The Creation of the Groningen Psychological Critical Thinking Task: A Pilot Study

Elisabeth van Nee

S3369781

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group number: BT2122_1a_09

Supervisor: drs. Marcella Fratescu

Second evaluator: Dr. Saleh Mohamed

In collaboration with: Laura Escudero Gimeno, Leonie van Jaarsveld, John Nagler, and Thomma Schröder

Month 02, 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

The goal of this project was to create a measure for psychological critical thinking (PCT) for the Psychology Bachelor at the University of Groningen. The Groningen Psychological Critical Thinking Task (GPCTT) is an essay task in which participants critically evaluate sources and come to a conclusion about the topic of resit exams. The essays are evaluated using a rubric consisting of five aspects that try to capture PCT: methodology, fallacy, assumption of authors, bias of participants, and synthesis. Eighteen first-year Psychology Bachelor students completed a survey consisting of the GPCTT and the Psychological Critical Thinking Exam (PCTE). A correlation analysis has been done to examine whether the scores of the GPCTT are positively correlated to the PCTE, which was unfortunately not the case. Suggestions for alterations of the task and the rubric have been made for future research.

Keywords: Psychological Critical Thinking, Creating a Measure

The Creation of the Groningen Psychological Critical Thinking Task: A Pilot Study

While conspiracy theories have been a common internet phenomenon for years, it seems even more relevant during the COVID-19 pandemic. Although reading up on conspiracy theories can be entertaining for some, they are unfortunately not harmless. Belief in political COVID-19 conspiracy has shown to predict less institutional trust, and less support for and adoption of regulations put in place by the government to stop the spread of the virus (Pummerer et al., 2021). A drive behind the belief in conspiracy theories could be groupthink, as described by Janis (1971). When one is part of a social circle with people who buy into certain beliefs, that person can feel a pressure to conform to the rest of the group. It can lead to the illusion that it is not possible that group is incorrect about their beliefs, which can make it difficult for members to voice their doubts about the common ideas as to not spoil the comfortable "we-feeling" atmosphere (Janis, 1971). Janis (1971) gives recommendations to combat the effects of groupthink, one of them being that members of the group should all take on the role of a critical evaluator. Fortunately, data suggests that critical thinking skills can be learned (Halpern, 1998; Kennedy et al., 1991).

Since critical thinking (CT) skills can be taught, and employers look for employees that possess these skills (Dwyer et al., 2015), the American Psychological Association (2012) has included it in the learning outcomes for the Psychology Bachelor. CT skills are of major importance for psychology students since these will help them assess and interpret claims and make them able to evaluate the quality of the source and evidence that support these claims (Lawson, 1999). Further, it is very important for future psychologists to be able to think critically, both in a research setting and for therapists. There is a need in both of these settings for scientific reasoning, avoiding bias, and evaluating claims (Lawson, 1999). Therefore, it is important to teach students these necessary skills.

In order to teach students CT skills, the University of Groningen (RUG) has implemented several courses that included CT in their learning goals. These include the Academic Skills course, the Theoretical Introduction to Research Methods course, the Research practicum, and the bachelor thesis (University of Groningen, n.d.). In order to evaluate these courses, students are mainly required to write papers, with the exception for the Theoretical Introduction of Research Methods course. Here, the students must pass a multiple-choice exam. However, situations in the daily working life of a psychologist will not be precisely mirror the problems they have encountered in the classroom. They have to apply learned skills to new settings. This process of applicating one's skills to a new context is called transfer (van Peppen et al., 2021). Transfer can be seen as a continuum, from near to far transfer. Near transfer entailing the transfer of knowledge or skill to a very similar situation, while far transfer refers to transferring between contexts that seem different from each other but might have some similar features (van Peppen et al., 2021). However, research has shown that, if not taught or trained, even transfer to similar tasks can be difficult (van Peppen et al., 2018).

Methods that are normally used to teach students content knowledge are not optimal for teaching how to transfer CT skills (Halpern, 1998). The empirically based model that Halpern (1998) proposes provides some direction on how to teach these skills and their transfer to new contexts. This model exists of four parts. The first states that certain dispositions are important for CT. One might have all the skills necessary to think critically but not have the motivation to put effort into using these skills. The educator can assist on this front by teaching students when it is worth to make the mental investment to think critically and when it is not. The second part emphasizes the importance of having a clearly identifiable set of skills that are important for CT so that is evident for students and their teachers what is

expected of them and what the learning goals are. Next, Halpern (1998) argues for the value of being able to recognize when to use the taught CT skills in new situations. By repeatedly using tasks that mirror the real-world and contain both relevant and less relevant information, students will learn how to transfer the CT skills to new contexts. Lastly, students should be aware of the level of their CT skills in relation to their learning goals. A useful tool to elicit reflection is to have students make a CT assignment and ask well-structured questions afterwards about their process. This can help them reflect on what they have learned and set learning goals for the future. In conclusion, CT skills and transfer of these skills can be taught.

It would be informative to use a measure that is based upon a 'real-life' problem that is ill-structured and not too similar to problems that students directly learn about (Bonk & Smith, 1998; Halpern, 1996). This way, one can evaluate whether the level of transfer of CT skills is acceptable at the end of the Psychology Bachelor. Hence, the goal of this project is to create a measure for CT that contains this 'real-life' quality that can be used to assess the Psychology Bachelor students at the RUG.

Defining CT

However, in order to create a measure, one should have a clear definition of the term in question. Unfortunately, there is no consensus in the literature about what CT exactly entails. For instance, the definition will vary from field to field as Lai (2011) describes in a research report. Differences and similarities between the philosophical approach and the cognitive psychological approach regarding CT are addressed. The philosophical approach focusing mostly on the critical thinker themselves and the ideal set of characteristics and personal qualities, while the cognitive psychologist will focus more on the actual way people think instead of the ideal way they could or should think (Lai, 2011). Further, for the

cognitive psychology approach, the types of actions or behaviors that are involved in CT are of greater interest than the characteristics of the critical thinker. But even within the field of psychology disagreements exist about the definition. Although there are researchers who agree that CT skills are specific to a certain domain (a psychologist will use CT differently than a physicist) (Mueller et al., 2020; Ennis, 1989), there are also those who argue that these are general skills that are not domain-specific (Halpern, 2001; Van Gelder, 2005). Another disagreement is regarding whether CT skills are transferable to new contexts. For example, Willingham (2008) found that students were able to show CT skills in one context but did not do so in another context. However, as Lai (2011) notes, there are researchers that believe that transfer from one context to the other can happen under the condition that students are taught how to do this this specifically. This is a point that is supported by research by Wittrock (2010). The study shows that through learning conditions that promote active and deep processing transfer of learning can occur.

Even though there are disagreements about what CT exactly entails, literary research by Petress (2004) shows there are also agreements about the concept. Researchers agree that evaluating evidence is a core aspect of CT (Paul & Elder, 2001; Petress, 2004; Scriven & Paul, 2003; Anderson et al., 2001). Moreover, exploring a problem, question or situation is an agreed upon aspect (Anderson et al., 2001; Warnick & Inch, 1994); just like the integration of available information (Anderson et al., 2001; Scriven & Paul, 2003; Warnick & Inch, 1994) and justifying one's position (Paul & Elder, 2001; Warnick & Inch; 1994). Halpern (1996) and Paul & Elder (2001) agree that evaluating ones thinking process and that CT is a skill rather than a disposition. Lastly, the ability to interpret an argument or claim (Halpern, 1998; Mueller et al., 2020; Willingham, 2008) and being able to come to a solution or making decisions (Halpern, 1998; Willingham, 2008) are also areas of agreement. Fortunately, this

shows that CT is not an undefinable concept and that there are certain facets of CT that are recognized and supported by evidence.

In conclusion, the definition that will be used for the GPCTT is that 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. This definition is derived from the definition used for the VALUE rubric (McConnell & Rhodes, 2017) which was used by Dibartolo et al. (2016) in the "Messy Problem", a measure for CT. However, the GPCTT is a measure for CT in the context of psychology, resulting in the addition of the term psychological critical thinking (PCT). This definition is in line with the one of another measure for PCT, which is the Psychological Critical Thinking Exam (PCTE) by Lawson (1999) which states that "*psychological critical thinking involves evaluating claims using the basic principles of psychological science*". Since the GPCTT aims to measure PCT as well, we specified in the definition for the GPCTT that it is based on principles of psychological science.

Now that a clear definition of PCT has been established for this project, the question remains as to how assess these skills. In a literature review Lai (2011) describes several suggestions made by researchers for assessments of CT skills. First, open-ended questions have been found to be more able to capture the concept of CT than multiple-choice measures are since they are more sensitive to the dispositional aspects of CT (Ku, 2009). Second, the task should reflect real-world problems, thus being ill-structured and having more than one defensible resolution. This gives the participants the opportunity to support multiple views (Moss & Koziol, 1991). Third, when making an assessment, one should be aware of the features of the tasks and the effects these can have on the natural inclination of the participant

to use CT skills. For example, Fischer and colleagues (2009) found that contradictions or inconsistencies in the given materials made it more likely for participants to use CT than material that was more consistent and coherent. Lastly, the focus of evaluation of the task should be on the quality of argument, not on the "correctness" of the answer (Moss & Koziol, 1991). The reason for this is that the task is not meant to quiz learned information, but instead to test the participant's CT skills. Being able to see the process of reasoning by the participants in order to assess the quality of the argument is important in this case.

A measure that fits the above-mentioned recommendations, is the Messy Problem by Dibartolo et al. (2016). This measure consists of an essay task about whether school time should be delayed and aims to measure CT. Participants were asked to imagine that they were helping a psychologist share her professional opinion on the topic by having them use a scientific perspective in examining the evidence to reach a conclusion. The given evidence included a fact-based newspaper article, an op-ed against the delay of schooltime, and the synopses of two real-life empirical scientific papers. This is a measure with open-ended questions, centered around a real-world problem and having more than one defensible possible standpoint, and with contradictory elements in the given material. In conclusion, this measure adheres to the recommendations previously discussed. As of now, there is no measure for PCT skills at the RUG. However, a measure similar to the Messy Problem that measures PCT might be useful to test whether student's these skills are up to par when they finish their Psychology Bachelors at the RUG. Therefore, the goal of this project is to create a measure for critical thinking for the RUG; the Groningen Psychological Critical Thinking Task (GPCTT).

Current Study and Creating the GPCTT

In order to test whether the task measures PCT, both the GPCTT and the PCTE will be administered, and a correlation test will be done between the scores on the GPCTT and the scores on the PCTE. Thus, the hypothesis is that there will be a significant positive correlation between the scores on the GPCTT and the PCTE. In addition, the interrater reliability will be tested, and an exploratory analysis will be conducted to assess the internal consistency. However, since the available sample is very limited, this is a pilot study.

The GPCTT assesses five aspects of the critical thinking process, namely methodology, the detection of fallacies, assumptions of authors, bias of participant, and synthesis. First off, the aspects of methodology and the detection of fallacies are aimed to assess if the psychology students are able to analyze and evaluate on principles of psychological science (Lawson, 1999), and if they are able to evaluate the provided evidence (Paul & Elder, 2001; Petress, 2004; Scriven & Paul, 2003; Anderson et al., 2001). Next, the assumptions of authors aspect was added to assess the ability to spot claims lacking supporting evidence (Mueller et al., 2020). The aspect of bias of participant was included to assess whether participants only used information and data provided in the task and argued their points without bias (Halpern, 1996; Paul & Elder, 2001). Lastly, to assess the ability of integrating the available information (Anderson et al., 2001; Scriven & Paul, 2003; Warnick & Inch, 1994), the aspect of synthesis was added. Thus, resulting in five aspects that aim to measure PCT.

In order to test these aspects, we used a similar task as Dibartolo et al. (2016), in which participants were asked to give a recommendation about a topic to the board at the RUG by writing an essay based on three provided sources. The topic was chosen to be resit exams, since this topic is relevant for most, if not all students. The participants were

instructed to imagine they were asked to advice the board on whether or not to abolish resit exams, and to critically evaluate the three given sources in an essay in which they would reach a conclusion about resit exams. The full task can be read in Appendix A.

Method

Participants

A total number of 18 first-year Psychology Bachelor students at the University of Groningen participated in the current pilot of the study. The students received course credit for participation. The initial sample size consisted of 22 participants, but four participants were excluded. Participants were excluded for responding in Dutch (n = 1), not completing the task (n = 3), and for filling in the PCTT and the GPCTT under 10 minutes (n = 0), since it takes around 10 minutes to read the assignments and clicking through the survey without filling in any answers. The 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 their best effort.

Materials

Groningen Psychological Critical Thinking Test

The GPCTT is test that aims at measuring Psychological Critical Thinking by having participants write an essay in which they critically evaluate several sources. Participants were presented with a fictional scenario in which they were asked to advise the Board of the RUG in a current discussion about abolishing or keeping resit exams. Subsequently they were

required to critically evaluate three sources on the topic of resits and write an essay about it, including a conclusion.

The three sources were summarized articles which included an opinion piece, a fact-based article in reaction to this opinion piece, and a research article. The first article was an opinion piece about resit exams (Boomsma, 2018), and primarily in favour of getting rid of resits. The provided summary included several statements on resit exams, some by professors and the mayor of Groningen, some made by the author without a provided source. These details were incorporated so that participants had opportunities to question assumptions and identify the appeal to authority fallacy. An extra paragraph was added that was not in the original article to provide a perspective that was in favour of resit exams to make the overall stance on resit exams less clear cut in one or the other direction. The second article was a reaction to the first and included a survey in which 450 students were asked questions like "Did you ever go to the exam just to see what was being asked of you?" (Boomsma & Siebelink, 2018). The content was altered in such a way that the article was predominantly in favour of keeping resit exams. For example, the number of students that participated in the survey was almost doubled to make the source feel more reliable. Lastly, the third summary was of a research article by Nijenkampt et al., (2016) in which an experiment was conducted to test the effect of resit exams on the amount of study time. The results of this experiment suggest that is would be better to abolish resit exam, since the investment of study time reduces once there is an opportunity for a resit exam. The participants were instructed to read the material thoroughly and base their conclusion on the provided sources. The complete task can be found in Appendix A.

Psychological Critical Thinking Exam.

Participants were presented with a shortened version of the Psychological Critical Thinking Exam (PCTE) (Lawson et al., 2015), consisting of seven instead of fourteen research-related scenarios (Appendix C). The use of a shortened version to assess PCT has been done before by Haw (2011) and Stark (2012). For each scenario a conclusion was reached, and the participants had to state the main problem with the conclusion in written form, if applicable. The goal of the PCTE is to assess psychological critical thinking by having participants evaluate whether the conclusion that is reached in each of these scenarios is the correct one, and if not, to state what the problem is with the conclusion. An example of one of the items is: "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" (Lawson, 2015). The seven PCTE items used for this project can be found in Appendix C.

Study Design and Procedure

The study is a within-subject correlational study; all participants had to complete both the PCTE and the GPCTT, and they were either first presented with the PTCE or the GPCTT at random to avoid a possible order effect.

Before the survey was distributed to the potential participants, the study was approved by the Ethics Committee of the University of Groningen. The first-year psychology students were able to access the study via the SONA system. Before the survey started, the participants saw a screen with information about the study, were informed about the amount of SONA credits they will receive, and were presented with the option to give their informed consent. Next, they were presented with either of the two measures. After finishing both tests, participants are 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

Scoring

For the PCTE, the scores for each question were combined into one final score. Accordingly, for the GPCTT, the scores from each aspect were combined into one final score. Both final scores were later used for statistical analysis and served as our variables. Both the answers to the PCTE and the essays for the GPCTT were scored independently by two blinded raters. The GPCTT essays were scored using the GPCTT-rubric (Appendix B) which was based on the VALUE rubric (McConnell & Rhodes, 2017) and contains the five previously discussed aspects. For the aspects of methodology, fallacy and assumption of authors, the participants can score on a scale including 0 (Subpar), 1 (Benchmark), 2 (Milestone), 3 (Capstone). For bias of participant and synthesis, participants can either score a 0 (Subpar) or 2 (Milestone). A score of zero would be given when a mistake is made (with an exception for the aspect of synthesis), and full points would be given when the aspects were applied correctly. To illustrate what the rubric entails two examples will be given. For the

aspect of methodology, the participant will receive full points if they evaluate the evidence they received regarding the methodology at least twice in their essay. As such, evaluating the research study on at least two accounts (e.g., the study's internal and external validity) would result in a score of three. Two points will be administered when the participant evaluates the evidence regarding methodology once, and one point will be administered when the participant does not mention anything regarding the methodology of any of the provided sources. Once the participant misinterprets the methodology of (one of) the sources (e.g., by saying the experiment has a high ecological value), they will receive zero points. The aspect of bias of participant is scored on a binary scale. So, either the participant exclusively uses information/evidence provided in the materials to support their conclusions and receives the maximum number of points, or the participant uses information/evidence that is not provided in the material and receives no points for this aspect.

At the end, any disagreements in scoring would be resolved between the two raters so that there would be an agreement on the final score for that aspect. The maximum score a participant could get is 13 and the minimum 0.

The answers to the PCTE were scored on a scale of 0 to 3. Zero 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 could be reached.

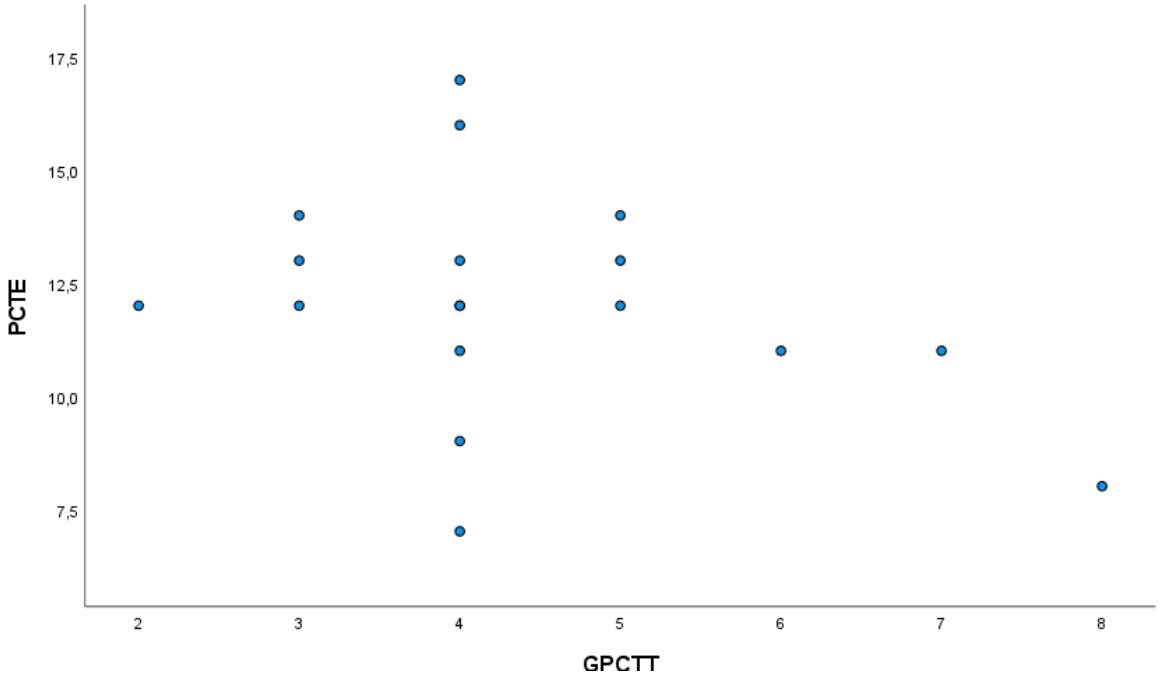
Analysis

Before conducting the planned correlation analysis, a Shapiro-Wilk test is done to test the assumption of normality (Shapiro & Wilk, 1965). For the PCTE the test was non-significant, but for the GPCTT the test showed a significant departure from normality (W =

.89, p = .038). Thus, a non-parametric test was required. Since the data is ordinal and not normally distributed, Spearman's rho coefficient was used to assess the relationship between the participants' scores on the PCTE and the GPCTT (Spearman, 1904). A small to moderate nonsignificant negative correlation was found between the two tests (r (16) = -.307, p = .215) (Cohen, 1992). The null hypothesis, which states that there is no monotonic relationship between the two variables, cannot be rejected since the correlation is not significant. Figure 1 shows a scatterplot depicting the relationship between the scores of the PCTE (on the y-axis) and the scores on the GPCTT (on the x-axis).

Figure 1

Correlation Between PCTE Scores and GPCTT Scores



Cohens Weighted Kappa was used to assess the interrater reliability (Cohen, 1968). The interrater reliability of the GPCTT was found to be weak with a Weighted Kappa of k_w = .42 (p = .002) (McHugh, 2012). As for the PCTE, which had a Weighted Kappa of k_w = .61 (p = .001), the interrater reliability was substantial (McHugh, 2012).

In order to test the internal consistency of the GPCTT which measures five aspects, the Cronbach's Alpha is used (Cronbach, 1951). The internal consistency for the five items is considered unacceptable (α = .20) (George & Mallery, 2003). The removal of the aspect of assumption of authors raised the Cronbach's Alpha to α = .46 (Appendix D, table 1). Similarly, the Cronbach's Alpha for the PCTE is unacceptable with α = .22 (George & Mallery, 2003). The sixth item was automatically removed by SPSS due to the lack of variation in scores, since every single participant scored 1 point for that question.

Discussion

The goal of this project was to create a measure for psychological critical thinking (PCT): the Groningen Psychological Critical Thinking Task (GPCTT). In order to test whether our task measures PCT, the scores on the GPCTT were correlated with the scores on the PCTE. Unfortunately, a Spearman's rho test showed that there was no significant correlation in the sample, which was negative, surprisingly. It should be considered, however, that this is the first time the GPCTT has been tested, and many alterations have to be made in order to make it a valuable measure. Besides, even though the PCTE and the GPCTT both aim to measure PCT, they have have a completely different approach. The PCTE consists of questions about research scenarios in the context of psychology, a way in which students are likely presented with in the classroom. Thus, there is a near transfer of skills (van Peppen et al., 2021). In contrast, the GPCTT is a task that mirrors an ill-structured, real-life problem that is not explicitly taught, thus requiring a far transfer of PCT skills (van Peppen et al., 2021). What could be interesting to assess in future studies is the correlation between the GPCTT and the Messy Problem by Dibartolo et al. (2016), since this task also requires the far transfer of CT skills. Another possible contribution to the unexpected correlation is the small and

unrepresentative sample size. The Spearman's rho becomes less reliable as the sample size gets smaller (Hackshaw, 2008), and only first year psychology students participated in this pilot study. For future research, a bigger sample size that includes psychology students from second and third year would be informative.

Besides the correlation test, the interrater reliability was computed. The Weighted Kappa was weak for the GPCTT, however it is worth to mention that the Weighted Kappa not only takes into account the number of disagreements between the two raters, but also the level of disagreement (Cohen, 1968). This means that the level of disagreement can be overrepresented due to the way two of the aspects are rated in the GPCTT rubric. For the aspects bias of participant and synthesis participants can score either 0 or 2 points, therefore a disagreement regarding these will always lead to a difference of two points, instead of the one point that is the case for the other three aspects. This can be settled by changing the scoring for these aspects by using the 0- and 1-point possibilities. Nevertheless, the weak interrater reliability could also have been caused by the lack of clarity on how to assess each of the aspects. The grading of the PCTE is very straightforward in comparison; there is a correct answer and a clear way of awarding points. Perhaps the descriptions in the GPCTT rubric are too ambiguous and not straightforward enough. A suggestion for the aspect of synthesis that solves both abovementioned problems is to step away from the binary scoring for this aspect and make this into a gradient, and to rephrase the description. Right now, the participant will score 0 points when "the participant does not show sufficient ability to weigh or combine evidence that is in line with, but also contradicting their position", and score 2 points when "the participant shows the ability to combine evidence and weigh contradictory evidence in taking their final stance". It was not clear what 'sufficient ability' exactly means in this context. Instead, the suggestion is that Subpar (0) would be the scored when the participant

only mentions one side of the discussion; the Benchmark (1) would be scored if both sides were mentioned but with no further elaboration on one or more of the sides of the discussion; and when both (or all) sides of the discussion are mentioned and elaborated upon, the participant would score the Milestone (2). This way, the number of times a participant has engaged in a certain behavior can be counted and the corresponding score can be given. This is a more straightforward way of scoring essays than when trying to conceptualize 'sufficient ability to weigh evidence'.

As an exploratory analysis, the internal consistency was explored by using the Cronbach's alpha. Both the GPCTT and the PCTE showed a very low internal consistency. But once again, there are some caveats that need to be mentioned. Cronbach's alpha could have been affected by the small sample size, the number of items in the test, and low variation in scores (Sijtsma, 2009).

Shifting the focus to the instructions of the GPCTT, some areas of improvement became evident when grading the essays. Particularly, the description of the task that instructs the participants on what is expected of them should be improved so that they are more inclined to use their critical thinking skills. An emphasis on the need for a professional opinion and the use of a scientific perspective on the provided evidence, much like Dibartolo et al. (2016) approached it, could be a good first step. We noticed that most essays did not include a critical analysis of the provided sources at all. Perhaps by emphasizing the need for a scientific approach to the task instead of a personal recommendation to the Board could prompt the CT skills better. Second, asking the participants to write down their thought process could be of help with evaluating the essays. It could be the case that the participants did evaluate the sources critically, came to a conclusion about them, and merely noted down

their recommendation without including their critical analysis. By making their thought process visible, a more accurate assessment of CT skills can be made.

More research is certainly needed for the GPCTT to be able to be used in the real world. Once the above-mentioned changes have been made, an interesting next step would be to test the GPCTT with bigger, more diverse samples. An insightful step would be to test whether there are differences in scores between first-year psychology students and second- or third-year psychology students. Lawson (2015) showed that the PCTE does distinguish between junior and senior Psychology majors, seniors scoring significantly higher. Lawson (1999) also found that Psychology majors scored significantly higher on the PCTE than other majors, thus it is interesting to test if psychology students score better on the GPCTT than non-psychology students.

In conclusion, once enough research has been done and the GPCTT has been improved and is ready to be put to practical use, it could be a useful tool for the educators at the RUG. Since this measure aims to assess the transfer of PCT skills to a context that is different from the ones students are used to in a classroom setting, professors can use the GPCTT to evaluate whether the transfer of PCT skills generally improved over the course of a particular course. Another use for this measure could be to monitor the quality of the Psychology Bachelor regarding the learning goal the APA set pertaining the ability to think critically (The American Psychological Association, 2012). By administering the task to first year students and students that are about to finish their Psychology Bachelor the improvement of CT skills can be examined. All in all, once more research has been done the GPCTT could have a promising future.

References

American Psychological Association. (2012). *APA Guidelines for the Undergraduate Psychology Major Version 2.0*.

Boomsma, C. (2018, January 11). *Get rid of resits*. UKrant.Nl. Retrieved October 1, 2021, from https://ukrant.nl/magazine/get-rid-of-resits/?lang=en

Boomsma, C., & Siebelink, R. (2018, January 17). *No resits? More stress*. UKrant.Nl. Retrieved October 1, 2021, from https://ukrant.nl/magazine/tentamen-enquete/?lang=en

Bonk, J. C., & Stevenson Smith, S. G. (1998). Alternative instructional strategies for creative and critical thinking in the accounting curriculum. *Journal of Accounting Education*, *16*(2), 261–293. https://doi.org/10.1016/s0748-5751(98)00012-8

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.

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

Datanovia. (2019, November 7). *Weighted Kappa in R: Best Reference*. Retrieved December 19, 2021, from https://www.datanovia.com/en/lessons/weighted-kappa-in-r-for-two-ordinal-variables/

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. https://doi.org/10.5325/jasseinsteffe.6.2.0191

Dwyer, C. P., Boswell, A., & Elliott, M. A. (2015). An Evaluation of Critical Thinking Competencies in Business Settings. *Journal of Education for Business*, *90*(5), 260–269. https://doi.org/10.1080/08832323.2015.1038978

Ennis, R. H. (1989). Critical Thinking and Subject Specificity: Clarification and Needed Research. *Educational Researcher*, *18*(3), 4–10. https://doi.org/10.3102/0013189x018003004

Fischer, S. C., Spiker, V. A., & Riedel, S. L. (2009). Critical thinking training for Army officers. Volume 2: A model of critical thinking.

Gelder, T. V. (2005). Teaching Critical Thinking: Some Lessons From Cognitive Science. *College Teaching*, *53*(1), 41–48. https://doi.org/10.3200/ctch.53.1.41-48

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. (1996). *Thought and Knowledge: An Introduction to Critical Thinking.* Psychology Press.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, *53*(4), 449–455. https://doi.org/10.1037/0003-066x.53.4.449

Halpern, D. F. (2001). Assessing the Effectiveness of Critical Thinking Instruction. *The Journal of General Education*, *50*(4), 270–286. https://doi.org/10.1353/jge.2001.0024

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.x

Janis, I. L. (1971). Groupthink. *Psychology today*, *5*(6), 43-46.

Kennedy, M., Fisher, M. B., & Ennis, R. H. (1991). Critical thinking: Literature review and needed research. In L. Idol & B.F. Jones (Eds.), *Educational values and cognitive instruction: Implications for reform (pp. 11-40)*. Hillsdale, New Jersey: Lawrence Erlbaum & Associates.

Ku, K. Y. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity*, *4*(1), 70–76. https://doi.org/10.1016/j.tsc.2009.02.001

Lai, E. R. (2011). *Critical Thinking: A Literature Review*. Pearson Education.

Lawson, T. J. (1999). Assessing Psychological Critical Thinking As a Learning Outcome for Psychology Majors. *Teaching of Psychology*, *26*(3), 207–209. https://doi.org/10.1207/s15328023top260311

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

McConnell, K. D., & Rhodes, T. L. (2017). *On Solid Ground. A Preliminary Look at the Quality of Student Learning in the United States*. Association of American Colleges and Universities.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 276–282. https://doi.org/10.11613/bm.2012.031

Moss, P. A., & Koziol, S. M. (1991). Investigating the Validity of a Locally Developed Critical Thinking Test. *Educational Measurement: Issues and Practice*, *10*(3), 17–22. https://doi.org/10.1111/j.1745-3992.1991.tb00199.x

Mueller, J. F., Taylor, H. K., Brakke, K., Drysdale, M., Kelly, K., Levine, G. M., & Ronquillo-Adachi, J. (2020). Assessment of Scientific Inquiry and Critical Thinking: Measuring APA Goal 2 Student Learning Outcomes. *Teaching of Psychology*, *47*(4), 274–284. https://doi.org/10.1177/0098628320945114

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

Paul, R., & Elder, L. (2001). *The miniature guide to critical thinking concepts and tools*. Rowman & Littlefield.

Petress, K. (2004). Critical thinking: An extended definition. *Education*, *124*(3).

Pummerer, L., Böhm, R., Lilleholt, L., Winter, K., Zettler, I., & Sassenberg, K. (2021). Conspiracy Theories and Their Societal Effects During the COVID-19 Pandemic. *Social Psychological and Personality Science*, *13*(1), 49–59. https://doi.org/10.1177/19485506211000217

Scriven, M., & Paul, R. (2003). *Defining critical thinking: a statement prepared for the National Council for Excellence in Critical Thinking Instruction*.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Sijtsma, K. (2008). 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

University of Groningen. (n.d.-a). *Ocasys: Academische Vaardigheden*. RUG. Retrieved December 19, 2021, from https://www.rug.nl/ocasys/rug/vak/show?code=PSBE1-25

University of Groningen. (n.d.-b). *Ocasys: A Theoretical Introduction to Research Methods*. RUG. Retrieved December 19, 2021, from https://www.rug.nl/ocasys/gmw/vak/show?print&iflang=NL&code=PSBE1-27

University of Groningen. (n.d.-c). *Ocasys: Bachelorthese*. RUG. Retrieved January 13, 2022, from https://www.rug.nl/ocasys/gmw/vak/show?print&iflang=NL&code=PSB3E-BT15

University of Groningen. (n.d.-d). *Ocasys: Research practicum*. RUG. Retrieved January 13, 2022, from https://www.rug.nl/ocasys/gmw/vak/show?print&iflang=NL&code=PSBE2-09

van Peppen, L. M., van Gog, T., Verkoeijen, P. P. J. L., & Alexander, P. A. (2021). Identifying obstacles to transfer of critical thinking skills. *Journal of Cognitive Psychology*, 1–28. https://doi.org/10.1080/20445911.2021.1990302

Warwick, B., & Inch, E. S. (1994). *Critical Thinking and Communication.* 2nd ed. New York.

Willingham, D. T. (2008). Critical Thinking: Why Is It So Hard to Teach? *Arts Education Policy Review*, *109*(4), 21–32. https://doi.org/10.3200/aepr.109.4.21-32

Appendix A

GPCTT task

You will now be presented with three articles on the topic of resits at the University 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 analyzing 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 analyze 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.

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.

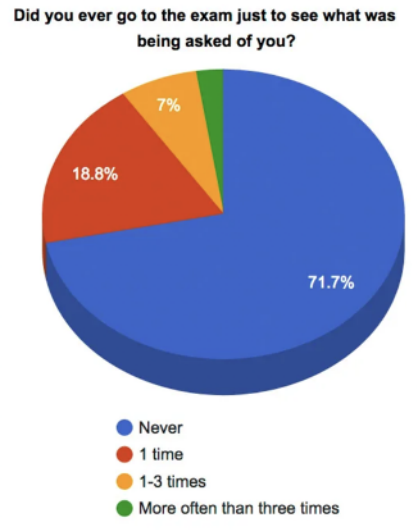
In conclusion, the tests should be used to steer education. Plan many, forcing students to keep studying. Offer students the opportunity to compensate for bad grades so they don't get hung up on a single failed test. Offer cumulative testing, to ensure that a later good grade makes up for an earlier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

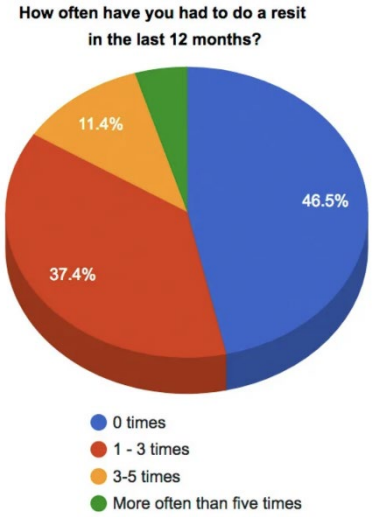
No more resits? More stress (A reaction to "Get rid of resits").

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.





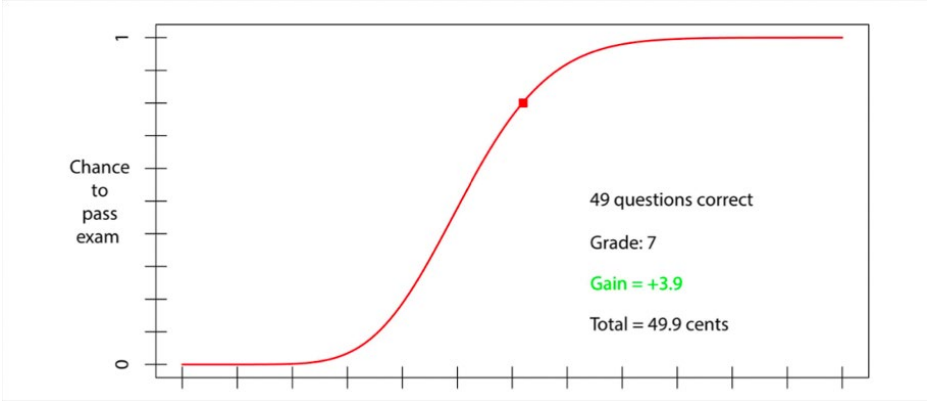
Then the main question: should resits be discouraged by scheduling them at unusual times? A fair number of students (27.1%) don't think the idea is too bad. The most used argument is that the increase in pressure will force students to start studying earlier and take exams more seriously.

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 B

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**:	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

	Option: The participant mentions that the 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).*			because it has always been like that" is a **valid** argument. **appeal to authority fallacy**: 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 (ability to spot claims lacking supporting evidence)	The participant considers at least 2 assumptions of the authors, including sources for statements and facts and considers them non-valid. **Example:**	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*	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.

	"It takes time, and the students might suffer delays but without this option students have a higher chance of dropping out. " *AND* *"When you fail an exam, you want a second chance as quickly as possible."*	"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.

Appendix C

The Psychological Critical Thinking Exam (PCTE) items (Lawson, 1999)

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 filmgoers 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 nonalcoholics, 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 the analysis

Table 1

Item-Total Statistics

Aspect	Cronbach's alpha if item deleted
Methodology	-.391
Fallacy	.233
Assumptions of authors	.455
Bias of participants	-.053
Synthesis	.266