universität wien

Diplomarbeit

Titel der Diplomarbeit

Multiple-Evaluation: Do personality variables influence students' response behavior?

Verfasserin

Kathrin Viola Stathis

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, 2014

Studienkennzahl: A 298

Studienrichtung: Psychologie

Betreut von: Dr.in Christine Hohensinn

Abstract

Due to inconsistent findings regarding the influence of personality variables on response behavior in a multiple-evaluation test and the notion Holmes introduces in 2002, to use such a testing system for certifying tests, it is found to be of considerable value to scrutinize these effects further. It is specifically thought to be of importance to take the unique circumstances such a test situation poses into account as well. Therefore active university students were chosen as a sample. Further, personality variables linked to achievement were utilized, in particular: achievement motivation (assessing hope and fear components), aspiration level, reflexivity and impulsivity, and decisiveness. In contrast to past research it was decided to employ personality inventories posing special tasks, from which inferences can be drawn in regard to the dimensions of interest, opposed to eliciting self-reports. Moreover, a possible influence due to the course of study was explored. Effects were measured in regard to “response certainty” and “realism”, constituting the dependent variables. Regression analyses were utilized showing that students’ response certainty was not influenced by personality variables. In regard to realism results indicate that ambitious students tend to be more realistic in their probability estimates or put differently are less prone to overestimate themselves. A slight trend was discernible, signifying that anxious students are also more likely to be realistic.

Abstract

Aufgrund uneinheitlicher Befunde bezüglich eines möglichen Einflusses von Persönlichkeitsvariablen auf das Antwortverhalten in einem Multiple-Evaluation-Test und der Idee, ein solches Antwortformat für Prüfungen zu verwenden, wie es Holmes 2002 vorschlug, erscheint es von großem Wert diese Einflüsse näher zu untersuchen. Im Vordergrund steht hierbei die Berücksichtigung der Besonderheiten, die Testsituationen mit sich bringen. Aus diesem Grund wurden aktive UniversitätsstudentInnen rekrutiert und Persönlichkeitsvariablen, die mit Leistung in Verbindung stehen für den Untersuchungsrahmen gewählt. Konkret handelt es sich um die Variablen: Leistungsmotivation (wobei sowohl die Hoffnungs- als auch die Furchtkomponente erfasst wird), Anspruchsniveau, Reflexivität und Impulsivität, sowie Entschlussfreudigkeit. Im Gegensatz zu früheren Arbeiten, wurden Persönlichkeitsverfahren gewählt, die besondere Aufgaben an die TeilnehmerInnen stellen, wodurch Rückschlüsse auf die interessierenden Dimensionen möglich sind, anstatt Selbsteinschätzungen heranzuziehen. Darüber hinaus wurde ein möglicher Einfluss der Studienrichtung exploriert. Es wurden Regressionsanalysen zur Hypothesenprüfung gewählt, wobei „Antwortsicherheit“ und „Realismus“ die abhängigen Variablen darstellten. Die Ergebnisse zeigen, dass Persönlichkeitsvariablen keinen Einfluss auf die Antwortsicherheit nahmen. In Bezug auf Realismus zeigte sich, dass StudentInnen mit hohem Anspruchsniveau ihr Wissen realistischer einschätzten. Auch konnte ein leichter Trend gefunden werden, der anzeigt, dass ängstliche StudentInnen ebenso dazu neigten ihr Wissen realistisch einzuschätzen bzw. sich weniger überschätzten.

Contents

I. Introduction

In the past years an increased research interest has been observed in regard to the response format multiple-evaluation. Preceding studies have mostly been conducted in comparison to the multiple-choice format (Jenner, 2012, Holmes, 2002). The dominance of the multiple-choice format in achievement tests, whether it is in a university context or in terms of a psychological assessment of ones abilities, is great but also questionable. On the one hand this is due to its properties, which can encourage people to guess, leading to a decrease in measurement accuracy and therefore making a larger number of items necessary. But on the other hand, it also dismisses the possibility to obtain valuable information on a person's present state of knowledge.

Moreover, multiple-choice is unconvincing in regard to the quality criteria *fairness* and *appropriateness* as described by Kubinger (2009) and this will be discussed later.

Abundant research has been performed in recent years, addressing the issue on how to mathematically eliminate or reduce the influence of guessing in achievement tests, utilizing the multiple-choice format (see Kubinger, 2009). Yet, multiple-evaluation might be an interesting alternative to cope with this problem. Literature has shown that multiple-evaluation has not only rendered guessing ineffective, motivating people to reflect their knowledge honestly and realistically, but also leads to an improvement of the quality criteria *reliability* and *validity* (Jenner, 2012 and Holmes, 2002 offer compact summaries of corresponding results).

However, some authors suppose that personality traits could have an impact on response behavior and argue that this would lead to a decrease of the validity of a test applying the response format multiple-evaluation. Yet, findings regarding the influence of personality traits are equivocal, some research showing effects, other findings supporting the claim that these effects disappear after practice (see Holmes, 2002 for an overview, p. 27). Some of the personality traits that have been connected to response behavior are risk taking and competence beliefs, and in relation to the often-found tendency towards overestimation of ones knowledge, extraversion seems noteworthy.

Due to the introduction of the notion to use the response format multiple-evaluation for certifying tests by Holmes (2002), it seems of value to further scrutinize a possible influence of personality traits, especially variables linked to achievement, on response behavior, which is precisely the intention of this study. Opposed to implementing questionnaires, eliciting self-reports, it was decided to employ personality inventories, which produce more objective results, as tasks are constructed in such a way that it is highly improbable for the participant to discover the underlying measurement intention. For the purpose of the presented study university students were recruited, as it was thought that this would contribute to realistic results, as it is likely that students will be the primary population of interest. Effects on response behavior are measured utilizing Hansen's (1971) mean item certainty and a measure of realism postulated by Dirkzwager (2003).

The theoretical part will illustrate the properties of the multiple-evaluation format as well as its scope. In addition, measures how to assess the quality of multiple-evaluation in regard to its supposed propensity to evoke responses that are influenced by personality variables are elaborated. Further, it will discuss previous findings on personality traits and their supposed influence on response behavior, specifically response confidence.

In the empirical part of this work the research aim and hypotheses addressing various personality traits are described in detail. Furthermore, a comprehensive illustration on how the study was conceptualized and realized is given. This is followed by a detailed report of results and their discussion, also in regard to future research.

II. Theoretical Part

The theoretical part is organized into two parts. The first half focusing on describing the properties of multiple-evaluation, differences to multiple-choice, benefits of multiple-evaluation as well as theoretical framework, such as methods to derive an item score. Besides, measures are presented, with which the effects of personality on response behavior can be assessed. The second half gives a comprehensive overview of past research examining the influences of personality variables on response behavior in confidence-based testing.

1. The response format multiple-evaluation

Holmes (2002) introduced the idea of using the response format multiple-evaluation for certifying tests; tests that determine whether a student passes or fails a course (p. 40). This notion was developed owing to the fact that multiple-evaluation tests seem to offer more reliable (accurate) results than multiple-choice tests.

In one way the multiple-choice and multiple-evaluation response format are basically identical. Both elicit answers to a posed question by offering the test taker several answer options. However, when taking a multiple-choice test one has to choose the answer option(s) one believe(s) to be correct. In contrast, multiple-evaluation requires the testee to evaluate every single answer option and state numerically, how probable one thinks the answer option to be correct. In the case of multiple-choice tests only the answer that was marked can be taken into account, signifying if the given answer was correct or wrong. Depending on this the testee either receives full points for the item or none (in some cases penalty points). However, when the test taker chooses an answer that is wrong, the testees' ideas and beliefs regarding the correct answer remain unknown (Holmes, 2002). Yet, as Holmes stated in 2002 test takers often find themselves in the predicament, that several answer options seem more/less/equally likely. Here multiple-evaluation offers an interesting approach. Due to the fact that an answer is elicited in the form of percentage points for every single answer option of an item, the testee has the possibility to give a highly differentiated response to an item and one acquires a holistic picture of the test takers momentary knowledge.

1.1. Probability measurement – the foundation of multiple-evaluation

In conventional multiple-choice testing no information is obtained, as to how a person chose the answer option s/he selected. A given response could be based on absolute certainty, or maybe several answer options were considered possible, perhaps a lucky guess was tried, or misconceptions regarding the subject matter lead to the response given (Holmes, 2002). Confidence weighting testing methods were one of the first attempts to make testing more precise and incorporated a second scale, where testees could further indicate their certainty in the answer option they chose (Dressel & Schmid, 1953, Ebel, 1965, Echternacht, 1972). However, these testing methods yield no valuable information when a wrong answer is selected, as no information on the testees' thoughts regarding the correct answer is obtained (Holmes, 2002).

In 1965 de Finetti suggested the use of subjective probabilities. As in multiple-choice testing, the testee is presented a number of answer options for every item. However, in contrast to multiple-choice the testee does not select the answer option s/he considers mostly likely, but rather evaluates every single answer option and indicates how (un)likely s/he believes each answer option to be. Hence, the testee distributes personal probabilities among all answer options, where estimates lie between 0 and 1 and summation over all answer options of an item equals 1.

In regard to scoring a linear scoring rule was often proposed (Rippey, 1970), where the probability assigned to the correct answer option constitutes the item score. Dirkzwager (2003) stated critically that this would take all incentive from honestly reflecting ones knowledge, since the expected score would reach a maximum when the testee always allocates 100 percent to the answer option considered most probable of being correct. Shuford, Albert and Massengill (1966) shared the opinion that the obtained item score should solely be based on the probability estimate assigned to the correct answer option, however, showed that only a logarithmic scoring function would pose a proper scoring rule. This scoring rule was then further refined by Dirkzwager (2003) and is described in detail in the following chapter.

1.2. Logarithmic scoring rule

In 1966 Shuford et al. introduced a scoring system that "… guarantees that any student at whatever level of knowledge or skill, can maximize his expected score if and only if he honestly reflects his degree-of-belief probabilities" (p. 1). When employing reproducing scoring systems, as discussed in Shuford et al. (1966), the score of an item is obtained by determining how the probability estimates were assigned among all answer options. To be specific, when all distractors seem equally unlikely the maximum item score is obtained, however, in doing so this rule minimizes the item score when the remaining probability (1 – r_c; r_c indicating the response given to the correct answer) is assigned to a single other answer option. It follows, that when an item has three or more answer options, testees receive lower scores when they are able to rule out answer options with absolute certainty (Holmes, 2002).

Shuford et al. (1966) further prove in their paper that the only scoring system, which assures that the score depends solely on the probability estimate assigned to the correct answer, is the logarithmic scoring system. Yet this has a great problem, it is unbounded, meaning if a person assigns a probability of zero to the correct answer option, s/he will receive an item score of minus infinity, therefore making this system unfeasible in practice.

In 2003 Dirkzwager proposed a solution to combat this issue by implementing a tolerance parameter, limiting the number of penalty points. Considering this, the score of an item (i) can be obtained with the following formula, t indicating the selected tolerance parameter, k signifying the number of answer options and $r_c(i)$ the probability estimate assigned to the correct answer option of the corresponding item (Dirzkzwager, 2003, p. 337):

$$s(r_c(i)) = \frac{\ln[(1 - tk) * r_c(i) + t] + \ln(k)}{\ln(1 - tk + t) + \ln(k)} \qquad (1)$$

If a tolerance parameter (t) of zero is selected and a person gives a maximally incorrect answer (zero percent assigned to the correct answer option) this results in a score of minus infinity, this rule is clearly too severe (Dirkzwager, 2003). A tolerance parameter of 1/k on the other hand is too tolerant. Depending on the desired severity the tolerance parameter, limiting the maximum number of penalty points to any intended value, is selected. For example, if we want to limit the score of a maximally incorrect answer to minus one we calculate $t = 1/(k*(k-1)$, for $k = 5$ this results in $t = .05$ (Holmes, 2002, p. 43). So, if a person assigns a probability of 100 percent to the correct answer option s/he will receive one point and if a person assigns a probability of zero percent to the correct answer option s/he will receive the maximum number of penalty points (T), depending on the tolerance parameter (t) being used. The choice of the tolerance parameter also determines how many items must be answered correctly and with absolute certainty, in order to compensate for one maximally incorrect item (Holmes, 2002, Dirkzwager, 2003). As a result, the tolerance parameter can be seen as a means to deal with misconceptions and wild guessing, as it limits the amount of penalty points when a person considers the correct answer option to be highly unlikely. Below a table from Holmes (2002, p.44) is reproduced, listing corresponding values of the tolerance parameter for several practical values of T and k, which were derived by numeric approximation.

TABLE 1
Tolerance parameters, t(T,k)

	k=2	**k=3**	**k=4**	**k=5**
ME1 (T=1)	0.4999	0.1667	0.0833	0.0500
ME2 (T=2)	0.1910	0.0447	0.0174	0.0086
ME3 (T=3)	0.0804	0.0134	0.0041	0.00016
ME4 (T=4)	0.0362	0.0043	0.0010	0.0003

For the purpose of this study a tolerance parameter of 0.05 (T=1, k=5) was chosen. This decision is based on Holmes (2002) assumption that a tolerance parameter of 0.0362 is sufficient for two-choice items. Further, as stated above, a tolerance parameter of 1/k would be too lax, for $k = 5$, as used in this work, this would result in $t = 0.2$, this also leads to the conclusion that $t = 0.05$ should suffice for the purpose of this study. Table 2 compares the scores that would be obtained by implementing a conventional linear scoring rule and scores that result from a logarithmic scoring rule (T=1, k=5), as intended when utilizing multiple-evaluation.

TABLE 2

Comparison of scores depending on the scoring rule (Jenner, 2012, p. 18)

% assigned to correct answer	Linear scoring rule	ME1
0 %	0	-1
10 %	0.1	-0.34
20 %	0.2	0
50 %	0.5	0.54
80 %	0.8	0.85
100 %	1	1

In addition Holmes (2002) has suggested correcting the score for lack of realism in certifying tests, since the results have a great effect on a students' progress, and students' success should primarily be based on their knowledge and not their ability to give perfectly accurate estimates of their knowledge, however, this correction only applies to overestimation.

1.3. The scope of multiple-evaluation

This chapter intends to give an insight to the possible benefits when utilizing multiple-evaluation for testing.

1.3.1. Quality criteria „reliability" and „validity"

Preceding studies have shown an increase in reliability and validity (Holmes, 2002, see p. 24 for a tabular overview of conducted research). However, Schaefer (1976) and others have noted skepticism regarding the results to validity assuming influences of personality, calling for further research regarding this aspect (chapter 3 discusses previous findings on this topic).

1.3.2. Guessing

It seems logical that guessing is rather undesirable in "testing", regardless of the context. Several strategies have been developed to cope with guessing effects when utilizing the multiple-choice format. Means to deal with this issue try to solve this matter from different angles, some focus on the scoring system, employing penalty points (Espinosa & Gardeazabal, 2010), others on the response format itself, increasing the number of distractors but also the numbers of distractors in relation to the number of correct answers options, therefore reducing the a priori guessing probability (Kubinger, Holocher-Ertl, Reif, Hohensinn, & Frebort, 2011). Further approaches include the 3-PL and the Difficulty plus Guessing PL model (see Kubinger, 2009, p. 130), which were developed from item response theory. However, these models are not able to acknowledge (un-)fortunate guessing, therefore these effects are obviously still included in the test score an individual obtains.

Though inherent to the response format multiple-choice to provoke guessing, when one is unsure of the correct answer, the properties of multiple-evaluation render guessing ineffective. A scoring rule introduced by Shuford et al. (1966) guarantees that a testee can only maximize her/his score when realistic probability estimates are made, details will follow in the elaboration of the derivation of a proper scoring system (see chapter 1.2.). Owing to this fact, testees are motivated to give accurate reflections of their knowledge, rather than just picking the most likely (in the case of an educated guess) or in fact any answer option (when no useful information is available).

1.3.3. Quality criteria „appropriateness" and „fairness"

As mentioned above the multiple-choice format is often criticized because it can lead some test takers to take a lucky guess. However, this puts those to a disadvantage, who principally refuse to guess (Kubinger, 2009), hence multiple-choice lacks in fulfilling the requirement of fairness. Further multiple-choice does not meet the specifications of the quality criterion appropriateness, since it can have a negative impact on a person's motivation and emotional state. This is due to the fact, that it is obvious that a person can obtain a score that is higher than would be expected, given that persons knowledge, merely because of the fact that this person was lucky at guessing. From the perspective of a person with a higher ability level, the notion that someone who is less capable could obtain the same or an even higher score merely because of guessing could be very frustrating and unnerving.

Yet, as mentioned above multiple-evaluation makes guessing unattractive, hence meets the standards as defined by the quality criteria "appropriateness" and "fairness".

1.3.4. Partial knowledge

Multiple-choice seems to overlook the fact that, on the one hand, there seem to be several levels of knowledge (Echternacht, Boldt, & Sellman, 1972), and on the other hand, problems can be encountered with other strategies, rather than merely distinguishing between right and wrong (Schaefer, 1976). Multiple-evaluation addresses this by giving testees the possibility to directly specify how much they know about the subject matter, making it possible to display varying degrees of knowledge, but also making it feasible to acquire information to complex matters by utilizing a probabilistic approach.

In reference to Mondak (2001) the extent of a persons' knowledge can be grouped into four categories:

1. Completely informed
2. Uninformed
3. Fallacy
4. Partially informed.

If an individual is completely informed Mondak (2001) claims that this results in a decision under perfect certainty. When one is partially informed this could perhaps lead to an educated guess as described by Nadeau and Niemi (1995). Under fallacy the testee picks a wrong answer and does this with subjective certainty (Mondak, 2001). Finally, a person could be induced to try a lucky guess, when s/he is uninformed in the respective subject matter.

Schaefer (1976) further distinguishes between two levels of partial information, on the one hand, if a person is able to rule out certain distractors, and on the other hand, when a person has a given preference as to the order of the likeliness of the available answer options, yet without being able to eliminate an answer option with certainty.

When an examination is realized in the multiple-evaluation format, a completely informed individual should answer with perfect certainty; that is s/he should allocate 100 percentage points to the correct answer option. However, if persons decisions are influenced by personality variables in the face of such a response format, it could be that someone who possesses the necessary knowledge and should therefore be absolutely certain, allocates too few percentage points and receives a lower score than would be expected given his/her knowledge. If a student is not fully knowledgeable in the matter of interest, but is able to eliminate one or more distractors with certainty, s/he has the possibility to indicate this under the multiple-evaluation format. Also, if the testee is not able to eliminate a false answer option, s/he can utilize the percentage points to communicate subjective preferences in respect to the likeliness of the presented answer options. When an individual is uninformed, s/he is likely to try a lucky guess under the multiple-choice format, whereas when facing an examination realized in the multiple-evaluation format, the testee has the possibility to specify this, by equally distributing the percentage points among the available answer options. Further, it is unwise to attempt a lucky guess, because one would receive the maximum amount of penalty points, if 100% were to be assigned to an incorrect answer option. Finally, if a person is misinformed and mistakenly believes an incorrect answer to be correct with absolute certainty, s/he will also receive the maximum amount of penalty points under multiple-evaluation just as in multiple-choice.

1.3.5. Feedback for educational staff

The unique way of eliciting responses through probability estimates further creates the opportunity to obtain information on how well material was understood by the students. In reference to Walker and Thompson (2001) students generally tend to be reluctant to admit to what they (do not) know or have not understood fully yet. Results obtained through multiple-evaluation testing could therefore provide valuable information for educational staff, shedding light on which aspects need to receive more attention in class. This potential benefit opposed to multiple-choice was also noted by Schaefer in 1976.

2. Measures to evaluate the response format multiple-evaluation

2.1. Hansen's mean item certainty

In 1971 Hansen introduced a measure called mean item certainty, C_T, which is assessed independent of ones knowledge and therefore a genuine indicator of response certainty. It can be evaluated for a test with any number of answer options (n) per item and over all items (N) and takes the probability assignments of all answer options into account, reflecting the tendency to be confident in ones response. It is calculated as follows (Hansen, 1971, p. 10):

$$C_T = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{n_i}{2\,(n_j - 1)} \right) \sum_{i=1}^{N} \left| \frac{1}{n_j} - P_{ij} \right| \qquad (2)$$

Given an item with n answer options, the mean item certainty is the sum of the absolute deviations from a 1/n uniform distribution. C_T is a maximum (=1) when a testee assigns the full 100 percent to one of the answer options, irrespective if the answer is actually correct and it is a minimum (=0) when a testee distributes the 100 percent equally among all answer options.

This measure seems of use for the research question, to investigate differing levels of confidence in individuals. Because the test score in a multiple-evaluation test is based on the response given to the correct answer option, this leads to the question of fairness, if two testees have the same knowledge but display varying degrees of certainty. Then again, one could also argue that response confidence is also an integral part of knowledge, which might justify the difference in item score.

2.2. Realism

The realism measure was postulated by Dirkzwager (2003) and implemented by Holmes in his research 2002 as a basis for feedback provided to the students. When examining a question and its answer options, students tend to select the option that seems most likely, which is an effective test-taking strategy when faced with the multiple-choice format. Since students are used to this testing method, some have difficulties assigning realistic probabilities and therefore tend to transmit their usual test-taking method (pick the most likely answer option) to multiple-evaluation, leading to the assignment of unrealistically high probability estimates (Holmes, 2002). Students are confronted with a very new task, when asked to give estimations of their knowledge. Considering the cases that a question is fairly easy for a student and the situation that one definitely does not know the correct answer, it should be rather simple to give realistic probability estimates (100% when one is sure and 1/k when one does not know the answer). However, when taking a knowledge test, one is presumably confronted with the situation that several answer options might seem more or less likely to be true and this is where it becomes trying. This present state of uncertainty must be transformed into an adequate numerical response, which accurately represents this feeling of uncertainty (Holmes, 2002). Basically, students have to learn to give responses that correspond to their personal probabilities and this can be assessed with the realism formula postulated by Dirkzwager (2003, p. 342, see also Holmes, 2002, p. 51).

$$A = \frac{\sum r_c - m/k - m/k + m*k/k^2}{\sum\sum r^2 - 2m/k + m*k/k^2} = \frac{k * \sum_{i=1}^{m} r_c(i) - m}{k * \sum_{i=1}^{m}\sum_{j=1}^{k} r(i,j)^2 - m} \qquad (3)$$

As Dirkzwager (2003) and Holmes (2002) have stated, this estimate gives the best least squares fit (m indicating the number of items and k the number of answer options) when the true probabilities are estimated with the linear formula:

$p = A*r + (1-A)/k$.

Values of this measure of realism can be interpreted as follows:

If A < 1 respondents are too confident, given their knowledge of the matter, and are assigning probabilities that are too extreme.

If A > 1 this signifies responses that are too modest, corresponding to underconfidence.

Finally, if A = 1 this indicates perfectly realistic responses.

3. Personality variables and multiple-evaluation

The notion that confident students, although this does not necessarily entail they are more knowledgeable, would display a greater response certainty opposed to students who are less confident (again, this does not have to mean they are less knowledgeable) and that such students with a lesser confident attitude would be more hesitant in their probability estimates, seems rather apparent. When thinking of the characteristics of the response format multiple-evaluation, this could be very critical, since students are required to give realistic estimates of their knowledge. Especially in the academic field, where certifying tests determine whether one passes or fails, the results of these tests should particularly reflect the actual knowledge of the testee, rather than personality traits (Holmes, 2002). Already Hansen (1971) emphasized that the responses elicited in such tests, should primarily be determined by what a student knows, otherwise it logically follows that the resulting score would be a less valid indicator of an individual's knowledge.

So far, conducted studies scrutinizing the possible effects of personality on response behavior in confidence testing sought relationships between personality variables and score (Echternacht et al. 1972), response confidence (Hansen, 1971, Koehler, 1974, Stankov & Crawford, 1997, Kleitman & Stankov, 2007, Jenner, 2012), and overconfidence (Schaefer, Williams, Goodie, & Campbell, 2004).

In 1972 Echternacht et al. explored the relationship between personality variables and the resulting score in a test in the multiple-evaluation format. Two groups of Air Force students completed a technical test, of which one group received the test in the multiple-evaluation format, the other group received the same test in the pick-one format and in addition rated their confidence in that choice. Both groups were asked to fill out several personality questionnaires. Measured variables included scales to: dogmatism, anxiety, rigidity, impulsivity, self-sufficiency, risk taking, and the F-Scale. Following a practice period where students were able to get better acquainted with the new testing format, students were tested again. For both testing periods and both testing formats no consistent relationships were found between the measured personality variables and the resulting score. It is also noteworthy that the questions were fairly easy; therefore confidence ratings below maximum were hardly used.

Hansen (1971) sought the relationship between response confidence and personality variables. Based on a finding reported by Massengill and Shuford (as cited in Hansen, 1971), that students must be accustomed to the confidence testing scheme for it to be effective, the author included a two-week "training" prior to testing. Testing was realized with a group of psychology students, once at midterm and then again at the final examination. After the midterm examination students received detailed feedback. Furthermore, the personality questionnaires were handed out after the Christmas vacation, rather than being distributed immediately after testing. Response confidence was operationalized with the mean item certainty measure the author developed. The personality variables explored in the study concerned risk taking, test anxiety, rigidity, and the F-Scale. Results showed that individuals, who displayed a preference for risky options, also received mean item certainty values that were higher than would be expected given their knowledge.

Similarly, Koehler (1974) explored the relationship between a persons mean item certainty (as postulated by Hansen, 1971) and disposition to risk taking. In addition to the usual items of a vocabulary test, the administered test comprised seven nonsense items. Koehler (1974) postulated an overconfidence measure on the basis of the derived mean item certainty values testees received on these nonsense items. He found a moderate relationship between this overconfidence measure and a persons risk preference. The author mentions critically that both risk preference and the overconfidence measure resulted from the responses made to the nonsense items and questions if the findings are genuine.

Stankov and Crawford (1997) examined the influence of academic self-concept on confidence ratings. For this purpose they administered a vocabulary test, Raven's Progressive Matrices (RPM) and accordingly an English self-concept measure and a Maths self-concept measure to 271 psychology students. Results showed significant relationships between vocabulary confidence ratings and English self-concept scores. Analogous results were produced for RPM confidence ratings and Maths self-concept scores. However, when controlling for the percentage of correct answers, by computing partial correlations, the correlations were reduced (.17 and .10) and were no longer significant. In this study Stankov and Crawford (1997) were also able to replicate findings regarding persons tendency towards overestimation of ones accuracy in achievement tests, yet could not attribute this tendency to individual differences.

Subsequent research showed noteworthy relationships between self-confidence and competence beliefs (Kleitman & Stankov, 2007). Self-confidence was measured during performance on cognitive tests by obtaining ratings of the participants' response certainty. Confidence ratings were evoked for four cognitive tests: a verbal reasoning task, nonsense syllogisms, esoteric analogies and general knowledge. 296 first-year psychology students completed the aforementioned tests. A Big Five measure was utilized to examine the influence of personality variables. Hierarchical regression analyses showed that beliefs of competence in reasoning contributed significantly to predicting self-confidence (i.e. confidence ratings) next to cognitive ability (i.e. performance) and awareness of metacognitive processes. Interestingly, none of the Big Five dimensions had a significant impact on predicting self-confidence.

In 2012 Jenner investigated the influence of personality variables on response behavior in a multiple-evaluation verbal intelligence test. For this purpose 10-12th grade students (the age ranged from 15 to 19 years) participated as subjects. As a measure of response certainty Hansen's postulated mean item certainty was utilized. Further, the magnitude of a participants' profit from a logarithmic scoring rule opposed to a linear scoring role was calculated und utilized for analyses. Measured personality variables comprised the dimensions: neuroticism, extraversion, openness, control, altruism and risk taking. However, neither Hansen's mean item certainty, nor profit were shown to have a noteworthy relation with the assessed personality variables.

Due to findings showing a tendency towards overconfidence in assessment of performance, Schaefer et al. (2004) decided to investigate the underlying causes, why certain people are more overconfident than others. They did this by examining the connection between overconfidence and the Big Five. A positive difference between confidence and accuracy signifies overconfidence, suggesting that people are more confident than they are accurate. 104 undergraduates taking Psychology courses participated in exchange for course credit and completed a general knowledge test. This test was made up of questions with two answer options and upon choosing required a rating, regarding ones confidence in that answer (seven categories were available for selection). Subsequently, participants completed a personality inventory measuring the Big Five. Extraversion, agreeableness and conscientiousness were to be found significantly positively correlated to overconfidence. However, when controlling for accuracy, and confidence only extraversion remained significantly correlated to overconfidence.

Overall it can be noted that former research either implemented common confidence weighting testing methods, which require individuals to indicate their certainty in a given response on a separate scale (Echternacht, 1972, Stankov & Crawford, 1997, Kleitman & Stankov, 2007, Schaefer et. al., 2004), or a probabilistic approach, eliciting subjective probability estimates, as in multiple-evaluation (Echternacht, 1972, Hansen, 1971, Koehler, 1974, Jenner, 2012). Further, most authors have focused on influences of personality on a persons (response) confidence. This is absolutely understandable, as it would be highly critical if such influences were to exist. Although there has been much effort to uncover such effects, findings have not been conclusive. Previous studies have explored the effects of general personality variables (such as the Big Five), risk taking and self-concept. However, it seems to have been neglected so far, that academic testing constitutes a very particular situation, in which other personal dispositions could be of great(er) relevance. When thinking of a test situation in a university context, it can be supposed that aspects, such as achievement motivation, test anxiety, the importance of the test, personal aspirations, etc., play a significant role in how a student approaches a test. Further, past research has employed questionnaires, in order to examine an effect of personality variables. Such an approach is precarious, as self-reports might contain an undue influence, as individuals could purposely distort results, because the underlying dimensions are mostly very transparent. On the other hand, persons, who find it difficult to accurately evaluate themselves, are prone to give unintentional misrepresentations. It would therefore be of significance to examine possible influences of personality, utilizing more objective measures, such as objective personality tests, which was realized in this study. Besides focusing on response certainty it seems relevant to scrutinize, whether there is a clear tendency towards overestimation, but also personal dispositions causing such an inclination must be considered. To examine this, Dirkzwagers' (2003) formula to assess how realistic the assigned probability estimates were, presents a suitable measure to evaluate this matter.

4. Response behavior under multiple-evaluation

A test situation poses a condition in which an individual most definitely experiences uncertainty. Aspects, such as developed and habituated test-taking strategies, risk preference, feeling of competence (influencing confidence), but also perhaps the course of study a student has chosen could have an impact on students' response behavior when facing a test situation.

Test-taking strategies

It is apparent that numerous test-taking strategies exist, some students develop their own strategies throughout their academic career and on the other side, books covering this topic circulate as well, not to mention all the websites that turn up when entering such search terms. However, test-taking strategies are not first activated in the test situation itself, in fact such strategies are already active during learning, influencing how a student approaches his/her learning material. In this context Scouller (1998) found, that when confronted with a test in the multiple-choice format, students show a tendency to use less elaborate learning approaches and do not go into the same depth, as when working on an assignment essay. Scouller (1998) also found that this phenomenon has an effect on students' performance in a multiple-choice test, disadvantaging those who employ deep learning strategies. Having this in mind, the question how students would approach multiple-evaluation examinations, rises. It is thinkable that the multiple-evaluation format could motivate students to employ deep learning strategies, because an evaluation of each answer option is elicited, making it sensible to gain a broad perspective while learning. At the same time, it might also contribute to a more thorough examination of the question and its answer options, and students would be less induced to jump on the first likely answer option.

Risk preference

In regard to risk taking, and taking the main characteristic of multiple-evaluation, the possibility to exhibit partial knowledge, into account, conducted research by Bereby-Meyer, Meyer and Budescu in 2003 appears to be of significance. Their work explored the effects of framing on choices participants make when answering multiple-choice questions, which enable testees to display partial knowledge. They accomplished this by employing two response formats, one where participants were asked to mark answer options that could be correct, the other format requiring testees to eliminate incorrect answer options. In both scenarios students knew that only one answer option was correct. Under the first rule, participants were informed that they would gain 3 points for identifying the correct answer and would lose 1 point for every false answer that was ticked (loss frame). Testees working under the second rule received 1 point for every correct elimination and lost 3 points when the correct answer was marked (gain frame). Results showed that students were more apt to risk when trying to identify possibly correct answer options, a finding which is in keeping with the prospect theory as postulated by Kahneman and Tversky in 1979, stating that sure gains (positive framing) lead to risk averse behavior, and sure losses (negative framing) result in risky decision making. When thinking of the characteristics of the multiple-evaluation format it could be assumed that students would be driven by the same strategy, namely to focus on identifying a possibly correct answer option and might therefore show a similar decision pattern, exhibiting a tendency towards risky decisions. Yet under multiple-evaluation, distributing percentage points among as many possibly correct answer options is quite the opposite of risky behavior, as one only has a total of 100% to distribute and therefore choosing multiple answers as a tactic, would always lead to a lower score. Risky decision making under multiple-evaluation would be to always allocate the full 100% to a single answer option. In contrast, when thinking back to chapter 1.2., where the logarithmic scoring rule was described, such a response style should be considered unattractive.

Further, when thinking of the age of typical test takers in an academic context, a finding by Dahl (2005) is of great importance. This research revealed that adolescents risk behavior (aged 12-26) did not adhere to findings postulated in predominant decision making theories. Instead, adolescents showed a general proneness to risk. Imagining the unique circumstance a test situation poses, it is thinkable that personality variables relating to work style become dominant, in such a context risk preference might therefore be reflected in an impulsive work manner.

Confidence

It is evident, that it is a good thing to be absolutely certain of an answer, however, when studying for an exam it is very difficult to acquire knowledge to the extent, that one is absolutely confident in every particular area of interest. When one goes further and tries to picture the test situation, the influence of other factors, such as stress, competence beliefs, perhaps also superstitions, etc., become apparent. Envisioning a test situation where the multiple-evaluation format is employed, it is conceivable that students with greater confidence could also give more extreme probability estimates, and contrarily, less confident students might be more hesitant, possibly resulting in lower test scores due to the more modest allocation of percentage points. When focusing on academic achievement, confidence might also be interpreted as the existence (or lack) of a positive academic self-concept, or motivation, especially when conceptualized in the sense of Schmalt, Sokolowski, and Langens (2000), where hope and fear components are distinguished.

Numerous findings showing a clear tendency towards overconfidence in ones own judgment (Budescu, Wallsten, & Au, 1997, Bradley, 1981, Lichtenstein, Fischhoff, & Phillips, 1982) seem relevant in regard to this study. Further, previous assumptions by Schaefer et al. (2004), predicting neuroticism to be negatively associated with overconfidence seem noteworthy, even though these could not be substantiated by their research. Wolfe and Grosch (1990) also found a negative correlation between negative affectivity and confidence (an example for a test item of the scale the authors used is “Things rarely work out the way I want them to.”). In relation to achievement, neuroticism and negative affectivity are likely to be manifested as test anxiety, in the sense that one is afraid to fail in a test situation. Such a disposition, and its possible effect on response behavior in a test situation, could be examined with the construct introduced by Schmalt et al. (2000), which takes the fear component of achievement motivation into consideration as well. In 2006 Eckert, Schilling and Stiensmeier-Pelster revealed that individuals, who had a poor academic self-concept and experienced failure, achieved lower results than people with a positive academic self-concept. This finding could potentially support the implementation of multiple-evaluation, because if the experience of failure has an effect on performance when one has a poorly developed academic self-concept, which is perhaps due to high anxiety levels, the possibility to express ones thoughts on a presented question more freely and differentiated, might make one feel more at ease. Thereby this effect could possibly be reduced or eliminated, resulting in more fairness.

As reported in the previous chapter, Kleitman and Stankov (2007) found interesting connections between competence beliefs and self-confidence. Yet it seems inherent to the matter, that response confidence should stand in relation with one’s competence beliefs in a certain field, since ones feeling of competence develops in situations where one experienced success, and can therefore be considered as a reflection of ones ability.

However, it would be interesting to see if students also give answers with greater confidence, independent of their beliefs concerning their ability in a given domain. On this more general level, competence beliefs might be interpreted in terms of motivation and the belief that one will succeed; again, the conceptualization of achievement motivation of Schmalt et al. (2000), acknowledging a hope component and a fear component, seems practicable. Further, Ehrlinger and Dunning (2003) found that self-views, meaning preexisting convictions of what one thinks one knows, sometimes have a greater impact on performance estimates than does the actual knowledge; a finding that could be very critical with respect to multiple-evaluation. Also Lindeman, Sundvik, and Rouhiainen (1995) observed that evaluations of performance were influenced by global self-views, such as self-esteem. Lindeman et al. (1995) further uncovered that overestimators were typically young (≤ 27 years), male, highly motivated, and had high self-esteem, on the other hand, underestimators showed lower self-esteem values. Such tendencies could be manifested in personality variables, such as impulsivity, reflexivity, decisiveness, or the ability to set realistic goals. Overestimators could be inclined to have an impulsive work style, and be less capable of estimating their performance realistically, therefore having difficulties forming goals, that are in keeping with their abilities. On the other hand, underestimators might have a tendency to ruminate, leading to high reflexivity values, but because of this, underestimators could also have difficulties reaching decisions, and as overestimators, be prone to form unrealistic goals.

Course of study

A further conceivable influence on response behavior is the course of study students applied for. Perhaps certain studies could nurture certain dispositions or even bring about new behavioral manners, leading to greater or more cautious response certainty, e.g. due to the nature of their future profession, medical students may be more likely to become accustomed to making decisions under pressure and uncertainty and yet be very certain of these. A multitude of scientific papers scrutinize the inevitability of uncertain decisions in a doctor's career (Hall, 2002, Hayward, 2006, Moore, 2011). Also, the importance of the ability to make informed decisions when confronted with an ambiguous situation receives great emphasis. Moore (2011) further points out an important change to come in regard to decision making in medicine. To date the aspiration was to be confident in the made decisions and moreover, that these decisions were to be reached quickly. However, the increasing complexity of medicine calls for a change of strategy, where doctors are proficient in the use of information resources and a further focus is to integrate this transformation into the medical curriculum. This shows that decision making is an integral part of medical education, even if this in not done overtly, and could be beneficial to medical students in the face of a test situation.

III. Empirical Part

4. Research aim and hypotheses

Past research has mainly focused on general aspects of personality when exploring influences on response behavior. However, achievement situations represent very unique circumstances under which other personal dispositions could be activated. The purpose of this study was to scrutinize, if response behavior in a knowledge-based multiple-evaluation test could be ascribed to certain personality variables linked to achievement. Specifically it is conceivable that students, who are more confident, would display greater certainty in the distribution of their subjective probability estimates. In order to examine this research question it was further considered of importance, to assess the personality variables of interest by utilizing objective personality tests, or respectively semi-projective tests, opposed to questionnaires, which rely solely on self-reports and therefore a persons ability to make adequate self-judgments. Following, two sets of hypotheses, the first focusing on students displayed response confidence, and the second examining students ability to produce realistic probability estimates, are presented.

4.1. Hypotheses in regard to response certainty and personality variables

These hypotheses will explore if exhibited response certainty is linked to personality. Response certainty is operationalized with Hansen's (1971) measure "mean item certainty". For the individual dimensions of the personality measures and course of study the following assumptions were made.

4.1.1. Hypothesis 1a: Multi-Motive-Grid - Hope of success

It is hypothesized that students, whose achievement motivation is based on hope of success, will display greater response certainty. This is in keeping with Kleitman and Stankovs' (2007) findings, but could also show that greater confidence, in the sense that one is confident that one will succeed, could lead to uncalled for overconfidence.

4.1.2. Hypothesis 1b: Multi-Motive-Grid - Fear of failure

It is further hypothesized that fear of failure will be associated with more cautious subjective probability estimates, which will be manifested in lower mean item certainty values. This assumption can be connected to past findings of correlations with negative affectivity and suppositions regarding neuroticism.

4.1.3. Hypothesis 1c: Attitudes towards Work - Aspiration level

In relation to aspiration level no predictions are made in regard to the direction of the correlation. This is because it is conceivable that high expectations in ones performance can lead to more careful response behavior, but it is also thinkable that when one displays low aspirations, in the sense that one has underestimated ones ability in the task of this scale, one tends to assign probability estimates hesitantly, resulting in low mean item certainty values.

4.1.4. Hypothesis 1d: Attitudes towards Work – Decisiveness

It is hypothesized that the ability to make decisions in ambiguous situations is linked to greater response confidence.

4.1.5. Hypothesis 1e: Attitudes towards Work - Reflexivity vs. Impulsivity

Furthermore, it is hypothesized that an impulsive work style is connected to greater response confidence.

4.1.6. Hypothesis 1f: Course of study

An additional interest was to explore the influence of the course of study. However, no predictions were made regarding which students would display greater response certainty or if certain studies correlate with especially cautious response behavior.

4.2. Hypotheses in regard to realism and personality variables

The second set of hypotheses examine potential connections between personality variables and the ability to reflect ones knowledge realistically and also convert this thought process into a numerical number. This is assessed with Dirkzwager's measure of realism (2003). The assumptions, which were made in regard to the specific dimensions of the utilized personality measures and course of study, are as follows.

4.2.1. Hypothesis 2a: Multi-Motive-Grid - Hope of success

It is assumed that hope of success is negatively correlated with realism (cave: realism values < 1 signify overconfidence, whereas values > 1 are indicative of underconfidence, perfect realism is reflected in a value of precisely 1). A negative relation therefore indicates, that people confident in their performance are less likely to be underconfident. In contrast, low scores suggest underconfidence.

4.2.2. Hypothesis 2b: Multi-Motive-Grid - Fear of failure

It is further hypothesized that underconfidence is connected to anxiety, in the sense that students are too modest when their achievement motivation is based on avoiding failure, therefore, a positive correlation is supposed.

4.2.3. Hypothesis 2c: Attitudes towards Work - Aspiration level

Again, no assumptions are made in regard to the direction of the relationship, since positive and negative relationships are plausible. Positive correlations indicating that, high ambitions are related to cautious response behavior and therefore reflecting realistic or perhaps even too modest responses. But perhaps also negative correlations are reasonable, as they would indicate that people with high hopes are inclined to be more confident and maybe even more willing to take greater risks and therefore tend to overestimate themselves.

4.2.4. Hypothesis 2d: Attitudes towards Work – Decisiveness

For a potential relationship between decisiveness and realism it is hypothesized that, people who are well capable of reaching a decision under uncertain conditions might have a tendency to overestimate themselves.

4.2.5. Hypothesis 2e: Attitudes towards Work - Reflexivity vs. Impulsivity

Similar goes for impulsivity and it is hypothesized that, people with high impulsivity scores will also be more likely to overestimate themselves, contrarily indicating that, people with a reflexive work manner will tend to be more realistic or perhaps even respond too modestly.

4.2.6. Hypothesis 2f: Course of study

This study also intends to explore if students of a certain course of study are more/less capable of estimating their knowledge realistically. However no speculations were made as to which students might be more/less realistic.

5. Method

5.1. Design

For the most part students are the population affected by knowledge tests. In the school context tests are commonly realized in an ordinary question answer scheme, where students have to produce the answers themselves. In contrast, when looking at the situation at university, the large number of students, especially in comparison to the number of professors, makes more efficient testing schemes necessary. For this reason the utilization of the multiple-choice format in knowledge examinations has become common practice. Taking this background into consideration it was chosen to recruit university students for the purpose of this study.

As Dirkzwager (1996, 2003) and Holmes (2002) have emphasized, it was decided to utilize a computer setting for testing, to ensure that responses to test items are made properly. This is important because the sum of the numerical responses to the answer options of an item must add up to one hundred. This is because the students are required to distribute one hundred percent across the available answer options, depending on their certainty in regard to the correctness of the given answer options. The use of a computerized test administration made it feasible to incorporate a feedback mechanism into the knowledge test used for this study, which prevented a loss of data due to inadequate distribution of percentages by either allocating too many or too few percentage points.

Knowledge was assessed with the Allgemeiner Wortschatztest, AWST – Version 1 (Hohensinn, unpublished), a measure to verbal competence and was programmed in the multiple-evaluation format with the open source survey application “LimeSurvey” (Version 2.00).

The personality inventories Multi-Motive-Grid, subsequently referred to as MMG (Schmalt et al. 2000) and Attitudes towards Work, subsequently referred to as AHA (Kubinger & Ebenhöh, 2007) were used to assess the personality variables of interest. In contrast to past research in this field, these inventories do not represent questionnaires eliciting self-reports. Instead the material is designed in a way, making it highly improbable for the participant to discover the underlying motive of the tasks. Therefore distorted results, as a result of purposeful faking or because the participant found it difficult to make realistic self-judgments, are less likely. A detailed description is found in the next chapter (5.2.).

In addition, the following data were also collected: age, gender, mother tongue, and course of study.

Scores for the AWST – Version 1 were derived utilizing the aforementioned logarithmic scoring rule (ME1).

To examine response certainty, Hansen's mean item certainty measure was used.

The extent to which participants display overconfident, underconfident or realistic subjective probability estimates was assessed with the realism measure, postulated by Dirkzwager (2003).

Finally, to examine a possible influence of personality variables, as assessed with the dimensions of the MMG (Schmalt et al., 2000) and AHA (Kubinger & Ebenhöh, 2007), on an individuals response behavior, regression analyses were utilized, in which the personality variables were entered as independent variables and the variables mean item certainty and realism as dependent variables.

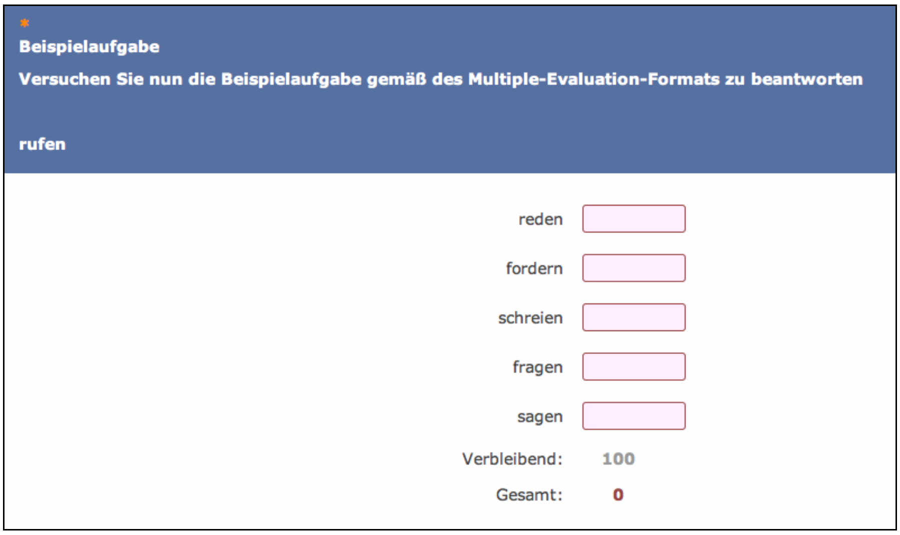
5.2. Materials

In this chapter details to the administered knowledge test and the utilized psychological assessment tools are given.

5.2.1. Allgemeiner Wortschatztest (AWST – Version 1)

The AWST – Version 1 (Hohensinn, unpublished) assesses verbal intelligence, in the sense that it is a measure of vocabulary and yields information on ones knowledge of the meaning of words. It was developed to precisely discriminate between individuals at an intermediate or advanced level regarding language ability; its use is therefore intended among high-school graduates and university students. The tasks of the AWST – Version 1 require the testee to find the synonym to a given word. It consists of 21 items of which each has 5 answer options that are to be evaluated according to the multiple-evaluation testing scheme and only one option is correct. Test completion takes approximately 10 minutes. Prior to the actual test, testees are required to solve the following practice item (in this study the practice item also functioned as a means for participants, to get acquainted with the multiple-evaluation format):

IMAGE 1
Practice task of the AWST – Version 1



5.2.2. Multi-Motive-Grid (MMG)

The MMG (Schmalt et al., 2000) is an instrument used for psychological assessment and determines achievement, affiliation and power motives. All variables are measured with respect to their hope and fear components. What makes this tool unique is its combination of projective measure and self-report questionnaire. The testee views an image that depicts a social context and is merely a schematic drawing. In doing so, the testee is supposed to identify him/herself with one of the figures in the picture and based on this identification is asked to respond to certain statements, representing important motivational states in a yes/no manner. The MMG allows for six motive scores to be calculated. For the achievement motive hope of success (HS) and fear of failure (FF) are distinguished. The affiliation motive examines the tendencies toward hope of affiliation (HA) and fear of rejection (FR). Finally, the power motive discriminates between hope of power (HP) and fear of power (FP). It has been shown that high achievement is associated with better performance and flow experience (Sokolowski, Schmalt, Langens, &, Puca, 2000). The MMG (Schmalt et al., 2000) as a diagnostic tool should only be administered among adults. Also, several reference tables are available and the reference population closest in age to the average age in the sample was chosen as a comparison group for this study.

Since the affiliation and power motives weren't key to the subject matter of the present study only the achievement motives were used for hypotheses testing. However, the entire test was administered due to the short duration of only 10 minutes. The achievement motives, which are of interest for the research question, are described below.

In general the achievement motive is aroused in situations where ones own accomplishment can be compared with either the result of others or a certain criterion.

Hope of success (HS)

Schmalt et al. (2000) state that, high results in the dimension hope of success are associated with a proneness to innovation, perseverance, perfectionism, and conscientiousness. It is noteworthy that people with high hope of success scores, seek pleasure in creating something new or having done something especially well and prefer tasks, which require self-reliant working. Because people with high results in this dimension have great confidence in their own abilities, they seek situations in which they have the possibility to compare themselves to the outside world.

Fear of failure (FF)

Contrarily, people with high scores in the dimension fear of failure are afraid to fail in achievement situations, especially when the accomplishment can be compared with the results of others (Schmalt et al., 2000). However, keen to reduce the perceived stress levels, people with high fear of failure scores tend to put a lot of effort in to their work and are often very meticulous and careful workers.

When taking both dimensions into account average results in both dimensions are indicative of good achievement motivation. Besides, two conflict types are apparent, on the one hand if a person scores high in both dimensions, constituting conflict type A, and on the other hand if low results are scored in both dimensions, conflict type B.

Conflict type A

When high scores are observed for both the hope and the fear component, initially the hope component is active and a person engages readily in a task/assignment. However, when a "real" test situation approaches, the fear component is triggered and takes over (Schmalt et al., 2000).

Conflict type B

Schmalt et al. (2000) describe the second conflict type, where a person scores low in hope of success and fear of failure, as a lack of spontaneous interest in the corresponding motives, in this case the achievement motives.

5.2.3. Attitudes towards Work (AHA)

The personality inventory Attitudes towards Work, referred to as AHA (Kubinger & Ebenhöh, 2007), yields information to a persons cognitive style, conceptualized as reflexivity/impulsivity, aspiration level, achievement motivation, and frustration tolerance. In order to measure the desired traits, testees complete different tasks, opposed to describing themselves through an ordinary questionnaire. What makes this inventory especially interesting is that these tasks were designed so that it is highly improbable for the person to discover the underlying dimensions during testing. This is why AHA belongs to the group of objective personality tests (see Kubinger, 2009). Specifically, the AHA (Kubinger & Ebenhöh, 2007) intends to explore a persons work conduct in an achievement situation and can be administered from an age of 14 years and up. The diagnostician can choose from a selection of reference tables for various populations to compare the results of a testee. In this study the population with the most similar age to the sample was chosen as a comparison group.

For the purpose of this study the variables aspiration level, reflexivity vs. impulsivity and decisiveness were utilized and a brief description of the aforementioned is given below. Because achievement motivation was already being assessed with the MMG the corresponding subtest of the AHA was omitted and therefore completion time was reduced to approximately 10 minutes opposed to circa 45 minutes.

Aspiration level

This dimension shows the tendency of a testee to set realistic or unrealistic goals. Testees are asked to complete a sorting task and after each round they receive feedback regarding the total amount of completed tasks, but also how many tasks others usually accomplish in the same time span. Then the testee is required to make an estimate on how many tasks s/he will complete in the next round.

Low results show a propensity to underestimation, whereas high results are to be seen as a proneness to overestimation. Realistic estimates are mirrored by average results.

Reflexivity vs. Impulsivity

This dimension yields information on a persons work conduct. It discriminates between people, who either tend to work slowly, with few mistakes (reflexivity) and people, who have a rapid work pace but therefore also have a higher error rate (impulsivity). This information is elicited by a task that requires the testee to compare two figures and decide which of the two has the larger area.

Low results reflect an impulsive work style, whereas high results indicate a reflexive work manner. Average results show that the testee neither has a tendency towards impulsivity nor to reflexivity.

Decisiveness

The basis of this value is the same task as described under impulsivity vs. reflexivity. It is an indicator of how easily someone is capable of reaching a decision in an uncertain situation.

People who find it more difficult to decide in ambiguous situations attain low results, whereas high results reflect a great capability to make decisions under uncertain circumstances.

5.3. Procedure

5.3.1. Preparation

Important steps in the preparation process included the programming of the AWST – Version 1, developing a standardized feedback for the personality inventories and the recruitment of participants. A brief description of how this was accomplished is given in this chapter.

Programming the multiple-evaluation format

The vocabulary test AWST – Version 1 was programmed in the multiple-evaluation format using "LimeSurvey – Version 2.00", an open source software for conducting surveys. For simplicity a very plain mask was chosen and a basic structure was maintained. First, socio-demographic information was yielded. Then the multiple-evaluation format was described in detail (see Appendix A). In addition to the provided information participants also completed a practice task (see chapter 5.2. Materials). This was then followed by the actual vocabulary test. After completion, testees were asked to beckon the test administrator so that the following assessment tools could be administered.

The vocabulary test itself was programmed that the given word was written on the top left of the computer screen. The five answer options were listed in the center and next to each word participants found an input field in which they were to enter their subjective certainty to the correctness of the corresponding answer option. This input field was programmed so that only numerical data was regarded as valid input and a maximum of three digits could be entered. Beneath the five data input fields an additional field kept count of the allocated percentages. If too few or too many percentage points were assigned, resulting the sum to differ from 100, a notification was given and testees were asked to check their input. Only if the sum equaled 100 exactly the next task was administered.

Standardized feedback

As an incentive for their participation, students were offered feedback to the administered personality inventories. This was especially useful and interesting for students because abilities associated with achievement were assessed (achievement motivation and work styles). Although only the main achievement motives of the MMG were utilized for hypotheses testing (i.e. hope and fear components) a comprehensive feedback giving students an overall insight in their achievement motivation, also taking conflict types into consideration, was delivered. Analogous stands for the AHA, where frustration tolerance, which was not relevant for hypotheses testing, was also included in the feedback provided. Participants received their personalized feedback via e-mail and also received the authors e-mail address and were encouraged to contact the author if the results should be unclear or unsettling in any way. The full, standardized feedback is attached in the appendix (Appendix B).

Recruitment of participants

In order to reach out to students and create awareness for the study, information leaflets were posted in highly frequented locations, such as libraries and other learning facilities of different universities in Vienna (Medical University of Vienna, Vienna University of Economics and Business Administration, University of Vienna). At the same time, students were approached directly and informed about the study and asked to participate. When students showed interest in participation, they gave their e-mail address and then received a link, which lead them to an online calendar with the available dates for testing. On the other hand, students interested by the information leaflet were asked to contact the author (students were able to tear off a paper slip with the e-mail address of the author) to arrange an appointment for testing and then also received the above-mentioned link. Through this calendar students were able to register themselves for available appointments.

5.3.2. Testing

Testing took place from 22.07.2013 through 04.10.2013 in the computer lab of the Division for Psychological Assessment at the Faculty of Psychology. A maximum of six participants were able to work simultaneously at individual computer workspaces. Since testing took place during the summer break it was very quiet contributing to a good work atmosphere. Also, the maximum capacity of six people was seldom reached, on the contrary mostly single sessions were arranged.

Upon arrival participants were greeted and seated at a computer of their choosing. Once they settled students signed a notice of informed consent. Thereafter, the attention was brought to the computer, where the programmed vocabulary test was already loaded and a start page, welcoming participants to the study, was visible. Participants were informed once more that completion would take approximately 30 minutes and that about halfway along they would be required to call for the test administrators' assistance, to load the following software. Following this notice testing began and the participants accessed the study. Depending on the participant the socio-demographic data were either entered together with the test administrator or independently. The following pages gave an instruction to the multiple-evaluation format, concluding with a practice task. Every workspace also had a printed version of the instruction. Before the first task of the vocabulary test, AWST – Version 1, was administered, students were reminded to try to be as realistic as possible in the estimation of their subjective probabilities. Once the knowledge test was completed the test administrator loaded the MMG and the AHA and they were administered in the same order.

It was decided to administer the vocabulary test before the personality inventories; this order was given preference so that these would not have an influence on the distribution of percentages. Also, a possible effect of the knowledge test on the results of the personality inventories seems less likely, since neither of the two are questionnaires, which would be based on self-reports and therefore be more prone to distortions, but to the contrary are tools that should be (more) resistant to (deliberate) misrepresentation given their characteristics as semi-projective or respectively, objective personality tests. After test completion, participants who wanted to receive feedback gave the test administrator their identification code and verified their e-mail address. This feedback was usually rendered within one week.

5.4. Sample

71 active university students (36 females, 35 males) participated in the study. The age ranged from 18 to 53 years (M = 25.56, SD = 6.38) and the predominant mother tongue was German (see table 3). In table 4 the applied courses of study are categorized into five groups, as used in the analyses. Which courses of study were grouped to which of the five categories, can be seen in the appendix (Appendix C).

TABLE 3
Students' mother tongue

Mother Tongue	Number of students
German	62
Serbian	3
Spanish	2
Albanian	1
Dari	1
Romanian	1
Hungarian	1

TABLE 4
Number of students as categorized in regard to their course of study

Course of study	Number of students
Humanities and Cultural Sciences	16
Law and Economic Studies	16
Social Sciences (Psychology, Pedagogy, Social Work)	20
Medical Studies (Human Medicine, Pharmaceutics)	12
Technical Studies	7

6. Results

The following chapter presents the obtained results from the statistical analyses. First, data collection and the calculation of the variables will be described, followed by an overview of descriptive data. Subsequently, the results for the hypotheses in regard to mean item certainty and realism will be presented. For regression analyses a significance level of $\alpha = .05$ was chosen.

6.1. Data collection and calculation of variables

Since all responses were entered directly into the computer and students could only proceed when responses were complete, no data was lost. The raw data of the AWST – Version 1 and the personality inventories MMG and AHA were imported into Excel 2004. Additionally, T-values and percentile ranks of the results of the MMG and AHA were imported. Subsequently, students' mean item certainty and realism values were calculated. The data were then imported into SPSS Version 20 and scores using various scoring rules were computed (linear scoring rule, ME1-4). Table 5 shows the descriptive statistics of the dependent variables (mean item certainty and realism) as well as the personality measures. Table 6 gives an overview of means in reference to socio-demographic data and shows that results are almost identical for gender, especially in regard to the values mean item certainty and realism. Results differ slightly when looking at the influence of mother tongue, showing higher mean item certainty rates when the mother tongue was German, and a greater tendency towards overestimation when students had a different mother tongue. In regard to age, the group of 41-53 year old students displayed the greatest response confidence. This is in part because one of the students responded to all questions of the AWST – Version 1 with absolute certainty and this age group merely comprises 3 participants. Otherwise most age groups are very similar with respect to response confidence. Also, the group of 41-53 year old students gave the most realistic responses, opposed to the group of 18-24 year old students, which were most prone to overestimation. Further, when looking at the results of the personality measures it becomes apparent, that the age group of 41-53 year old students not only has the lowest average decisiveness score, but this result is also below average, indicating that this age group has more difficulty to decide when facing uncertain conditions. Considering the results depending on students' course of study, results are comparable in regard to displayed mean item certainty. Students of technical studies showed the greatest tendency towards overestimation, whereas medical and social science students gave the most realistic probability estimates, however also showed a tendency towards overestimation. Interestingly, medical students had the highest decisiveness values, on the upper border of average results, which indicates that students of medical studies find it easier to reach a decision in ambiguous situations.

TABLE 5
Descriptive statistics

	Mean	SD
Mean item certainty: AWST – Version 1	.799	.148
Realism: AWST – Version 1	0.687	.250
Personality measures (T-values)		
MMG: Hope of success (HS)	49.62	8.636
MMG: Fear of failure (FF)	48.35	10.221
AHA: Aspiration level (AL)	52.93	9.128
AHA: Reflexivity vs. Impulsivity (ImRe)	50.59	9.882
AHA: Decisiveness (DE)	51.21	10.383

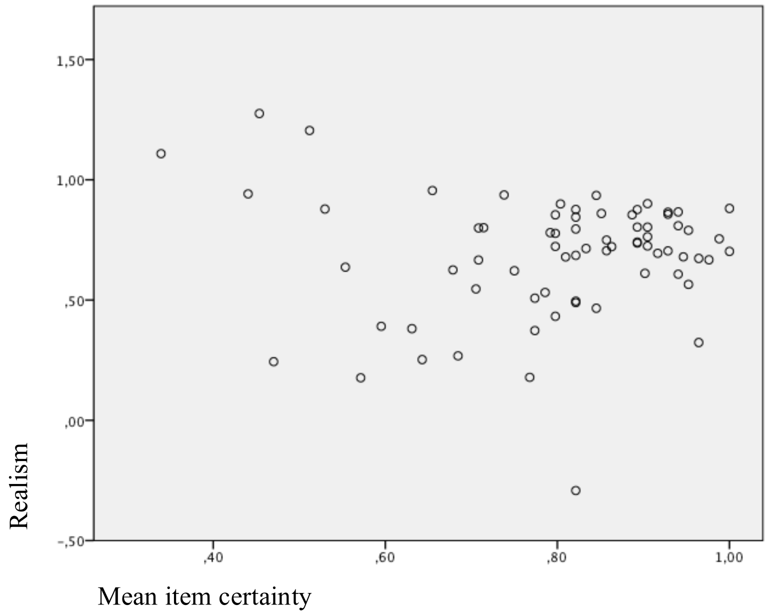
TABLE 6
Means in reference to socio-demographic data

	IC	Real.	AHA			MMG		N
			AL	DE	ImRe	HS	FF	
Gender								
Male	.80	0.68	55	51	51	51	46	35
Female	.79	0.70	51	52	50	49	50	36
Mother Tongue								
German	.82	0.72	53	51	50	50	48	62
Non-German	.67	0.47	51	54	53	47	50	9
Age								
18-24	.79	0.66	52	53	51	49	49	36
25-30	.79	0.70	55	51	49	50	47	25
31-40	.77	0.71	51	52	48	48	46	7
41-53	.99	0.78	47	38	63	54	48	3
Course of study								
Humanities and Cultural Sciences	.76	0.70	51	50	52	49	48	16
Law and Economic Studies	.82	0.63	51	52	46	54	47	16
Social Sciences	.83	0.74	54	48	54	50	50	20
Medical Studies	.80	0.73	55	57	48	46	50	12
Technical Studies	.76	0.56	56	51	53	46	46	7

Note. IC (Hansens' mean item certainty) is bounded (0,1); 1 indicating maximum certainty. Perfectly realistic results are reflected in a value of 1, underestimation is signified through values > 1, and values < 1 indicate overestimation (see p. 15). The results of the AHA and MMG are presented in t-values (average values range from 43-57).

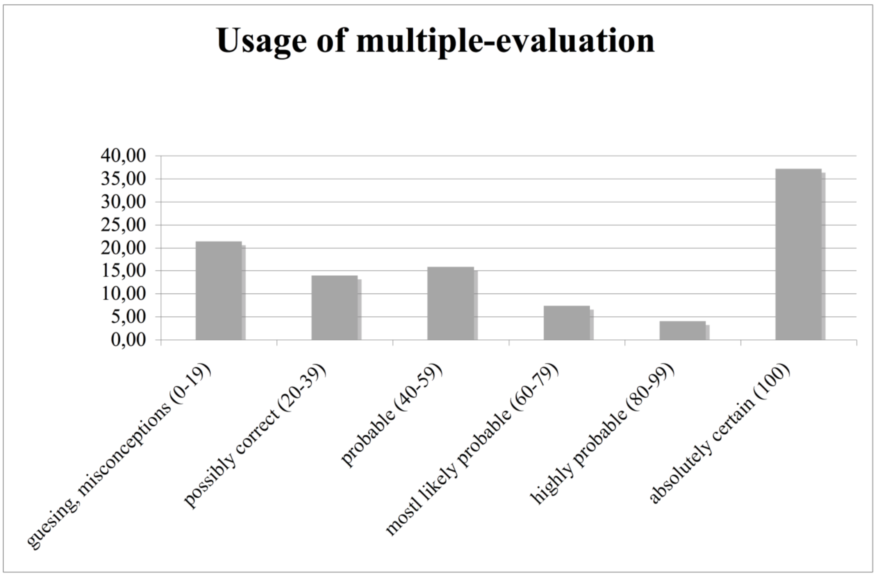
Figure 1 shows the dispersion of the dependent variables, realism and mean item certainty. It can be seen that the majority of students gave rather certain responses (values ≥ .80). Further the scatter plot reveals, that merely two students gave responses with absolute certainty to all questions (this is shown by mean item certainty values of 1); one of these two students was nevertheless very realistic in his/her responses, which is mirrored in a realism value very close to 1. This finding shows that students in general made use of the response format and gave differentiated answers, which contributed to the quality of the data. This observation also indicates that the questions were of differing levels of difficulty, making the use of the possibilities of the response format sensible and useful. In regard to realism it seems as though students made an effort to give realistic probability estimates, this is mirrored in realism values close to 1, although a clear tendency towards overestimation is still evident.

FIGURE 1



How students used the multiple-evaluation format can be seen in figure 2. The bar chart shows how many percentage points were allocated to the correct answer option on average over all 21 questions of the AWST – Version 1. When students were relatively certain they mostly indicated absolute certainty, by allocating 100 percentage points to the correct answer option (37%), less than 5% gave percentage points between 80 and 99. However when students considered the correct answer option to be likely or possible, they tried to voice their tendency by giving differentiated responses and allocated percentage points ranging from 20 to 79 (37%). It can be also seen that guessing and misconceptions were nevertheless also an issue (21%).

FIGURE 2



It is further noteworthy that prior to testing, the test administrator got the impression that participants were doing the conductor of the study a favor. This might be because there was a considerable time lapse between recruitment of participants and testing, possibly leading to a decrease of the initial interest in the research topic. Yet after testing, most participants were eager to engage in a conversation with the test administrator about the multiple-evaluation format and many wanted to see it realized in the academic field as soon as possible, emphasizing the freedom and being given the opportunity to more precisely share their thoughts on the subject matter. Therefore, it is likely that students' motivation, to contribute to the progression in this field, was reawakened in the test situation.

Chapters 6.2. and 6.3. are devoted to presenting the results of the hierarchical multiple regression analyses utilized for hypotheses testing.

6.2. Results in regard to the influence of personality variables on response certainty (C_T, as postulated by Hansen)

In order to examine the influence of personality variables, as assessed with the AHA and MMG, and course of study on students mean item certainty values, a beta regression was employed, since this technique is able to model continuous variables that assume values in the standard unit interval (0,1) (Cribari-Neto & Zeileis, 2010). Further, four cases were excluded from analysis due to standardized residuals greater than 3. This decision is based on criteria for excluding cases, summarized by Field (2009). Further likelihood ratio tests were computed to assess the improvement of the established models. Table 7 shows the stages of each hierarchy and Table 8 gives information to the models precision parameters (ϕ), where higher estimates indicate greater precision, this is because as phi increases, the variance of the dependent variable (mean item certainty) decreases (Cribari-Neto & Zeileis, 2010).

Model 1 shows that students' ability contributes significantly to response certainty, explaining 33% of its variance ($z = 5.659$, $p < .001$). It can be seen from model 2 that the dimensions of the AHA (aspiration level, decisiveness and reflexivity vs. impulsivity) do not contribute significantly to students' exhibited response certainty ($\Delta R^2 = .05$, $p > .05$). Adding the variables hope of success and fear of failure as assessed with the MMG in model 3 also did not contribute significant additional variance to the prediction of response certainty ($\Delta R^2 = .01$, $p > .05$). Model 4, incorporated students' course of study. Looking at model 4 it can be seen that Law and Economy students, as well as Social Science students tend to be more certain in their responses ($p < .05$). However, increasing the explained variance by 5% this model still does not significantly contribute to a better prediction of students' response certainty.

TABLE 7
Model summary of the hierarchical multiple regression analysis with mean item certainty as a criterion (N=67)

Model	Model predictors	B	SE B	z-value
1	Constant	0.663	0.137	4.831***
	ME1	0.092	0.016	5.659***
2	Constant	1.046	0.919	1.138
	ME1	0.098	0.016	6.175***
	Aspiration level	-0.014	0.009	-1.553
	Decisiveness	0.013	0.010	1.308
	Reflexivity vs. Impulsivity	-0.007	0.010	-0.717
3	Constant	1.349	1.202	1.123
	ME1	0.097	0.016	6.158***
	Aspiration level	-0.015	0.009	-1.653
	Decisiveness	0.012	0.010	1.123
	Reflexivity vs. Impulsivity	-0.006	0.010	-0.656
	Hope for success	0.000	0.009	0.028
	Fear of failure	-0.005	0.008	-0.574
4	Constant	0.787	1.143	0.689
	ME1	0.106	0.016	6.617***
	Aspiration level	-0.016	0.009	-1.861
	Decisiveness	0.018	0.010	1.908
	Reflexivity vs. Impulsivity	-0.003	0.010	-0.245
	Hope for success	-0.005	0.009	-0.586
	Fear of failure	-0.003	0.008	-0.450
	Humanities and Cultural Sciences vs. Law and Economic Studies	0.532	0.249	2.135*
	Humanities and Cultural Sciences vs. Social Sciences	0.439	0.215	2.045*
	Humanities and Cultural Sciences vs. Medical Studies	0.038	0.248	0.155
	Humanities and cultural Sciences vs. Technical Studies	0.533	0.283	1.887

Note. R^2 = .33 for model 1; ΔR^2 = .05 for model 2 (χ^2 = 6.65, *ps* > .05); ΔR^2 = .01 for model 3 (χ^2 = 0.32, *ps* > .05); ΔR^2 = .05 for model 4 (χ^2 = 8.01, *ps* > .05)
*** $p < .001$, ** $p < .01$, * $p < .05$

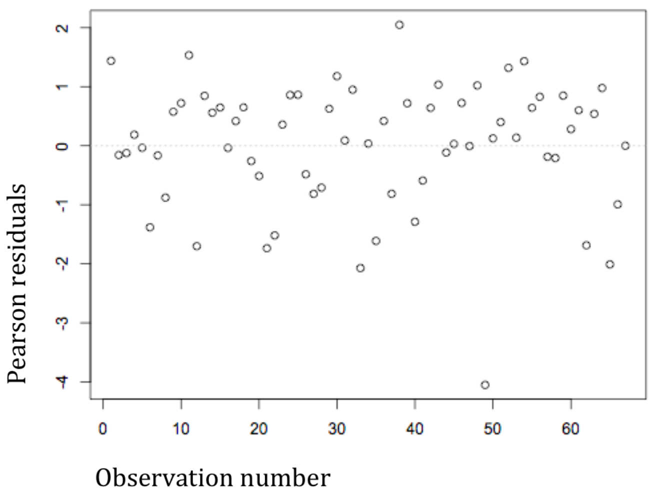
TABLE 8
Phi coefficients

Model	Estimate	SE	z-value
1	11.37	1.94	5.87***
2	12.64	2.16	5.85***
3	12.68	2.17	5.86***
4	14.48	2.48	5.84***

*** $p < 0.001$

The distribution of the pearson residuals, which are the residuals recommended in beta regression, since the deviation of an observation from the mean is not useful due to the heteroscedasticity inherent to the model (Cribari-Neto & Zeileis, 2010), can be seen in figure 3 and can be considered rather evenly dispersed, merely one case has a relatively large residual of -4.

FIGURE 3



6.3. Results in regard to the influence of personality variables on realism

A hierarchical multiple regression analysis was employed to examine a possible impact of personality variables on students' ability to estimate their knowledge realistically. Also, four cases were excluded from analysis due to standardized residuals greater than 3. This decision is based on criteria for excluding cases, summarized by Field (2009). The obtained results for each stage of the hierarchy are presented in table 9.

It is evident in model 1 that students' capability to estimate their knowledge realistically is based on their ability ($\beta = .951$, $p < .001$) and explains 90.5% of the variance. Model 2 shows that the dimension aspiration level of the AHA contributes significantly to predicting students' realism ($\beta = .141$, $p < .001$), explaining further 1.6% of the variance ($p < .05$). In model 3 the scales hope of success and fear of failure (assessed with the MMG) were incorporated, fear of failure showing an effect on students' ability to accurately estimate their knowledge ($\beta = .076$, $p < .05$). Yet, this does not improve the overall prediction of the model significantly ($\Delta R^2 = .005$, $p > .05$). In model 4 students' course of study was added, however, this leads to no further significant increase in predicting students' realism ($\Delta R^2 = .002$, $p > .05$).

TABLE 9
Model summary of the hierarchical multiple regression analysis with realism as a criterion (N=67)

Model	Model predictors	B	SE B	β
1	Constant	.390	.014	
	ME1	.036	.001	.951***
2	Constant	.188	.092	
	ME1	.036	.001	.935***
	Aspiration level	.003	.001	.119**
	Decisiveness	.000	.001	.014
	Reflexivity vs. Impulsivity	.001	.001	.026
3	Constant	.061	.116	
	ME1	.036	.001	.936***
	Aspiration level	.004	.001	.141***
	Decisiveness	.000	.001	.014
	Reflexivity vs. Impulsivity	.000	.001	.021
	Hope for success	.000	.001	.016
	Fear of failure	.002	.001	.076*
4	Constant	.063	.120	
	ME1	.036	.001	.932***
	Aspiration level	.003	.001	.138**
	Decisiveness	.001	.001	.022
	Reflexivity vs. Impulsivity	.000	.001	.016
	Hope for success	.000	.001	.015
	Fear of failure	.002	.001	.072
	Humanities and Cultural Sciences vs. Law and Economic Studies	-.003	.027	-.005
	Humanities and Cultural Sciences vs. Social Sciences	.021	.025	.041
	Humanities and Cultural Sciences vs. Medical Studies	-.003	.028	-.006
	Humanities and Cultural Sciences vs. Technical Studies	.001	.033	.002

Note. R^2 = .905 for model 1; ΔR^2 = .016 for model 2 (*ps* < .05); ΔR^2 = .005 for model 3; ΔR^2 = .002 for model 4
*** $p < .001$, ** $p < .01$, * $p < .05$

As can be seen from the P-P plot (figure 4) residuals can be considered normally distributed. The scatter plot, showing the standardized residuals and standardized predicted values (figure 5), further shows that the assumptions of random errors and homoscedasticity have been met, as points are relatively evenly dispersed, merely one observation has an extreme value.

FIGURE 4

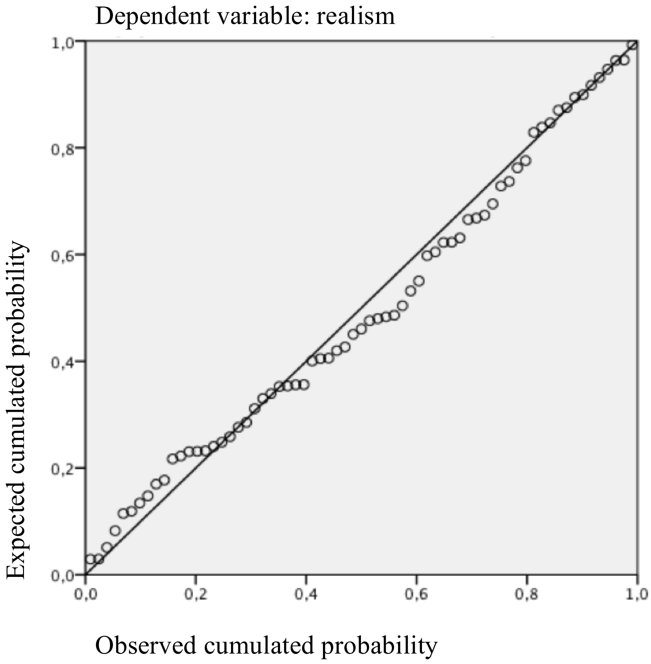
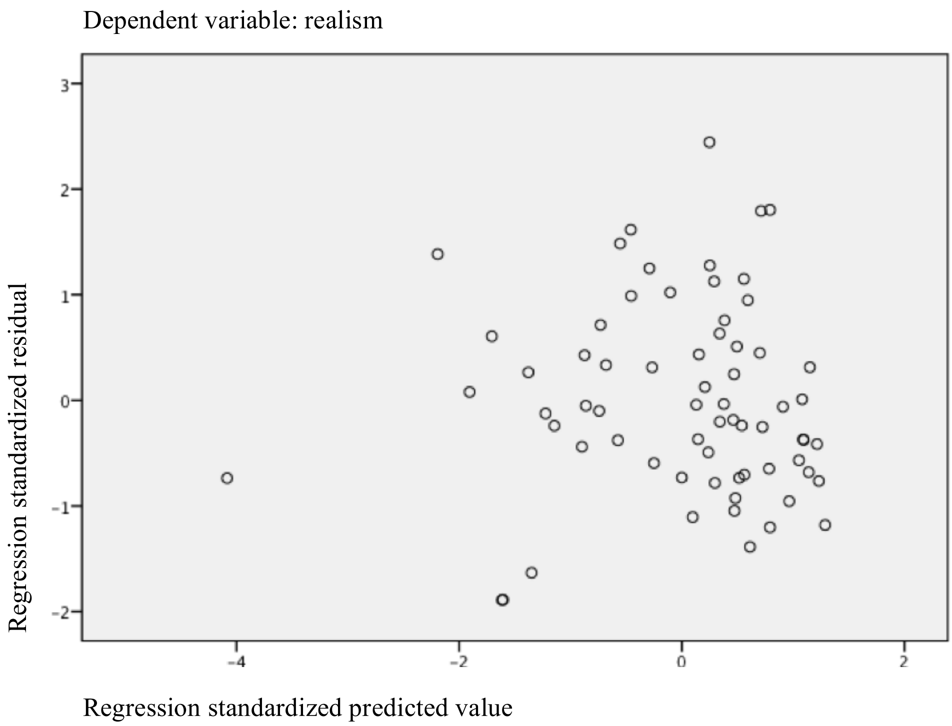


FIGURE 5



7. Discussion

Independent of the results of the regression analyses employed two findings are noteworthy, that are also in keeping with previous research. First, detailed analyses of descriptive data showed that younger students (age 18-24 years) had the greatest difficulty to accurately evaluate their present state of knowledge. This can be inferred from the obtained high mean item certainty values and the low realism results, reflecting overestimation. Secondly, descriptive data revealed that medical students are most able in making decisions in the face of uncertainty. However, this circumstance did not show to have a significant impact on displayed response confidence nor on the ability to make realistic probability estimates.

The following sections discuss the results of the regression analyses presented in the previous chapters.

7.1. Findings in regard to students' response certainty

Hierarchical multiple regression analysis showed that none of the assessed personality variables had a significant impact on students' response certainty. The sole factor, which had an effect, was students' ability. Therefore these results can be seen as in favor of the response format multiple-evaluation, because students' response certainty is not biased by individual differences. Taking students' course of study into account, a slight tendency of Law, Economy, and Social Science students to display greater certainty in their answers was shown. However, this did not prove to be of significant impact and is therefore negligible.

Although new personality variables, assessed through objective personality measures, were implemented in this work, which focused on personal dispositions in achievement situations, results are in general in keeping with past research showing no significant influences of individual differences (see also Echternacht et al., 1972, Stankov & Crawford, 1997, Jenner, 2012). Interestingly, the suspicion that achievement motivation would be connected to students' response certainty – which was founded on previous findings, linking competence beliefs to self-confidence (cf. Kleitman & Stankov, 2007) – could not be substantiated by this research. This could be a consequence of the means of measurement. In Kleitman and Stankovs' work (2007) self-confidence was assessed through a separate scale, where individuals indicated how certain they were, that their answer was correct. Whereas in multiple-evaluation the response format itself evokes probability estimates for every single answer option of a given question – intertwining knowledge and certainty in the given response – and therefore reflects individuals' certainty and also depth of knowledge in the respective domain. Further, Kleitman and Stankov (2007) utilized an overall measure of confidence, by calculating an average value over all test items. In contrast Hansens' mean item certainty is based on the distribution of percentage points per item. The difference in outcome could be explained by taking these distinctions into account.

7.2. Findings in regard to students' ability to reflect their knowledge realistically

In regard to realism, an effect of students' aspiration level was found at all stages of analysis. This significant positive relationship between aspiration level and realism shows that students with high performance expectations tend to be more realistic in their judgments. This seems logical since it can be considered likely, that if a student has high ambitions, s/he will also want to ensure that these are realized and when utilizing the response format multiple-evaluation this is done best, when one realistically reflects ones knowledge.

Further, results seem to indicate that students who are afraid to fail are less prone to overestimate themselves (as shown in model 3, see table 9). However, in stage 4 of the model, where students' course of study was included, this effect did not prove stable and was no longer significant. These findings perhaps indicate a trend, showing a slight tendency that students with greater anxiety levels are more realistic and cautious when trying to represent their present knowledge, by translating their thoughts hereto into a numerical value. Consequently it seems likely that students, who do not experience such anxiety would benefit more from a training, helping students to get better acquainted with multiple-evaluation format or put more strongly, are more dependent on such a training in order to obtain valid results.

However, the main influence being students' ability indicates that, the more able students are in the domain of interest, the more realistic their judgments will be.

These findings seem to support the claims by Dirkzwager (1996) and Holmes (2002), that students are able to give realistic probability estimates and also learn to improve the quality of these estimates through practice. This also seems of importance since students' tendency towards overestimation in their responses was also evident in this study.

Overall the main limitations of the study can be seen in the small sample size of merely 71 students and the fact that the study presented a low-stakes test situation. Due to the modest sample size drawn conclusions should be treated with care. It is further likely to be possible, that when students are confronted with the multiple-evaluation format in a high-stakes test, response behavior might differ from what was seen in this study; students most probably showing more extreme answers displaying greater response confidence, and less accurately assessing their present state of knowledge (see Sieber, 1974).

7.3. Future research

In the future it would be of interest to examine, whether findings could be replicated if testing would be realized in a "real", high-stakes test situation, or if perhaps such a situation would amplify personality traits and therefore produce effects. On the other hand it is conceivable that in such a situation, students might stick to strategies they have already developed. Further, it would be interesting to observe how well students would adjust to this response format, as most will be used to multiple-choice and might react similarly to a new, yet similar response format, exhibiting response behavior that would be expected when using multiple-choice. This could be done by giving short examinations in a seminar throughout a semester or over the course of an academic year. At the same time such a test might also be an interesting way to observe personality influences, as it might pose a situation that is neither a low- nor high-stakes test. A further interesting question regards the applicability of multiple-evaluation, meaning if it is appropriate for a high-stakes test or perhaps more practicable for self-assessments. A possible approach to explore this matter, would be to realize the above mentioned examination, but also compare two groups, one were the obtained results are integrated (to a small extent) into the semester grade, and in the other results would be used as means of self-evaluation.

8. Summary

The intention of this study was to shed light on the ongoing debate, if personality traits influence response behavior in a multiple-evaluation test. Especially because of the limitations of the multiple-choice format, it seems of invaluable worth to look for interesting alternatives. The main assets of multiple-evaluation in comparison to multiple-choice seem to be an increase in reliability, it further poses an effective method to exorcise guessing, it is able to take partial knowledge into account, it fulfills the requirements of the quality criteria "appropriateness" and "fairness", and besides could potentially be a valuable tool for instructors to obtain information in regard to students' understanding of discussed topics.

In contrast to multiple-choice, the multiple-evaluation format gives the test taker the possibility to give highly differentiated responses. Opposed to just picking the most likely answer, the respondent evaluates each answer option and distributes 100% over the presented answer options, according to his/her certainty regarding the correct answer. Besides, a logarithmic scoring rule is implemented, which ensures that the score is maximized when the testee gives realistic responses. Due to this property, it is possible that a testee receives a negative score. However, negative scores represent serious misconceptions or guessing. Therefore this aspect and the fact, that the score is maximized when one gives realistic probability estimates, motivate students to reflect their knowledge honestly and refrain from guessing.

When using multiple-evaluation, testees are required to consider their present state of knowledge, but further also have to translate this thought process into numerical values, that are then assigned to the available answer options of a question. However, it is conceivable that personality variables have an influence on how this task is fulfilled.

In the present study 71 active university students were recruited (36 females and 35 males) and completed the vocabulary test, AWST – Version 1, which was programmed in the multiple-evaluation format. Further, students completed the personality inventories Multi-Motive-Grid and Attitudes towards Work, measuring personality variables linked to achievement. In contrast to past research, this study utilized objective personality measures, opposed to obtaining results through self-reports. It was scrutinized if these variables as well as students' course of study had an impact on response behavior. Effects on response behavior were operationalized with the measures mean item certainty (as postulated by Hansen, 1971) and realism (as postulated by Dirkzwager, 2003).

Hypotheses were formed in regard to the effects of personality on mean item certainty and realism. A hierarchical beta regression was employed to examine the effects on mean item certainty, as these values cannot exceed 1. It was found that none of the measured personality variables had an influence on students' response certainty. In order to evaluate students' ability to reflect their knowledge realistically, a hierarchical multiple regression was used as a means of analysis. Results indicate that students with high ambitions are more able, when it comes to estimating their state of knowledge. Further, a trend signifying that anxious students also tend to be more cautious and realistic in the distribution of their probability estimates was noticeable.

Overall, results are in favor of the response format multiple-evaluation, showing that response behavior is predominantly based on a person's ability, opposed to personal dispositions.

It would be interesting to examine if these results would be replicated, if data were to be collected in a "real" test situation. Additionally, it could potentially be of value to observe response behavior over a longer period of time, in order to evaluate if testees require time to adapt to the requirements of this new testing system, as most students will be used to taking multiple-choice tests.

References

Bereby-Meyer, Y., Meyer, J., & Budescu, D. V. (2003). Decision making under internal uncertainty: the case of multiple-choice tests with different scoring rules. *Acta Psychologica, 112*, 207-220.

Bradley, J. V. (1981). Overconfidence in ignorant experts. *Bulletin of the Psychonomic Society*, *17*, 82-84.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997) On the importance of random error in the study of probability judgment. Part II: Applying the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making*, *10*, 173-188.

Cribari-Neto, F., & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, *34*, 1-24.

Dahl, M. J. (2005). Adolescent decisions in situations of uncertainty: The impact of risky choice framing and decision making competency. *Dissertation Abstracts International, 66*, 12B. (UMI No. 3199273)

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, *18*, 87-123.

Dirkzwager, A. (1996). Testing with personal probabilities: eleven year olds can correctly estimate their personal probabilities. *Educational and psychological measurement*, *56*, 957-971.

Dirkzwager, A. (2003). Multiple Evaluation: A New Testing Paradigm That Exorcizes Guessing. *International Journal of Testing*, *3*, 333-352.

Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and psychological measurement, 13*, 574-595.

Ebel, R. L. (1965) Confidence Weighting and Test Reliability. *Journal of Educational Measurement, 2*, 49-57.

Echternacht, G. J. (1972). The Use of Confidence Testing in Objective Tests. *Review of Educational Research, 42*, 217-236.

Echternacht, G. J., Boldt, R. F., & Sellman, W. S. (1972). Personality Influences on Confidence Test Scores. *Journal of Educational Measurement*, *9*, 235-241.

Eckert, C., Schilling, D., & Stiensmeier-Pelster, J. (2006). Einfluss des Fähigkeitsselbstkonzepts auf die Intelligenz- und Konzentrationsleistung. *Zeitschrift für Pädagogische Psychologie, 20*, 41-48.

Ehrlinger, J., & Dunning, D. (2003). How Chronic Self-Views Influence (and Potentially Mislead) Estimates of Performance. *Journal of Personality and Social Psychology, 84*, 5-17.

Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology, 54*, 415-425.

Field, A. (2009). *Discovering Statistics Using SPSS* (3rd rev. ed). London: Sage.

Hall, K. H. (2002). Reviewing intuitive decision-making and uncertainty: the implications for medical education. *Medical Education, 36*, 216-224.

Hansen, R. (1971). The Influence of Variables Other than Knowledge on Probabilistic Tests. *Journal of Educational Measurement*, *8*, 9-14.

Hayward, R. (2006). Balancing certainty and uncertainty in clinical medicine. *Developmental Medicine and Child Neurology, 48*, 74-77.

Hohensinn, C. (unpublished). *Allgemeiner Wortschatztest – Version 1*.

Holmes, P. (2002). *Multiple Evaluation versus Multiple Choice as Testing Paradigm - Feasibility, Reliability and Validity in Practice*. Published thesis, University of Twente, Enschede, Netherlands.

Jenner, L. M. (2012). *Multiple-Evaluation - Pilotstudie zu einem neuen Antwortformat*. Unpublished thesis, University of Vienna, Vienna, Austria.

Kahneman, D., & Tversky, A. (1979). *Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47*, 263-292.

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*, 161-173.

Koehler, R. A. (1974). Overconfidence on Probabilistic Tests. *Journal of Educational Measurement*, *11*, 101-108.

Kubinger, K. D. (2009). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens* (2nd rev. ed.). Göttingen: Hogrefe.

Kubinger, K. D., & Ebenhöh, J. (2007) Arbeitshaltungen – Kurze Testbatterie: Anspruchsniveau, Frustrationstoleranz, Leistungsmotivation, Impulsivität/Reflexivität [Computer software]. Mödling: Schuhfried.

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.

Lichtenstein, S., Fischhoff, G., & Phillips, L. D. (1982). Calibration of probabilities. The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: heuristics and biases* (pp. 306-334). New York: Cambridge University Press.

Lindeman, M., Sundvik, L., & Rouhiainen, P. (1995). Under- or overestimation of self? Person variables and self-assessment accuracy in work settings. *Journal of Social Behavior & Personality, 10*, 123-134.

Mondak, J. J. (2001). Developing Valid Knowledge Scales. *American Journal of Political Science*, *45*, 224-238.

Moore, M. (2011). Teaching Physicians to Make Informed Decisions in the Face of Uncertainty: Librarians and Informaticians on the Health Care Team. *Academic Medicine, 86*, 1345.

Nadeau, R., & Niemi, R. G. (1995). Educated guesses: The process of answering factual knowledge questions in surveys. *Public Opinion Quarterly*, *59*, 323-346.

Rippey, R. M. (1970). A Comparision of Five Different Scoring Functions for Confidence Tests. *Journal of Educational Measurement*, *7*, 165-170.

Schaefer, P. S., Williams, C. C., Goodie A. S., & Campbell, W. K. (2004). Overconfidence and the Big Five. *Journal of Research in Personality*, *38*, 473-480.

Schaefer, R. E. (1976). Eine Alternative zur konventionellen Methode der Beantwortung und Auswertung von Tests mit Mehrfachantworten. *Diagnostica*, *22*, 49-63.

Schmalt, H. D., Sokolowski, K., & Langens, T. (2000). Das Multi-Motiv-Gitter für Anschluß, Leistung und Macht (MMG) [Computer software]. Frankfurt/M.: Swets.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*, 453-472.

Shuford, E. H., Jr., Albert, A., & Massengill, H. E. (1966). Admissible Probability Measurement Procedures. *Psychometrika, 31*, 125-145.

Sieber, J. E. (1974). Effects of decision importance on ability to generate warranted subjective uncertainty. *Journal of Personality and Social Psychology, 30*, 688-694.

Sokolowski, K., Schmalt, H. D., Langens, T. A., & Puca, R. M. (2000). Assessing Achievement, Affiliation, and Power Motives All at Once: The Multi-Motive Grid (MMG). *Journal of Personality Assessment*, *74*, 126-145.

Stankov, L., & Crawford, J. D. (1997). Self-Confidence and Performance on Tests of Cognitive Abilities. *Intelligence*, *25*, 93-109.

Walker, D. M., & Thompson, J. S. (2001). A Not on Multiple Choice Exams, with Respecht to Students’ Risk Preference and Confidence. *Assessment & Evaluation in Higher Education*, *26*, 261-267.

Wolfe, R. N., & Grosch, J. W. (1990). Personality correlates of confidence in one’s decisions. *Journal of Personality*, *58*, 525-534.

List of tables

Appendix A: Instruction to the AWST – Version 1 with multiple-evaluation as response format

Bei den folgenden Aufgaben geht es um die Bedeutung von Wörtern. Bei jeder Aufgabe wird zunächst 1 Wort vorgegeben. Sie sollen zu diesem Wort das Synonym finden, also ein Wort, das **dieselbe** oder eine **sehr ähnliche Bedeutung** hat. Dazu sind pro Aufgabe jeweils 5 weitere Wörter als Antwortmöglichkeiten angeführt. Finden Sie aus diesen 5 Wörtern das Synonym bzw. jenes Wort, das die ähnlichste Bedeutung zum Vorgabewort hat.

Die Aufgaben werden im neuen Antwortformat "Multiple Evaluation" vorgegeben:- Im Multiple-Choice-Format müsste man sich für eine Antwortmöglichkeit entscheiden und bekommt dementsprechend 0 Punkte oder 1 Punkt (meistens).- Bei „Multiple Evaluation“ hingegen hat man die Möglichkeit sein Wissen differenziert anzugeben. Das wird dadurch ermöglicht, dass man für jede Frage 100 % zur Verfügung hat und diese dann – entsprechend der persönlichen Sicherheit hinsichtlich der richtigen Antwort – auf die Antwortmöglichkeiten verteilt. Die Summe der verteilten Prozentwerte muss bei jeder Aufgabe 100 betragen.Beispiele zur Prozentvergabe:
rufen

a) reden	20 %	a) reden	0 %	a) reden	0 %
b) fordern	20 %	b) fordern	0 %	b) fordern	0 %
c) schreien	20 %	c) schreien	70 %	c) schreien	100 %
d) fragen	20 %	d) fragen	30 %	d) fragen	0 %
e) sagen	20 %	e) sagen	0 %	e) sagen	0 %
	100 %		100 %		100 %

Person X kann keine Antwortmöglichkeit ausschließen, daher verteilt Sie 20 % auf jede Antwortalternative.

Person Y schwankt zwischen den Antwortmöglichkeiten c und d und schätzt ihre Sicherheit hinsichtlich der richtigen Antwort so ein, dass Sie sich zu 70 % sicher ist, dass Antwortmöglichkeit c die richtige Lösung ist und zu 30 % Antwortmöglichkeit d.

Person Z ist sich sicher das bedeutungsähnlichste Wort zu kennen und vergibt daher die gesamten 100 % auf Antwortmöglichkeit c.

Wie sieht die **Punktevergabe** bei Multipler Evaluation aus?
Wenn Sie, ganz sicher sind, die richtige Antwort – also das bedeutungsähnlichste Wort – zu kennen, setzen Sie die gesamten 100 % auf diese Antwort und 0 % auf alle übrigen Antwortalternativen. So erzielen Sie die **volle Punktezahl**, sofern Ihre Antwort richtig ist.
Wenn Sie zwischen zwei oder mehreren Antwortmöglichkeiten schwanken, versuchen Sie Ihre Tendenz mit Hilfe der Prozentverteilung **so genau wie möglich** anzugeben.Sobald Sie **mehr** als 20 % auf die **richtige** Antwortmöglichkeit setzen, bekommen Sie **Pluspunkte**.Sobald Sie **weniger** als 20 % auf die **richtige** Antwortmöglichkeit setzen, bekommen Sie **Minuspunkte**.

Wenn Sie auf eine Aufgabe die Antwort nicht wissen, verteilen Sie 20 % auf jede Antwortmöglichkeit. Dann bekommen Sie in jedem Fall **0 Punkte**.

Versuchen Sie Ihr Wissen möglichst realistisch einzuschätzen, denn nur so können Sie Ihr Ergebnis maximieren!

Appendix B: Standardized Feedback

Du hast vor kurzem an einer Studie zum Antwortformat „Multiple Evaluation“ teilgenommen. Im Rahmen der Überprüfung dieses Antwortformats wurden auch Persönlichkeitsverfahren verwendet. Nach der Bearbeitung hast du angegeben, dass du gerne eine Rückmeldung zu deinen Ergebnissen hättest. Das Feedback resultiert aus dem Vergleich deiner Angaben und Einschätzungen mit den Ergebnissen von Personen, die in etwa in deinem Alter sind. Ergebnisse über Persönlichkeitseigenschaften stellen in besonderem Maße eine Momentaufnahme dar, wobei unterschiedliche Faktoren, wie z. B. die Tagesverfassung, einen Einfluss auf die Bearbeitung und damit das Ergebnis nehmen können. Viele Persönlichkeitseigenschaften können sich auch über die Zeit verändern, sodass die nachfolgenden Ergebnisse ein derzeitiges Bild widerspiegeln, das sich über die Zeit aber ändern kann.

Motivation wird in die Komponenten „Hoffnung auf Erfolg“ und „Furcht vor Misserfolg“ unterteilt:

- Textabschnitt XY für die Skala „Hoffnung auf Erfolg“ und
- Textabschnitt XY für die Skala „Furcht vor Misserfolg“
- Textabschnitt XY für MMG insgesamt bzw. Konflikttyp

- Textabschnitt XY für die Skala Entschlussfreudigkeit
- Textabschnitt XY für die Skala Impulsivität vs. Reflexivität
- Textabschnitt XY für die Skala Frustrationstoleranz
- Textabschnitt XY für die Skala Anspruchsniveau

MMG (definierter Durchschnittsbereich: PR 25-75)

1) Hoffung auf Erfolg
   a) Unterdurchschnittlich
   Die Ergebnisse der Dimension „Hoffnung auf Erfolg“ weisen darauf hin, dass Leistungssituationen für dich weniger mit Spaß verbunden sind. Den Vergleich mit anderen oder eigenen Leistungen suchst du weniger. Langfristige Ziele konsequent zu verfolgen und momentane Impulse zu unterdrücken fällt dir eher etwas schwer. Eine Möglichkeit die Leistungsmotivation zu fördern, ist z. B. ein größeres Ziel in mehrere kleine aufzuteilen, wodurch die Teilerfolge einen Boost-Effekt für die nächstfolgenden Etappen auslösen können.

   b) Durchschnittlich
   Die Ergebnisse der Dimension „Hoffnung auf Erfolg“ weisen darauf hin, dass Leistungssituationen für dich auch mit Spaß verbunden sind. Du hast Freude daran deine Leistung zu steigern und traust dich auch neue Wege zu gehen, um ein Ziel zu erreichen. Du schaffst es meistens deine langfristigen Ziele im Blick zu behalten und du bist bereit dafür momentane Impulse zu unterdrücken. Du setzt dir meist realistische Ziele und bevorzugst Aufgaben, die eigenverantwortliches Handeln verlangen.

   c) Überdurchschnittlich
   Die Ergebnisse der Dimension „Hoffnung auf Erfolg“ weisen darauf hin, dass dir Leistungssituationen Freude bereiten. Du weißt gerne, wie du abgeschnitten hast und probierst auch neue Wege, um ans Ziel zu kommen. Aufgaben, die eigenverantwortliches Handeln verlangen, werden von dir bevorzugt. Du arbeitest außerdem sehr sorgfältig und setzt dir meistens realistische Ziele und verfolgst diese auch konsequent. Es kann auch sein, dass du manchmal eine Tendenz zu Perfektionismus zeigst. Bei einfachen Aufgaben könnte das jedoch zu wenig

effizienten Strategien führen.

2) Furcht vor Misserfolg

a) Unterdurchschnittlich
In Bezug auf Leistungssituationen zeigen die erzielten Ergebnisse eine Tendenz zu Sorglosigkeit in Hinblick auf mögliche Fehler und deren Konsequenzen. Häufige Ursachen dafür können u. a. ein geringes Interesse sein, aber auch ein ungünstiger Aufgabentyp, etwa Aufgaben, die mit zu wenig Verantwortung verbunden sind oder Aufgaben, die als monoton erlebt werden, auch Aufgaben, die einen über- oder unterfordern wirken sich ungünstig auf die Leistungsmotivation aus.

b) Durchschnittlich
Die erzielten Ergebnisse zeigen, dass du über eine arbeitsförderliche Balance von Sicherheit und Unsicherheit verfügst. Dies wirkt sich zusätzlich positiv auf deine Neigung aus, gewissenhaft mit verantwortungsvollen Aufgaben umzugehen.

c) Überdurchschnittlich
Die Ergebnisse zeigen eine Sorge auf, sich in Leistungssituationen nicht zu bewähren, v.a. wenn die eigene Leistung mit der Leistung anderer verglichen werden kann. Dies führt bei vielen Personen zu einer sehr gründlichen und sorgfältigen Arbeitsweise. Dadurch wird ermöglicht, die Unsicherheit bezüglich der eigenen Leistung zu reduzieren. Kleine Belohnungen und Bestärkung durch Lob wirken hier besonders positiv auf die Leistungsmotivation.

3) MMG Insgesamt / Konflikttyp

a) HE PR>75, FM PR 25-75
Betrachtet man beide Dimensionen gemeinsam, sprechen die Ergebnisse für eine gut ausgeprägte Leistungsmotivation. Kennzeichnend für leistungsmotivierte Personen ist deren intrinsische Motivation, d.h. die Neugierde, das Interesse und die Freude an der Handlung an sich, sind Belohnung genug.

b) HE PR <25, FM PR>75
Betrachtet man beide Dimensionen gemeinsam, sprechen die Ergebnisse für eine eher gering ausgeprägte Leistungsmotivation. Daraus zu schließen, dass diese Personen weniger hart arbeiten, wäre jedoch falsch, denn diese zeigen sehr häufig besonders hohe Anstrengungen, um die Unsicherheit in Bezug auf Leistungssituationen zu kompensieren. Lob und Belohnungen können in diesem Falle einen besonders motivationsförderlichen Einfluss haben.

c) Konflikttyp: Hohe Werte in beiden Dimensionen
Betrachtet man beide Dimensionen gemeinsam, sprechen die Ergebnisse dafür, dass es zu Beginn einer Leistungssituation typischerweise zu einer Aktivierung der Hoffnungskomponente kommt. Die Motivation kann auch lange Zeit aufrechterhalten werden. Sobald jedoch eine „echte“ Prüfung bevorsteht, tritt die Furchtkomponente hinzu. Das kann sich in Form von Selbstzweifeln und/oder einer erhöhten physiologischen Aktivierung (z. B. Schwitzen, Anstieg der Herzfrequenz) zeigen. Damit einhergehend fällt es Personen mit solchen Ergebnissen manchmal schwer, das eigene Können in Prüfungssituationen umzusetzen. Diese Situation findet sich häufig im Leistungssport, weshalb sich dort dafür der Begriff „Trainingsweltmeister“ verbreitet hat.

d) Konflikttyp: Niedrige Werte in beiden Dimensionen
Betrachtet man beide Dimensionen gemeinsam, sprechen die erzielten Ergebnisse dafür, dass in Hinblick auf Leistungssituationen ein geringes spontanes Interesse besteht.

AHA (definierter Durchschnittsbereich: PR 25-75)

4) AHA Entschlussfreudigkeit

a) Unterdurchschnittlich
In Hinblick auf deine Entschlussfreudigkeit zeigen die Ergebnisse, dass du im Vergleich zu anderen weniger gut dazu in der Lage bist, in uneindeutigen Situationen Entscheidungen zu treffen.

b) Durchschnittlich
In Hinblick auf deine Entschlussfreudigkeit zeigen die Ergebnisse, dass du genauso gut wie andere dazu in der Lage bist, in uneindeutigen Situationen Entscheidungen zu treffen.

c) Überdurchschnittlich
In Hinblick auf deine Entschlussfreudigkeit zeigen die Ergebnisse, dass du im Vergleich zu anderen besser dazu in der Lage bist, in uneindeutigen Situationen Entscheidungen zu treffen.

5) AHA Impulsivität/Reflexivität

a) Unterdurchschnittlich
Die Ergebnisse zeigen, dass du im Arbeitsstil eher impulsiv, d.h. eher spontan und rasch vorgehst.

b) Durchschnittlich
Die Ergebnisse zeigen, dass du im Arbeitsstil weder besonders impulsiv noch besonders reflexiv vorgehst, d.h. du weder besonders stark spontan und rasch, noch besonders stark geplant vorgehst.

c) Überdurchschnittlich
Die Ergebnisse zeigen, dass du im Arbeitsstil eher reflexiv, d.h. sehr überlegt und geplant, vorgehst.

6) AHA Frustrationstoleranz

a) Unterdurchschnittlich
Von wiederholten negativen Rückmeldungen lässt du dich eher beeinflussen als andere, insofern dass du dir aufgrund dessen weniger zutraust.

b) Durchschnittlich
Von wiederholten negativen Rückmeldungen lässt du dich nicht wesentlich beeinflussen, das bedeutet, dass du über eine angemessene Frustrationstoleranz verfügst.

c) Überdurchschnittlich
Von wiederholten negativen Rückmeldungen lässt du dich weniger beeinflussen als andere, das bedeutet, dass du über eine gut ausgeprägte Frustrationstoleranz verfügst.

7) AHA Anspruchsniveau

   a) Unterdurchschnittlich
      Die Ergebnisse zeigen weiters, dass du bei deinen Angaben zu unrealistischen Zielsetzungen geneigt hast, wobei du dich in deinen Einschätzungen unterschätzt hast.
   b) Durchschnittlich
      Die Ergebnisse zeigen weiters, dass du realistische Ansprüche an deine eigene Leistung stellst und somit vernünftige und realisierbare Ziele setzt.

   c) Überdurchschnittlich
      Die Ergebnisse zeigen weiters, dass du bei deinen Angaben zu unrealistischen Zielsetzungen geneigt hast, wobei du dich aus irgendwelchen Gründen (z. B. Ermüdung, Misserfolgserwartung, udgl.) in deinen Einschätzungen überschätzt hast.

An dieser Stelle möchte ich mich abschließend noch einmal herzlich für deine Teilnahme bedanken! Wenn die Ergebnisse für dich wenig nachvollziehbar sind, und du gerne in einem Gespräch genauer darauf eingehen möchtest, kannst du mich unter der E-Mail-Adresse kathrin.stathis@univie.ac.at erreichen. In jedem Falle wünsche ich dir noch einen schönen Sommer!

Mit lieben Grüßen

Kathrin Stathis

Appendix C: Course of Study

Humanities and Cultural Sciences

Philosophie
Geschichte
Kunstgeschichte
Germanistik
Klassische Philologie
Politikwissenschaft
Finno-Ugrische Sprachwissenschaften
Theater-, Film- und Medienwissenschaften
Publizistik- und Kommunikationswissenschaft
Musik
Violine
Kultur- und Sozialanthropologie
Internationale Entwicklung

Social Sciences

Psychologie
Lehramt Englisch und Spanisch
Lehramt Biologie und Sport
Lehramt Deutsch und Biologie
Volksschullehramt
Soziale Arbeit

Medical Studies

Humanmedizin
Pharmazie

Law and Economic Studies

Rechtswissenschaften
Betriebswirtschaft
Betriebswirtschaftslehre mit Finance u. Accounting
Volkswirtschaft
Wirtschaftsgeschichte
Wirtschaftspädagogik
Wissensmanagement
Wirtschaftsrecht
Wirtschafts- und Sozialwissenschaften

Technical Studies

Elektrotechnik
Informatik
Bauingenieurwesen
Physik
Raumplanung und Raumordnung
Architektur
Software Design

Appendix D: Summary (German)

Das Ziel der vorliegenden Studie war zu überprüfen, ob Persönlichkeitsfaktoren einen Einfluss auf das Antwortverhalten in einem Test zur Wissensüberprüfung haben, welcher im Multiplen-Evaluations-Format gestaltet ist. Die Untersuchung dieser Frage dient vor dem Hintergrund uneindeutiger Befunde auch dazu, ein einheitliches Bild hinsichtlich dieser Debatte zu erlangen. Speziell wegen der Einschränkungen des Multiple-Choice-Formats erscheint es besonders relevant Alternativen dazu zu untersuchen. Die Vorteile des Multiple-Evaluation-Formats liegen verglichen mit dem Multiple-Choice-Format insbesondere in der Erhöhung der Reliabilität, der Minimierung der Ratewahrscheinlichkeit, der Berücksichtigung von Teilwissen und der Tatsache, dass es weiters den Gütekriterien „Zumutbarkeit" und „Fairness" gerecht wird. Außerdem könnte es für Lehrende ein wertvolles Werkzeug darstellen, um Informationen darüber zu erlangen, inwieweit die Studierenden die behandelten Inhalte verstanden haben.

Im Gegensatz zum Multiple-Choice-Format bietet Multiple-Evaluation der Testperson die Möglichkeit differenzierte Antworten zu geben. Anstatt lediglich die wahrscheinlichste Antwort zu wählen, gibt die Testperson für jede Antwortmöglichkeit an, für wie wahrscheinlich er/sie diese hält, dazu werden 100% über die gebotenen Antwortmöglichkeiten verteilt. Die Verrechnung erfolgt gemäß einer logarithmischen Auswertung, welche gewährleistet, dass das Testresultat maximiert wird, wenn von der Testperson realistische Einschätzungen vorgenommen werden. Aufgrund dieser Verrechnungsart kann es auch zu negativen Item- als auch Testscores kommen, wobei jedoch negative Resultate ausschließlich aufgrund von Fehlwissen oder Raten zustande kommen können. Dieser Umstand und auch, dass das Testresultat durch eine realistische Einschätzung des eigenen Wissens maximiert wird, motiviert Testpersonen ihr Wissen zu reflektieren und dieses auch ehrlich wiederzugeben.

Bei einem Test im Multiple-Evaluation-Format wird von den Testpersonen verlangt, dass diese ihren Wissenstand einschätzen und darüber hinaus diese Überlegungen in entsprechende numerische Werte ausdrücken. Es ist jedoch plausibel, dass Persönlichkeitsfaktoren diesen Vorgang beeinflussen könnten.

In der vorliegenden Studie nahmen 71 aktive Studierende verschiedener Fachrichtungen teil (36 Studentinnen und 35 Studenten). Im Gegensatz zu früheren Arbeiten wurden zur

Erfassung der Persönlichkeitsfaktoren anstatt von Fragebögen, die auf Selbsteinschätzungen beruhen, ein objektiver Persönlichkeitstest (AHA) und ein semi-projektives Verfahren (MMG) verwendet. Die Studierenden bearbeiteten einen Wortschatztest, AWST – Version 1, welcher im Multiple-Evaluation-Format programmiert wurde, und die Persönlichkeitsverfahren AHA und MMG, welche Persönlichkeitseigenschaften erfassen die mit Leistung in Verbindung stehen. Es wurde untersucht, ob die erfassten Persönlichkeitsvariablen, wie auch die Studienrichtung, einen Einfluss auf das Antwortverhalten der Studierenden haben. Dieser wurde anhand der Maße „Antwortsicherheit", welcher 1971 von Hansen postuliert wurde, und „Realismus" (postuliert von Dirkzwager, 2003) operationalisiert.

Hypothesen hinsichtlich der Einflüsse auf die Antwortsicherheit als auch Realismus wurden aufgestellt und anhand von Regressionsanalysen untersucht. Im Hinblick auf die Einflüsse auf die Antwortsicherheit wurde eine hierarchische Beta-Regression angewandt, da die Werte dieses Maßes ausschließlich zwischen 0 und 1 liegen können. Es zeigte sich, dass keine der erfassten Persönlichkeitseigenschaften einen Einfluss auf die Antwortsicherheit der Studierenden hatte. Die Fähigkeit der Studierenden ihr Wissen realistisch zu reflektieren, wurde anhand einer hierarchischen multiplen Regression geprüft. Die Ergebnisse zeigen auf, dass Studierende mit einem hohen Anspruchsniveau, also hohen Erwartungen an ihre eigene Leistung stellen, ihr Wissen realistischer einschätzten. Außerdem deuten die Ergebnisse auf einen Trend hin, dass misserfolgsorientierte Studierende bedachter in der Verteilung der Prozente vorgehen, welches sich in realistischeren Einschätzungen widerspiegelte.

Die vorliegenden Ergebnisse sprechen dafür, dass das Antwortformat Multiple-Evaluation als geeignete Alternative zu Multiple-Choice gesehen werden kann, da vornehmlich die Fähigkeit der Studierenden einen Einfluss auf das Antwortverhalten hat, und nicht von Persönlichkeitseigenschaften geleitet ist.

In Zukunft wäre es relevant zu überprüfen, ob diese Ergebnisse in einer echten, high-stakes, Testsituation repliziert werden würden. Außerdem könnte es von Bedeutung sein, das Antwortverhalten über einen längeren Zeitraum zu untersuchen, um Information darüber zu erhalten, ob und gegebenenfalls wie viel Zeit Testpersonen benötigen, um sich an das neue Antwortformat und dessen Eigenschaften zu gewöhnen, insbesondere deshalb, weil die meisten Studierenden an das Multiple-Choice-Format gewöhnt sein werden.

Curriculum vitae

PERSÖNLICHE DATEN

Name:	Kathrin Viola Stathis
Geboren am:	21. August 1984, in Wien
Staatsbürgerschaft:	Österreich
Kontakt:	kathrin.stathis@univie.ac.at

BERUFSERFAHRUNG

09/2013 – 02/2014	Universität Wien, Fakultät für Psychologie Studienassistenz im Projekt “Supervised Orientation Tutorium”
04/2013 – 06/2013	Österreichisches Institut für Sexualpädagogik freiwilliges Praktikum
01/2013 – 03/2013	AKH Wien, Universitätsklinik für Kinder- und Jugendpsychiatrie 6-Wochen-Praktikum
09/2012 – 02/2013	Universität Wien, Fakultät für Psychologie Studienassistenz im Projekt “Supervised Orientation Tutorium”
06/2011 – 08/2011	Ludwig Boltzmann Institut für Menschenrechte Quantitative Analyse – “Access to Justice – A sociological study on cases of discrimination in the EU”
Seit 09/2009	ZARA – Zivilcourage und Anti-Rassismus-Arbeit Statistische Auswertung von Workshops der Amnesty Academy
06/2006 – 03/2009	Firma Karin Gertner Managementcoaching, Persönlichkeits- und Personaltraining Unterstützung bei der Aufbereitung, Durchführung und Organisation von Seminaren

AUSBILDUNG

Seit 10/2006	Diplomstudium der Psychologie an der Universität Wien
09/1998 – 06/2003	HLTW 13 Bergheidengasse, Wien Ausbildungsschwerpunkt: Fremdsprachen und Wirtschaft
09/1995 – 06/1998	BRG IV Waltergasse, Wien 2. – 4. Klasse
09/1993 – 06/1995	German School New York White Plains, N.Y. 4. Klasse Volksschule, 1. Klasse Gymnasium
09/1990 – 06/1993	P.S. 81 (amerikanische Volksschule) Riverdale, N.Y. 1. – 3. Klasse Volkschule