter values for the target item are allowed to vary by groups. The test statistic for the null hypothesis of *c*-DIF is then calculated as the difference between the -2 log-likelihood values of the constrained and unconstrained models, as described in general above.

The null hypothesis being tested in IRT-D2 and IRT-LR *c*-DIF analyses is essentially the same: that the specified item *c*-parameter does not differ between the groups. Wald (1943) and Rao (1973, pp. 416–418) showed that the two tests are asymptotically equivalent. Thissen et al. (1988, p. 154) stated that "to the extent that the likelihood is normal and is estimated well, both methods should perform identically".

A common motivation for DIF detection is to support test fairness and validity. In this study, *c*-DIF was used to assess the scale for sex differences in the *c*-parameter, conceived traditionally as guessing practices. *C*-DIF analyses may not be reliable as 3PLM is considered to yield technical and theoretical problems (Hambleton et al., 1991; Holland, 1990; Kolen, 1981; Lord, 1974; 1975; 1980). Despite these arguments, the authors believe that *c*-DIF is a valuable in assessing the causes for sex gap in standardised tests.

The aim of the study

The main goal of this study was to determine whether two response strategies (guessing and item omission), claimed by theories to differentiate between boys and girls, hold true for high-stake lower-secondary Polish schools leaving exams. The following research questions were specified:

- Are systematic differences between boys and girls in guessing and item omission salient and to what extent they can be associated with sex?
- Which sex tends to omit more items?
- Can *c*-parameter be explained by item difficulty and item omission rate differently according to sex?

Data

Results from standardised compulsory external exams from 2012 to 2014 administered at the end of lower-secondary school were analysed. Item omission rates as well as *c*-parameter differences in the four tests comprising the exam were analysed: humanities (history and civic education), language (Polish reading and writing skills), maths, science. A closer look was given to maths and language tests from 2012. These tests were selected to serve as an example for *c*-DIF analysis. The data were collected by the Central Examination Board (Centralna Komisja Egzaminacyjna, CKE) and cover the entire student population (ca. 400K per year). Please see Table 1 for exact numbers of examinees population and percentage of girls.

Table 1
Number of examines

Year	Exam	Number of students	% of females
2012	Humanities	393 849	49.19
	Language	393 836	49.19
	Maths	393 715	49.19
	Science	393 723	49.19
2013	Humanities	379 756	48.98
	Language	379 752	48.97
	Maths	379 633	48.97
	Science	379 634	48.97
2014	Humanities	362 752	48.96
	Language	362 755	48.96
	Maths	362 749	48.95
	Science	362 749	48.95

Methods

To measure omission, separately by sex, the simplest indicator suggested by Ben-Shakhar and Sinai (1991) was chosen, i.e. the total number of items omitted by an examinee. The percentage of omitted items was computed for convenience for comparison between tests, since the tests had unequal numbers of items. To analyse differences in the scale of item omission between boys and girls, negative binomial regression model was used. Number of omitted items was regressed on sex with control for ability level. The negative binomial regression model was chosen since the dependent variable was item count data. Since variance is higher than the mean, the negative binomial regression model is more appropriate than the Poisson type (Hilbe, 2011; Long, 2001). The appropriateness of the chosen approach was confirmed with the likelihood ratio (LR) test, in which the base hypothesis was tested for that alpha equals zero. If this hypothesis is rejected then negative binominal regression would be reduced to a Poisson regression model. The results of the LR tests (Table 3), indicated that alpha was non-zero for all models and the negative binominal model is the preferred one.

In order to assess the scale of student guessing, a three-parameter (3PL) Item Response Theory (IRT) model was used. This allowed estimation of the *c*-parameter (Birnbaum, 1968). The *c*-parameter is technically a lower asymptote of the logistic curve. Therefore in 3PLM a nonzero performance on MCQs by examinees with low ability is modelled. The 3PLM can be described with the following equation:

$$P(Y_{pi}=1\,|\,\theta_p,b_i,a_i,c_i)=c_i+(1-c_i)\frac{\exp(a_i(\theta_p-b_i))}{1+\exp(\mathrm{a}_i(\theta_p-b_i))}, \quad (1)$$

where: b_i is difficulty parameter of item i; a_i is discrimination parameter of item i; θ_p is ability level of person p and c_i is the guessing parameter.

In this article, we do not consider the nature of the guessing parameter in 3PLM. Some introduction to the complexity of this issue has been provided at the beginning of this article. For convenience, guessing was attributed to the *c*-parameter regardless whether test-takers guessed at random, took an educated guess, or imposed a different strategy to provide an answer to the item beyond their ability.

The method used in this study for *c*-DIF detection on 2012 maths and language test items was the item response theory log-likelihood ratio (IRT-LR) approach. IRTLRDIF v.2.0b software was used (Thissen, 2001). Only MCQs with 4 answer options available were filtered for analysis. From the maths test 14 out of 23, and for language 18 out of 27 such items were selected. For computational convenience a random sample of 5000 examinees was selected out of the total of ca. 400K lower-secondary school leavers population in 2012. The unidimensional 3PL models were fitted to response data separately for maths and language tests.

Since studies have showed that another psychometric property of an item, i.e. difficulty, may mediate the differentiate probability of correct answers between boys and girls (Bielinski and Davison, 2001; Penner, 2003), *c*-parameters were regressed on the percentage omission and the difficulty parameters for all MCQs within the domain tested over the analysed years. The result was four linear regression analyses: all MCQs from tests on maths from 2012–2014, language ability, science, humanities.

To sum up, four analyses were launched: (a) comparison of percentage omission by boys and girls, (b) negative binomial regression to asses differences in number of omitted items by sex, (c) *c*-DIF was assessed via IRT-LR test on selected tests (maths 2012 and language 2012) and a random sample of test-takers (n = 5000), (d) linear regression using item attributes, to assess whether the *c*-parameter were determined by the percentage omission with control of item difficulty.

Results

A small proportion of students omitted at least one MCQ (Table 2). In percentage terms, this was around 2–3% students. Since the analysis covered the entire population data there were ca. 8–12K examinees. Because there were no penalty points for wrong answers, therefore, motivation to omit items did not lie in scoring procedures and the volume of students omitting at least one item was a surprise.

Results presented in Table 2 show that the mean percentage of omitted items was lower for girls than for boys on all exams. It means that boys more often leave questions unanswered, which is contrary to the theoretical claims that boys are more likely to take risks (Byrnes, Miller and Schafer, 1999). Since the variance is higher among boys, this group is more diverse in tendency to omit items. To verify the obtained results with a more robust procedure we ran a negative binomial regression to asses differences in the number of omissions between sexes.

Table 3 presents the twelve negative binomial regression models. The number of items omitted in each test in a given year is regressed on examinee sex and ability level. Regression results were presented as the incidence rate ratios for more convenient interpretation. Results shows that boys are more likely to omit items than girls even when controlling for ability. For example, for the humanities test in 2012, males compared to females, while holding ability level constant, would be expected to have an omission rate 1.4 times higher. There was also a negative relationship between item omission and ability. For the test mentioned, if a student were to increase his ability level by one standard deviation, his omission rate would be expected to decrease by a factor of 0.45 for the same sex.

Results were very similar for other subjects and exam years, except for the language test in 2014, in which no significant difference was observed between boys and girls.

C-DIF analysis was performed for analysis of differences in guessing, operationalised as the *c*-parameter. Only maths and language tests from 2012 were analysed as an example and introduction to further analysis. Only two MCQs out of 14 exhibit *c*-DIF in maths (two with larger *c*-parameter for boys) and another two out of 18 in the language ability exam (one with larger *c*-parameter for boys and one with larger *c*-parameter for girls). The hypothesis that parameters are equal for the reference and focal groups should be rejected if the *G2* test exceeds 3.84 (the $\alpha = 0.05$ critical value of the χ^2 distribution for one degree of freedom).

The following columns in Tables 4 and 5 contain, as labelled in headers: the item number, the hypothesis being tested ("all equal"), the value of the *G2* statistic and its degrees of freedom, the item parameters for the reference group and then the focal group when they are estimated with no equality constraints.

In rows named "c-", "a-", or "b equal" the item parameter estimates are presented from which the single *df* tests were derived. The first row under "all equal" shows the item parameters for the reference and focal groups with the lower asymptote (c) parameter constrained to be equal, the second shows the item parameter estimates with both the asymptote and slope (a) parameters constrained equal, and the final line shows all item parameters constrained equal (Thissen, 2001).

Figure 1 shows item characteristic curves (ICCs) for items detected as *c*-DIF on 2012 maths test, while Figure 2 shows ICCs for items detected as *c*-DIF on 2012 language test.In order to demonstrate whether variation of the *c*-parameter was determined by sex, *c*-parameters were regressed on percentage omission and the difficulty parameters for all MCQs within domain tests over the

Table 2
The tendency to omit items – summary statistics

				Summary statistic for "% of items omitted in the test"					
				Total		Female		Male	
Year	Exam	Number of MCQ	% of students who omit at least one MCQ	Mean	Variance	Mean	Variance	Mean	Variance
2012	Humanities	33	3.80	0.202	1.457	0.170	1.315	0.233	1.581
	Language	20	2.13	0.129	1.097	0.102	0.838	0.156	1.299
	Maths	20	2.18	0.127	1.033	0.121	0.994	0.133	1.071
	Science	26	2.25	0.108	0.893	0.092	0.812	0.124	0.964
2013	Humanities	33	2.28	0.139	1.355	0.112	1.198	0.164	1.489
	Language	20	1.53	0.091	0.930	0.072	0.804	0.108	1.035
	Maths	20	2.16	0.123	0.987	0.116	0.889	0.130	1.072
	Science	28	2.03	0.090	0.787	0.073	0.654	0.107	0.897
2014	Humanities	33	2.81	0.169	1.467	0.138	1.351	0.200	1.569
	Language	21	1.40	0.084	1.057	0.069	1.021	0.098	1.091
	Maths	20	1.89	0.113	1.134	0.102	1.016	0.123	1.236
	Science	28	2.59	0.129	1.010	0.108	0.889	0.149	1.113

Table 3

The tendency to omit items between girls and boys. Results of the negative binomial regression model

No. of items omitted		2012			2013			2014		
		IRR[a]	*SE*	*z*	IRR[a]	*SE*	*z*	IRR[a]	*SE*	*z*
Humanities	Sex[b]	1.407*	0.028	17.10	1.456*	0.040	13.83	1.430*	0.036	14.40
	Ability level	0.446*	0.005	-77.80	0.438*	0.006	-59.75	0.440*	0.005	-67.43
	Cons	0.027*	0.001	-109.84	0.017*	0.001	-91.70	0.021*	0.001	-94.95
	Ln alpha	2.888	0.014		3.585	0.018		3.333	0.017	
	Alpha	17.95	0.258		36.06	0.641		28.014	0.463	
	LR test of alpha = 0: chibar2(01)	G2 = 5.2e+04 *p* < 0.001			G2 = 4.9e+04 *p* < 0.001			G2 = 6.3e+04 *p* < 0.001		
	No. of obs.	393 838			379 756			362 752		
Language	Sex[b]	1.100*	0.027	3.92	1.133*	0.033	4.28	1.050	0.033	1.57
	Ability level	0.521*	0.006	-60.16	0.564*	0.007	-44.34	0.492*	0.006	-53.97
	Cons	0.016*	0.001	-103.98	0.012*	0.001	-93.73	0.011*	0.001	-88.55
	Ln alpha	2.325	0.030		2.658	0.037		2.770	0.037	
	Alpha	10.230	0.310		14.263	0.528		15.951	0.583	
	LR test of alpha = 0: chibar2(01)	G2 = 6 147.16 *p* < 0.001			G2 = 4 650.72 *p* < 0.001			G2 = 6 265.53 *p* < 0.001		
	No. of obs.	393 831			379 752			362 755		
Maths	Sex[b]	1.108*	0.026	4.44	1.131*	0.026	5.25	1.272*	0.033	9.22
	Ability level	0.662*	0.008	-35.06	0.706*	0.008	-30.38	0.612*	0.008	-37.52
	Cons	0.020*	0.001	-105.31	0.019*	0.001	-104.57	0.013*	0.001	-100.03
	Ln alpha	2.423	0.031		2.299	0.035		2.637	0.033	
	Alpha	11.284	0.354		9.963	0.345		13.975	0.459	
	LR test of alpha = 0: chibar2(01)	G2 = 5 513.82 p < 0.001			G2 = 4 245.63 *p* < 0.001			G2 = 6 310.74 *p* < 0.001		
	No. of obs.	393 700			379 633			362 749		
Science	Sex[b]	1.343*	0.031	12.58	1.382*	0.035	12.91	1.374*	0.033	13.43
	Ability level	0.639*	0.007	-42.43	0.598*	0.007	-45.19	0.662*	0.007	-38.89
	Cons	0.016*	0.001	-107.73	0.013*	0.001	-105.09	0.020*	0.001	-101.22
	Ln alpha	2.697	0.026		2.677	0.029		2.979	0.022	
	Alpha	14.835	0.388		14.549	0.419		19.668	0.423	
	LR test of alpha = 0: chibar2(01)	G2 =9 444.39 *p* < 0.001			G2 = 7 940.71 *p* < 0.001			G2 = 1.7e+04 *p* < 0.001		
	No. of obs.	393 704			379 634			362 749		

* Coefficient statistically significant at the 0.05 level.

[a] IRR – These are the incidence rate ratios for the negative binomial regression model.

[b] Reference group: female.

analysed years. Item parameters and regression models were estimated separately for data from male and female students. The results are presented in Table 6. It can be seen that the *c*-parameter value was positively related to item difficulty, and also that the relationship was slightly stronger for boys. The positive relationship between *b*- and *c*-parameter is intuitive, the more difficult the item is, the less examinees are able to answer according to their knowledge, and some of them succeed in guessing the answer correctly. However, the relationship between the *c*-parameter and percentage of omissions is not so clear. The *c*-parameter value was negatively related to percentage item omission for girls in language and humanities exams, but there was no significant relation for maths and science exams. Furthermore, there was no significant relation between the *c*-parameter and percentage of item omissions for most exams for boys (a negative

Table 4

Results of the c-DIF analysis on 2012 maths test. Results presented only for items identified as biased

					Reference group (boys)			Focal group (girls)		
Item	Hypothesis test	G2	*df*	G2/*df*	*a*	*b*	*c*	*a*	*b*	*c*
k_1048	All equal	95.20*	3.00	31.73	1.23	-0.70	0.28	1.36	-0.23	0.24
	c equal	50.60*	1.00	50.60	1.65	-0.27	0.48	2.13	0.10	0.48
	a equal	7.40*	1.00	7.40	1.65	-0.29	0.44	1.65	-0.05	0.44
	b equal	37.20*	1.00	37.20	1.22	-0.51	0.24	1.22	-0.51	0.24
k_1053	All equal	8.30*	3.00	2.77	1.74	0.76	0.22	1.64	0.61	0.16
	c equal	7.70*	1.00	7.70	1.65	0.70	0.20	1.73	0.68	0.20
	a equal	0.50	1.00	0.50	1.65	0.68	0.19	1.65	0.68	0.19
	b equal	0.10	1.00	0.10	1.64	0.68	0.19	1.64	0.68	0.19

* Coefficient statistically significant at the 0.05 level.

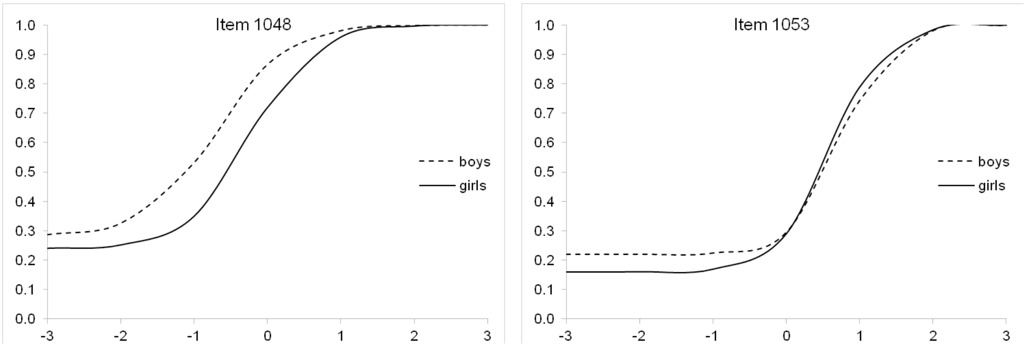

Figure 1. Item characteristic curves for items detected as *c*-DIF on 2012 maths test.

relation only for the humanities exam). This showed that the *c*-parameter was clearly not the result of not omitting items.

Discussion and future directions for research

The starting point for this article was to define test-taking strategy as a process of adaptation to a testing situation. In such a theoretical framework, guessing and item omission were viewed as risk-taking/ /risk-avoiding strategies – sex specific processes of adaptation to a testing situation.

Theories of risk behaviour were grouped into the personological and experimental. The first emphasizes examinee dispositional attributes, the second focuses on purely situational factors. Obviously, analyses on sex differences in item omissions as well as guessing should be carried out within the personological framework. The hypothesis

Table 5
Results of the c-DIC analysis on 2012 language test. Results presented only for items identified as biased

					Reference group (boys)			Focal group (girls)		
Item	Hypothesis test	G2	*df*	G2/*df*	*a*	*b*	*c*	*a*	*b*	*c*
k_528	All equal	7.00*	3.00	2.33	0.40	-0.66	0.21	0.49	-0.25	0.23
	c equal	7.00*	1.00	7.00	0.43	-0.39	0.25	0.43	-0.38	0.25
	a equal	0.00	1.00	0.00	0.43	-0.39	0.25	0.43	-0.38	0.25
	b equal	0.00	1.00	0.00	0.42	-0.51	0.22	0.42	-0.51	0.22
k_539	All equal	11.80*	3.00	3.93	0.39	-1.07	0.25	0.46	-0.66	0.21
	c equal	11.50*	1.00	11.50	0.41	-0.82	0.26	0.41	-0.82	0.26
	a equal	0.00	1.00	0.00	0.42	-0.72	0.29	0.42	-0.70	0.29
	b equal	0.50	1.00	0.50	0.40	-0.92	0.24	0.40	-0.92	0.24

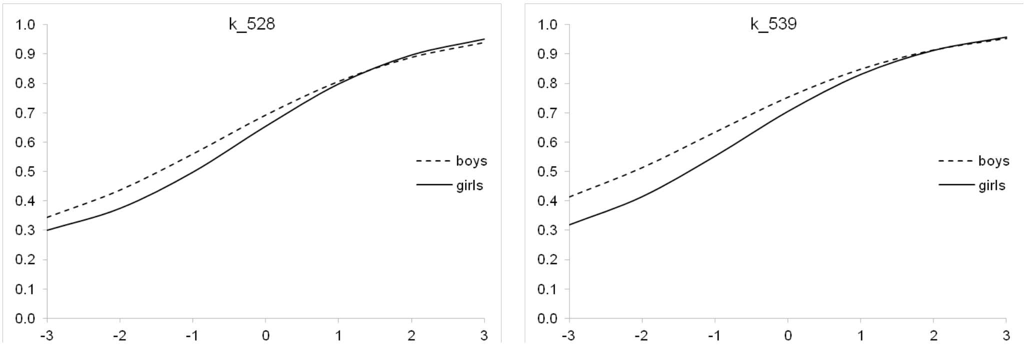

Figure 2. Item characteristic curves for items detected as *c*-DIF on 2012 language test.

Table 6

Relationship between c-parameter and percentage omission with control for item difficulty. Results of linear regression

		Female		Male	
Exam (2012–2014)	Regression results	Coef.	*SE*	Coef.	*SE*
Humanities No. of items = 96	Item difficulty	0.013*	0.006	0.018*	0.005
	% of item omission	-0.257*	0.104	-0.126*	0.062
	Constant	0.247*	0.021	0.242*	0.019
	R-squared	0.091		0.130	
Language No. of items = 59	Item difficulty	0.074*	0.028	0.092*	0.018
	% of item omission	-1.452*	0.387	-0.331	0.176
	Constant	0.499*	0.046	0.308*	0.032
	R-squared	0.242		0.325	
Maths No. of items = 60	Item difficulty	0.037	0.029	0.067*	0.026
	% of item omission	-0.068	0.235	-0.100	0.255
	Constant	0.203*	0.030	0.203*	0.034
	R-squared	0.028		0.117	
Science No. of items = 82	Item difficulty	0.056*	0.010	0.091*	0.012
	% of item omission	-0.092	0.208	-0.129	0.116
	Constant	0.213*	0.024	0.202*	0.019
	R-squared	0.283		0.409	

* Coefficient statistically significant at the 0.05 level.

posed in this article was that differences in item omission and guessing may be associated with sex regardless of different situations. This was shown partly correct, since systematic sex differences in item omission were found over three years' results (2012, 2013, 2014) on different student populations of four tests to measure different domains. Negligible differences in *c*-parameters between boys and girls were found in a random sample of test-takers ($n = 5000$) sitting the 2012 maths and language tests.

Despite that systematic sex differences in item omission were observed, no clear data-driven interpretations were provided to explain these differences. Contrary to the available literature (e.g. Byrnes et al. 1999) indicating that girls omit items more often, both analysis of omission counts and negative binomial regression models showed that boys omitted more frequently than girls. For all tests analysed, boys on average omitted 0.04% more items on tests than girls, which is about 0.01 item per test. Negative binomial regression suggested that boys are expected to have an omission rate 1.29 times greater than girls (significant effects ranging from 1.10 to 1.46). The sex effects observed were statistically significant for 11 of the 12 regression analyses. This result is in line with the works of Slakter (1967; 1968a; 1968b), as well as Zhu and Thompson (1995), however contradictory to the vast literature indicating that item omission is a phenomenon associated with girls (e.g., Pekkarinen, 2014; Von Schrader and Ansley, 2006).

Moreover, it seems that this result should be interpreted with caution, since it could be caused by specific test administration procedures in Poland. After solving paper-and-pen tests, examinees mark answers on special response-cards, which are later scanned for scoring. Since boys may be less careful in doing this, or postpone it to the very end of the examination, some items for which they actually provided answers on the test sheet may have been omitted on the response-card or were incorrectly placed.

Another potential explanation would be that boys have a different test-solving strategy than girls, that is – they first answer items to which they know the answers, and then return to deferred items. To provide answers to those items might be impossible for them due to lack of time, or they might simply be unintentionally omitted (not seen).

These alternative explanations are potentially interesting topics for future investigation of sex response differences to uncertainty.

Another explanation for this phenomenon may lie in sex differences in test-taking motivation, since girls tend to have higher motivation than boys (DeMars, Bashkov and Socha, 2013). Further investigations should also account for interactions of sex differences in item omission with differences in test completion times, e.g. by using a hybrid model (Boughton and Yamamoto, 2007). Therefore complex explanations for sex gap in test item omission should include interaction between test-taking motivation and speedy response between boys and girls.

The results regarding sex differences in guessing proved that such differences were rather unlikely. Although, analysis of *c*-DIF on selected items showed that low-performing boys guessed slightly more often than girls, differences were rather minor. Only two out of 14 MCQs on the language test, and two out of 18 MCQs on the maths test, were flagged as biased according to *c*-DIF. It is worth mentioning that the higher *c*-parameter for boys on some items should not only be interpreted as guessing. According to modern interpretation of the *c*-parameter concept (Han, 2012, Meijer 1996a; 1996b) boys may have better problem solving skills than girls, may provide creative responses, or may cheat more often or more effectively than girls, which should also be subject to future study.

Literature

Angoff, W. A. and Schrader, W. B. (1981). *A study of alternative methods for equating rights scores to formula scores* (ETS RR-81-8). Princeton: Educational Testing Service.

Arnett, J. (1992). Reckless behavior in adolescence: a developmental perspective. *Developmental Review*, *12*, 339–373.

Ben-Shakhar, G. and Sinai, Y. (1991). Gender differences in multiple-choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement*, *28*(1), 23–25.

Bielinski, J. and Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple--choice mathematics items administered to national probability samples. *Journal of Educational Measurement*, *38*(1), 51–77.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (eds.), *Statistical theories of mental test scores* (pp. 17–20). Reading: Addison--Wesley.

Boughton, K. A. and Yamamoto, K. (2007). A hybrid model for test speededness. In M. von Davier and C. H. Carstensen (eds.), *Multivariate and mixture distribution Rasch models* (pp. 147–156). New York: Springer.

Byrnes, J. P. (1998). *The nature and development of decision-making: a self-regulation model.* Hillsdale: Erlbaum.

Byrnes, J. P., Miller, D. C. and Schafer, W. D. (1999). Gender differences in risk taking: a meta-analysis. *Psychological Bulletin*, *125*(3), 367–383.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Coombs, C. H., and Lehner, P. E. (1981). Evaluation of two alternative models of a theory of risk: I. Are moments of distributions useful in assessing risk? *Journal of Experimental Psychology: Human Perception and Performance, 7(1)*, 1110–1123.

DeMars, C. E., Bashkov, B. M. and Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, *8*(2), 69–82.

Finch, W. H. and French, B. F. (2014). The impact of group pseudo-guessing parameter differences on the detection of uniform and nonuniform DIF. *Psychological Test and Assessment Modeling*, *56*(1), 25–44.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park: Sage.

Han, K. T. (2012). Fixing the c parameter in the three--parameter logistic model. *Practical Assessment, Research & Evaluation*, *17*(1), 1–24.

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge–New York: Cambridge University Press.

Holland, P. W. (1990). The Dutch identity: a new tool for the study of item response models. *Psychometrika*, 55(1), 5–18.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kelling, G. W., Zirkes, R. and Myerowitz, D. (1976). Risk as value: a switch of set hypothesis. *Psychological Reports*, *38*, 655–658.

Kolen, M. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, *18*(1), 1–11.

Long, J. S. (2001). *Regression models for categorical dependent variables using Stata*. College Station: Stata Press.

Lopes, L. L. (1987). Between hope and fear: the psychology of risk. In L. Berkowitz (ed.), *Advances in experimental social psychology* (vol. 20, pp. 255––295). New York: Academic Press.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*(2), 247–264.

Lord, F. M. (1975). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters.* (Research Bulletin RB-75--33). Princeton: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. London Routledge.

Meijer, R. R. (guest ed.). (1996a). Person fit research: theory and applications. *Applied Measurement In Education* [Special Issue], *9*(1).

Meijer, R. R. (1996b). Person-fit research: an introduction. *Applied Measurement in Education*, *9*(1), 3–8.

Pekkarinen, T. (2014). *Gender differences in strategic behaviour under competitive pressure: evidence on omission patterns in university entrance examinations*. (IZA Discussion Paper No. 8018.) Retrieved from http://ftp.iza.org/dp8018.pdf

Penner, A. (2003). International gender X item difficulty interactions in mathematics and science achievement tests. *Journal of Educational Psychology*, *95*(3), 650–655.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed). New York: John Wiley and Sons.

San Martin, E. Pino, G. del and De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*(3), 183–203.

Slakter, M. J. (1967). Risk taking on objective examinations. *American Educational Research Journal*, *4*(1), 31–43.

Slakter, M. J. (1968a). The effect of guessing strategy on objective test scores. *Journal of Educational Measurment*, *5*(3), 217–221.

Slakter, M. J. (1968b). The penalty for not guessing. *Journal of Educational Measurement*, *5*(3), 141–144.

Slakter, M. J., Koehler, R. A., Hampton, S. H. and Grennell, R. L. (1971). Sex, grade level, and risk taking on objective examinations. *The Journal of Experimental Educational*, *39*(3), 65–68.

Swineford, F. (1941). Analysis of a personality trait. *Journal of Educational Psychology*, *32*(6), 438–444.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., Crane, P. K., Gibbons, L. E., Morales, L. S., Orlando-Edelen, M. and Cell, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. *Quality of Life Research*, *16*(Supplement 1), 43–68.

Thissen, D. (2001). *IRTLRDIF v2.0b: software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning.* 2.0b ed. Chapel Hill: L.L. Thurstone Psychometric Laboratory, University of North Carolina.

Thissen, D., Steinber, L. and Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer and H. Braun (eds.), *Test Validity* (pp. 147–170). Hillsdale: Erlbaum.

Von Schrader, S. and Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-

-choice tests: 1980–2000. *Applied Measurement in Education*, *19*(1), 41–65.

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.

Wilson, M. and Daly, M. (1985). Competitiveness, risk taking and violence: the young male syndrome, *Ethology and Sociology*, *6*(1), 59–73.

Zhu, D. and Thompson, T. D. (1995). *Gender and ethnic differences in tendencies to omit responses on multiple-choice tests using number-right scoring*. (ERIC Document Reproduction Service No. ED 382 689). Retrieved from http://files.eric.ed.gov/fulltext/ED382689.pdf

Ziller, R. C. (1957a). A measure of gambling response-set in objective tests. *Psychometrika*, *22*(3), 289–292.

Ziller, R. C. (1957b). Vocational choice and utility for risk. *Journal of Counseling Psychology*, *4*(1), 61–64.

Zuckerman, M. (1991). *Psychobiology of personality.* Cambridge: Cambridge University Press.