Psychological Test and Assessment Modeling, Volume 63, 2021 (4), 458-468

Note: Reducing the risk of lucky guessing as well as avoiding the contamination of speed and power in (paper-pencil) group-testing – illustrated by a new test-battery

Klaus D. Kubinger[1]

Abstract

This Note illustrates how two typical problems can be solved when psychological test administration shall be carried out in a group of testees simultaneously, rather than only individually. That is, group-testing (by paper and pencil) commonly uses both items with a multiple-choice response format as well as time limits for working on the items. The test-battery AID-G (Intelligence Diagnosticum for Group administration; Kubinger & Hagenmüller, 2019) firstly shows that multiple-choice response formats which reduce the probability of lucky guessing, actually work in practice. Moreover, even the use of a free-response format is occasionally manageable, though this is hardly established in other tests for group administration. Secondly, it shows that two IRT- (item response theory-) based options are actually realizable in practice to avoid measurement of "power" being contaminated with "speed": Only the items the testee actually worked on are scored and, optionally, the completion time for a test is restricted to the time the slowest testee of the group needs until he/she has worked on a defined minimum number of items. Incidentally, this test-battery also allows the application of various test versions with different levels of item difficulties when a respective adaption, testee by testee, is desirable within the group.

Keywords: Rasch model, speed and power, multiple-choice, lucky guessing, group-testing

[1] *Klaus D. Kubinger, PhD., Professorial Research Fellow, University of Vienna, Faculty of Psychology, Liebiggasse 5, 1010 Vienna, Austria. email: klaus.kubinger@univie.ac.at*

Introduction

Psychological (paper-pencil) tests are often presented in a group setting, i.e. the presentation of the items to more than one testee simultaneously, as this is a more economical approach. However, apart from content-based problems (see Kubinger, Deimann, and Kastner-Koller, 2012), such a setting entails serious well-known psychometric consequences. Group-testing commonly uses both items with a multiple-choice response format and time limits for working on the items. A risk of the former is the testee's chance of lucky guessing, while the latter contaminates the measurement of "power" with "speed". The former reduces the validity of the test (as well as its reasonableness, as testees may be worried they will be unluckier in their guessing than others). The latter involves the very likely danger that the speed-and-power-combined achievements do not reflect uni-dimensional measurements, and therefore do not reach validity.

Although some suggestions to minimize the probability of lucky guessing (e.g. Kubinger, 2015) are at a test author's disposal, most of the published psychological tests using a multiple-choice response format only apply the type "1 of 5": that is, there are five response options, one of which is correct and the others are distractors (i.e. wrong responses) – even if there are six or even eight response options, only a single option is correct in the majority of cases. Obviously, the so-called item *a-priori* probability of lucky guessing (that is, the probability of passing an item before any ability is used to do so, but any response option is only chosen by chance) is then relatively high. It amounts to 1/5 = .20, or 1/6 = .1667, or 1/8 = .125; hence, it is likely that even a testee with a very low ability will score a hit if he/she chooses the correct answer by chance. Instead, multiple-choice response formats with more than a single solution are to prefer. For instance, there are the types "2 of 5" (exactly two of five response options are correct and the item is only scored as mastered if both correct options and none of the distractors are chosen) and "x of 5" (the testee is informed upfront that either none, one, two, three, four, or even all five answer options might be correct for any item, but an item is only scored as mastered if all correct options but no distractors are chosen). Item *a-priori* probability of lucky guessing amounts to $\binom{5}{2} = 1/10 = .10$ for the response format "2 of 5" and to $(½)^5 = 1/32 = .03125$ for the format "x of 5". Kubinger, Holocher-Ertl, Reif, Hohensinn, and Frebort (2010) proved empirically that the response format "2 of 5" indeed nearly solves the problem: While this format revealed almost the same (Rasch model) item difficulty parameters as the free-response format, the response format "1 of 6" disclosed much lower difficulty parameters, which indicates relevant guessing effects. Even more convincing is the result of an experiment by Kubinger and Gottschall (2007), where items with exactly the same content but different response formats were used: The (Rasch model) item difficulty parameters for the response format "1 of 6" differ not only significantly but with a remarkable effect size (i.e. the items being easier) from those of the response

format “x of 5”, while the latter do not significantly differ from the difficulty parameters of the free-response format.[2]

When a time limit for working on the items is given, the issue is whether quick solution finding indeed indicates a higher ability than slower solution finding; or to say, whether too slow solution finding indicates the same (low) ability as finding no solution. In fact, many widely used tests in consulting practice proceed according to this assumption without any empirical evidence that the resulting test-scores actually grade the testees along a single (intended) ability dimension. In particular, such tests score all items that a testee did not attempt because of reaching the time limit as not solved. Kubinger (1983) already demonstrated a considerable bias in the (Rasch model) ability parameter estimation when non-attempted items at the end of a test were scored as not solved rather than defining such cases as missing data. Hohensinn and Kubinger (2011) support this result most notably through a simulation study: The scheduled ability parameters become systematically underestimated the slower a testee is. That is, taking speed in addition to power into account most likely results in unfair scoring: testees who show slow but deliberate processing will be discriminated as their score could be higher with (almost) no time limit.

The test-battery AID-G (*Intelligence Diagnosticum for Group administration*; Kubinger & Hagenmüller, 2019) now serves to illustrate how reducing the risk of lucky guessing actually succeeds in practice. And it serves as a demonstration that two IRT- (*item response theory*-) based options are also realizable in practice in order to avoid contamination of speed and power as a result of (too strict) time limits for working on the items.

[2] Of course, some IRT (*item response theory*) models that estimate every testee’s ability parameter by taking the possibility of lucky guessing into account (e.g. Kubinger & Draxler, 2006) would somehow manage to deal with the problem of lucky guessing. However, they do not offer a sufficient statistic for the looked-for ability parameter, as a consequence of which their use is not functional in practice when paper-pencil (group-) administration applies: Estimation of the ability parameter works only computerized, based on the testee’s almost unique pattern of solved and not solved items out of $\sum_{0=1}^{k}\binom{n}{k}$ possible patterns – n ... the number of items, k ... the number of solved items. For instance, for $n = 20$ items the number of different patterns amounts to 1 048 576 (calculated by https://www.wolframalpha.com/input/?i2d=true&i=Sum%5BBinomial%5C%2891%296700%5C%2844%29k%5C%2893%29Divide%5BBinomial%5C%2891%293300%5C%2844%291000-k%5C%2893%29%2CBinomial%5C%2891%2910000%5C%2844%291000%5C%2893%29%5D%2C%7Bk%2C570%2C770%7D%5D)

Suggestion I: Reducing the risk of lucky guessing

As a matter of fact, seven subtests of the test-battery AID-G prove that even items with a free-response format are occasionally manageable, though hardly established in other tests for group administration. At least in the standardization sample (N = 6461 testees), no problems occurred with administration nor with scoring. Figure 1 shows an item as an example for each subtest – the task as well as the solution are included.

a) A part of a compounded word is missing; by adding a second part, the sentence shall be made meaningful!
After three weeks of vacation the letter is full.

Solution, handwritten in a blank field: *letterbox*

b) *Pascal does sport daily for 60 minutes, Kevin just the half. How many minutes does Kevin do sport every day?*
Solution, handwritten in a blank field: *30*

c) Presented words are to be ordered according their logical relation!
winter – spring – autumn – summer

Solution, handwritten in four blank fields below the words: *4 – 1 – 3 – 2*[3]

d) Eleven words have to be learnt by heart in the given sequence!
bus – door – ...

Solution, handwritten in eleven blank fields: *bus – door – ...*

e) The antonym of a word is to be found!
wet

Solution, handwritten in a blank field: *dry*

f) For a presented word, all its letters have to be marked in a list of the alphabet
ALP

Solution: ~~A~~ B C D E F G H I J K ~~L~~ M N O ~~P~~ Q R S T U V W X Y Z[4]

[3] This response format might rather be seen as a multiple-choice response format than a free-response format as there are 4! = 24 possibilities to arrange 4 objects. Hence the item *a-priori* probability of lucky guessing amounts to 1/24 = .0417.

[4] this response format might rather be seen as a multiple-choice response format than a free-response format as there are $\binom{26}{3}$ = 2600 possibilities to mark 3 out of the 26 letters. By the way, in this case, the item *a-priori* probability of lucky guessing amounts to 1/2600 = .0004.

g) A pattern, compounded by different given sub-patterns, has to be appropriately decomposed by drawing respective lines

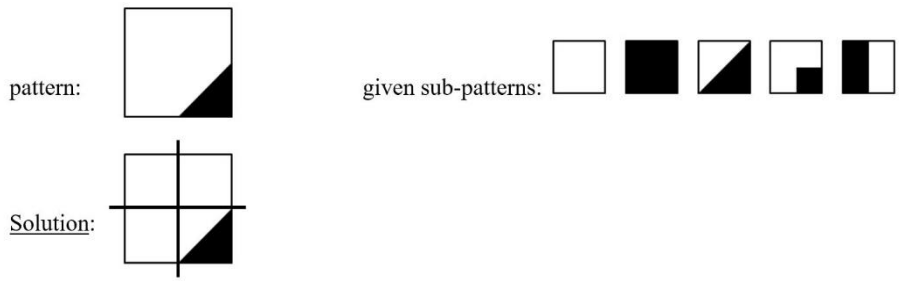

Figure 1: Examples of applying a free-response format for seven subtests of AID-G (Kubinger & Hagenmüller, 2019), an intelligence test-battery for paper-pencil group administration (translation by the author).

In addition to these seven subtests, there are five more that apply a multiple-choice response format. Two of them have the format "1 of 6", one has the format "2 of 5", one "1 of 5", and the last "1 of 16". Figure 2 gives an item as an example for each case – again, the task as well as the solution are included.

a) *Which day follows Sunday?*
☐ Tuesday
☐ Saturday
☐ Wednesday
☒ Monday ← Solution!
☐ Friday
☐ Thursday

b) Two of the five objects, which have something in common or fulfil the same function, shall be marked.
book – journal – movie – theatre play – computer game

Solution: *book; journal*

c) *Why do many people go to the theatre? Because …*
☐ friends of theirs perform there
☐ they have received tickets as a gift
☐ they are actors themselves
☐ they like to make themselves public
☒ they like theatre ← Solution!

d) Out of 16 geometric shapes, that one which continues a given series of such shapes in a logical manner shall be marked.

↑ Solution!

Figure 2: Examples of applying different multiple-choice response formats for four subtests of the AID-G (translation by the author).

Admittedly, using the format "1 of 6" and the more using the format "1 of 5" illustrated by the examples a) and c) in Figure 2 goes directly against the empirical findings cited above and consequently the suggestion that it is better to apply response formats like "2 of 5" and "x of 5". That shows: there are probably always certain psychodiagnostic informative tasks for which more than a single solution is not possible (cf. example a) in Fig. 2). Moreover, the use of more than four or five distractors can then be impossible too (cf. again example a) in Fig. 2). On the other hand, example b) shows that applying the response format "2 of 5" instead of the response format "1 of 5" or the like is sometimes very easy.

Suggestion II: Measuring power without contamination with speed

The test-battery AID-G implements two IRT-based options in order to exclusively measure the testees' power although the administration of a test is speeded. Both options are covered by the adaptive testing approach, which has been well recognized for a long time (cf. e.g. Kubinger, 2016).

In the first instance, AID-G's subtests score only those items the testee actually worked on. In the second optional instance, its subtests simply allow the amount of time the slowest testee of the group needs until he/she has worked on a defined minimum number of items.

The issue with both options is to get a test-score that represents a testee's degree of measured ability ("power"), even if several testees did not work on (all) the same items. Within IRT, the modelled ability parameter works this out. For this, AID-G applies the Rasch model (Rasch, 1960/1980), or to say the 1-PL model. As it is well established, this model postulates a specific probabilistic function between solving an

item ("+") – and not solving the item ("-") – on the one side, and the testee's ability-parameter ξ_v and the item's difficulty-parameter σ_i, on the other side:

$$P\left(+\middle|\xi_v,\sigma_i\right)=\frac{e^{\xi_v-\sigma_i}}{1+e^{\xi_v-\sigma_i}} \qquad (1)$$

Given the model holds for the items of a respective test and given the item parameters are known from a large calibration sample, then the Maximum Likelihood estimation approach can be applied in order to get an estimated value $\hat{\xi}_v$ of a certain testee *v*'s ability parameter ξ_v. That is, the first derivative (as a function of $\hat{\xi}_v$) of the model-specific likelihood of the actual data must be set to zero in order to find an extremum (maximum):

$$L_v=\prod_{i=f_1(v)}^{f_{k_v}(v)}\left(\frac{e^{(\hat{\xi}_v-\sigma_i)}}{1+e^{(\hat{\xi}_v-\sigma_i)}}\right)^{x_{vi}}\cdot\left(\frac{1}{1+e^{(\hat{\xi}_v-\sigma_i)}}\right)^{1-x_{vi}}=max \qquad (2)$$

with x_{vi} = 1 if testee *v* has correctly responded to item *i* and x_{vi} = 0 if *v* did not; the labels $f_1(v), f_2(v), \ldots f_{k_v}(v)$ of the k_v items which were administered to testee *v* are due to the fact that several testees did not work on (all) the same items. Numerical mathematics delivers appropriate algorithms for iterative solutions to the extremum problem (cf. e.g. the R-package PP, Person Parameter estimation; Reif, 2012).

By applying this approach, any testees can be compared with respect to their test performances in a fair manner – in particular, disregarding whether some of them proceeded faster than others and therefore worked on more items. Generally, no matter which items a testee worked on, the estimation procedure for his/her ability parameter is always the same. Take, for instance, a testee *v* who solves item number 1 and item number 9, but does not solve item number 5. In this case, the ability parameter estimation $\hat{\xi}_v$ succeeds accordingly as

$$L_v=\prod_{i=f_1(v)}^{f_{k_v}(v)}\left(\frac{e^{(\hat{\xi}_v-\sigma_i)}}{1+e^{(\hat{\xi}_v-\sigma_i)}}\right)^{x_{vi}}\cdot\left(\frac{1}{1+e^{(\hat{\xi}_v-\sigma_i)}}\right)^{1-x_{vi}}=$$

$$P\left(1^+,5^-,9^+\middle|\hat{\xi}_v;\sigma_1,\sigma_5,\sigma_9\right)=\frac{e^{\hat{\xi}_v-\sigma_1}}{1+e^{\hat{\xi}_v-\sigma_1}}\cdot\frac{1}{1+e^{\hat{\xi}_v-\sigma_5}}\cdot\frac{e^{\hat{\xi}_v-\sigma_9}}{1+e^{\hat{\xi}_v-\sigma_9}}=max \qquad (3)$$

For another testee *w*, responding correctly to items number 2 and 3, but not to item number 4, the looked-for ability parameter estimation $\hat{\xi}_w$ results analogously – bear in mind, that $\hat{\xi}_w - \hat{\xi}_v$ discloses a difference in the target ability dimension in an interval-scaled manner.

Apart from a single subtest (for which the so-called Rasch-*Poisson*-model applies) every AID-G subtest proved to fit the Rasch model (see for details Kubinger & Hagenmüller, 2019). Therefore, the outlined suggestion above, for avoiding any confounding of power with speed, indeed works. This is demonstrated in the following.

In the first instance, there are actually set time limits. In subtest 1, the time limit amounts to eight minutes (after four minutes, the testees are instructed that half of the time still remains). When the administration is stopped after eight minutes, the testees are advised to mark which item they last worked on by drawing a line below to it. Only the items up to that line are taken into account for scoring, and thus assigning a final test score. For each number of worked on items, the standardization tables provide the transformation of the number of thereof solved items into the respectively ability parameter (estimation).

In the second, optional, instance, the scheduled time limits will be shortened when the slowest testee of the tested group has worked on a given minimum number of items. In this case, the testees are also advised to indicate which item they worked on last by drawing a line below to it, after administration has ended. The assignment of a test-score occurs in the same way as described above. In common practice with no more than eight testees in a tested group, no problems should arise to determine when the slowest testee worked on enough items.

Irrespective of which of the two instances apply, this approach is contrasted to the traditional one in the following through a numerical example.

In subtest 1, there are 20 items altogether and the slowest testee of a group must work on at least six items. Now imagine three testees of the same age. Testee A has worked on eight items and has solved every one of them. Testees B and C both worked on all the 20 items, but while B also solved eight items, C solved ten. If this were the case in a traditional test, then A and B reached the same test-score, i.e. “8”, and are therefore assessed as having an equal level of ability; and C is evaluated to have the highest level. However, if only the power of the testees is of interest and the speed effect is not taken into account, this order completely changes, as illustrated by the transformation table given in Figure 3. There, traditional determination of the test-score would only take the last column (i.e. 20 items) into account; irrespective of how many items the testee has worked on, the number of solved items definitely determines the looked-for ability parameter. However, if only those items are scored that the testee actually worked on, testee A proves to have the highest ability (ability parameter 2.5; see the third column in Fig. 3 referring to all cases with eight worked on items), C the next highest (0.5), and B the lowest ability (0.0). As a matter of fact (cf. the standardization tables in the manual, Kubinger & Hagenmüller, 2019), the respective *T*-score for

testee A amounts to 72, for B to 47, and for C to 52, which corresponds to the percentile ranks 99, 38, and 58, respectively.

PS	number of worked-on items															PS
	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
0	-3.4	-3.5	-3.5	-3.6	-3.6	-3.6	-3.7	-3.7	-3.7	-3.7	-3.7	-3.7	-3.7	-3.7	-3.8	**0**
1	-2.4	-2.5	-2.6	-2.6	-2.7	-2.7	-2.8	-2.8	-2.8	-2.9	-2.9	-2.9	-2.9	-2.9	-2.9	**1**
2	-1.4	-1.6	-1.7	-1.8	-1.9	-1.9	-1.9	-2.0	-2.0	-2.0	-2.1	-2.1	-2.1	-2.1	-2.1	**2**
3	-0.7	-0.9	-1.1	-1.2	-1.3	-1.3	-1.4	-1.4	-1.5	-1.5	-1.5	-1.5	-1.6	-1.6	-1.6	**3**
4	0.0	-0.3	-0.5	-0.7	-0.8	-0.9	-1.0	-1.0	-1.0	-1.1	-1.1	-1.1	-1.2	-1.2	-1.2	**4**
5	1.0	0.3	0.0	-0.2	-0.4	-0.5	-0.6	-0.6	-0.7	-0.7	-0.8	-0.8	-0.8	-0.8	-0.8	**5**
6	2.0	1.3	0.7	0.3	0.1	-0.1	-0.2	-0.3	-0.3	-0.4	-0.4	-0.5	-0.5	-0.5	-0.5	**6**
7		2.2	1.5	0.9	0.6	0.4	0.2	0.1	0.0	-0.1	-0.1	-0.2	-0.2	-0.2	-0.2	**7**
8			2.5	1.8	1.1	0.8	0.6	0.4	0.3	0.2	0.1	0.1	0.1	0.1	0.0	**8**
9				2.7	2.0	1.4	1.0	0.8	0.6	0.5	0.4	0.4	0.3	0.3	0.3	**9**
10					2.8	2.2	1.5	1.2	1.0	0.8	0.7	0.6	0.6	0.6	0.5	**10**
11						3.1	2.4	1.8	1.4	1.2	1.0	0.9	0.9	0.8	0.8	**11**
12							3.2	2.6	2.0	1.6	1.4	1.2	1.2	1.1	1.0	**12**
13								3.4	2.8	2.1	1.8	1.6	1.5	1.4	1.3	**13**
14									3.6	2.9	2.3	2.0	1.8	1.7	1.6	**14**
15										3.7	3.1	2.5	2.2	2.1	1.9	**15**
16											4.0	3.3	2.8	2.5	2.3	**16**
17												4.1	3.6	3.1	2.7	**17**
18													4.5	3.9	3.2	**18**
19														4.8	4.1	**19**
20															4.9	**20**

Figure 3: A numerical example of determining the test-score when speeded test administration applies, using the transformation table of subtest 1 from the test-battery AID-G (Kubinger & Hagenmüller, 2019, p. 86), whose approach is compared to the traditional one. While the former only considers those items the testee really worked on, the latter scores the hits for all items, whether the testee worked on (all of) them or not. Simulating the latter case, only the last column of the AID-G transformation table would be of relevance (pretending all

20 items have been worked-on). In the intended case of AID-G, the other columns have to be taken into account due to the number of items the testee actually worked on (PS stands for the sum of points, i.e. the number of solved items). Three testees are considered in this example, with A solving all eight items he/she worked on, B solving eight items as well and C solving ten, both the latter having worked on (all) 20 items of the subtest. The given transformation table refers to the respective (Rasch model) ability parameters as the test-score in question.

The example discloses how traditional psychological tests determine the looked-for test-score in an extremely unfair manner when time of test processing is restricted, although speed is fundamentally not intended to be measured. That is, if only the power of a testee is of relevance, the test-score resulting from such traditional tests is completely unsuitable. The test-battery AID-G, on the other hand, determines the test-scores in a fair manner.

Discussion

Evidently, the use of a multiple-choice response format is barely avoidable for (paper-pencil) group-testing on some occasions. This is also true for the test-battery AID-G; even the format "2 of 5" (or "x of 5" and the like) are often not feasible at all, as demonstrated by the AID-G. Nevertheless, the latter proves that the use of a free-response format is probably more often manageable than actually established in other tests for group administration. Several such efforts for this test-battery stood the test in practice.

Although the applied IRT-based approach(es) for avoiding the contamination of speed and power are perfectly obvious since the establishment of adaptive testing, it is embarrassing that they are not commonly applied in psychological assessment – be aware that the option of terminating the test administration when the slowest testee has worked on a certain given number of items has already been offered many decades ago: see the test 3DW of spatial imagination (Gittler, 1990).

Yet not mentioned is the fact that both approaches entail different measurement errors (i.e. standard error of estimation) if a different number of items were worked on. That is, the test-scores will be more accurate if that number is large but less accurate if it is small – hence the option of terminating the test administration early is constrained to a certain minimum of administered items. In the example above, the standard error of estimation of the ability parameter for testee A is almost twice as high as for B and C. Consequently, if a practitioner cannot deal with this fact, he/she must not apply any speeded test but rather a test that is conceptualized for individual administration (almost) without any time limit for working on the items.

Incidentally, if the adaptive testing approach is once already being applied, then the advantages of offering various test versions with different levels of item difficulties can be utilized, too. Within a group, the difficulty can then be individually adapted per testee based on some upfront information about their ability. The test-battery AID-G offers this, indeed.

References

Gittler, G. (1990). *Dreidimensionaler Würfeltest (3DW)* [Three-dimensional cube test, 3DW]. Weinheim: Beltz.

Hohensinn, C. & Kubinger, K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling*, *53*, 380-393.

Hohensinn, C. & Kubinger, K. D. (2017). Using Rasch model generalizations for taking testees' speed in addition to their power into account. *Psychological Test and Assessment Modeling*, 59, 93-108.

Kubinger, K. D. (1983). Anhang: Einige besondere testtheoretische Belange [Appendix: some special psychometric concerns]. In K. D. Kubinger (ed.), Der HAWIK – Möglichkeiten und Grenzen seiner Anwendung [The German WISC – potentials and short-comings of its application] (pp. 229-235). Weinheim: Beltz.

Kubinger, K. D. (2014). Gutachten zur Erstellung „gerichtsfester" Multiple-Choice-Prüfungsaufgaben [expert opinion for the construction of court-proved multiple choice examinations tasks]. *Psychologische Rundschau, 65,* 169-178.

Kubinger, K. D. (2016). Adaptive testing. In K. Schweizer & C. DiStefano (eds.), *Principles and methods of test construction* (pp. 104-119). Göttingen: Hogrefe.

Kubinger, K. D., Deimann, P., & Kastner-Koller, U. (2012). Der diagnostische Mehrwert von Einzeltestsituationen [The psycho-diagnostic added value of individual settings]. In K. D. Kubinger & S. Holocher-Ertl (eds.), *Fallbuch AID* [Reader: Case Studies of AID] (pp. 21-28). 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. & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats – An experiment in fundamental research on psychological assessment. *Psychology Science, 49,* 361-374.

Kubinger, K. D. & Hagenmüller, B. (2019). *Gruppentest zur Erfassung der Intelligenz auf Basis des AID (AID-G)* [Intelligence diagnosticum for group administration, AID-G]. Göttingen: Hogrefe.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C. & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment, 18*, 111-115.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Reif, M. (2012). *PP*: *Person Parameter estimation. R package version 0.2.*: http://cran.r-project.org.