Developing the Groningen Psychological Critical Thinking Task: A pilot study

T. J. Schröder

S3797007

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group Number: BT2122_1a_09

Supervisor: Marcella Fratescu

Second evaluator: Dr. Saleh Mohamed, PhD

In collaboration with:

Laura Escudero Gimeno, John Nagler, Leonie van Jaarsveld and Elisabeth van Nee

Month 01, 2022

A thesis is an aptitude test for students. The approval of the thesis is proof that the student has sufficient research and reporting skills to graduate, but does not guarantee the quality of the research and the results of the research as such, and the thesis is therefore not necessarily suitable to be used as an academic source to refer to. If you would like to know more about the research discussed in this thesis and any publications based on it, to which you could refer, please contact the supervisor mentioned.

Abstract

Critical thinking is regarded a crucial learning outcome by the APA. It is an important skill for psychologist and part of the curriculum of the psychology bachelor of the RUG. There is currently no assessment for it at the RUG. This within-subject correlational pilot study aims at contributing an open-ended measure for psychological critical thinking for psychology bachelor students at the RUG. The Groningen Psychological Critical Thinking Task (GPCTT) is an open-ended essay task, in which participants are asked to critically evaluate different materials before arriving at a conclusion. We administered the Psychological Critical Thinking Exam (PCTE) and the GPCTT, hypothesizing a strong positive correlation between the scores as both measures aim at measuring psychological critical thinking. We found no support for our hypothesis. Instead, we found a non-significant negative correlation between the scores. Possible improvements for the GPCTT are discussed.

Keywords: critical thinking, PCTE, psychological critical thinking

Developing the Groningen Psychological Critical Thinking Task: A pilot study

With the start of the COVID-19 crisis, conspiracy theories seem to have become more prominent in the media (McCarthy, Murphy, Sargeant & Williamson, 2021). People on social media platforms started to spread their ideas or beliefs and other people believing the ideas and theories about secret agendas and conspiracies. When the vaccines became available, many people started to spread misinformation about the side effects of vaccines. One might wonder why people believe these claims without questioning them or taking a more critical attitude (McCarthy et al., 2021). It illustrates that there is a need to critically evaluate the information we are presented with each day. Besides being important in everyday life, critical thinking is also crucial for psychologists (Lawson, 1999). The American Psychological Association regards critical thinking a crucial skill and learning outcome for psychology students (American Psychological Association, 2013). The relevance of critical thinking skills in the context of psychology becomes imminent when looking at the application in later work life. According to Lilienfeld et al. (2012), it is important to use critical thinking for evaluating research and adapting assessment to prevent harm due to mistreatment that is based solely on intuition instead of evidence. Even though it is an important skill there is a reasonable number of psychologists that do not utilize it in their practice (Bramlett, Murphy, Johnson, Wallingsford, & Hall, 2002).

At the University of Groningen, critical thinking has been implemented into different courses in the psychology bachelor programme as one of the desired learning outcomes, including Academic Skills and Theoretical Introduction to Research Methods (Ocasys: Academic skills, n.d.; Ocasys: A Theoretical Introduction to Research Methods, n.d.). Even though the University of Groningen considers critical thinking an important learning outcome, it is only assessed as part of different assignments throughout courses in the bachelor programme. There is currently no single assessment used to evaluate whether

teaching critical thinking skills was successful and students acquired this important skill in terms of a learning outcome of whole bachelor programme. We therefore developed a measure for psychological critical thinking for psychology bachelor students at the University of Groningen, called the Groningen Psychological Critical Thinking Task.

In general, the field of Critical Thinking identifies two main different approaches to defining critical thinking, the philosophical and cognitive psychological approach (Lewis & Smith, 1993). The philosophical approach views critical thinking as an ideal way of thinking, including qualities and characteristics, that relies on the rules of logic (Lewis & Smith, 1993; Sternberg, 1986). The cognitive psychological approach sees critical thinking more as a process consisting of steps or procedures (Lewis & Smith, 1993). While the philosophical approach is criticized for not being an accurate display of reality, the cognitive psychological approach is criticized for trying to reduce the complex concept of critical thinking to a list of steps or procedures (Sternberg, 1986).

Several definitions for critical thinking exist in the literature, differing mainly in what approach to critical thinking was taken. Critical thinking is for example defined by Ennis (1985, p.45), taking a philosophical approach, as “reflective and reasonable thinking that is focused on deciding what to believe or do” or by Facione (1990, p.3) as “judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or conceptual considerations upon which that judgment is based”. Using a cognitive approach, critical thinking is defined as “the use of those cognitive skills or strategies that increase the probability of a desirable outcome (Halpern, 1998, p. 450) or as “the mental processes, strategies, and representations people use to solve problems, make decisions, and learn new concepts” (Sternberg, 1986, p. 3). These examples show the differences in definitions depending on the approach taken. While definitions stemming from the philosophical approach emphasize criteria and judgment, the

definitions from a cognitive approach focus on critical thinking consisting of processes and skills.

As described in the course catalogue of the course Academic Skills at the RUG, students that successfully complete the course should be able to use critical thinking skills for recognizing fallacies and justifying statements. The literature provided for the course includes the book by Chatfield (2018) in which critical thinking is defined through using skills like reasoning and evaluating evidence. There seems to be consensus that critical thinking involves analysing, evaluating and coming to a conclusion or making a judgment. For the purpose of this study, critical thinking was defined as a "habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion" (Rhodes, 2010). This definition was previously used for the Critical Thinking VALUE Rubric (Rhodes, 2010) and combines the most common similarities between different definitions. However, since this is a definition for general critical thinking, we decided to extend the definition after the example of Lawsons definition of psychological critical thinking used for the Psychological Critical thinking test, one of the few existing measures for psychological critical thinking (Lawson, 1999): "Psychological critical thinking involves evaluating claims using the basic principles of psychological science." The definition therefore is: "Psychological Critical Thinking is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events based on principles of psychological science before accepting or formulating an opinion or conclusion."

Although the definitions and respected abilities that critical thinking entails differ depending on the research and approach, there is common ground both between approaches as well as within. For the purpose of our task, we incorporated five common aspects that critical thinking entails into the conceptualization. Analysing arguments, evidence or claims

(Ennis, 1985; Facione, 1990; Halpern, 1998; Paul, 1992), evaluating and judging (Case, 2005; Ennis, 1985; Facione, 1990) as well as solving problems and making decisions (Ennis, 1985; Halpern, 1998; Willingham, 2007) are identified as abilities or processes crucial to critical thinking in conceptualizations of the cognitive and philosophical approach. Identifying assumptions is identified as part of critical thinking by Ennis (1985) and Paul (1992) and considering both sides of an issue is an ability found in the work by Willingham (2007) as part of critical thinking process.

We incorporated the recommendations in the GPCTT Rubric (Appendix A), after the example of the Critical Thinking VALUE Rubric by the Association of American Colleges and Universities (AAC&U) (2010): The aspects of methodology and fallacies account for principles of psychological science that psychology students should have knowledge of (Lawson, 1999) and use to analyse evidence or claims (Ennis, 1985; Facione, 1990; Halpern, 1998). Assumption of authors accounts for identifying assumptions (Ennis, 1985; Paul, 1992) and synthesis was created to assess the forming of a conclusion and weighing evidence (Ennis, 1985; Facione, 1990; Halpern, 1998; Willingham, 2007). Bias of participants was included to evaluate whether students adhered to the instructions to only base their analysis on the provided evidence.

The number of different definitions and conceptualizations of critical thinking illustrate the complexity of the construct and the difficulty in defining it. This difficulty is reflected in the number of assessments of critical thinking as well. Some of the most commonly used assessments include the Cornell Critical Thinking Tests (Ennis & Millman, 2005), the California Critical Thinking Skills Test (Facione, 1990) and the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980). However, due to their design, those tests do not offer the opportunity for multiple views and limit the possibilities of analysis for the participant. According to Norris (1989), these traditional assessments might not be suited

for testing a construct like critical thinking, for example assessments using multiple-choice formats. This kind of assessment identifies right and wrong answers, which are determined by the developers and are likely to be influenced by the personal, political or cultural beliefs of the people developing the measure. This is the case for the PCTE as well, as it requires to come to "right" conclusion (Lawson, 1999). In this assessment a full score can only be achieved when identifying what the creator considers the main problem in the conclusion. Any further critique or identification of other issues besides the main one results in lower scores.

In the literature, multiple recommendations for developing more appropriate assessments for critical thinking can be identified. First, Assessment tasks should reflect real life problems (Bonk & Smith, 1998; Halpern, 1998) and require students to apply knowledge in a new context (Lewis & Smith,1993). Second, there should be sufficient information and evidence to take multiple views and there should be more than one possible and defensible position (Moss & Koziol, 1991). In concordance with that, evaluation of the responses should be based on the quality, not the correctness of the response (Moss & Koziol, 1991). Further, prompting critical thinking is more likely with inconsistent materials, compared to given materials with cohesive information (Fischer, Spiker, and Riedel, 2009). Also, tasks asking to judge information proved more useful in measuring critical thinking than those asking to only understand a given material (Fisher et al., 2009). Lastly, another suggestion is that critical thinking assessments should make student reasoning visible. For example, one recommendation for accomplishing this is to require students to provide a justification or explain the reasoning for their choice (Norris, 1989).

One measure for psychological critical thinking that incorporates these recommendations is described in a paper by DiBartolo, Duncan, Ly, & Rudnitsky (2016), called the "Messy Problem" It was the base and inspiration for our research. As the literature

recommends for assessments of critical thinking, the “Messy Problem” requires participants to use their skills on a novel and close to real-life problem, in this case the possible delay of school start time (Bonk & Smith, 1998; Halpern, 1998; Lewis & Smith,1993). It is an essay task that requires participants to evaluate different materials, including a fact-based newspaper article, and a corresponding opinion piece. In addition, they receive one page with two summaries of research articles. Some parts of the materials were altered so that they for example overstated the strength of their data in support of delaying start time of the school. The materials differ in their stance on the debate and therefore more likely to prompt critical thinking (Fischer, Spiker, and Riedel, 2009) and allow to take multiple different positions on the topic as recommended by Moss & Koziol (1991). Participants were asked explicitly to come to a conclusion, explain all of their results and put the limitations and strengths of the material in writing, which is an effective way to make participants reasoning visible (Norris, 1989). The scoring for this task, like the GPCTT Rubric, is based on the Critical Thinking VALUE Rubric (Rhodes, 2010) and was modified to fit with their assessment, creating a new scoring rubric. It allows for a quality assessment of the answers instead of scoring right and wrong answers, which is a better way of assessing a complex construct like critical thinking (Moss & Koziol, 1991).

The Groningen Psychological Critical Thinking Task

Incorporating the recommendations for assessment, we created the Groningen Psychological Critical Thinking Task, an open-ended assessment for psychological critical thinking, measuring transfer of skills to a less structured, more realistic environment. For the topic we chose resit exams as it is relevant and we assume that all students at least know about it, which might increase the chance that participants will be motivated enough to do their best on the test. In the GPCTT, participants are asked to imagine that they have to advise the Board of RUG on the decision to keep or abolish resits. They are presented with

three different sources and are required to critically evaluate them. An opinion article (Boomsma, 2018), a fact-based article, including a survey, (Boomsma & Siebelink, 2018) and the summary of a research article (Nijenkamp, Nieuwenstein, de Jong & Lorist, 2016). Similar to the paper by DiBartolo et al. (2016), we altered the material so that each aspect included in our own scoring rubric could be evaluated. We presented the participants with two shortened articles and a summary of a research paper. The participants are asked to give their advice in form of a final conclusion after critically analysing the material.

In this pilot study, we are aiming at developing a measure for psychological critical thinking specifically for psychology students that can be used to assess psychological critical thinking in other contexts and domains, with issues that are less obvious and closer to what they might encounter during their work later on. We are trying to answer the following question: To what extent does the Groningen Psychological Critical Thinking Task (GPCTT) measure the Psychological Critical Thinking skills in psychology bachelor students from the University of Groningen? We will use participants' scores for the PCTE as a comparison to the scores for the GPCTT to see if the scores show a correlation since they both aim at measuring the same concept. This will provide an initial understanding of the measure´s convergent validity.

Accordingly, our main hypothesis (H1) is that there will be a significant positive correlation between participant´s scores for the PCTE and the GPCTT.

Method

Participants

A total number of 22 Bachelor students of the University of Groningen participated in the current pilot of the study. Data was collected from first-year psychology students who received course credit for participating in the study. Of the initial number of participants, we excluded 4. Participants were excluded for filling in the PCTT and the GPCTT under 10

minutes (n = 0), identified as the minimum amount of time to read all provided material without starting on the task, for responding in Dutch (n = 1) and not completing the task (n = 3). Our sample consisted of 8 males (44.4%) and 10 females (55.6%). Participants' age ranged from 17-20 years (n = 14 (77.8 %)), 21-24 years (n = 3 (16.7%)) and 25+ (n = 1 (5.6%)). All participants were non - native English speakers. From our sample, 12 participants were from a western country (66.7 %), 2 from another country (11.1 %) and 4 did not answer the question (22.2%). All participants indicated that they put in their best effort.

Materials

Groningen Psychological Critical Thinking Task (GPCTT)

The GPCTT is an essay test that aims at measuring Psychological Critical Thinking. Participants were presented with a fictional scenario in which they are asked to advise the Board of the RUG in a current discussion about abolishing or keeping resit exams. Subsequently they are required to critically evaluate three sources on the topic of resits (Appendix B) and write an essay justifying their conclusion. The three sources consisted of one opinion piece (Boomsma, 2018), a reaction to this article which includes a survey on resits (Boomsma & Siebelink, 2018) and a summary of a research article about resits. The opinion-based article was from the newsletter Ukrant (Boomsma, 2018) and originally in favour of abolishing resits. The article was shortened and altered it so it includes statements by different professors and the mayor of Groningen, as well as statements by the author that did not include any reference. Those details were included to provide options to question assumptions and identifying fallacies. Further, we added a section that is in favour of keeping resit exams to give students the opportunity to take either stance. The fact-based article was a reaction to the opinion-article mentioned above and from the Ukrant as well (Boomsma & Siebelink, 2018). We shortened, altered and rewrote it so it would support resit exams as

there was limited existing material in favour of resit exams and we needed to give students the option to take either stance with sufficient evidence, including a survey which with first-year bachelor students. The research article is an existing research paper by Nijenkamp et al. (2016), in favour of abolishing resits as well. It contains an experiment on resits and spend study time. The presented experiment is about study investment with and without having the options for resits. It shows support to abolish resits as it appears that a resit opportunity reduces investment of study time (Nijenkamp et al., 2016). Both the second and third source were included to provide participants with different research designs and statistics to analyse. Participants are specifically instructed to read the material thoroughly and base their decision on the provided evidence.

Psychological Critical Thinking Exam (PCTE)

The PCTE measures psychological critical thinking by assessing participants ability to evaluate claims based on principles of psychological science, like falsifiability or causation vs. correlation (Lawson, 1999). Participants were presented with a shortened version of the Psychological Critical Thinking Exam (PCTE) (Lawson, Jordan-Fleming & Bodle, 2015) consisting of seven instead of fourteen research-related scenarios, as used in previous studies (Haw, 2011; Stark, 2012). Each scenario relates to one of seven questions developed by Lawson relating to principles of psychological science. In the PCTE, for each described scenario a conclusion was reached (Appendix C) and the participants had to state the main problem with the conclusion in written form, if applicable. For example: “Dr. Jones is testing a new treatment for cancer. He administered the treatment to a large sample of patients and kept track of who lived and who died after receiving the treatment. For each person who lived, he attributed the success to the treatment. For each person who died, he attributed the death to the severity of the person’s cancer. He concluded that his treatment was effective.”

(Lawson, 2015). For this example, a maximum score would be achieved by stating that the researcher does not make his claims falsifiable.

Study design & Procedure

This study is a within-subject correlational study. All participants had to complete both the PCTE and the GPCTT. The order of the tests was randomized to avoid a possible order effect. The study was approved by the Ethics Committee of the University of Groningen. First-year psychology students could access the study via the SONA system. Before the survey started, the participants would see a screen with information about the study and are informed about the amount of SONA credits they will receive and about the option to participate in a lottery with a chance to win 15 euros if they are non-SONA participants. Participants were asked for their informed consent before the start of the survey. They were then presented with either of the two critical thinking measures. The participants are instructed to state whether there is a problem with the conclusions being drawn in those scenarios, and to explain the problem.

The GPCTT starts with an instruction which states that participants are going to be presented with three articles about resit exams at the University of Groningen. The task is to imagine they are an adviser of the Board, tasked with analysing research about resits and to advise the Board on their final conclusion on whether or not to abolish resit exams. Participants are instructed to read the articles thoroughly and to write an essay consisting of an introduction, body, and conclusion in which they critically evaluate the articles and come to a conclusion about resits. The articles presented next include an opinion-based article, a fact-based article, and a research article respectively (Appendix B). After finishing both tests, participants were asked to indicate if they did or did not put their best effort into the tasks. They were also asked about some demographic information (age, gender, major, native language, ethnicity).

Training

A pilot study was conducted, which also served as training for the raters. It provided the opportunity to gather feedback, clarify the task and adjust the rubric. The pilot study contained 6 participants that were recruited by the research team. Each rater independently scored the participant´s answers for the GPCTT. Differences in score were then discussed until consensus was reached. In addition, the raters familiarized themselves with the scoring of the PCTE.

Results

In total, 22 people participated in this study. The data analysis was based on 18 participants. The participants completed both the PCTE and GPCTT. For both measures, the individual scores on each question/aspect were combined into a final score for each test.

Scoring

For the PCTE, the answers were scored independently by two blinded raters. Participants were scored on a scale of 0 to 3. A score of 0 was given for not identifying a problem, 1 for mentioning a problem but misidentifying it, 2 for mentioning more than just the main problem and 3 for only identifying the main problem with the conclusion. Afterwards, disagreement in scoring was resolved so that one final score for each question was given. Hence, for this task a maximum score of 21 and a minimum score of 0 could be reached.

For the GPCTT, each essay was scored independently by two blinded raters, based on the GPCTT-rubric (Appendix A) that includes the aspects Methodology, Fallacy, Assumption of Authors, Bias of Participant and Synthesis. Under the aspect of Methodology, participants were expected to evaluate internal or external validity, the research design or the statistics. For the aspect of Fallacy, participants were required to identify fallacies of reasoning in the material. For example, the status quo bias and appeal to authority fallacy. Under the aspect of

Assumption of authors, the participants were supposed to question assumptions made by the authors. For example, they had to examine whether a source for a claim was given. Under Bias of Participants, it was evaluated whether participants only used the information from the material and for the aspect of synthesis participants had to combine and weigh evidence they gathered from the provided sources.

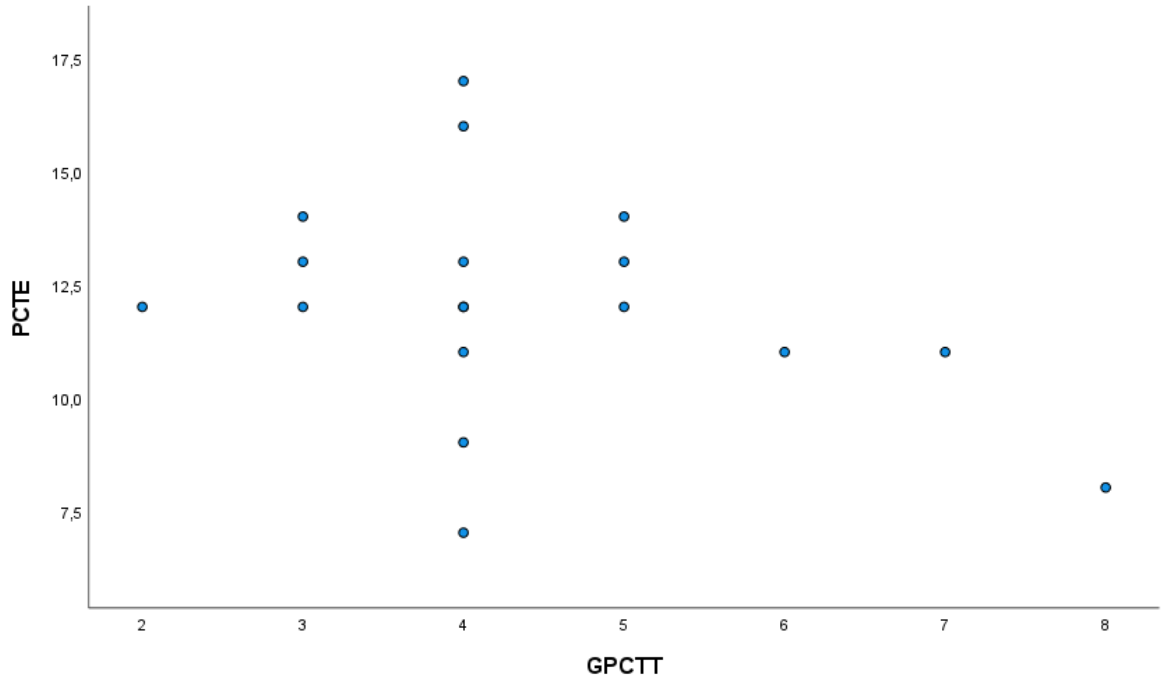
For the aspects Methodology, Fallacy and Assumption of Authors, the participant could score on a scale including 0 (Subpar), 1 (Benchmark), 2 (Milestone), 3 (Capstone). A score of 0 was given to participants for mentioning the aspect but making a mistake, 1 point was awarded for neither mentioning nor misinterpreting information regarding an aspect, 2 points were given for mentioning information regarding an aspect at least once and 3 points were given for mentioning an aspect at least twice. For Bias of Participant and Synthesis, each participant could either score a 0 (Subpar) or 2 (Milestone). For Bias of Participants, a 0 was given for including information that was not part of the provided material and a 2 was given if the content of the essay only contained information from the material. Therefore, the maximum score a participant could get is 13 and the minimum 0. For both measures the scores of each aspect/question were combined into a final score so that each participant has a final score for the PCTE and GPCTT, which was used for analysis.

Analysis

The assumption of normality of the distribution for both measures was assessed with the Shapiro-Wilk Test (Shapiro & Wilk, 1965). The test was non-significant for the PCTE. Hence, we cannot reject the null hypothesis that the data comes from a normally distributed population (Appendix D, Figure 1). However, for the GPCTT the Shapiro-Wilk Test was significant (W=.89, p= .038). We therefore found evidence that the data is not normally distributed (Appendix D, Figure 2). Further, the data is ordinal and there is a monotonic relationship between the scores.

Figure 1

Correlation between PCTE and GPCTT scores.



Note. Scatterplot representing the correlation between the total scores of each participant for the GPCTT (x-axis) and the PCTE (y-axis).

As a result of that, Spearman´s rank correlation (Spearman, 1904) was chosen as the measure of correlation because it is a non-paramedic test and therefore does not assume normality and our data assumptions of ordinal data and monotonicity fort this measure. Spearman´s rho ($r(16) = -.307$, $p = .215$) showed a small to moderate negative correlation between the scores on the PCTE and GPCTT (Cohen, 1992). This correlation is not statistically significant, which means that we cannot reject the null hypothesis for this test, which hypothesises that there is no monotonic relationship between the two variables.

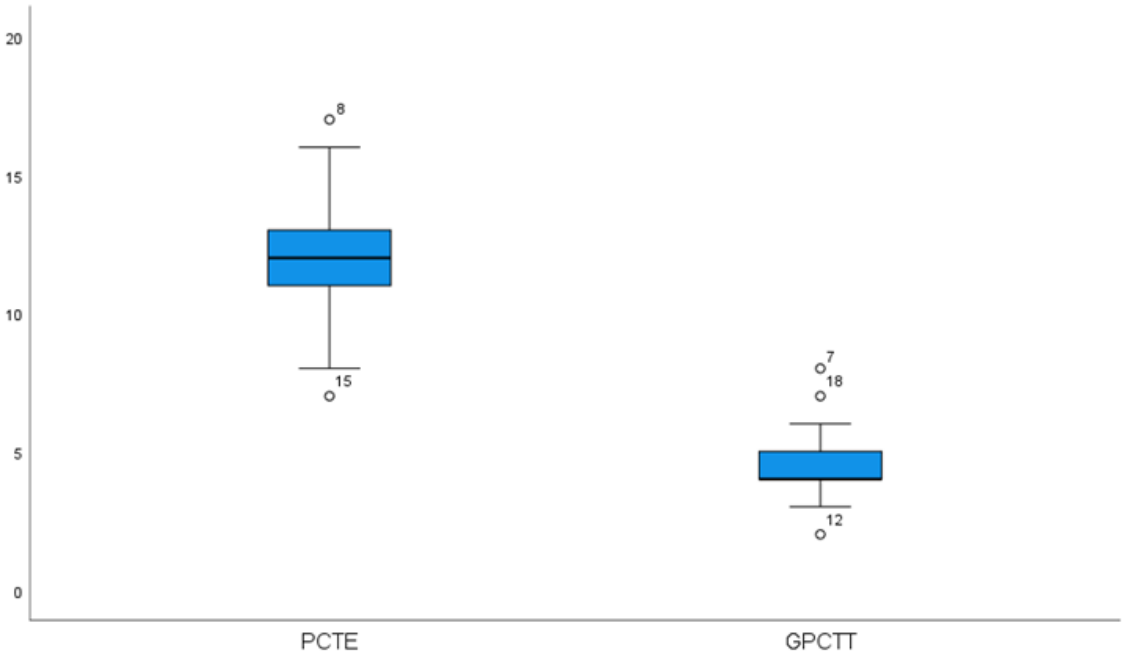
Interrater-reliability was first assessed by percentage agreement between rater 1 and rater 2. For the PCTE both raters agreed on 99 out of 126 responses (78,57%).

For the GPCTT, the raters agreed on 76 out of 90 responses (84,44%). In addition to that Cohens Weighted Kappa was used for a more nuanced assessment (Cohen, 1968). In total, for the PCTE interrater-reliability was good with kw=.61 (p<.001) (Fleiss, Levin & Paik, 2003). For the GPCTT, the was fair with kw=.42 (*p*=.002) (Fleiss, Levin & Paik, 2003)

Internal consistency for the five items of the GPCTT was assessed through Cronbach´s Alpha (Cronbach, 1951) and is considered unacceptable (α=.20) (George & Mallery, 2003; Tavakol & Dennick, 2011). After removing the aspect of Assumption of Authors, Cronbach´s Alpha would be α=.46 (Appendix D, Table 1). In comparison, the Cronbach´s Alpha (Cronbach, 1951) for the six items of the PCTE was unacceptable as well (α =.22) (George & Mallery, 2003; Tavkol & Dennick, 2011). Item 6 of the PCTE was automatically removed by the SPSS software for this calculation due to zero variance in the scores for this item.

Figure 2

Variance in PCTE and GPCTT scores.



Note. Boxplot of the variance in the total scores for the PCTE and GPCTT. Outliers are denoted with numbers representing the assigned number of the participant with the respected score.

Discussion

The aim of this study was to develop an open-ended measure for critical thinking for psychology bachelor students of the University of Groningen. We hypothesized that there will be a significant positive correlation between the PCTE and GPCTT scores as they both aim at measuring the same concept. However, our results did not support this hypothesis. We found a moderate negative relationship between the scores. Due to its statistical insignificance, we cannot exclude that this correlation is by chance. It also has to be considered that our sample size is fairly small and correlation coefficients like Spearman´s rho become less reliable the smaller the sample is (Hackshaw, 2008). The rest of the statistical analysis should be put into context too. First, Cronbach´s Alpha is also influenced by sample size and small variances in scores. Since this pilot study has a limited sample size and the variance in scores was low (see figure 2), it could have affected the Cronbach's alpha (Sijtsma, 2009). Further, the alpha can be influenced by length of a study or number of items on a test. It should also be noted that Cronbach´s Alpha does not test for dimensionality and can therefore not be used to determine whether our measure does only measure critical thinking. But for factor analysis the sample is not sufficient (Hackshaw, 2008; Sijitsma, 2009). Next, the weighted kappa statistic might have been influenced by our scoring rubric. For the weighted kappa, not only the number of disagreements but also the level of disagreement is taken into account (Cohen, 1986). However, for two of our aspects in the GPCTT Rubric, participants could only score between 0 and 2 points. That means that for those two aspects, the difference in disagreement would always be 2 points. These differences in scoring could potentially influence the weighted kappa statistic.

Comparison of assessment methods

There are multiple possible reasons as to why we found a negative correlation between the PCTE and GPCTT scores. First, the PCTE and the GPCTT measure two different kinds of transfer of critical thinking skills. The PCTE presents participants already with the most important information they need and points them into what they have to examine, assessing near transfer. In opposite to that, our measure is ill-structured and measures far transfer of knowledge by using a less organized, more realistic problem that students generally do not encounter in the classroom. This is a novel situation to the participants and could be part of the reason that they score lower on the GPCTT. Second, both tests use a different approach to assessing psychological critical thinking, both in terms of the task as well as the scoring. In the PCTE, PCT is assessed by making participants state the problem with a given conclusion while in the GPCTT, they are required to analyse materials and form a conclusion themselves. For the scoring, the PCTE identifies right and wrong answers. Participants have to identify one main problem to score the maximum point for one question. If they identify more, they score lower. In comparison, the GPCTT rewards extensive and comprehensive answers and does not define one correct position or conclusion that has to be reached, incorporating the recommendations for critical thinking assessment. Further, the PCTE and the GPCTT conceptualize psychological critical thinking differently. The PCTE focuses on principles of psychological science, like falsifiability (Lawson, 1999) while for the GPCTT, in addition to evaluating the evidence, skills like withholding subjective judgment and synthesizing the results of analysis are assessed as well. In addition, the quality of analysis and evaluation is taken into account by scoring participants higher for more thorough analysis and evaluation. It is therefore plausible that we could not establish a positive correlation between the two.

Revising the GPCTT

Analysing the responses for the GPCTT**,** there are multiple things that can be improved about our measure and the corresponding rubric. Students might have performed better on the GPCTT because the scoring system is different for both tests. The first three aspects of the GPCTT Rubric are scored with one point for not mentioning and not misinterpreting information regarding an aspect. That means that even if a participant for example writes a few sentences containing no analysis at all, they would still score at least three points. For the PCTE, not identifying a problem equals no points. Therefore, the GPCTT rubric should be adjusted, so that not mentioning an aspect at all would score a 0. Accordingly, if a participant would for example misinterpret the statistics or commit a fallacy themselves, one point would be subtracted.

In addition to the scoring, the wording of the aspect synthesis should be revised too. It was not clear for the raters what exactly would be defined as a "sufficient" way of weighing and combining evidence. It was then agreed during the grading for this pilot, that the quality of the synthesis should not be considered but if the participant weighed and combined evidence at all. In order to avoid this in the future, the aspect of synthesis could be changed from a dichotomous item to one that differentiates between at least three scores. Awarding 0 points for not synthesising at all, 1 point for doing a basic or insufficient synthesis and 2 for doing a sufficient synthesis. "Sufficient" and "insufficient" in this context then have to be defined in detail in the revised rubric. A sufficient synthesis could, for example, entail weighing and combining evidence for both their own and opposite position. In comparison, not connecting the evidence or only taking into account one position could be ruled an insufficient synthesis.

Throughout the scoring of the GPCTT, issues with the task itself became evident as well. Both raters reported that the essays written by the participants did not reach our

expectations. First of all, many participants included their own opinions or arguments which had no backing in the provided material, although they have been instructed to exclusively use the information from the articles. Second, the level of evaluation of the provided material was insufficient. From all 18 participants, only five evaluated the methodology at all. For the aspect fallacy, 16 participants scored a 1 for neither using nor mentioning the fallacies. The other two scored a 0 for using one or more fallacies as valid arguments. For the aspect of assumption of the authors, 14 participants scored a 0 for committing a fallacy themselves, three scored 1 point for not using but also not evaluating them and one participant questioned an assumption made by an author. This is also reflected in the distribution of the final scores. Roughly 44% of the participants scored 4 out of 13 points for the test. There are multiple possible explanations for these findings. All of our students are first year students and have not yet finished the course Academic Skills. Hence, they have not yet finished the courses aiming at teaching critical thinking skills. Perhaps in a replication of the study possible differences between third- and first-year students´ scores for specific aspects could be explored to identify whether the low scores in our study are due to flaws in the GPCTT or lack of knowledge. Another possible explanation could be that the instructions might have not been clear enough. We could clarify the instructions after the example of the instructions given for the task by DiBartolo et al. (2016). Here participants are asked to explain their analysis and arguments (DiBartolo et al., 2016). This would be in line with Norris´ (1989) recommendation for instructions that allow to follow the thought process of participants better.

Further, we found that some essays included an emotional component. Participants were passionate about the relevance of resits and discussed topics like mental health and the effects of stress on the body, relating to resit exams. A similar finding was reported by DiBartolo et al. (2016). In their essays they did find people responding in an emotional matter

but described the phenomenon as rare. We assume that this kind of response has been prompted by asking participants for their advice. We received the feedback that this was misunderstood as the request to give a personal opinion. Therefore, it might be useful to change the wording and remove the instructions asking participants to advise the board.

For the future, this study should be replicated which ideally would include a bigger, more heterogenous sample. It would be interesting to see whether the GPCTT can differentiate between different groups. Since the GPCTT aims at measuring critical thinking in psychology students, the main goal should be to have it successfully differentiate between psychology and non-psychology students as well as different study years. This could potentially help establish construct validity.

Conclusion

The goal of this pilot study was to create a new open-ended measure for psychological critical thinking for bachelor psychology students. However, it was not enough to establish convergent validity for the Groningen Psychological Critical Thinking Task. It does however illustrate the potential of the GPCTT as a measure for psychological critical thinking. It could be a useful tool to assess PCT as a learning outcome on a university level. As an essay task, the GPCTT can be used to assess PCT in a more nuanced way and give insight into common mistakes among students and patterns in thinking. It should be noted that neither the “Messy problem” by DiBartolo et al. (2016) nor the original Critical Thinking VALUE Rubric (Rhodes, 2010) was meant for grading students or to be a standardized measure. Instead, they are meant for quality assessment of critical thinking skills and to gain an insight into students learning process (DiBartolo et al, 2016; Rhodes, 2010). A similar view could be taken for the GPCTT. The pilot study was useful as it provides us with insights into possible areas of improvement for both the task and scoring of our measure that can be used to further develop this task.

References

American Psychological Association. (2013). Guidelines for the undergraduate psychology major: Version 2.0. *American Psychologist, 71*(2), 102111. https://doi.org/10.1037/a0037562

Boomsma, C. (2018). “Get rid of resits” *Ukrant*

Boomsma, C., Siebelink, R. (2018). „No resits? More stress.” *Ukant*

Bramlett, R. K., Murphy, J. J., Johnson, J., Wallingsford, L., & Hall, J. D. (2002). Contemporary practices in school psychology: A national survey of roles and referral problems. *Psychology in the Schools,* 39, 327–335.

Chatfield, T. (2018). *Critical thinking: your guide to effective argument, successful analysis & independent study* (First). Sage.

Cohen, J. (1968). “Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit.” *Psychological Bulletin* 70 (4): 213—220. doi:10.1037/h0026256.

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

Cronbach, Lee J. (1951). "Coefficient alpha and the internal structure of tests". *Psychometrika*. Springer Science and Business Media LLC. 16 (3): 297–334.

DiBartolo, P. M., Duncan, L. E., Ly, M., & Rudnitsky, A. N. (2016). Using a “Messy” Problem as a Departmental Assessment of Undergraduates’ Ability to Think Like Psychologists. *Journal of Assessment and Institutional Effectiveness*, *6*(2), 191–211. https://doi.org/10.5325/jasseinsteffe.6.2.0191

Douglas, K. M. (2021). COVID-19 conspiracy theories. *Group Processes & Intergroup Relations*, *24*(2), 270–275. https://doi.org/10.1177/1368430220982068

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. Educational Leadership, 43(2), 44–48.

Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Millbrae, CA: The California Academic Press

Fleiss, J.L., Levin, B. and Paik, M.C. (2003). The Measurement of Interrater Agreement. In Statistical Methods for Rates and Proportions. *Wiley Series in Probability and Statistics,* 598 626. https://doi.org/10.1002/0471445428.ch18

George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. *11.0 update* (4th ed.). Boston: Allyn & Bacon.

Hackshaw A. (2008). Small studies: strengths and limitations. *The European respiratory journal*, *32*(5), 1141–1143. https://doi.org/10.1183/09031936.00136408

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist, 53(4).*

Halpern, D. F. (2001) Assessing the effectiveness of critical thinking instruction. *The Journal of General Education, 50(4*), 270–286.

Haw, J. (2011). Improving psychological critical thinking in australian university students. Australian Journal of Psychology, 63(3), 150–153. https://doi.org/10.1111/j.1742 9536.2011.00018.

Lawson, T. J. (1999). Assessing psychological critical thinking as a learning outcome for psychology majors. *Teaching of Psychology*, *26,* 207–209. doi:10.1207/S15328023TOP260311

Lawson, T. J., Jordan-Fleming, M. K., & Bodle, J. H. (2015). Measuring Psychological Critical Thinking: An Update. *Teaching of Psychology*, *42*(3), 248–253. https://doi.org/10.1177/0098628315587624

Lewis, A., & Smith, D. (1993). Defining higher order thinking. *Theory into Practice, 32(3)*, 131–137.

Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology*, *50*(1), 7-36.

Paul, R. W. (1992). Critical thinking: What, why, and how? New Directions for Community Colleges, 1992(77), 3–24.

McCarthy, M., Murphy, K., Sargeant, E., & Williamson, H. (2021). Examining the relationship between conspiracy theories and COVID-19 vaccine hesitancy: A mediating role for perceived health threats, trust, and anomie? *Anal Soc Issues Public Policy*.

Nijenkamp, R., Nieuwenstein, M. R., de Jong, R., & Lorist, M. M. (2016). Do Resit Exams Promote Lower Investments of Study Time? Theory and Data from a Laboratory Study. *PloS one*, *11*(10), e0161708. https://doi.org/10.1371/journal.pone.0161708

Norris, S. P. (1989). Can we test validly for critical thinking? Educational Researcher, 18(9), 21-–26.

Ocasys: A Theoretical Introduction to Research Methods. (n.d.). Rug.nl. Retrieved January 6, 2022, from https://www.rug.nl/ocasys/gmw/vak/show?code=PSBE1-27

Ocasys: Academic skills. (n.d.). Rug.nl. Retrieved January 6, 2022, from https://www.rug.nl/ocasys/gmw/vak/show?code=PSBE1-25

Rhodes, T. (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics.* Washington, DC: Association of American Colleges and Universities.

Shapiro, S.S. and Wilk, M.B. (1965) An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52, 591-611. https://doi.org/10.1093/biomet/52.3-4.591

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101

Stark, E. (2012). Enhancing and assessing critical thinking in a psychological research methods course. Teaching of Psychology, 39(2), 107–112. https://doi.org/10.1177/0098628312437725

Sternberg, R. J. (1986). Critical thinking: Its nature, measurement, and improvement National Institute of Education. Retrieved from http://eric.ed.gov/PDFS/ED272882.pdf.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Watson, G. & Glaser, E. M. (1980) Watson-Glaser critical thinking appraisal. *The Psychological Corporation.*

Willingham, D. T. (2007). Critical thinking: Why is it so hard to teach? American Educator, 8–19.

Appendix A

The GPCTT Rubric

Aspect of CT	Capstone 3	Milestone 2	Benchmark 1	Subpar 0
Methodology	The participant takes into account methodology at least twice in their essay. **Example: Internal validity**: The participant mentioned that the experiment has a higher internal validity than the survey. **Ecological validity:** The participant mentioned that the ecological validity of the experiment is lower due to an artificial setting.	The participant takes into account methodology at least once in their essay.	The participant does not take into account any items relating to methodology but also does not make an invalid argument regarding the methodology.	The participant misinterprets items relating to methodology. **Example:** The participant mentioned a high ecological validity for the experiment.
Fallacy	At least both status-quo bias and appeal to authority fallacy are identified. **status-quo bias**: Option: The participant mentions that the	Either the Status-quo bias or appeal to authority fallacy is identified.	Identification of 0 fallacies of reasoning mentioned below and do not use them.	Usage of at least one of the fallacies as valid arguments. **status-quo bias**: Option: The participant mentions that the argument of "keeping the resits because it has always been like that" is a **valid** argument. **appeal to authority fallacy**:

	argument of "keeping the resits because it has always been like that" is a non-valid argument. **appeal to authority fallacy**: Option: The participant mentions that the mayor of Groningen has the opinion to keep the resits, but identifies this as a not valid argument, *(because the mayor is not an expert).*			Option: The participant mentions that the mayor of Groningen has the opinion to keep the resits, and identifies this as a **valid** argument.
Assumptions of authors	The participant considers at least 2 assumptions of the authors, including sources for statements and facts and considers them non-valid. **Example:** "It takes time, and the students might suffer delays but without this option students	The participant considers at least one of the assumptions of the authors as non-valid. **Example:** *"It takes time, and the students might suffer delays but without this option students have a higher chance of dropping out. "* *OR* "When you fail an exam, you want a second chance as quickly as possible."	The participant does not mention the possible bias at all and does not use it as a valid argument.	The participants use assumptions of the authors as a valid argument.

	have a higher chance of dropping out. “ *AND* *“When you fail an exam, you want a second chance as quickly as possible.”*			
Bias of participants		The participant only uses information/evidence provided in the materials to evaluate and support their conclusions.		The participant uses information/evidence not provided in the materials in their essay.
Synthesis		The participant shows the ability to combine evidence and weigh contradictory evidence in taking their final stance.		The participant does not show sufficient ability to weigh or combine evidence that is in line with, but also contradicting their position.

This rubric was created on the basis of the Association of American Colleges and Universities (AAC&U) Critical Thinking VALUE Rubric. Retrieved from https://www.aacu.org/value-rubrics

Appendix B

The GPCTT instructions, task and provided material

You will now be presented with three articles on the topic of resits at the University of Groningen (RUG). Currently, there is an ongoing discussion among Board Members of the University about whether resits should be kept or abolished. Imagine you are a representative of the Board, tasked with analysing research on this topic. Based on this research, you need to advise the Board on their final decision. So, after thoroughly reading the articles on this topic, please write an essay (introduction, body, conclusion) in which you critically analyse the articles and come to a final conclusion about whether resits should be kept or abolished at the University of Groningen. This task does not have a time limit, however it should take you about 60 minutes.

The University of Groningen is a university in the Netherlands with approximately 32 thousand students. Each student receives at least one resit opportunity for each course. For most faculties at the RUG the resits take place at the end of each block.

The University of Groningen is a university in the Netherlands with approximately 32 thousand students. Each student receives at least one resit opportunity for each course. For most faculties at the RUG the resits take place at the end of each block.

Get rid of resits

Nelly McTally, 2020 in the Ukrant

When you fail an exam, you want a second chance as quickly as possible. Educational experts say the RUG should stop offering these second chances. Scheduling a second chance before the first one has passed is asking for trouble, Jansen says. ‘It leads to students getting way too strategic about their exams. They figure that if at first they don’t succeed, they’ll just take the test again.’

‘We shouldn’t underestimate the psychological effect’, says Nienke Renting, from the Faculty of Economics and Business. ‘If students only get one chance, they’ll actually work harder. They’ll do everything they can to pass, which they don’t do when they get a second chance.’

On the other hand, this is an incredibly efficient system. It takes time, and the students might suffer delays but without this option students have a higher chance of dropping out. Even though it takes time for the teachers to create the tests, without resit exams many students who did not pass the first exam due to unforeseen circumstances suffer even more delay. One spokesperson for resit opportunities is the Mayor of Groningen: ‘I used to love resits during my time at the university. They are useful and needed. Besides, doesn't everyone deserve a second chance?’, he said during an interview.

Resits are best planned at the end of the year, which allows students to focus solely on studying for them. It’s annoying for people who’ve planned vacations, but it should be annoying. ‘We have to make passing the norm. Right now, failing is the norm’, says Cohen-Schotanus.

In conclusion, the tests should be used to steer education. Plan many, forcing students to keep studying. Offer students the opportunity to compensate for bad grades so they don’t get hung up on a single failed test. Offer cumulative testing, to ensure that a later good grade makes up for an earlier poor grade. And finally, make taking a resit as unappealing as possible.

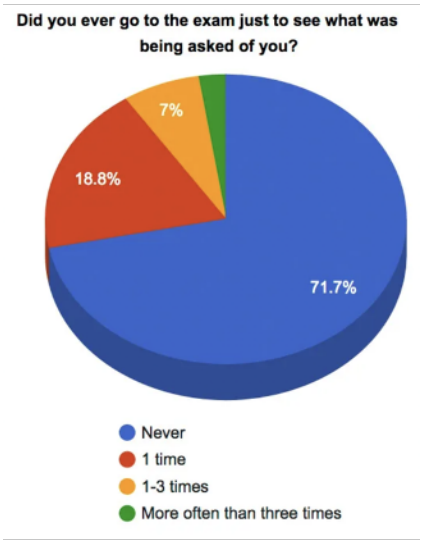
No more resits? More stress (A reaction to “Get rid of resits”)

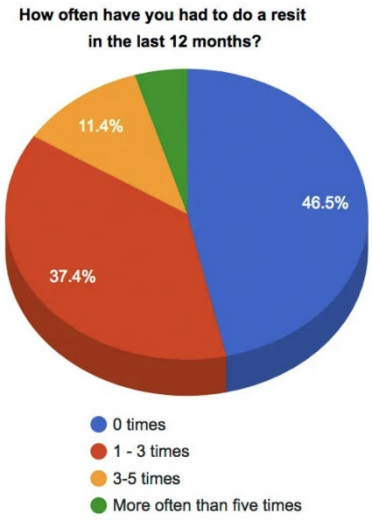
Julian Weber, 2020 in the Ukrant

Is it true that students are ‘abusing’ the resits? Are they indeed using exams to scope out what is being asked of them? And do they think it’s a good idea to discourage students

from banking on resits?

The UKrant asked 820 first-year students about their experience with an attitude to resits. The following graphs show the results.





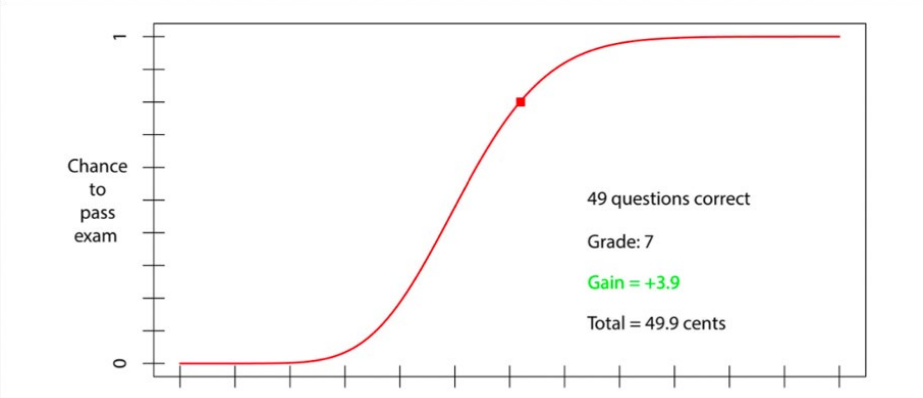
Then the main question: should resits be discouraged by scheduling them at unusual times? A fair number of students (27.1%) don't think the idea is too bad. The most used argument is that the increase in pressure will force students to start studying earlier and take exams more seriously.

Nevertheless, almost three out of four students are against the measure. 'It would only cause more stress, and the pressure to perform is high enough already', many of them argue. Or: an exam is just a snapshot. Failure happens. Quite a few students argue that they shouldn't be punished for unforeseen circumstances, such as illness, accidents, or blackouts. Also, taking resits has always been like this, so why should we change it now?

Do Resit Exams Promote Lower Investments of Study Time?

Author: Rob Nijenkamp, et al. (2012)

In 2012, Nijenkamp and colleagues did an experiment to test the effect of resit exams on the amount of study time. Participants were asked to invest fictional study time for a fictional exam, 50 psychology students for the University of Groningen participated. The students would sit behind computers and were shown the graph below which depicts the relationship between the study time investment (x-axis) and the probability of passing a 60-item multiple choice exam (y-axis).



In the task, the participants had to indicate their choice of study-time investment for passing an exam. To select the desired amount of study time, participants had to move a cursor along the curve in the graph (like the red dot in the figure).

The availability of a resit exam was manipulated within-subjects in a blocked design, such that each participant completed 6 blocks of 60 trials. During a trial the participants would be shown the graph to indicate how much time they wished to invest, then the screen would show whether or not they passed the exam. When a passing grade was obtained, the participants would move on to the next trial, and only in the resit condition they would move on to the resit exam when receiving a failing grade.

Three blocks included the option for a resit exam, whereas for the other three blocks they were granted only the first exam. The resit and no-resit conditions were alternated throughout the blocks. In addition, participants were informed that they could earn real money such that they would obtain a reward of 10 cents if they passed the exam, with the

cost of study time being 1 cent per time unit invested. If they did not pass the exam, they would not get a reward. The results confirmed the hypothesis of the researchers; the prospect of a resit exam was found to promote lower investment of study time for the first exam.

Appendix C

The PCTE (Lawson, 1999; Lawson et al, 2015)

For the following examples, state whether or not there is a problem with the person's conclusions and explain the problem (if there is one). Think of how it may violate a rule of research. The following questions take approximately 20 minutes.

1. A researcher located 100 pairs of identical twins who had been reared apart and reunited them. The twins discovered that they had an extraordinary number of things in common. For example, one set discovered that, among other things, both have a daughter named Cindy, a workshop where they restore old cars, cocker spaniels, and they both crush their beer cans with their left hands. The other pairs of twins also had numerous similarities. The researcher concluded that these stories are evidence that our personalities are influenced by genetics.

2. A researcher tested a new drug designed to decrease depression. She gave it to 100 clinically depressed patients and discovered that their average level of depression, as measured by a standardized depression inventory, declined after 4 months of taking the drug. She concluded that the drug reduces depression.

3. A survey research company hired by the Democratic party contacted a large, representative sample of Americans to examine their beliefs about new legislation designed to reduce crime. They asked the respondents, "Would you agree that this new legislation that will reduce crime and make our streets safer is a good piece of legislation for America?" Close to 92% of the sample answered "yes." The research company concluded that most Americans support the legislation.

4. An animal advocacy group studied the effects of animal ownership on owners' health. They studied a large, representative sample of older adults and obtained their medical records. Their findings showed that adults who had owned pets (i.e., dogs or cats) for a longer period of time had fewer medical problems than did adults who never owned pets or

owned them for a shorter time period. They concluded that owning pets decreases the likelihood of developing health problems.

5. Researchers randomly assigned male juvenile offenders to conditions where they watched either violent or nonviolent films. They discovered that those in the violent film group were less likely to go for help when they witnessed a later real-life violent episode than those in the nonviolent film group. On that basis, the researchers concluded that violent films harden all film-goers to real-life aggression.

6. Dr. Jones is testing a new treatment for cancer. He administered the treatment to a large sample of patients and kept track of who lived and who died after receiving the treatment. For each person who lived, he attributed the success to the treatment. For each person who died, he attributed the death to the severity of the person's cancer. He concluded that his treatment was effective.

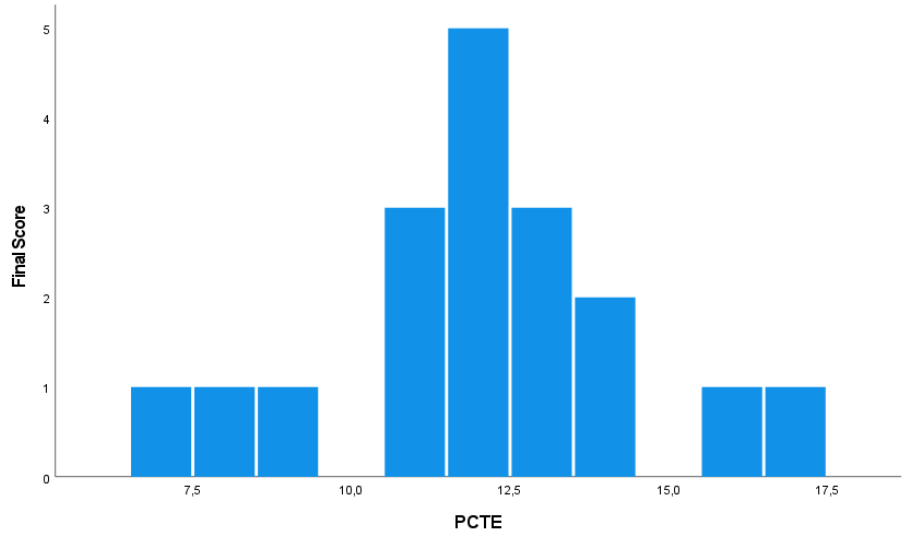
7. A group of biological researchers concluded that they have found THE cause of alcoholism. They discovered that alcoholics do not have a small cluster of cells, common to non-alcoholics, located near the hypothalamus. They have also demonstrated that destroying this area of the brain in normal rats caused them to develop a preference for alcohol in their water. Moreover, in another study they found that normal humans who had this part of the brain damaged in accidents later became alcoholics.

Appendix D

Results of statistical analysis

Figure 1

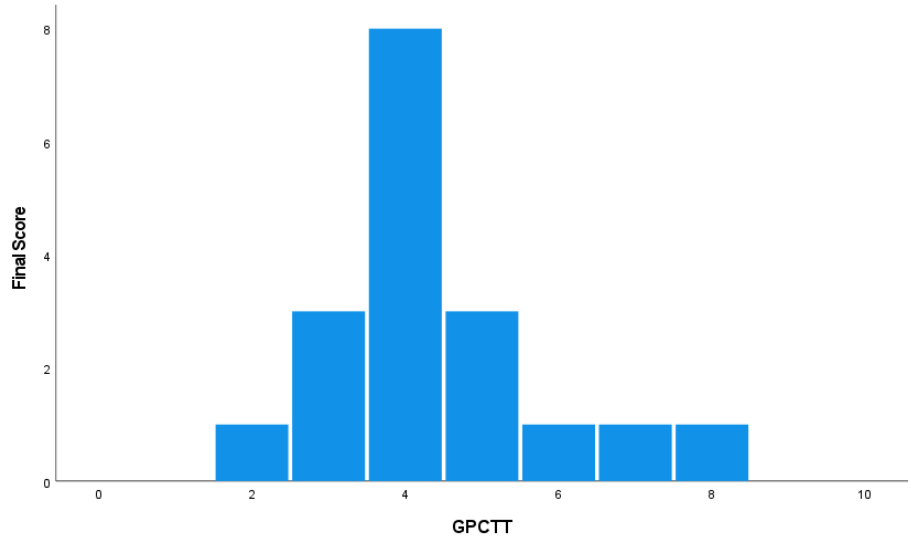
Histogram of PCTE scores



Note. Possible scores range from 0 to 21.

Figure 2

Histogram of GPCTT scores



Note. Possible scores range from 0 to 13.

Table 1

Item-Total Statistics

Aspect	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach´s alpha if item deleted
Methodology	3.1	1.1	.532	-.391
Fallacy	3.5	2.0	.000	.233
Assumption of Authors	4.1	2.2	-.244	.455
Bias of Participants	4.2	1.3	.264	-.053
Synthesis	2.7	1.6	.021	.266