Psychology Science, Volume 49, 2007 (4), p. 361-374

Item difficulty of multiple choice tests dependant on different item response formats – An experiment in fundamental research on psychological assessment

KLAUS D. KUBINGER[1] & CHRISTIAN H. GOTTSCHALL

Abstract

Multiple choice response formats are problematical as an item is often scored as solved simply because the test-taker is a lucky guesser. Instead of applying pertinent IRT models which take guessing effects into account, a pragmatic approach of re-conceptualizing multiple choice response formats to reduce the chance of lucky guessing is considered. This paper compares the free response format with two different multiple choice formats. A common multiple choice format with a single correct response option and five distractors ("1 of 6") is used, as well as a multiple choice format with five response options, of which any number of the five is correct and the item is only scored as mastered if all the correct response options and none of the wrong ones are marked ("x of 5"). An experiment was designed, using pairs of items with exactly the same content but different response formats. 173 test-takers were randomly assigned to two test booklets of 150 items altogether. Rasch model analyses adduced a fitting item pool, after the deletion of 39 items. The resulting item difficulty parameters were used for the comparison of the different formats. The multiple choice format "1 of 6" differs significantly from "x of 5", with a relative effect of 1.63, while the multiple choice format "x of 5" does not significantly differ from the free response format. Therefore, the lower degree of difficulty of items with the "1 of 6" multiple choice format is an indicator of relevant guessing effects. In contrast the "x of 5" multiple choice format can be seen as an appropriate substitute for free response format.

Key words: multiple choice response format, guessing effect, item difficulty, Rasch model, psychometrics

[1] Klaus D. Kubinger, PhD, Head of the Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria, Europe, email: klaus.kubinger@univie.ac.at

1. Introduction

Computer administration of psychological tests forces test authors to use items with multiple choice response formats instead of a free response format. And for economical reasons, even when a free response format would do, multiple choice tests are usually preferred. This raises the question of whether the same or different measurement dimensions are being used – experiments on learning have shown, that examination of learned materials becomes much easier if simply recognition rather than reproduction of learned material is required –, but this point will be neglected in the following. However, the problem with which we deal in this paper is the guessing phenomenon which applies if a multiple choice response format is used for the items of a test.

In other words, a multiple choice response format is problematical because items are often scored as solved even though a test-taker has a poor level of the ability that is intended to be measured. Empirical experience demonstrates that a prototypical test-taker chooses any one of the response options (i.e. suggested solutions which are offered for choice) by chance if he/she does not know the correct answer – given that refusing to respond does not seem a fair option. Hence, even when referring to a specific test-taker with an ability level of zero and minus infinite, respectively, there is always a larger than zero solution probability for every item. For this probability the term "*a priori* guessing probability" is used. This probability is conventionally $1/k$, with k being the number of response options. Very often this probability equates to 1/5, as there is only one correct among five given response options – the other four options being so-called 'distractors'. The guessing problem worsens as test-takers with a moderate ability are immediately able to rule-out certain distractors from serious consideration and consequently the actual guessing probability of a certain test-taker is larger than $1/k$, sometimes as high as $1/[k-(k-2)] = 1/2$.

Of course, such guessing effects diminish the reliability as well as the validity of a test. Even more seriously they indicate unfair consulting. Bearing in mind that according to binominal distribution a test-taker with a very poor ability level, given $h = 10$ items – with an *a priori* guessing probability of 1/5, and any selection criteria with a minimum score of 5 –, has a probability of .0328 for passing the test. This result, though often of no relevance, indicates that test-takers with a moderate ability pass or fail the test merely because they are lucky or unlucky at guessing. This can hardly be acceptable from a scientific and ethical point of view of psychological assessment: a test-taker with a moderate ability may rightfully claim the assessment was unfair because other candidates with lower ability levels may have passed the test while he/she has not.

Item Response Theory (IRT) has developed different approaches aimed to overcome guessing effects in a fair manner. Firstly, pertinent models, such as the well-known 3-PL model (Birnbaum, 1968), provide a person ability parameter and an item difficulty parameter, like the Rasch model (1-PL model; Rasch, 1960/1980) does, as well as an item discrimination parameter and an item guessing parameter. Then there is the recently recommended Difficulty plus Guessing-PL model (Kubinger & Draxler, 2006), which is simply the 3-PL model without an item discrimination parameter. In order to estimate the person ability parameter, both models take into account that any correct response to an item could be due to an item specific guessing effect. From the psychometric perspective this is of course the optimal approach. Another option are certain person fit indices (cf. for instance Ponocny & Klauer, 2002), that use the respective IRT model in order to analyze to what extent a particu-

lar response pattern of a test-taker is likely, given the test-taker's actual estimated ability parameter. The logic is that if a test-taker fails to correctly answer relatively easy items but selects the correct response option for rather difficult items, the value of the person's index becomes suspicious. Of course, there is no guarantee that every lucky guesser will be discovered by this means.

In the following we will not be dealing with these IRT approaches, but will rather present a pragmatic approach to reducing the *a priori* guessing probability of an item. In doing so, we of course agree, from a content point of view, with the recently refined guidelines for the construction of multiple choice items (cf. Moreno, Martinez & Muniz, 2006), however in the following we deal with an approach from a formal point of view.

Besides increasing the number of distractors so that k is six, occasionally seven, and often eight and consequently decreasing the *a priori* guessing probability to 1/6 or even 1/8, a proper means is, for instance, to increase the number of correct response options among the given ones. In the case of tests with five response options, of which two are correct and three are distractors, an item is only then scored as mastered if both correct response options and none of the distractors are marked. The *a priori* guessing probability then amounts to $1/\binom{5}{2} = 1/10$. In this case the test-taker is informed about both, the number of correct response options among the given ones and the scoring rule. Another concept again uses, for example, five response options but either none or one, two, three, four or even all five of them are correct; once more an item is only scored as mastered if all the correct response options and none of the wrong ones are marked – the test-taker is informed about this as well. The *a priori* guessing probability then amounts to $(½)^5 = 1/32$.

2. Aim of the research

The calculation of the *a priori* guessing probability serves for good argumentation as to how to improve the conceptualization of multiple choice response formats. However calculation *per se* does not offer the necessary empirical evidence that would make it worth the effort. Research on this topic is lacking; however a pilot study by Kubinger, Holocher-Ertl, and Frebort (2006) inspired us to attempt a systematical investigation.

This pilot study was based on a paper pencil mathematics test for large scale assessment with data from more than 6000 pupils. It resulted in four Rasch model fitting subtests with altogether 81 items. Three different response formats were used: A free response format (i.e. the numerical result of an item had to be written into the provided field; 47 items), a multiple choice format with five response options (two correct response options and three distractors: "2 of 5"; 21 items), and a multiple choice format with six response options (a single correct response option and five distractors: "1 of 6"; 13 items). Although the analysis of variance with respect to the item difficulty parameters did not yield a significant result ($p = .058$), a clear trend was evident: The difficulty of the response format "2 of 5" is almost equal to the difficulty of the free response format, while the response format "1 of 6" is much easier. This implies the presence of relevant guessing effects in the multiple choice response format "1 of 6" – bear in mind that the authors of the pilot study emphasized that designing the sample size would need $k = 44$ items from each response format in order to test a relevant effect with an adequate type-I- and type-II-risk. In the meantime this result at least upbraids the commonly used response format "1 of 5", and even more so the format "1 of 4".

However, one should also bear in mind that this pilot study did not apply item pairs with identical content but only different response formats. So the objection may arise that the kind of response format and the item contents depend on each other and therefore the established effects are not a matter of response format in principle but a matter of a used sub-universe of item contents – of course some item contents have an unequivocal solution so that no other solution is feasible.

The aim of our research is to analyze the effect of item difficulty depending on different item response formats, however, now in part using pairs of items with identical content.

3. Method

Hence an experimental approach has to be taken. For instance one half of the sample is administered the first half of an item pair, which has a certain response format, and the other half of the sample is administered the second half of this item pair, which has a different response format. The test that was used was a check list of an introductory course to Psychological Assessment at university level. This check list has subsequently been published in a text book (Kubinger, 2006), however was administered before publication. It consists of the following six subtests: *Fundamental Assessment Knowledge*, *Statistics and Psychometrics, Knowledge of the Psychological Test Pool, Theories of Intelligence and Personality, Special Knowledge about pertinent Tests,* and *Up-To-Date Information.* The first subtest contains 30 items, all the others contain half this amount (15 items), which makes a total of 105 items. There are three different response formats:

1) A free response format, of which the given answers were scored, according to a catalogue of solutions, as either right or wrong.
2) The multiple choice response format "1 of 6" as described above.
3) A multiple choice format with five response options, with none or one, two, three, four or all five of them being correct (we call this response format "x of 5") – as already indicated an item is only then scored as mastered if all correct response options and none of the wrong ones are marked; the test-takers were informed about this specific scoring rule.

Table 1 gives an item example for every subtest. Altogether there are 66 items with the free response format, 18 items with the "1 of 6" format, and 21 items with "x of 5" format. Regarding the "x of 5" format x was realized as follows: x = 0 twice, x = 1 six times, x = 2 seven times, x =3 four times, x =4 twice, and – rather arbitrarily – x = 5 never. The check list was of course not originally published with item pairs, containing the same content and different response formats. For this purpose a second test had to be developed, for which approximately half the items were changed from one response format to another (cf. the last column in Table 1 presents the modified item when applying another response format). In effect, 45 items were changed. Hence, 60 items are identical in both tests: the test with the original 105 items is called test A in the following and the test with the partly changed items constitutes test B. Consequently a cumulative item pool of 105 + 45 = 150 items resulted, with both tests regarded as two different but linked booklets of that item pool. In detail, the first subtest of the item pool consists of 46 items, the second of 26 items, the third of 23, the fourth of 19, the fifth of 17, and the sixth of 19 items.

The experiment was designed with the intent to administer test A and B in a randomized manner. The test-takers were all those students of psychology who had passed the introductory course on Psychological Assessment and had on the day of testing started with a second level seminar on that topic. They were told that the test would serve as feedback on how well trained the students were for this seminar, which is mandatory to obtain a Psychology degree at the particular university (Vienna, Austria). Additionally they were told that a good test result may help them to achieve a better grade at the end of the seminar. From the population of $N = 175$ students $n = 152$ students were tested at the beginning of the seminar (there were 7 parallel seminars being held). Furthermore, the experiment was designed so as to test practicing psychologists in the field of Psychological Assessment; they differ with no respect from the students but with regard to their professional practical experience, in particular they have undergone the same university curriculum. The pool of such psychologists was $N = 94$. Although online testing was arranged for them and anonymity guaranteed, only $n = 21$ finally participated in the study.

As the check list in question had not yet been psychometrically analyzed and we intended to use the scoring rule of simply adding up the number of solved items, a Rasch model analysis was provided for every subtest – for this scoring rule to be fair the Rasch model must hold (cf. for instance Fischer, 1995). Respective analyses had to be done according to the standards provided by Kubinger (2005): Andersen's Likelihood Ratio Test has to be applied, which in addition to the graphical model check determines whether the Rasch model holds or not. Eventually, a few items had to be eliminated in order to achieve an *a posteriori* model fit. As Andersen's Likelihood Ratio Test needs pertinent criteria to divide the sample of test-takers into two sub-samples the following four criteria were chosen (nominal $\alpha = .01$): i) low *vs.* high score, ii) younger than age 24 *vs.* older than 24, iii) low *vs.* high (self reported) grade achieved in the introductory course on Psychological Assessment, and iv) planned field of specialization for ones master thesis being Psychological Assessment or Methodology or Personality Psychology *vs.* all other subjects.

Given that the model fits *a posteriori*, the resulting item difficulty parameters – standardized to a sum of zero for each of the subtests – are used to analyze the following hypotheses (nominal $\alpha = .05$):

a) H_0: There is no mean difference between the item difficulty parameters of paired items (with identical content but different item response formats) with respect to
    1. free response format *vs.* multiple choice "1 of 6"
    2. free response format *vs.* multiple choice "x of 5"
    3. multiple choice "1 of 6" *vs.* multiple choice "x of 5".

    H_1: Due to established guessing effects there is a mean difference between the item difficulty parameters of paired items; in particular
    1. multiple choice "1 of 6" generates lower item difficulty parameters than the free response format does
    2. multiple choice "x of 5" generates lower item difficulty parameters than the free response format does
    3. multiple choice "1 of 6" generates lower item difficulty parameters than multiple choice "x of 5".

b) H_0: There is no mean difference among the item difficulty parameters of items with different response formats (taking only the original items into account).

Table 1:

Item examples for every check list's subtest. In the left column the item in the original response format is given, in the right column the modified item in another response format – for the multiple choice format the correct responses are underlined

original item	modified item
Fundamental Assessment Knowledge	
The reliability of a psycho-diagnostic instrument is necessary for …….	The reliability of a psycho-diagnostic instrument is necessary for • the calculation of the standard error • the determination of the usefulness • the determination of the credibility of the test-taker • the calculation of the probability of a correct prediction • the calculation of the confidence interval of the true score
Statistics and Psychometrics	
Item difficulty is determined by • the absolute frequency of item solutions • the relative frequency of item solutions • the probability of lucky guessing • the length of an item's text • the item parameter	Item difficulty is determined by ○ the absolute frequency of item solutions ○ the probability of lucky guessing ○ the length of an item's text ○ the item parameter ○ the number of previously administered items ○ the person parameter
Knowledge of the Psychological Test Pool	
TAT is based on Murray's personality theory. Which personality questionnaire is also based on this theory?	TAT is based on Murray's personality theory. Which personality questionnaire is also based on this theory? ○ Sceno-Test ○ LMI ○ TIPI ○ OLMT ○ MMG ○ PRF

Theories of Intelligence and Personality	
The construct *locus of control of reinforcement* refers to a person's attitude to ○ who or what controls his/her life ○ whom or what he/she controls him-/herself ○ whom he/she may trust ○ possessing a great deal of power ○ control being of more importance than trust ○ the necessity of delayed gratification	The construct *locus of control of reinforcement* refers to a person's attitude to .....
Special Knowledge about pertinent Tests	
FAIR and *Test d2* differ with respect to the fact that • FAIR uses a „circle" and „square" instead of „d" and „p" • *Test d2* prevents a test performance that is non-conform with the instruction while FAIR does not • FAIR prevents a test performance that is non-conform with the instruction while *Test d2* does not • FAIR exclusively uses symmetric symbols • FAIR is a computer test while *Test d2* is not	FAIR and *Test d2* differ with respect to the fact that ○ *Test d2* prevents a test performance that is non-conform with the instruction while FAIR does not ○ FAIR prevents a test performance that is non-conform with the instruction while *Test d2* does not ○ FAIR contains extra standardization tables for the Austrian population while *Test d2* does not ○ *Test d2* needs at most half the administration time to complete in comparison to FAIR ○ FAIR measures concentration and *Test d2* measures alertness ○ FAIR contains extra standardization tables for men as well as women while *Test d2* does not
Up-To-Date Information	
The current version of the German *Wechsler* intelligence test-battery for children is	The current version of the German *Wechsler* intelligence test-battery for children is ○ HAWIK-R ○ HAWIK ○ HAWIK-III ○ HAWIE-R ○ WIP ○ W-ITB

H_1: Due to established guessing effects there is a mean difference among the item difficulty parameters of items with different response formats. In particular multiple choice "1 of 6" generates lower item difficulty parameters than multiple choice "x of 5" and multiple choice "x of 5" by itself generates lower item difficulty parameters than the free response format does.

Although the first null-hypothesis is much more conclusive (see above), the latter one is of interest because of a possible comparison with the results of the cited pilot study.

4. Results

Altogether there were $n = 152 + 21 = 173$ test-takers. The software *LPCM-WIN* (Fischer & Ponocny-Seliger, 1998) was used for the Rasch model analyses.

The applied check list indeed proved to fit the Rasch model *a posteriori*, after several items from the first subtest and a few items from every other subtest had been eliminated. In the case of the first subtest, 15 out of 46 items did not fit the model with respect to all criteria, but 28 fitted *a posteriori* – 3 other items could not be analyzed because they had been either solved or not solved by every test-taker in at least one of the two sub-samples with respect to every partition criteria. Amongst the other subtests 7 of 26, 3 of 23, 5 of 19, 3 of 17, and 6 of 19 items were disclosed as unfitting and eliminated. Table 2, as an example, summarizes the results of Andersen's Likelihood Ratio Test for criteria i) with respect to each subtest, first analyzing all items and then again after elimination of the unfitting items. Figures 1 and 2 additionally illustrate the graphical model check of the first subtest, again first analyzing all items and then again after elimination of the unfitting items. Besides a non significant Likelihood Ratio Test after item elimination, the graphical model check demonstrates model fit, as all the dots representing item difficulty parameters come close to the 45-degree line.

Table 2:

Results from Andersen's Likelihood Ratio Tests with respect to the partition of the sample into test-takers with low *vs.* high scores for all six subtests. The results of an analysis of all items as well as after elimination of the unfitting items; $n = 173$ ($\alpha = .01$)

	Fundamental Assessment Knowledge			***Statistics and Psychometrics***			***Knowledge of the Psychological Test Pool***		
	χ^2(LRT)	*df*	$\chi^2(\alpha=.01)$	χ^2(LRT)	*df*	$\chi^2(\alpha=.01)$	χ^2(LRT)	*df*	$\chi^2(\alpha=.01)$
all items	85.35	40	63.71	29.93	25	44.34	27.92	20	37.59
after item elimination	31.00	26	45.66	11.08	18	34.83	17.46	18	34.83

	Theories of Intelligence and Personality			***Special Knowledge about pertinent Tests***			***Up-To-Date Information***		
	χ^2(LRT)	*df*	$\chi^2(\alpha=.01)$	χ^2(LRT)	*df*	$\chi^2(\alpha=.01)$	χ^2(LRT)	*df*	$\chi^2(\alpha=.01)$
all items	32.09	16	32.03	12.94	15	30.60	24.70	16	32.03
after item elimination	9.76	11	24.75	11.10	13	27.72	5.62	11	24.75

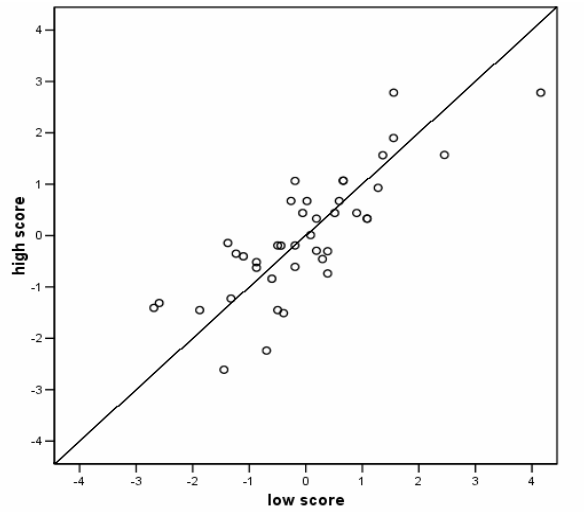


Figure 1:
Graphical Rasch model check of 41 items from the first subtest. Item difficulty parameters were opposed as estimated within the sub-samples of test-takers with low scores *vs.* high scores. Due to the fact that 5 items were either solved or not solved by every test-taker within the respective sub-sample the number of analyzed items is less than 46

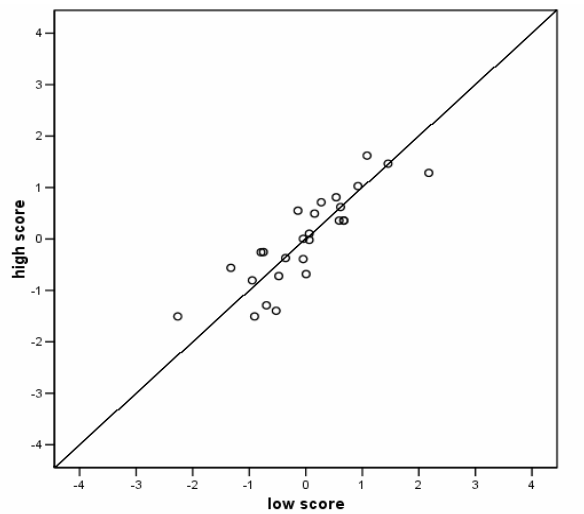


Figure 2:
Graphical Rasch model check of the first subtest after the elimination of 15 items. Item difficulty parameters were opposed as estimated within the sub-samples of test-takers with low scores *vs.* high scores. Due to the fact that 4 items were either solved or not solved by every test-taker within the respective sub-sample the number of analyzed items is less than 31.

Items which had to be eliminated are distributed according to their response format as given in Table 3. Unfortunately a lot of items lost their counterpart because one of the pair was disclosed as not fitting the Rasch model. 3 of the 22 paired items with a free response format were found not to be fitting, 5 of the 31 paired items with the multiple choice format "1 of 6", and 9 of the 37 paired items with the multiple choice format "x of 5" were found as unfitting. In 3 cases both items of a pair failed to fit the Rasch model. Consequently the following number of Rasch model fitting pairs of items resulted: 4 times the free response format in combination with "1 of 6", 10 times the free response format in combination with "x of 5", and 11 times "1 of 6" in combination with "x of 5".

Therefore, testing the null-hypothesis a), the first combination is not at all conclusive. Nevertheless the paired *t*-test resulted in significance: $t = 4.134$, $df = 3$, $p = .013$ (the mean item difficulty parameters are 0.7318 and -0.8300, which goes in the expected direction). The second and the third combination of the results are however rather conclusive and are presented in detail in Table 4.

In order to test the null-hypothesis b), the item difficulty parameters of 45 items with a free response format, 13 items with the multiple choice format "1 of 6", and 14 items with the multiple choice format "x of 5" were at our disposal – these being only those (Rasch model fitting) items, which originally constituted the check list and none of their counterpart. Again, the estimations of the item difficulty parameters are based on $n = 173$ test-takers. Descriptive analyses of the resulting histograms instantaneously disclosed heterogeneous variances due to the surprising empirical fact that some items with a free response format have a comparatively very low difficulty. As empirically expected, it was also ascertained that a lot of items with a free response format do indeed have very high difficulties (cf. Fig. 3a-3c). The Levene-test confirms: $p = .047$. Hence, a multiple *t*-test analysis was approached, that is the comparison of the means between the item difficulty parameters of multiple choice items "1 of 6" and "x of 5", on the one hand, and the comparison of the means between the item difficulty parameters of multiple choice items "x of 5" and items with a free response format, on the other hand. The results are given in detail in Table 5.

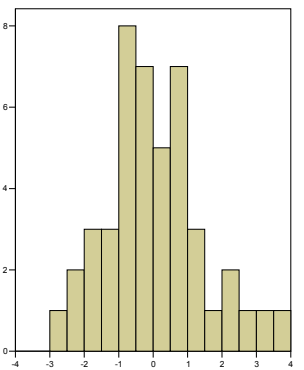
Table 3:
Number of items which had to be eliminated according to Rasch model analyses with respect to their response format

Number of items with free response format		Number of items with multiple choice format "1 of 6"		Number of items with multiple choice format "x of 5"											
				"0 from 5"		"1 from 5"		"2 from 5"		"3 from 5"		"4 from 5"		"5 from 5"	
overall pool	after elimination	overall pool	after elimination	overall pool	after elimination	overall pool	after elimination	overall pool	after elimination	overall pool	after elimination	overall pool	after elimination	overall pool	after elimination
74	51	36	29	7	5	10	7	15	10	6	4	2	2	-	-

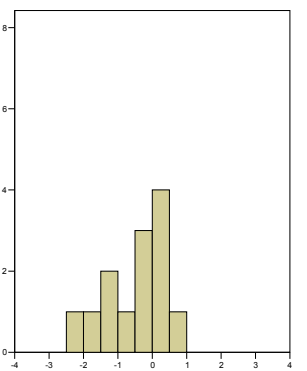
Table 4:
Results of the paired *t*-test with respect to the item difficulty parameters of items with different response formats ($\alpha = .05$, one-sided)[2]

	multiple choice "x of 5"		free response format				
	mean	standard deviation	mean	standard deviation	*t*	*df*	*p*
$n = 10$	0.2078	1.1177	-0.2055	1.6160	0.757	9	[.234]

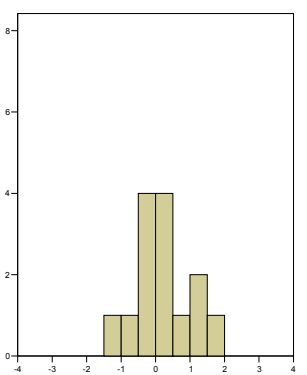
	multiple choice "1 of 6"		multiple choice "x of 5"				
	mean	standard deviation	mean	standard deviation	*t*	*df*	*p*
$n = 11$	-0.5505	0.9417	0.5633	0.7301	-5.418	10	.000



a



b



c

Figure 3a-3c:
Histogram of the item difficulty parameters of the items with a free response format (a), the items with the multiple choice format "1 of 6" (b), and the items with the multiple choice format "x of 5" (c).

[2] As in the first case the sign of the difference of means contradicts to the sign of the alternative hypothesis H_1, the respective *p*-value is given in brackets.

Table 5:
Results of the *t*-test for independent samples with respect to the item difficulty parameters of items with different response formats ($\alpha = .05$, one-sided)[3]

multiple choice "x of 5"			free response format					
	mean	standard deviation		mean	standard deviation	*t*	*df*	*p*
$n = 14$	0.2940	0.8000	$n = 45$	0.0401	1.4218	0.635	57	[.264]

multiple choice "1 of 6"			multiple choice "x of 5"					
	mean	standard deviation		mean	standard deviation	*t*	*df*	*p*
$n = 13$	-0.4555	0.8263	$n = 14$	0.2940	0.8000	-2.395	25	.012

5. Interpretation

Although neither the test-taker sample size nor the number of item pairs with identical content but different item response formats are very large, our results demonstrate that item difficulty varies significantly dependant on the conceptualization of different multiple choice response formats. The difference is, however, not just significant, but very relevant as the relative effect of the paired items, for instance, in the case of the response formats "1 of 6" vs. "x of 5" amounts to 1.63: (0.5505 + 0.5633)/0.6818 – the denominator is the standard deviation of the differences of the item difficulty parameters.

The results of our analysis without item pairing confirm the results of the cited pilot study. Using the multiple choice format "x of 5", instead of the format "2 of 5", the analysis even yields significance as concerns the lower difficulty of items with the multiple choice format "1 of 6".

6. Discussion

The problem this paper deals with is based on a prototypical test-taker who chooses any one of the given response options of a multiple choice item by chance if he/she does not know the correct answer. Admittedly there is no evidence on how frequently this prototype is represented within the population and above all, this frequency surely depends on the conditions under which the testing occurs. Very anxious test-takers who don't know the solution may not guess often and rather leave out an item than respond incorrectly. Furthermore, it is likely that guessing behavior is culturally determined, as in some countries people are acculturated not to guess if they do not know the correct answer. In consideration of these arguments, the results of our experiment do not generalize absolutely. However, there is no viable reason to neglect these results, because the question is not how often and to what extreme the problem occurs but rather that the problem may actually occur at all: By apply-

[3] As in the first case the sign of the difference of means contradicts to the sign of the alternative hypothesis H_1, the respective *p*-value is given in brackets.

ing an alternative approach, like the one indicated here, it is not necessary to risk the distortion of results due to a test-taker being a lucky guesser.

In particular, the paper deals with the multiple choice format "x of 5". From a naïve point of view, the scoring rule does not look fair, as an item is only then scored as mastered if all the correct response options and none of the wrong ones are marked. So the question may be raised: What about a test-taker who has marked two of three correct response options and none of the wrong ones? Should he/she not be given any credit for proving to be partially competent? A content-based, yet slightly lofty counter-argument could be: No one would judge a car with three perfectly installed wheels as road-worthy as long as the forth wheel is missing. A more factual counter-argument is, that analysis according to the Rasch model has factually proven this scoring rule is indeed fair; as the Rasch model holds, there is no empirical support that a certain group of test-takers is systematically discriminated against.

Due to our experimental approach there is no other explanation for the large differences in item difficulty parameters, than the one we give: the differences are clearly due to the different applied response formats. – As discussed above, unlike our experiment the cited pilot study can be criticized in that the established effects are perhaps not a consequence of the different response formats *per se* but rather a result of response formats' conditioned sub-universes of item contents.

Furthermore, the considerable differences between the multiple choice format "1 of 6" and, both, multiple choice format "x of 5" as well as the free response format, are interpreted as being the result of guessing effects. One may, however, argue that a lower item difficulty in the multiple choice response format, in comparison to the free response format, is due to the fact that free response formats require reproduction while a multiple choice response format only requires recognition. Yet, this argument does not hold, because the multiple choice format "x of 5" is equally as difficult as the free response format is.

Consequently it can be concluded, the multiple choice format "1 of 6" should be avoided and if a free response format is not practical, the multiple choice format "x of 5" should preferably be used. Be aware that applying the format "1 of 6" instead of "x of 5" at times greatly enhances the probability of an item being credited as having been solved: According to the Rasch model formula, the probability of correctly solving an item increases from .50 to .7528 in the case of a test-taker with a medium ability (0.00), because of the given shift of an item with medium difficulty (originally 0.00) – instead of $e^{0-0}/(1+e^{0-0})$ this probability is then $e^{0-(-0.5505-0.5633)}/(1+e^{0-(-0.5505-0.5633)})$. Furthermore, by generalizing our results, we anticipate that the situation would be even worse when using the multiple choice format "1 of 5", which is usually the case, never mind the commonly used multiple choice format "1 of 4", which we did not even consider. Finally, it can be stated that the multiple choice format "x of 5" is indeed an equivalent substitute for a free response format. Whether the, previously mentioned, multiple choice format "2 of 5" is a suitable alternative as indicated in the cited pilot study, is to be clarified in forthcoming fundamental research on psychological assessment. Another research topic, to be further investigated, would be to examine the effect of x = 0 and x = 5 within the multiple choice format "x of 5" in detail, specifically dependant on different test-taker populations.

As our pragmatic approach of re-conceptualizing multiple choice response formats is of comparatively less effort for test authors than it is for practitioners to apply an IRT model with a guessing parameter, and moreover almost no guessing effects occur by using this approach, we conclude that this will be the future solution for psychological assessment.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison- Wesley, 395-479.

Fischer, G.H. (1995). Derivations of the Rasch Model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models* (pp. 15-38). New York: Springer.

Fischer, G. H. & Ponocny-Seliger, E. (1998*). Structural Rasch modeling. Handbook of the usage of LPCM-WIN 1.0.* Groningen: ProGAMMA.

Kubinger, K.D. (2005). Psychological Test Calibration using the Rasch Model - Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5*, 377-394.

Kubinger, K.D. (2006). *Psychologische Diagnostik – Theorie und Praxis psychologischen Diagnostizierens* [Psychological Assessment – Theory and Application]. Göttingen: Hogrefe.

Kubinger, K.D. & Draxler, C. (2006). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier & C.H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models - Extensions and Applications* (pp. 295-312). New York: Springer.

Kubinger, K.D., Holocher-Ertl, S. & Frebort, M. (2006). Zur testtheoretischen Qualität von Multiple Choice-Items: 2 richtige aus 5 vs. 1 richtige aus 6 Antwortmöglichkeiten [Psychometric quality of multiple choice items: 2 correct out of 5 and 1 correct out of 6 suggested solutions]. In B. Gula, R. Alexandrowicz, S. Strauß, E. Brunner, B. Jenull-Schiefer & O. Vitouch (Eds.), *Perspektiven psychologischer Forschung in Österreich. Proceedings zur 7. Wissenschaftlichen Tagung der Österreichischen Gesellschaft für Psychologie* [Proceedings of the 7th Meeting of the Austrian Society of Psychology in Klagenfurt/Carynthia] (pp. 459-464). Lengerich: Pabst.

Moreno, R., Martinez, R.J. & Muniz, J. (2006). New Guidelines for Developing Multiple Choice Items. *Methodology, 2*, 65-72.

Ponocny, I. & Klauer, K.C. (2002). Towards identification of unscaleable personality questionnaire respondents: The use of person fit indices. *Psychologische Beiträge* (laterly: *Psychology Science*)*, 44,* 94-107.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Education Research (Expanded Edition, 1980. Chicago: University of Chicago Press).