Psychological Test and Assessment Modeling, Volume 57, 2015 (2), 147-161

Which person variables predict how people benefit from True-False over Constructed Response items?

Stella Bollmann[1], Eva Böbel[2], Moritz Heene[2] & Markus Bühner[2]

Abstract

The aim of this study was the investigation of the variable *Benefit from TF*, which we assumed to be additionally measured when using True-False instead of Constructed Response tests. Subjects who benefit from True-False have an advantage over other subjects in answering Multiple Choice or True-False exams. We expected it to be related to partial knowledge and examined its relation to other personal abilities and traits in a total of n = 106 psychology students. They completed a statistics exam in Constructed Response and True-False format and benefit items were defined as those to which the associated constructed response answer was not correct. Additionally, verbal intelligence and Big 5 measures were obtained. Results confirm the existence of the person variable *Benefit from TF* and its relation to partial knowledge. Furthermore, benefiters differed from others in conscientiousness and openness to experience variables. However, contrary to expectations, they did not differ in verbal IQ.

Key words: Multiple Choice, True-False, partial knowledge

[1] *Correspondence concerning this article should be addressed to:* Stella Bollmann, PhD, Ludwig-Maximilians University Munich, Department Psychologie, Leopoldstr. 13, 80802 Munich, Germany; email: stella.bollmann@psy.lmu.de

[2] Ludwig-Maximilians University Munich, Germany

Multiple Choice (MC) and True-False (TF) tests have previously been investigated in the early 20th century and their diagnostic quality has been a controversial topic among researchers. Nonetheless, their popularity among examiners continues to grow. An important advantage is their high economy both in administration and analysis. Furthermore, in comparison to Constructed Response (CR) questions, where examinees are asked to give a short answer, the analysis of MC or TF questions is more objective in evaluation. On the other hand, it is a well-known fact that MC tests encourage guessing what is known to be a huge diagnostic problem. Since different strategies might lead to the correct answer to MC items, they face the possible loss of unidimensionality (Kubinger, 2014). The main question that is addressed in this paper is whether people benefit from answering MC (or TF) items over CR items. More precisely we are interested in a person specific variable, which we call *Benefit from TF* that indicates to what extent someone benefits from answering TF over CR items.

In the current paper TF items are used which means "each item is a statement to be judged true or false", whereas MC means "items involving a single choice from 3 or more answer options" (Burton, 2002, p.805). If the test type is not specified in the literature, we call it 'MC test'. In the following section we present an overview of former research, which leads to our hypotheses (1) to (3).

When it comes to research on MC tests, studies have focused mainly on the effect of guessing, which is one aspect of what happens when using MC instead of CR exams. Those studies mostly involve some kind of punishment for false answers (e.g. subtracting incorrect answers from the total score) in order to examine guessing behavior related to risk-taking. Nevertheless the present paper does not aim to examine exclusively guessing. We rather intend to find out, which characteristics a person must exhibit in order to benefit from TF over CR items irrespective of his guessing behavior.

Benefit from TF as a systematic variable. The main question of interest in guessing research was, whether examinees benefit from answering items on which they guess. The main methodology for measuring guessing was to ask people to answer items they would have omitted under penalty for guessing (e.g. Bliss, 1980; Cross & Frary, 1977; Lord, 1975; Sherriffs & Boomer, 1954). Although many of these studies revealed that people benefit from guessing, also some of them showed the contrary. These results suggest that not all people benefit from guessing in the same way. Therefore guessing cannot only be seen as a constant contribution to error variance but should rather be seen as a systematic variable on which people differ. For this reasons, we believe that also *Benefit from TF* is a systematic variable. In the following we are going to define the precise measurement of this variable.

Measurement of Benefit from TF. In some former studies researchers measured guessing by comparing sum scores in CR with those of MC tests concerning the same subject matters (e.g. Kennedy & Walstad, 1997; Kuechler & Simkin, 2010). Thereby it is possible to quantify an overall *Benefit from TF* in a certain subject. However, in the present study the aim was to be able to tell for every item whether the answer was based on the *Benefit from TF*, as the aim was to separate achievement with MC items completely from

achievement in general. Given the fact that this is not possible with any of the previously used methods, it was necessary in our case to formulate a completely new method.

We decided to operationalize *Benefit from TF* by designing a CR item to each of the TF items, which contains the same knowledge. An example of a TF item is: “Is the following statement true or false? The binomial random variable is a continuous random variable.” And its corresponding CR item is: “Give an example of a) a discrete and b) a continuous random variable”. If the student does not know the answer to the CR item his answer on the MC item, which we then call benefit item, is seen to involve *Benefit from TF*. This may be any ability or skill that helps finding the right answer to an MC item beyond random chance, where the knowledge is not sufficient to find the correct answer to the CR item. Next, we had to divide people into benefiters and non-benefiters. Benefiters should be only those test-takers who have higher than random chance of giving the right answer to a TF item to which they did not know the answer in the CR item. Using the 5-%-level of the binomial distribution would have meant different cut off values for different test-takers since they all have different number of trials. In TF items the expected value of number of correct when answering completely at random is 50%. We decided when test-takers also give a correct answer to the remaining 50% of the questions, their answer is considered as being above chance. This means we chose the relatively high cut off of > 75 % for distinguishing between benefiters and non-benefiters.

Nevertheless, it is not clear whether this additional variable measures something that MC exams intend to measure or whether it should rather be seen as a confounding variable. While some of the researchers proposed a test-dependent confounding variable, namely testwiseness, others focused on a knowledge-dependent variable, which is partial knowledge (e.g. Albanese, 1988; Angoff, 1989; Boldt, 1968; Votaw, 1936).

Testwiseness describes ‘a subject’s capacity to utilize the characteristics and formats of the test and/or test-taking situation, to receive a high score.’ (Millman, Bishop, & Ebel, 1965, p.707). Millman et al. (1965) additionally noted, "Test-wiseness is logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures." (S. 707) In this paper we are more interested in a positive aspect of the benefit from TF, which is in a large part independent from the quality of item-construction, although the effect of testwiseness can never be eliminated entirely. This positive aspect is the examinees' partial knowledge.

Partial knowledge. Partial knowledge refers to knowledge in the subject matter, which is not sufficient for a confident answer, but raises the possibility of a correct guess (Burton, 2002). In several studies it was found that especially high achievement test-takers receive even better results when they answer items they do not know the answer for sure and low scoring examinees even lower their scores (e.g. Albanese, 1988; Angoff, 1989; Boldt, 1968; Votaw, 1936), which means that their scores on MC items are influenced by the so called variable partial knowledge. Given the fact that there is no more recent work on this specific topic that proves these rather old findings to be wrong, we decided to relate our hypothesis to this research. Therefore we assumed benefiters to have higher partial knowledge.

In this study we used the two variables grade in the last statistics exam (grade) and CR sum score as proxy variables for partial knowledge. We believe that test-takers with higher overall achievement (i.e. higher CR sum score) have higher chances to answer correctly on a benefit item. Grade was chosen because it provides additional information about ones' ability in the required knowledge. Unfortunately, the definition of a benefiter being someone who answered more than 75% of the benefit items correctly implies that someone who has a higher CR sum score automatically has a higher baseline probability of being a benefiter – just by chance. This can only explain small effect though. However, in order to account for this effect, correlational analyses were included.

Our first hypothesis (1) was that students who benefit from True-False would have higher partial knowledge than those who do not. This means that benefiters are not able to recall the right answer on an CR item but they can use their partial knowledge to interpret the TF item the right way. We therefore divided examinees into benefiters and non-benefiters, with benefiters meaning that at least 75 % of the benefit items are correct.

To get a better understanding of what we are measuring by using TF tests, in the next step we want to examine whether this variable *Benefit from TF* items is not only related to acquired skills as partial knowledge but maybe also to personal abilities and traits.

Intelligence. Concerning cognitive abilities, only relations with testwiseness have been investigated. Only small relations to intelligence have been found (e.g. Diamond & Evans, 1972; Dunn & Goldstein, 1959), although it has been shown that verbal skills have a positive influence on testwiseness (Rowley, 1974; Sarnacki, 1979).

Furthermore, intelligence is known to be a predictor of general achievement in educational settings (Laidra, Pullmann, & Allik, 2007; Ziegler, Knogler, & Bühner, 2009) which suggests a positive relation to *Benefit from TF* as well. In this study we are particularly interested in abilities that are not related to the subject matter (statistics), since we want to know what ability enables test-takers to benefit from TF items in general, what we assume to be much more related to verbal intelligence than to numeric intelligence.

Derived from these findings, our hypothesis (2) was that benefiters have higher scores on verbal intelligence (verbal IQ) than non-benefiters.

Personality. As far as we know, there has not been any previous investigation concerning the relation between *Benefit from TF* or success in guessing and personality variables. Former studies only investigated the relation between guessing habits and anxiety and the level of relations they found was low (Bliss, 1980; Cross & Frary, 1977; Sherriffs & Boomer, 1954). However, there is broad evidence for the negative influence of impulsivity on educational achievement (e.g. Barratt & White, 1968, November; Ziegler, Danay, Schölmerich, & Bühner, 2010). So we used the personality factor neuroticism, consisting in 6 different facets of which one is anxiety and the other impulsivity and we expected especially these two to be negatively related to *Benefit from TF*. Furthermore, we analyzed the relation between *Benefit from TF* and the personality variables openness to experience and conscientiousness, as these variables are known to be related to achievement in education and certain learning styles in general (e.g. Busato et al., 1999; Furnham, 1996; de Raad & Schouwenburg, 1996; Ziegler et al., 2009) and conscientiousness is related to motivation to learn (Colquitt & Simmering, 1998).

Thus, hypothesis (3) assumes that benefiters are distinct to other people when it comes to Big 5 factors. For the facets anxiety and impulsivity we expect a negative effect on *Benefit from TF*. Since previous research does not deliver sufficient indications as to the precise direction of influence of openness to experience and conscientiousness *Benefit from TF*, we only analyze these variables in an exploratory way.

Item construction. Former studies suggest that the reliability of tests depends on the number of items keyed false and items keyed true (Cronbach, 1942; Ebel & Frisbie, 1991; Grosse & Wright, 1985). In order to avoid favoring students who were either more or less acquiescent, we keyed exactly half of our items true and the other half false.

Furthermore, we had to decide whether to choose MC or TF items. Concerning the reliability of the tests, it was found that a set of TF items has at least as much reliability as the same number of options on MC item (Grosse & Wright, 1985; Ebel & Frisbie, 1991). Moreover, TF items are faster to answer, easier to write (Ebel & Frisbie, 1991) and of course, a greater number of answers can be obtained with fewer items covering a greater range of content (Frisbie & Sweeney, 1982). But the principal reason why we chose TF items was that in MC tests the distractors may be sources of clueing and most of the testwiseness principles are based on badly constructed distractors (Albanese & Sabers, 1988, Diamond & Evans, 1972). In this study we were interested in partial knowledge and not testwiseness, thus we had to minimize the effect of testwiseness as much as possible. For this reason, in the present study we decided to use TF items.

Minimizing the effect of guessing behavior. Former studies revealed that the extent to which someone guesses on MC items differs between test-takers and is related to different trait variables (Ben-Shakhar & Sinai, 1991; Hassmén & Hunt, 1994; Swineford & Miller, 1953). In our study we wanted to eliminate this effect of guessing behavior. Therefore, we instructed students to respond to every single item even if they did not know the answer. Thereby, we forced people to guess once they did not know the answer for sure. This guarantees that differences in guessing behavior could not have any impact on sum scores.

Based on the findings summarized above the following hypotheses were made:

1. Benefiters possess more partial knowledge.
2. Benefiters have a higher verbal IQ.
3. Benefiters distinguish themselves from others in Neuroticism, openness to experience and Conscientiousness.

Method

Participants. The sample consisted of n = 106 psychology students attending a compulsory statistics seminar for the second semester at the Ludwig-Maximilian University of Munich. The mean age of the final group, which consisted of 14 males and 92 females, was 21.83 years with a standard deviation of 3.13 and a range from 18 to 32 years. All participants were German native speakers.

In order to guarantee sufficient motivation to participate, subjects were informed a few weeks in advance about how the enquiry was going to take place. For the same reason, we used questions about basic knowledge subject matters, which they would need for their next semester and promised students to provide them feedback about their results afterwards. Additionally, this first lesson was already counted as a regular compulsory lesson.

The Tests. During the first week, all of them filled out the statistics exam. First, all students took the CR exam and as soon as they had completed it, they could start individually to answer the TF exam. The order of first administering the CR and then the TF exam was chosen because the question of the TF exam may contain hints for the right answer to the CR exam but not the other way round. Participants were instructed to act as if it was a real academic exam and they were given as much time as they wanted. Of course by using this procedure the possibility that some students may employ strategies such as answering the CR question one way and the TF question the other way to obtain at least one point cannot be avoided entirely. However this occurrence is not very probable as students were primarily interested in getting a personal feedback, which was not published.

One week later, verbal IQ was assessed using the verbal part of the I-S-T 2000 R. 92 students from the final sample were present. In between these two lessons they had to fill out the NEO-PI-R at home. 77 of the students completed the NEO-PI-R. Thus, these variables had the highest percentage of missing values with 13.21 % (verbal IQ variables) and 27.36 % (personality variables). Furthermore, there were values missing in the variable grade (7.55 %), which were imputed using expectation maximization (EM) by SPSS 15.

Constructed response exam: Students were asked to answer 40 questions. Each CR item was formulated so that it retrieved the same knowledge as its corresponding TF item. Attention was paid to formulating items in such a way so as to allow answers to be given briefly in short sentences or headwords.

True False Exam: Construction of the TF questions was based on the prior exam that the students had taken at the end of the first semester. Each question consisted of a statement, which was to be identified as being true or false.

Test of verbal intelligence: For the assessment of verbal IQ, the second edition of the multidimensional intelligence structure test I-S-T 2000 R by Amthauer, Brocke, Liepmann and Beauducel (2007) was used in a group-testing. For the present study, we only used the verbal part, which consists of a total of 60 items. It assesses competence in the handling of verbal material, namely vocabulary, as well as the ability to establish relations between terms. It is available in two versions, A and B, which only differ in their item sequence. Half of the students completed version A and the other half version B.

The inventory is divided into three subtests: completing sentences, verbal analogies and similarities with each of them containing 20 items. For statistical analysis the raw score verbal IQ and the raw score of its three subtests was used. Means, standard deviations and reliability estimates can be obtained from table 1. The mean of the verbal

IQ was high in our sample compared to the related norm sample and standard deviations of the tests were low, which may be the reason for the lower reliability estimates (see table 1).

Table 1:
Means, standard deviations and reliability estimates for NEO-PI-R and I-S-T 2000 R variables

Variables	M	SD	α
I-S-T 2000 R			
Completing sentences	14.70	2.48	.60
Verbal analogies	14.11	2.49	.54
Similarities	12.76	2.65	.62
Verbal IQ	41.57	5.56	.77
NEO-PI-R			
Anxiety	17.78	6.50	.89
Angry Hostility	14.36	5.41	.20
Depression	13.95	6.34	.89
Self Consciousness	16.66	4.91	.76
Impulsiveness	17.31	4.98	.73
Vulnerability	13.44	4.35	.20
Neuroticism	93.51	24.66	.71
Fantasy	22.55	11.66	.09
Aesthetics	22.74	5.43	.80
Feelings	24.34	3.42	.63
Actions	17.55	4.58	.69
Ideas	21.29	4.09	.67
Values	21.59	2.97	.41
Openness	130.01	18.13	.58
Competence	22.99	4.81	.33
Order	20.12	4.78	.75
Dutifulness	22.75	4.23	.76
Achievement striving	21.13	4.08	.63
Self-discipline	19.55	5.44	.21
Deliberation	18.13	4.83	.78
Conscientiousness	124.66	19.65	.71

Note. α = Cronbach's alpha; domain-scales of the NEO-PI-R are italicised.

Test of personality: The German version of the psychological personality inventory NEO-PI-R (Ostendorf & Angleitner, 2004) was used in a group-testing to measure personality of participants in five major domains: Neuroticism, extraversion, openness to experience, agreeableness and conscientiousness, whereas for this paper we only used neuroticism, openness to experience and conscientiousness. Each domain scale is divided into six facets and each facet is operationalized by eight items.

For this study, the self-report form was used in which the participant has to provide self-report about typical behaviors or reactions on a five point Likert scale, ranging from 0 = 'strongly disagree' to 4 = 'strongly agree'. Validity for all questionnaires was shown by Ostendorf and Angleitner (2004). Except for some facets, Cronbach's alpha values indicate good reliability of the measurement (see table 1).

Statistical analyses. For all statistical analyses SPSS 15.0 for Windows and the program G*Power by Faul, Erdfelder, Lang and Buchner (2007) were used. First, *Benefit from TF* was calculated by dividing the number of correct benefit items by the total number of benefit items for each person. Examinees were then classified as either benefiters or non-benefiters, using the fixed criterion of ≥.75 for benefiters.

In the next step, the relation between *Benefit from TF* and the relevant abilities and skills was analyzed. Therefore, t-tests were conducted in which we compared the benefiters with *Benefit from TF* above .75 with the non-benefiters. Additionally, correlations between *Benefit from TF* and the most important predictor variables were computed. This was done in order to examine more deeply the relation between those variables. Still, one has to bear in mind, that the main focus of this paper was not the examination of a linear relation between *Benefit from TF* and the other variables since we believe that only those people differ from others that have a particularly high (i.e. above chance) probability of giving the right answer to a benefit item. Finally, intercorrelations of relevant predictor variables were computed and in cases where we assumed it to be reasonable, a covariance analysis was conducted. Alpha was set to be 5 % and furthermore Hedges g effect sizes were calculated for each t-test. With help of the G*Power program, we conducted a compromise power analysis. Since there was not sufficient evidence in previous research for a confirmatory data analysis, we decided to evaluate our hypothesis descriptively according to Abt (1987). We used p-values as descriptive inferential weights for the effect differences and drew conclusions derived from the recognition of patterns of descriptive significances associated with meaningful effect sizes.

Results

Partial knowledge. 8 students were assigned to the benefiter group (*Benefit from TF* >.75). Table 2 displays means and standard deviations of independent variables separately for the groups and the results of the t-tests. Between the two groups there were large effect sizes for grade and CR sum score (M = 19.48; SD = 7.30). Benefiters had better grades (lower values because of inversed coding) and higher CR sum scores. Correlations between the variable *Benefit from TF* and grade or CR sum score respectively

Table 2:
Means and standard deviations of predictor variables for each group and t-test

Variables	M	SD	t	df	g
Grade	2.76	1.28	2.38	104	1.02**
	1.66	.82			
CR sum score	18.84	7.11	3.33	104	1.41**
	27.38	4.81			
NEO-PI-R					
Anxiety	17.54	6.50	1.22	75	.58
	21.20	6.06			
Angry hostility	14.13	5.27	1.48	75	.59
	17.80	6.98			
Depression	13.54	6.12	2.19	75	.93*
	19.80	7.26			
Impulsiveness	17.56	4.98	1.65	75	.84
	13.80	3.90			
Fantasy	22.99	11.82	1.26	75	.70
	16.20	6.94			
Aesthetics	23.00	5.22	1.61	75	.61
	19.00	7.58			
Feelings	24.38	3.51	.36	75	.21
	23.80	1.79			
Ideas	21.49	4.02	1.65	75	.72
	18.40	4.51			
Values	21.65	3.05	3.53	45.07	.67***
	20.20	.45			
Openness	131.18	17.98	2.20	75	1.20*
	113.20	11.30			
Order	19.88	4.84	4.87	18.64	1.06***
	23.60	1.14			
Dutifulness	22.56	4.29	3.18	7.63	.92*
	25.60	1.82			
Achievement striving	20.97	4.15	1.29	75	.75
	23.40	1.95			
Self-discipline	19.22	5.47	5.73	18.24	1.25***
	24.20	1.30			
Deliberation	17.90	4.87	1.58	75	.86
	21.40	3.05			
Conscientiousness	123.53	19.79	1.96	75	1.20
	141.00	5.67			

Note. g= hedges g; upper line: M and SD for Benefit from TF≤.75, lower line: M and SD for Benefit from TF > .75; group sizes: for grade and CR sum score: n(≤ .75) = 98, n(> .75) = 8; for NEO-PI-R-variables: n(≤ .75) = 72, n(> .75) = 5; asterisks indicate a descriptive significance: *:α = .05, **: α = .01, ***: α = .001; domain-scales of the NEO-PI-R are italicised.

can be obtained from table 3. They both show substantial correlations above .20 (grade in a negative sense because of invers coding). Therefore the rejection of the null hypothesis of no real differences between benefiters and non-benefiters in partial knowledge seems to be supported.

In a univariate analysis of variance the difference between the two groups in grade was also found to be significant with partial eta² = .05. But when the domain scale conscientiousness was included as a covariate the effect diminished to .03 and did not reach statistical significance anymore. This effect was not found for the variable CR sum score.

Verbal intelligence. Table 1 displays descriptive statistics of verbal IQ subtests and sum score. The t-test yielded no significant difference between benefiters and non-benefiters in verbal IQ and none of the subscales. The effect sizes were all below .40. To get a clearer interpretation of this result, we subsequently computed correlations between verbal IQ and *Benefit from TF* as well as other exam variables: TF sum score, CR sum score and grade. According to the results of the t-test, there was no correlation with *Benefit from TF* (r = .07) while all other correlations were descriptively significant on a 1 %-alpha-level and small to moderate in size (TF sum score: r = .29, CR sum score: r = .33, grade: r = -.37).

Personality. Table 1 shows descriptive statistics of the NEO-PI-R domain-scales Neuroticism, Openness to Experience and Conscientiousness and their facets are listed. In the following, t-tests for each facet are presented separately. Table 2 displays means, standard deviations and results of the t-test of predictor variables. Variables are included which reached significance or yielded effect sizes of above .50.

Power for most tests was low because of the small sample size in this test. This explains insignificance for moderate or even large effect sizes.

Neuroticism: In previous research, negative correlation between anxiety and guessing behavior was found to be low. In this study, benefiters exhibited significantly higher

Table 3:
Correlations between important study variables

	Benefit from TF	Verbal IQ	Grade	TF sum score	CR sum score	Impulsiveness	Values	*Openness*	Order	Dutifulness	Self-discipline
Benefit from TF		.01	-.23*	.62***	.23*	-.09	-.07	-.19	.07	.03	.14
Verbal IQ			-.37**	.29**	.33**	.01	.25*	-.05	-.09	.03	-.17

Note. Asterisks indicate a descriptive significance: *:α = .05, **: α = .01, ***: α = .001; domain-scales of the NEO-PI-R are italicised.

values with a large effect size for the depression facet but not significantly higher values for the facets anxiety (1-ß = .34) and angry hostility (1-ß = .35), both of which had moderate effect sizes. Impulsiveness did yield a lower value in the benefiter group, which was not significant (1-ß = .56).

Openness to Experience: The domain scale openness to experience differed significantly between the two groups with a large effect size. Furthermore, the facet values showed a significant effect. For the ideas facet, the high effect size could only be supported on an alpha level of 10 % but not on 5 % (1-ß = .47), with benefiters showing lower values, while the high effect size for fantasy (1-ß = .45) and the moderate effect size for aesthetics (1-ß = .37) did not reach any level of significance. Correlations with verbal IQ revealed a negative relation to feelings (r = -.26*) and a positive relation to values (r = .25*) while the domain-scale did not correlate.

Conscientiousness: The t-test yielded significant differences with high effect sizes between the two groups for the domain-scale as well as on facet-level for order, dutifulness and self-discipline. Benefiters had higher values. Though the effect for deliberation was large, it was not significant (1-ß = .60). Here we did not conduct an analysis of covariance because variances in both groups were not homogenous.

The correlations between *Benefit of MC* and all personality variables were below .20 with p-values >.05. Some examples are displayed in table 3.

Discussion

The results of this study suggest that *Benefit from TF* is systematic. Benefiters differ from non-benefiters in partial knowledge and certain other personal abilities and traits which we analyzed in a descriptive way according to Abt (1987). Results will be discussed separately in the following for each hypothesis before discussing limitations of the study. A general conclusion will then be drawn.

Partial knowledge. A significant difference between the groups for both variables (grade and CR sum score) could be found. Based on these findings one can conclude that *Benefit from TF* is a systematic personal variable, which depends on partial knowledge of the subject matter.

However, different possible conclusions of these findings must be considered. In discussing the relation between grade and *Benefit from TF* one can also assume that this relation can be reversed, and that benefiting from MC has led to higher grades already in the last TF statistics exam. Still, this conclusion would only explain the relation between *Benefit from TF* and grade but not to CR sum score.

Verbal intelligence. The expected effect that verbal IQ should be the crucial intelligence factor to have an influence on *Benefit from TF* could not be supported by our results. Thus, we conclude that consistent with previous findings, neither general intelligence nor verbal IQ influence *Benefit from TF*. This is particularly interesting as there is good evidence for the positive influence of cognitive abilities on performance in educational settings in general (Laidra, Pullmann, & Allik, 2007; Ziegler, Knogler, & Bühner, 2009).

These results are supported by the correlation of verbal IQ we found with TF sum score and CR sum score. However, this influence does not appear to account for achievement on benefit items. The formerly found relation between verbal skills and testwiseness may also have been caused by items containing clues on which testwiseness principals apply. However, we tried to avoid these items and therefore did not find any relation.

Personality. Comparing the two groups we found significant differences in some of the facets. In the following, possible interpretations are presented separately for each major domain though, for definite conclusions to be drawn, these findings need to be replicated by further research. Interestingly, many facets had significant or large effect sizes although only five students, who had filled in the NEO-PI-R, were classified as benefiters, what leads to a low power.

No overall negative relation between the facets of neuroticism and *Benefit from TF* was found. For the facet anxiety even higher values for the benefiters were shown. For the facet impulsiveness, a lower value was found for the benefiters with a large effect size, which did not reach significance. This leads us to the conclusion that neuroticism does not affect *Benefit from TF* in a negative way in general.

For the domain-scale openness to experience, a significant difference was found with benefiters showing lower values. This is a surprising result as openness is known to be a positive predictor of performance rather than a negative one, although the impact is limited (for an overview see de Raad & Schouwenburg, 1996). Therefore, if these results can be replicated by further research, openness can be seen as a Big 5 variable, which negatively predicts *Benefit from TF* independently to general performance. In other words, people who are more open to experience have more difficulties in responding to MC items.

Conscientiousness is known to be a predictor of achievement (e.g. de Raad & Schouwenburg, 1996). In our study a large effect size for the difference between benefiters and non-benefiters was found. Also, several sub scales yielded significant differences or at least large effect sizes. Still, conscientiousness was neither related to TF sum score nor CR sum score, which means that the results noted above showing correlations between conscientiousness and general achievement could not be replicated. Additionally, the relation between grade and *Benefit from TF* was found to diminish when conscientiousness was introduced as a covariate.

Limitations and further research. The most critical limitation of this study is the sample of psychology students. However, since the main purpose of the study was to draw conclusions about academic MC or TF tests the sample was deemed to be sufficient. For generalization on performance in academic exams, subjects other than statistics should be analyzed. Also, the sample size of the benefiters group was particularly low. This leads to lower power of the t-tests. Furthermore, the nominal alpha level of the t-test was compromised by the heterogeneity of error variance and the extremely unequal n's.

Moreover, reliability estimates obtained for some of the Big 5 variables were suboptimal. Hence, considering that low reliabilities can lead to a decrease of t-values and therefore reduced power, more significant effect sizes would still be possible.

The most important part of this study was the definition of *Benefit from TF*, which was undertaken using a completely novel method of measurement. Please note that this measurement is restricted to a certain kind of guessing, which we defined as the difference in achievement between CR and MC tests. Therefore, for our purpose it worked out very well, although it is a quite involved method, as it requires very careful matching of CR and MC items and of course we can never be absolutely sure that the CR item exactly measures the same knowledge as the TF item.

Practical implications. One of the most important conclusions drawn in this paper is that TF items measure an additional variable and this variable seems to be related to knowledge. This means that also in the benefit some people get from MC items there is some information about one's knowledge or rather partial knowledge. Thus, the *Benefit from TF* in parts includes something MC exams intend to measure. It should be a desirable goal of tests to be able to measure this partial knowledge. Furthermore, considerations about the appropriateness of Multiple Choice tests seem to depend on what the examiner wants to measure. For example, does the examiner wish to favour conscientiousness over intelligence or vice versa? According to the findings of this article, intelligence seems to be only related to success on CR items while *Benefit from TF* is more related to conscientiousness.

References

Abt, K. (1987). Descriptive Data Analysis: A Concept between Confirmatory and Exporatory Data Analysis. *Methods of Information in Medicine, 26*(2), 77-88.

Albanese, M. A. (1988). The Projected Impact of the Correction for Guessing on Individual Scores. *Journal of Educational Measurement, 25*(2), 149-157.

Albanese, M. A., & Sabers, D. L. (1988). Multiple True-False Items: A Study of Interitem Correlations, Scoring Alternatives and Reliability Estimation. *Journal of Educational Measurement, 25*(2), 111-123.

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2007). *Intelligenz-Struktur-Test 2000 R - IST 2000-R (2. erw. Auflage)*. Göttingen: Hogrefe.

Angoff, W. H. (1989). Does Guessing Really Help? *Journal of Educational Measurement, 26*(4), 323-336.

Barratt, E. S. (1985). Impulsiveness subtraits: Arousal and information processing. In J. T. Spence & C. E. Itard (Eds.), *Motivation, emotion and personality*. Amsterdam: Elsevier.

Barratt, E. S., & White, R. (1968, November). *Impulsiveness and Anxiety Related to Medical Students' Performance and Attitudes*. Paper presented at the 79th Annual Meeting of the Association of American Medical Colleges.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies. *Journal of Educational Measurement, 28*(1), 23-35.

Bliss, L. B. (1980). A Test of Lord's Assumption regarding Examinee Guessing Behavior on Multiple-Choice Tests using Elementary School Students. *Journal of Educational Measurement, 17*(2), 147-152.

Boldt, R. F. (1968). Study of Linearity and Homoscedasticity of Test Scores in the Chance Range. *Educational and Psychological Measurement, 28*(1), 47-60.

Burton, R. F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education, 36*(9), 805-811.

Busato, V.V., Prins, F.J., Elshout, J.J., Hamaker, C. (1999). The relation between learning styles, the Big Five personality traits and achievement motivation in higher education. *Personality and Individual Differences*. 26 (1), 129-140.

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *The Journal of Educational Psychology, 33*(6), 401-415.

Cross, L. H., & Frary, R. B. (1977). An Empirical Test of Lord's Theoretical Results regarding Formula Scoring of Multiple-Choice Tests. *Journal of Educational Measurement, 14*(4), 313-321.

de Raad, B. d., & Schouwenburg, H. C. (1996). Personality in learning and education: a review. *European Journal of Personality, 10*(5), 303-336.

Diamond, J. J., & Evans, W. J. (1972). An Investigation of the cognitive Correlates of Test-wiseness. *Journal of Educational Measurement, 9*(2), 145-150.

Dunn, T. F., & Goldstein, L. G. (1959). Test Difficulty, Validity and Reliability as Functions of Selected Multiple-Choice Item Construction Principles. *Educational and Psychological Measurement, 19*(2), 171-179.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement (5th ed.)*. New Jersey: Prentice-Hall.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G Power 3: A flexible statistical power analysis program for the social, behavioural, and biomedical sciences. *Behavioural Research Methods, 39*(2), 175-191.

Frisbie, D. A., & Sweeney, D. C. (1982). The Relative Merits of Multiple True-False Achievement Tests. *Journal of Educational Measurement, 19*(1), 29-35.

Furnham, A. (1996). The FIRO-B, the Learning Style Questionnaire, and the Five-Factor Model. *Journal of Social Behavior and Personality, 11*(2), 285-299.

Grosse, M. E., & Wright, B. D. (1985). Validity and Reliability of True-False Tests. *Educational and Psychological Measurement, 45*(1), 1-13.

Hassmén, P., & Hunt, D. P. (1994). Human Self-Assessment in Multiple-Choice Testing. *Journal of Educational Measurement, 31*(2), 149-160.

Kennedy, P., & Walstad, W. B. (1997). Combining Multiple-Choice and Constructed-Response Test Scores: An Economist's View. *Applied Measurement in Education, 10*(4), 359-375.

Kubinger, K.D. (2014). Gutachten zur Erstellung "gerichtsfester" Mulitple-Choice-Prüfungsaufgaben. Psychologische Rundschau, 65, 169-178.

Kuechler, W. L., & Simkin, M. G. (2010). Why Is Performance of Multiple Choice Tests and Constructed Response Tests Not More Closely Related? Theory and an Empirical Test. *Decision Sciences Journal of Innovative Education, 8*(1), 55-73.

Laidra, K., Pullmann, H., & Allik, J. (2007). Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. *Personality and Individual Differences, 42*(3), 441-451.

Lord, F. M. (1975). Formula Scoring and Number-Right Scoring. *Journal of Educational Measurement, 12*(1), 7-11.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An Analysis of Test-Wiseness. *Educational and Psychological Measurement, 25*(3), 707-726.

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung*. Göttingen: Hogrefe.

Rowley, G. L. (1974). Which Examinees are most favoured by the use of Multiple Choice Tests? *Journal of Educational Measurement, 11*(1), 15-23.

Sarnacki, R. E. (1979). An Examination of Testwiseness in the Cognitive Test Domain. *Review of Educational Research, 49*(2), 252-279.

Sherriffs, A. C., & Boomer, D. S. (1954). Who is penalized by the penalty for guessing? *Journal of Educational Psychology, 45*(2), 81-90

Swineford, F., & Miller, P. M. (1953). Effects of Directions regarding Guessing on Items Statistics of a Multiple-Choice Vocabulary Test. *Journal of Educational Psychology, 44*(3), 129-139.

Votaw, D. F. (1936). The effect of do-not-guess-directions upon the validity of true-false or multiple choice tests. *Journal of Educational Psychology, 27*(9), 698-703.

Ziegler, M., Danay, E., Schölmerich, F., & Bühner, M. (2010). Predicting Academic Success with the Big 5 Rated from Different Points of View: Self-Rated, Other Rated and Faked. *European Journal of Personality, 24*(4), 341-355.

Ziegler, M., Knogler, M., & Bühner, M. (2009). Conscientiousness, achievement striving, and intelligence as performance predictors in a sample of German psychology students: Always a linear relationship? *Learning and Individual Differences, 19*(2), 288-292.