Creating a Measure for Psychological Critical Thinking

John Clemens Nagler

S3796582

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group number: 2122_1a_09 EN

Supervisor: Marcella Fratescu

First evaluator: dr. J.A.M Heesink

Second evaluator: dr. S.M.H. Mohamed

In collaboration with: Laura Escudero, Thomma Schröder, Leonie van Jaarsveld, Elisabeth van Nee

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

Critical Thinking (CT) has grown into one of the central educational goals around the globe. Nevertheless, testing critical thinking is in its infancy. Notably, the creation of "authentic" tests that can assess CT skills as near to real-world situations as possible, has been scarce. However, these in particular would help educators understand weather current teaching practices translate into later benefits when it comes to being active in the field. The current study aims to create an authentic test to assess psychological CT in students, the Groningen Critical Thinking Task (GPCTT). To test the convergent validity, we correlated participants scores on the GPCTT with scores on the Psychological Critical Thinking Exam (PCTE) (Lawson et al., 2015). The PCTE is a widely used test for CT in the field and suggested as assessment instrument by the American Psychological Association (APA, 2013). Based on research by Dibartolo et al. (2016) and Shavelson and colleagues (2019), we created an authentic essay-based exam that was subsequently administered to students at the University of Groningen (*N* = 78). The results showed no correlation between the GPCTT and PCTE. The GPCTT also failed to distinguish between psychology students and non-psychology students. While students' inability to transfer their CT to an authentic test might have been at fault for the non-correlation, the later evaluation also showed deficits in the newly created measure. As an experiment it shows successful and less successful strategies, which will form considerations that can be taken into account when developing further versions of the GPCTT.

Keywords: psychological critical thinking, measure, authentic

Creating a Measure for Psychological Critical Thinking

The roots of critical thinking (CT) reach deeply back in human history. Socrates described 2500 years ago that one could not depend on those with "authority" to have judicious knowledge; individuals should ask penetrating questions, piercing the fog of empty rhetoric (Paul et al., 1997). Surprisingly, given its long history, CT was only depicted as an educational goal by Dewey in the 20th century for the first time (Hitchcock, 2020). Still, the importance of this idea has now infiltrated modern academia, where 95 % of surveyed chief academic officers from the USA consider CT as one of the essential educational outcomes for their college graduates (AAC&U, 2011). CT skills have found their way into learning goal guidelines of universities (e.g., University of Groningen) and are listed as fundamental to Psychologists by the American Psychological Association (American Psychological Association, 2013; Facione, 1990).

CT has progressively grown into one of the most emphasised learning goals in modern education (Facione, 1990; Lai, 2011), and that is for good reason. It is vital for people in general, but students in particular, to appraise questionable scientific claims and misinformation (Smith, 2011). Mitchell (2001) even suggests that science devoid of CT may subvert science as a whole. CT can be seen as the guarding shield against humans' inclination to use fallacies, heuristics and drawing inappropriate conclusions (Resinick, 1987). Students skilled in CT show more sophisticated reasoning and are also more likely to make improvements to their work (Holmes et al., 2015). Therefore, it is no surprise that the American Council in Trustees and Alumni found CT to be one of the most attractive skills for employees, so that employers start requiring it (Liu et al., 2014).

Definition

What is it about CT that makes it paramount to scientific inquiry and essential in many domains? While the concept of critical thinking has existed for millennia, there is prevailing disagreement over what CT actually entails (Mulnix, 2012). Finding a unitary definition for higher-level constructs has its challenges, and critical thinking is not exempted from this. While searching the literature for this project, it became apparent that no two academic papers agreed on the exact definition. Moreover, different intellectual disciplines define critical thinking in various ways. The field of Philosophy hereby focuses more on the critical thinker – on dispositions (Lai, 2011). Such thinkers are characterised by being fair-minded, inquisitive, open-minded and their willingness to hold out with their judgment until other perspectives are considered (Facione, 1990). Psychologists, on the other hand, especially those painted by behaviourist traditions, tend to define critical thinking by a list of skills the thinker has to master and procedures they have to act out (Lewis & Smith, 1993). These skills include comprehending, analysing, evaluating, and synthesising information (Ennis, 1993). Understandably, the habit of reducing such complex high-level skills to a mere list of procedures and steps is questionable (Bailin, 2002; Sternberg, 1986). Hence, critical thinking is more than the sum of its parts (Van Gelder, 2005) – one could go through all steps of critical thinking without actually engaging in critical thought (Bailin, 2002). Liyanage and colleagues (2021) argue similarly, stating that dutifully reproducing a set of criteria said to be associated with critical thought, is insufficient to be classified as exhibiting "proper" critical thinking.

In the end, CT has to be simplified into something that is parsimonious and can be observed, so that scholars can research, teach and evaluate it with some degree of consensus. As Tiruneh and colleagues (2014) pointed out, the lack of overarching definition might have led to the development of measures that diverge in terms of scope, psychometrics and general format (e.g., multiple-choice or essay based). Since the measures differ in their conceptualisation of critical thinking and thus the nature of items, different results across samples and studies are partly explained by this incoherence (Tiruneh et al., 2014). To illustrate, measures like the PCTE and Ennis-Weir Critical Thinking Essay Test (Ennis & Weir, 1985) differ in fundamental design choices. The Ennis-Weir measuring general CT with the means of essay-like answer formats, and the PCTE measuring domain specific CT based on short written answers. In a comprehensive review Tiruneh and colleagues (2014) found that the type of measurement used affected the results and contributed to conflicting findings even in the same samples.

The University of Groningen includes CT as part of its learning goals, with emphasis on the scientific method in research and basing conclusions on scientific aspects (see Appendix D). Furthermore, the cognitive and philosophical views outlined earlier agree on a few grounds, e.g., the ability to evaluate, analyse arguments, make decisions and solve problems (Lai, 2011).

In order to operationalise CT for our measure, we have come up with the following definition:

PCT is a habit of mind characterised by the comprehensive exploration of issues, ideas, artefacts, and events based on principles of psychological science before accepting or formulating an opinion or conclusion.

Why domain-specific?

Coherent definitions that span research groups and academic boarders, would help educators with their challenging task to teach and assess critical thinking. However, there is more clarification needed before curricula can be drafted and students taught. One integral part that academia has to investigate is the disagreement surrounding CTs domain-specificity. Researchers like Lipman (1988) state that the fundamental meaning of CT stays the same, even if the specific criteria might differ; CT might even be general in nature (Van Gelder, 2005). On the other hand, researchers argue that domain-specific knowledge is imperative to the success of CT. Accordingly, the types of evidences, explanations and evaluations differ across the domains (Bailin et al., 1999). Although most authors believe that background knowledge is essential for the application of critical thinking, some do not see it as a sufficient argument that CT is solely domain-specific (Lai, 2011). McPeck (1990) would disagree with this notion, claiming the more general a thinking skill is, the more it loses its usefulness. Each discipline has its own gold standards (statistics in social science or randomised control trials in the medical sciences), and identifying the values and standards is vital for effective CT (Ennis, 1989; Paul, 1992). Consequently, much as any other domain, psychology also has unique aspects that make its specific critical thinking different from general CT skills (Lawson, 1999; Lawson et al., 2015). One would not by any means claim that individuals from outside a particular domain, and who had trained in CT, would be unable to reflect critically on psychological material, but one would expect them to show different performances.

To consider these differences in more detail, Lehman and Nisbett (1990) found that training in the social sciences tends to have a more significant effect on statistical and methodological reasoning in undergraduate students than does training in the natural sciences. The here developed GPCTT is designed to measure critical thinking that relies on these types of reasoning. Measures like the PCTE have been effective in differentiating between majors, psychology majors outscoring students from biology and art (Lawson et al., 2015) Since our measure aims to assess psychological critical thinking, and different domains teach different types of critical thinking to their students, we hypothesise that

The scores on the GPCTT will be significantly different between psychology students and non-psychology students.

How to Measure

As with any other learning goal, CT skills have to be assessed to judge whether CT has been taught successfully. The noble goal of teaching high-level concepts such as this falls flat if no evaluation is conducted. However, research indicates problems related to both the validity and reliability of current measures (Lai, 2011). Moreover, the uncertainty around the domain-specificity makes the assessment all the more challenging (Norris, 1989). Irrespective of that disagreement, most accepted assessments, like the California Critical Thinking Skills Test (Facione, 1990), tend to assess general critical thinking (Kennedy et al., 1991; Lai, 2011; Watson & Glaser, 2019). Many measures also rely on multiple-choice questionnaires, which we judged to be ill-suited, as they more often also reflect test-makers political and empirical judgments and beliefs (Liu et al., 2016; Norris, 1989). We decided to use an open-ended task to assess critical thinking as it is highly recommended, so that students are not forced to simply to restate information (Lai, 2011).

Both Shavelson and colleagues (2019) and Dibartolo et al. (2016) laid essential groundwork for creating measures that test the real-world applicability of students' critical thinking skills. Existing CT measures often require students to reapply their skills in situations matching closely the contexts they have experienced in the classroom and thereby neglected the need for students to test their skills in new contextual circumstances (Dibartolo et al., 2016). According to Dibartolo and colleagues (2016), the use of interpretive knowledge in particular is imperative for thinking like a psychologist and enables students to make sense of "messy problems". Interpretive knowledge refers to the classification, prediction and inference made by the individual (Broudy, 1977). Messy problems are characterised by having multiple sources, all presenting information from different viewpoints, partly contradicting one another. An individual assessing the information has to discriminate between the validity and reliability of the content under consideration. Hence, instead of identifying a statement that includes one single CT issue, individuals have to independently find and weigh up information and then come to a conclusion based on the entirety of content presented. In practice, participants in messy problems are expected to (Shavelson et al., 2019):

1) assess the trustworthiness of the information and sources.

2) recognise the relevancy of the presented information.

3) spot biases and judgemental biases as well as avoiding such biases themselves.

4) reach a weighted conclusion.

On the basis of the aforementioned recommendations, we developed an open-ended assessment designed to assess CT skills in authentic scenarios that would mimic the work of psychologists (Appendix A). The rubric we used to score the participant's performances was an adaptation of the VALUE Rubric (Association of American Colleges and Universities, 2017). The VALUE rubric had been developed by the Association of American Colleges and Universities (AAC&U) to assess critical thinking in students based on authentic contexts. It is meant to present a framework for evaluating CT, and thus be translated and adjusted to individual courses, disciplines and campuses (AACU&U, 2017). We made adjustments to the VALUE Rubric based on the suggestions by Shavelson et al. (2019) and alterations to fit the materials we used to create the texts. The GPCTT rubric (Appendix B) also gave specific examples to make the assessment for alternating raters clearer. Our objective while creating the rubric was to evaluate students’ ability to spot methodological flaws (*Methodology*), their ability to assess the use of fallacies in the texts (*Fallacy*), to identify unsupported assumptions by the authors (*Assumptions*), their ability to solely argue with information that they could support with sources (*Bias*), and the proficiency to weigh, summarise and come to a conclusion (*Synthesis*).

To summarise, to address the aforementioned problems and build on the referred recommendations, the GPCTT will be characterised by the following aspects:

1. A clear definition that operationalises Critical Thinking in the domain of psychology

2. A messy problem specifically constructed to assess CT in authentic contexts

3. An essay based answering schema instead of multiple choice.

To validate the GPCTTs ability to measure CT, we compared the results with those to a different CT measure that had been established before. For this we chose the Psychological Critical Thinking Exam (PCTE) (Lawson et al., 2015). Several other studies have verified, established and supported its validity (Stark, 2012; Williams et al., 2003). In the revised PCTE, participants would get seven statement texts that would include a mistake. The task is to find and describe the faulty reasoning that is exhibited in those texts. Each of the tasks alludes to one principle of psychological critical thinking as described by Lawson (1999) (e.g., spotting fallacies, causation vs correlations, falsifiability). To ensure that our measure is able to assess CT accurately, we will compare it to PCTE. Hence, we will hypothesise that

The GPCTT and PCTE scores will have a significant positive correlation .

Methods

Participants

A total number of 78 Bachelor degree students of the University of Groningen (52 females (67%), 26 males(33%)) participated in the study. The age of the participants was measured in ranges from 17-20 years (n = 49 (63%)), 21-24 years (n = 26 (33%)) and 25+ (n = 3 (4%)) . Participants were excluded for responding in Dutch (n = 3) as not all the raters are familiar with Dutch. Another exclusion criterion was finishing the task in under ten minutes, which is the time it approximately takes to read the whole task. However, no participant needed to be excluded for this. The study was only available in English, hence sufficient English skills were essential to complete the study. The sample consisted of 67 psychology students (54 first-years, 2 second-year, 11 third-year or above, in total 86%) and 11 non-psychology students (14%). 9% of participants were native English speakers, and 91% non-native speakers. 1% came from an Asian country, 87% came from a Western country, and 12% from other countries. 82% indicated that they put in their best effort, 18% did not put their best effort in. First-year psychology students were recruited through SONA; any participants from a higher semester or different bachelor degree course were recruited by the researchers, thus making this a convenience sample.

Research design and 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 randomised 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. First-year psychology students could access the study via the SONA system, while all other participants received access via a link. Before the survey started, the participants were shown a screen with information about the study and were informed about the amount of SONA credits they would receive and about the option to participate in a lottery with a chance to win 15 euros if they were non-SONA participants.

Participants had to provide their informed consent in order to proceed with the survey. The GPCTT started with an instruction that stated that participants would be presented with three articles about resit exams at the University of Groningen. Participants were instructed to write an essay in which they critically evaluate the articles and come to a conclusion about the practicality of resits (Appendix A).

After finishing both tests, participants were asked about demographic information (age, gender, major, native language and ethnicity). Finally, first-year psychology students were granted SONA credits for participation, and the remaining participants could choose to participate in a lottery to win 15 euros.

Materials

Groningen Psychological Critical Thinking Task (GPCTT)

The GPCTT is an essay test that aims to measure Psychological Critical Thinking. Participants are presented with a fictional scenario in which they are asked to advise the board of the University 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 A) and write an essay about it, including an introduction, body and conclusion. Resits are a persistent a topic at the University of Groningen, exemplified by the multiple articles by the University Press Ukrant (Ukrant, 2018a; Ukrant, 2018b). The topic was chosen to increase engagement and motivate participants to put effort into completing the task. The three sources they are presented with include an opinion-based article, a fact-based article, and a research article, respectively. The first two articles are based on published articles from the Ukrant (2018), but have been slightly modified by us for grading purposes. To give an example, we added in the status-quo bias: “Taking resits has always been like this, so why should we change it now?" to test participants on their ability to recognise this Fallacy. The last article we included in our assessment packet is a synopsis of a research article, derived from a real-life experimental study in the literature (Nijenkamp et al., 2016). Each essay was scored based on the GPCTT-rubric (Appendix B) that includes the aspects Methodology, Fallacy, Assumption of Authors, Bias of Participants and Synthesis. Our grading scheme was as follows: For the aspects Methodology, Fallacy and Assumption of Authors, the participant can score on a scale from 0 (Subpar – participant misinterprets the aspect), 1 (Benchmark - participant does not consider the aspect at all), 2 (Milestone - participant interprets the aspect correctly once), 3 (Capstone - participant interprets at least two of the aspects correctly). For Bias of Participant and Synthesis, each participant can either score a 0 (Subpar) or 2 (Milestone). Therefore, the total scores could range from 0 to 13 points.

The rubric also includes examples of what the participant is expected to find and mention for each aspect. For instance, we mentioned two methodology threats in the rubric:

internal validity (“The participant mentioned that the experiment has a higher internal validity than the survey”) and ecological validity (“The participant mentioned that the ecological validity of the experiment is lower due to an artificial setting”). An example of how the aspect of Fallacy would be scored could include the following options: The participant mentions that the Mayor of Groningen has the opinion to keep the resits, but identifies this as a non-valid argument (because the mayor is not an expert). A participant whose essay states that the RUG-board should keep the resits because the Mayor of Groningen thinks so would score a 0 on the fallacy aspect, but a participant that states that the Mayor of Groningen thinks the resits should be kept, but next concludes this is a non-valid argument because the mayor is not an expert, would receive a 2.

Psychological Critical Thinking Exam

Participants were also presented with a shortened version of the Psychological Critical Thinking Exam (PCTE) (Lawson et al., 2015). We used seven of the fourteen research-related scenarios because of time constraints of participants. This version has been developed and validated by Lawson and colleagues (2015). For each scenario, a conclusion was reached and the participants had to state the main problem with the conclusion in written form, if applicable. Participants were scored on a scale of 0 to 3. 0 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. Hence, for this task, a maximum score of 21 could be reached (Appendix C).

Results

Internal Validity

To assess the quality and interpretability of the GPCTT rubric, we assessed the interrater reliability. Since we had alternating/non-unique raters, we used the Fleiss Kappa to assess the interrater reliability. A pilot study was conducted, serving as training for the raters. Each rater individually and independently scored the participant's answers for the GPCTT. Differences in scores were then discussed until consensus was reached. As seen in Table 1, we achieved moderate agreement in three categories: Fallacy, Methodology, and Synthesis. Less than moderate agreement was shown for Bias and Assumptions (interpretation based on Landis & Koch, 1977). Overall inter-rater reliability across all items was calculated with an averaged Fleiss Kappa (as done before in De Vries, 2008), $\kappa_{ave} = .466$. Hence, we achieve moderate interrater reliability across all five categories of the rubric (Synthesis, Bias, Assumption, Fallacy and Methodology).

Table 1

Fleiss Kappa Scores for each of the five Categories of the GPCTT

	Kappa	Asymptotic Standard Error	Z	P Value	Lower 95% Asymptotic CI Bound	Upper 95% Asymptotic CI Bound
Synthesis	,505	,113	4,458	,000	,283	,727
Bias	,309	,113	2,728	,006	,087	,531
Assumption	,341	,096	3,554	,000	,153	,529
Fallacy	,592	,092	6,451	,000	,412	,772
Methodology	,584	,077	7,598	,000	,433	,735

To assess the internal consistency of the GPCTT, we calculated the Cronbach's alpha, $\alpha = .495$. The Cronbach’s alpha gives us the extent to which the items in a test measure the same construct, critical thinking in this case. A score of .495 can be classified as below-par

(Tavakol & Dennick, 2011). As seen in Table 2, deleting any of the categories does not lead to a significant increase in α; therefore, no items were deleted.

Table 2

Item-Total Statistics and Cronbach's Alpha if Item Deleted

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Synthesis	3,45	2,822	,246	,151	,465
Bias of participants	4,42	2,689	,298	,113	,422
Assumptions	4,26	3,414	,198	,182	,481
Fallacy	3,97	3,636	,332	,159	,437
Methodology	3,64	3,012	,352	,160	,386

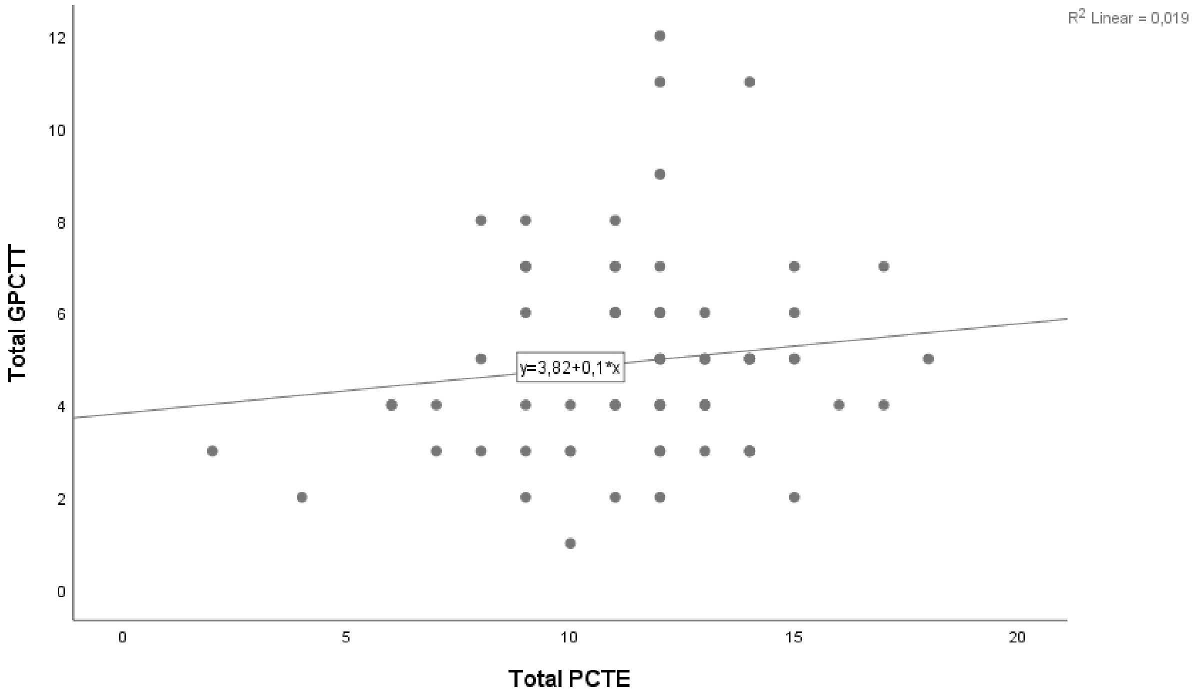
Hypothesis Testing

The normality for the scores on both the GPCTT and PCTE was checked with the Shapiro-Wilk test for normality, it showed a significant divergence from normality for both tasks, with $W(78) = .915$, $p = 0.00$ for the GPCTT, and $W(78) = .967$, $p = 0.038$ for the PCTE. Hence, we reject the null hypotheses since we have found evidence that the data on both tests is non-normally distributed.

We hypothesised (H1) that the GPCTT and PCTE scores will show significant positive correlation. Since the normality assumption was violated, we opted to use Kendall's tau (as suggested by Croux & Dehon, 2010), the assumptions for Kendall's tau have been met; the data is ordinal and the variables exhibit a monotonic relationship (see Figure 1). Kendall's tau showed a non-significant positive correlation between the two measures, $r_\tau(78) = 0.068$, $p = .433$.

Figure 1

Scatterplot between Scores on the PCTE and GPCTE



In H2, we hypothesised that the scores on the GPCTT will be significantly different between psychology students and non-psychology students. Since the normality assumption was violated, we used the Kruskal-Wallis Test. The data was grouped by *major* and the test variable *GPCTT Score*. All assumptions (independence and scale type) of the Kruskal-Wallis test were met. There was no statistically significant difference between the scores on the GPCTT by students with different majors (*H*(1) = .291, *p* = .988), with a mean rank of 39.41 for non-psychology students and mean rank of 39.51 for psychology students.

Discussion

This study aimed to create a measure for psychological critical thinking and to gather evidence of its validity. We assumed that the newly created measure (GPCTT) should exhibit convergent validity by correlating positively with an already established measurement for psychological critical thinking, the PCTE (Lawson et al., 2015). In line with previous research (Lawson et al., 2015) and the predominant view in the literature that CT has domain-specific aspects (Bailin et al., 1999; Ennis, 1989; Lai, 2011; Lawson, 1999; Lawson et al., 2015; Paul, 1992), together with previous studies finding higher scores of CT in psychology than other majors (Lawson et al., 2015), we also hypothesised that the GPCTT should be able to discriminate between psychology and non-psychology students.

Contrary to our expectations, we did not find a positive association between the score of the GPCTT and PCTT. There was also no support for the hypothesis that psychology students score higher than non-psychology students on the GPCTT; both groups performed similarly.

Since we could not establish the validity of the measure, one cannot use it to make inferences for the general discussion on whether CT is domain-specific. Furthermore, the sample was hugely unbalanced, with only 14.1% of students not following a psychology major. This significantly limited our power to detect any significant effects. Moreover, we did not control for study-year for the analysis. Since previous research detected differences between first-year students and students further along in their studies (see e.g., Holmes et al., 2015; Lawson et al., 2015; Williams et al., 2003), this might limit the validity of our results. Therefore, either sample might have been skewed more towards specific scores based on the amount of education received.

Across all five categories, the interrater reliability can be described as moderate (Landis & Koch, 1977). We ran a pilot study and discussed the scoring afterwards to weed out unclarities in the rubric we created. However, this practical test was limited to a few essays and contrary to other measures (e.g., Lawson et al., 2015), we did not practice the scoring until high inter-rater reliability was achieved. In our opinion, too much practice would have taken away from the independence of the raters. Independence took precedence over interrater reliability to allow us to see what possible struggles and inconsistencies first-time raters would have to deal with. It also gave us the ability to understand better which parts of the rubric were insufficiently phrased and constructed since all scoring differences had been discussed between the raters.

In the following review we will enumerate a few of the most influence flaws that were found:

For example, a participant who avoids using any outside information that could not be appropriately referenced would receive a *2* in the rubric. On the other hand, a participant who actively pointed out a methodical flaw in the given material would score a *2*. Therefore, the amount of critical thinking that would be needed to complete either task might not have been in relation to the scoring. In future iterations, we would suggest an attempt to make scoring categories more coherent. We recommend that researchers put sizable effort and thought into how much CT each task requires and how to put this into relation to a fair scoring. One way to achieve this is by entirely abstaining from using dichotomous scoring schemas. Using multilevel rating would allow scoring to be more comparative across categories.

Further supporting this proposal comes from analysing how the category Synthesis was scored, where the dichotomous scoring did not work well. Having read the participants essays, it became clear that this rating was not appropriately measuring the difference in contributions. At this point, a student that would show a minimum effort of weighing evidence would score in the same bracket as a participant integrating and weighing the evidence exhaustively. This flaw becomes even more apparent when looking back to the definition created for this project and when glancing back to what researchers previously stated; the ability to synthesise and form conclusions is an essential part of CT (Ennis, 1993; Lai, 2011). In hindsight, having such an unsophisticated scoring for this category does not seem appropriate. In the future we suggest using an ordinal scale rating, similarly to what was done for “methodology” or “fallacy”. Thus, making the assessment of this CT aspect more refined and enabling a more nuanced decision on students’ performances.

It also became clear that the strict way we set up the scoring may have been unfit to assess CT. As suggested by Shavelson and colleagues (2019), we assed judgmental errors in our categories for measuring CT. However, the way the scoring was structured was not congruent with the authentic context that the GPCTT was supposed to assess. For instance, as one of the criteria, we judged the use of fallacies, where A) using a fallacy would score a *0* B) neither using nor mentioning a fallacy a *1* C) pointing out one fallacy a *2* D) pointing out more than one fallacy a *3*. Therefore, explicit mentioning the fallacies would result in a higher score than not making use of them. However, our instructions only made it clear to "critically evaluate the articles and come to a final conclusion". There was no unequivocal instruction to point out (name) the fallacies explicitly. Consequently, participants who would correctly detect and sidestep the fallacies would be punished unfairly for their omission in their essay. While it is clear that a critical thinker should spot fallacies and make correct assessments based on that information, the way the instruction was phrased did not allow for fair scoring.

Cronbach's alpha showed questionable consistency across all items. Deleting any of the items from the analysis did not show improvements in the alpha score. Hence, while the total agreement requires improvement, the results showed that none of the items stood out as particularly unfitting. Multiple issues and areas for improvement have been laid out before as examples of why the rubric may have been inadequate in capturing the concept. Therefore, improving the rubric, e.g., as suggested above, might affect GCTT scores in a way that is not only fairer for participants, or more aligned with the literature, but also might improve interrelatedness between the items.

Setting aside the measure's shortcomings, participants' ability (or lack thereof) to transfer the learned CT skills to a more authentic setting might also explain the differences in scores between the PCTE and GPCTT, and hence lack of correlation. In contrast to the GPCTT, the PCTE more closely resembles the context within the classroom, given its design (Dibartolo et al., 2016)**.** Generally, evidence on the transfer of CT skills is mixed (Nickerson, 1988). Willingham (2007) stated that while students might be able to show CT skills in one context, transferring those skills to other contexts is highly challenging. Lehman & Nisbett (1990) found that undergraduate training is able to influence the way people approach real-life events, after assessing CT skills at T1 and then investigating students' abilities years later. Kennedy and colleagues (1991) noted on this topic that CT transfer is possible if the skills have been taught with the transfer in mind; practice in different contexts and domains is a prerequisite. Others like McPeck (1990) agree that authentic learning activities can help to transfer CT abilities to real-world contexts. Generally, the literature on the transfer of CT is still very contradictory, pertaining to the ambiguity around the distance of such a transfer (Bailin, 2002). While some studies investigate CT transfer in the light of different domains, others do so in the light of different contexts. We are still convinced that the use of authentic scenarios is preferable, however, anther methods to establish validity may need to be considered. A different measure that also makes use of authentic scenarios might show convergent validity with the GPCTT. Furthermore, using experts to judge students' ability and comparing it to compare it to GPCTT scores might be time consuming, but perhaps be a superior approach to validate the measure.

All in all, it becomes clear that more literature is needed on CT transfer to new contexts. Current literature is often limited and focused on domain transfer (Lai, 2011). Transfer within a particular domain but to new contexts seems to be equally if not more important, based on the practical implications for education and the workplace. What would be the use of educational institutions teaching CT if students cannot transfer the learned skills even within their own field?

Conclusion

Critical thinking has significant practical relevance and is growing as an educational goal, and so should the effort to teach and assess it correctly. Thus, we suggest so-called "authentic" assessments to be at the forefront of research since they use interpretive knowledge and the ability to evaluate messy problems is closest to the real-world application of CT (Broudy, 1977; Dibartolo et al., 2016). The goal of educational institutions to prepare students for their later academic or work careers will therefore depend highly on their ability to teach CT in a way that transfers those contexts. The struggles and experiences made in this study, trying to create an authentic measure for CT, can here be informative for more development in this direction.

References

Association of American Colleges and Universities. (2011). *The LEAP Vision for Learning: Outcomes, Practices, Impact, and Employers' View*. Washington, DC: Association of American Colleges and Universities.

Association of American Colleges and Universities. (2017). *On Solid Ground: VALUE Report 2017.* AAC&U. https://www.aacu.org/publication/on-solid-ground-value-report-2017

American Psychological Association. (2013). APA guidelines for the undergraduate psychology major: Version 2.0. Washington, DC. Retrieved from http://www.apa.org/ed/precollege/about/psymajor-guidelines.pdf

Bailin, S. (2002). Critical thinking and science education. *Science & Education, 11*(4), 361–375.

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualising critical thinking. *Journal of Curriculum Studies, 31*(3), 285–302.

Broudy, H. S. 1977. "Types of Knowledge and Purposes of Education." In *Schooling and the Acquisition of Knowledge*, edited by Richard C. Anderson, Rand J. Spiro, and William E. Montague, 1–17. Hillsdale, NJ: Lawrence Erlbaum.

Dibartolo, P., Duncan, L., Ly, M., & Rudnitsky, A. (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

Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18*(3), 4–10.

Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice, 32*(3), 179–186. https://doi.org/10.1080/00405849309543594

Ennis, R. H., & Weir, E. (1985). *The Ennis-Weir critical thinking essay test*. Midwest Publications.

Facione, P. A. 1990. *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instructions.* Research Findings and Recommendations. Millbrae: California Academic Press.

Hitchcock, D. (2020). Critical Thinking. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.). Stanford University. https://plato.stanford.edu/archives/fall2020/entries/critical-thinking/

Holmes, N. G., Wieman, C. E., & Bonn, D. A. (2015). Teaching critical thinking. *Proceedings of the National Academy of Sciences of the United States of America, 112*(36), 11199–11204. https://doi.org/10.1073/pnas.1505329112

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

Lai, E. R. (2011). Critical Thinking: A Literature Review Research Report. London: Parsons Publishing.

Landis, J.R., & Koch G.G. (1977). The measurement of observer agreement for categorical data. Biometrics. Mar;33(1):159-74. PMID: 843571.

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.

Lehman, D., & Nisbett, R. (1990). A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology*, *26*(6), 952-960. https://doi.org/10.1037/0012-1649.26.6.952

Lewis, A., & Smith, D. (1993). Defining higher-order thinking. *Theory Into Practice*, *32*(3), 131-137. https://doi.org/10.1080/00405849309543588

Lipman, M. (1988). Critical thinking—What can it be?. *Educational Leadership, 46*(1), 38–43.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, *2014*(1), 1–23. https://doi.org/10.1002/ets2.12009

Liu, O., Mao, L., Frankel, L., & Xu, J. (2016). Assessing critical thinking in higher education: the HEIghten™ approach and preliminary validity evidence. *Assessment & Evaluation In Higher Education*, *41*(5), 677-694. https://doi.org/10.1080/02602938.2016.1168358

Liyanage, I., Walker, T., & Shokouhi, H. (2021). Are we thinking critically about critical thinking? Uncovering uncertainties in internationalised higher education. *Thinking Skills And Creativity*, *39*(100762). https://doi.org/10.1016/j.tsc.2020.100762

McPeck, J. E. (1990). Critical thinking and subject specificity: A reply to Ennis. *Educational Researcher, 19*(4), 10–12.

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist, 36*(3), 211–218. https://doi.org/10.1080/00050060108259657

Mulnix, J.W. (2012). Thinking critically about critical thinking. *Educational Philosophy and Theory, 44*(5), 464–479. https://doi.org/10.1111/j.1469-5812.2010.00673.x

Nickerson, R. S. (1988). On improving thinking through instruction. *Review of Research in Education, 15*(1988–1989), 3–57.

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

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

Paul, R., Elder, L., & Bartell, T. (1997). *California Teacher Preparation for Instruction in Critical Thinking: Research Findings and Policy Recommendations*. California Commission on Teacher Credentialing, State of California.

Resnick, L. (1987). Education and learning to think. Washington, DC: National Academy Press.

Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. P. (2019). Assessment of university students' critical thinking: next-generation performance assessment. *International Journal of Testing, 19*(4), 337–362. https://doi.org/10.1080/15305058.2018.1543309

Smith, J. C. (2011). *Pseudoscience and Extraordinary Claims of the Paranormal: A Critical Thinker's Toolkit*. New York, NY: John Wiley and Sons.

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

Tiruneh, D. T., Verburgh, A., & Elen, J.. (2014). Effectiveness of Critical Thinking Instruction in Higher Education: A Systematic Review of Intervention Studies. *Higher Education Studies, 4*(1). https://doi.org/10.5539/hes.v4n1p1

Van Gelder, T. (2005). Teaching critical thinking: Some lessons from cognitive science. *College Teaching, 53*(1), 41–48.

Watson, G., & Glaser, E. M. (2019). *Watson-Glaser III critical thinking appraisal: User's guide and technical manual*. Pearson.

Williams, R. L., Oliver, R., Allin, J. L., Winn, B., & Booher, C. S. (2003). Psychological critical thinking as a course predictor and outcome variable. *Teaching of Psychology, 30*(3), 220-223.

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

Appendix A

The GPCTT

In the following, you will find the instructions and materials of the GPCTT that had been presented to the participants.

Instruction for Participants

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.

Introduction to Materials for Participants

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.

Materials GPCTT

Get rid of resits

Author: 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.

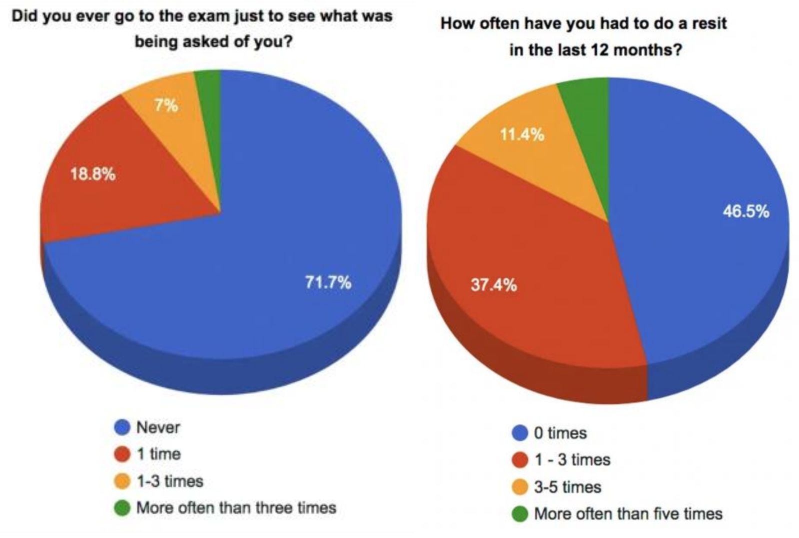
In conclusion, the tests should be used to steer education. Plan many, forcing students to keep studying. Offer students the opportunity to compensate for bad grades so they don’t get hung up on a single failed test. Offer cumulative testing, to ensure that a later good grade makes up for an earlier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



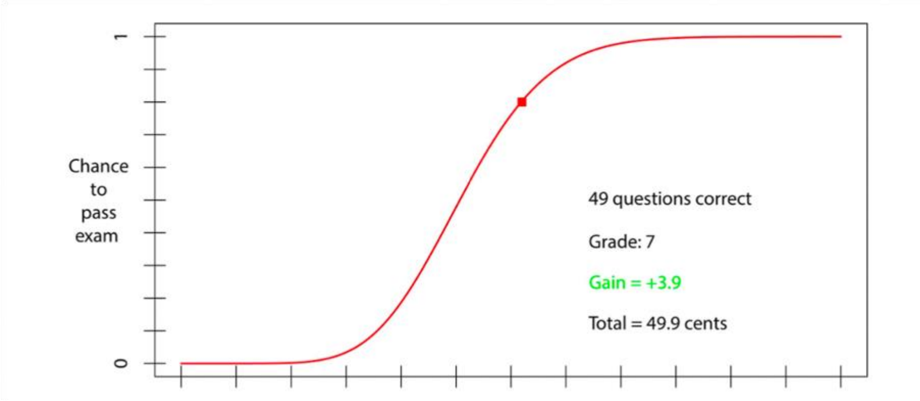
Then the main question: should resits be discouraged by scheduling them at unusual times? A fair number of students (27.1%) don't think the idea is too bad. The most used argument is that the increase in pressure will force students to start studying earlier and take exams more seriously.

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

Scoring Rubric of the GPCTT

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.	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

	Status-quo bias: 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).*			“keeping the resits 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)	The34articipant 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	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.

	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.

Note. This rubric was created on the basis of the Association of American Colleges and Universities (AAC&U) Critical Thinking VALUE Rubric (2017). Retrieved from https://www.aacu.org/value-rubrics

Appendix C

PCTE Scoring

In the following, you will find the Coder Training Sheet by the Mount St. Joseph University, that was used to assess participant performance on the PCTE.

Scoring Manual PCTE

Scoring Scale: *0 = didn't identify a problem; 1 = mentioned there was a problem but misidentified it; 2 = mentioned the main problem but also mentioned less relevant problems; 3 = mentioned only the main problem. The Sum of all scores is the final score.*

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.

Sample Answers (with a score in parentheses)

1. *These similarities are by chance (3)*
2. *Yes, I would agree that researchers can conclude our personalities are influenced by genetics, but I do not think that they can make these conclusions based on these specific case studies (1)*
3. *A limited set of evidence, not taking into account any other factors, selection biased (1)*

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.

Sample Answers (with a score in parentheses)

1. *The sample was not representative (1)*
2. *No control group (3)*
3. *The drug reduced depression after 4 months in those 100 cases. I feel that the research has not tested the drug enough to support her conclusions (1)*
4. *There is no control group to compare those who took the drug to those who didn't. And the sample was not representative of the general population (2)*
5. *Placebo effect (3)*

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.

Leading Question

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.

Correlation NE causation

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.

Unrepresentative sample (male juvenile offenders not the same as all film goers)

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.

1. *He did not make his findings falsifiable (3)*
2. *Biased, accuracy issues (1)*
3. *He did not take into account the 3rd variable problem. Something else, other than the treatment, may have impacted the number of people who lived or died (1)*
4. *Problem: Need for a control group; made impossible to falsify (2)*

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.

Sample Answers (with a score in parentheses)

1. *Correlation not equal to causation. There is not only one factor/variable leading to alcoholism. (2)*
2. *There may be more than one cause of alcoholism (3)*
3. *Stating they found THE cause isn't falsifiable (1)*

Appendix D

The final qualifications of the Psychology Bachelor students at the University of Groningen

1. The learning outcomes

1.1 Final Qualifications

Upon completion of the Bachelor's program, students must be able to conduct supervised research for which they must possess a broad knowledge of the subject matter, possess adequate skills, and be able to reflect critically. This leads to the following final qualifications:

- The student has knowledge of and insight into relevant and current concepts and theories in the main fields of psychological science and is able to reflect on psychological practice on the basis of this knowledge and insight;
- The student has knowledge of and insight into a broad spectrum of current techniques and methods in social scientific research, and is able to apply these to a number of research topics;
- From an academic attitude the student can analyse data from scientific research within the psychological domain, report the results and reflect on them;
- The student makes a start with applying knowledge and insight into the theories and concepts from psychology in an ethically responsible manner;
- The student has started to specialise.

1.2 General final attainment levels of the programme

The Bachelor's programme in Psychology is discipline-oriented and in most cases prepares students for the Master's programme in Psychology. The bachelor's course should therefore provide an adequate foundation for a seamless transition to the master's programme that is considered a minimum requirement for independent professional practice as an academic psychologist. To this end, the Bachelor's programme in Psychology offers students a broad basic training in named sub-fields of Psychology in which theoretical knowledge and academic training play a central role. A comprehensive training in theories, statistics, skills, methods and techniques of both fundamental and applied social science research takes place in this bachelor's phase. The programme offers students the opportunity to take responsibility for their own development and to increase, broaden or deepen the programme as they wish. Quality assurance (and quality control) is an inherent part of the programme with the aim of optimising the student's development into a professional.

In accordance with the guidelines of the Psychology Chamber, the Bachelor's programme in psychology includes the following components:

- Introductions to the following subfields of psychology: psychological function theory, biopsychology, developmental psychology, personality theory, occupational and organisational psychology, social psychology and psychopathology;
- Teaching and practice of the methods of psychological science: method theory, data analysis and statistics;
- Education and practice in the skills for professional practice, creating a basis that will enable the Master's phase to meet the entry requirements of the GZ study programme, for example;
- Global knowledge of the most important fundamental and application areas;

- The start of a specialisation in at least one of these, in preparation for a career in practice or as a researcher;
- Cross-domain courses such as Overview of Psychology, History of Psychology, Theory of Science;
- Integration of the above aspects in the form of a Bachelor's thesis, being a report of an empirical and/or theoretical research (whereby literature may be the source of the data).

1.3 Relation between learning outcomes and Dublin descriptors

The Bachelor's programme is based on the five Dublin descriptors. Below is a translation of each descriptor into the learning outcomes of the Bachelor's programme in Psychology:

1. Descriptor: Knowledge and understanding

The student has knowledge and understanding of the theories and findings of sub-disciplines of psychology, their interrelationships and their applications. Has knowledge and insight into the main fields and activities of a psychologist. Has knowledge and understanding of the process of experimental and field research. Has knowledge of and insight into the theoretical presuppositions of psychological research in comparison with other scientific disciplines.

2. Descriptor: Application of knowledge and insight

The student can identify, acquire and use knowledge to systematically solve problems. Can apply scientific knowledge to set up and carry out simple research. Is able to integrate knowledge from different fields. Can apply scientific knowledge from different fields to social situations. Can participate in social debates about policy that affects the field of study.

3. Descriptor: Judgment

The student is able to set up and carry out simple research; interprets the data and forms an opinion about the conclusions of the research based on considerations of relevant social, scientific and ethical aspects. Is able to assess knowledge sources and scientific publications. Is able to assess and justify choices made.

4. Descriptor: Communication

The student is able to communicate research findings and conclusions, both orally and in writing, to peers and third parties.

5. Descriptor: Learning skills

The student is able to actively and independently acquire and apply knowledge and insight in a research context and has the motivation to master the knowledge, insights and skills in psychology at an academic bachelor's level.