Creating a Measure for Psychological Critical Thinking

Laura Escudero Gimeno

s3668843

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group number: 2122_1a_09

Supervisor: Marcella Fratescu

Second evaluator: (dr.) Saleh Mohamed

In collaboration with: L. Jaarsveld, J. Nagler, E. van Nee, and T. Schröder

February 13, 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 role of critical thinking is increasingly relevant in education. Research shows that critical thinking can be domain-dependent or general. This creates a disagreement about the way it can be taught. The American Psychological Association and the University of Groningen consider critical thinking an important learning outcome for undergraduate psychology students. The goal of this study was to create and validate a measure for psychological critical thinking. We created a measure for psychological critical thinking and compared participants' scores on this measure with another, validated measure – the Psychological Critical Thinking Exam. A within-subject correlational study (N=78)was conducted. We considered the difference between the aspects determined by us as part of psychological critical thinking and those determined by Lawson (1999) as a possible explanation for our results. Results showed that there is no statistically significant relationship between the scores from the validated PCTE and the created GPCTT. The findings suggest that future research is needed for a validated domain-specific measure and on the quality of the teaching methods used for critical thinking. Further practical and theoretical implications are discussed.

Keywords: Psychological critical thinking, psychology students, Psychological Critical Thinking Exam, Groningen Psychological Critical Thinking Task

Creating a Measure for Psychological Critical Thinking

Critical Thinking (CT) requires complex cognitive skills like evaluation, synthesis, and analysis of information for outcomes such as problem-solving or scientific reasoning (Clifton et al., 1996; APA, 2012). CT is considered a high-level construct, and its role is increasingly relevant in education. Education institutions such as the University of Groningen (UG) are starting to introduce topics related to it, for instance, avoiding biases and empirical reasoning (Lawson, 1999).

There are disagreements within the definition of CT between the two primary disciplines, philosophy and psychology, where some aspects are included in one approach and excluded in the other (Petress, 2004). For instance, in the philosophical approach, CT is defined as "the reflective and reasonable thinking that is focused on deciding what to believe or do" (Ennis, 1985, p.45). Philosophers refer more to logical thinking and to general qualities and characteristics, such as appropriate self-reflection, that a critical thinker might have rather than behaviors they might perform (Lewis et al., 1993; Lamont, 2020). Conversely, psychologists focus more on behaviors a critical thinker should engage in, including skills and tasks such as decision making, inferring a conclusion, and problem-solving (Halpern, 1998; Stenberg, 1986; Willingham, 2007). The psychological approach defines CT as "the use of those cognitive skills or strategies that increase the probability of a desirable outcome" (Halpern, 1998, p.450). Most of the definitions agree that CT involves thinking, but there is no consensus on the type of thinking involved. Despite the differences among the researchers, there are some aspects they all agree on. Judging or evaluating, identifying assumptions, seeing both sides of an issue; analyzing arguments, claims or evidence; making inferences using inductive or deductive reasoning; interpreting and explaining; or solving problems (Ennis, 1985; Facione, 1990; Halpern, 1998; Lipman, 1988; Nolet & Tindal, 1995; Paul, 1992; Willingham, 2007) are some of the relevant abilities and behaviors to critical thinking.

Since there are numerous controversies in the definition of CT depending on the domain, there is an ongoing debate concerning whether CT should be considered general or domain specific (Petress, 2004; Lamont, 2020; Lai, 2011). This means that depending on the researcher's field of interest they use different aspects to define CT. Some researchers think that the possibility of generalization to other contexts is unlikely and that it is harder to learn in a generic way than given a domain (Willinga, 2007). Others think that for students to transfer those skills, they must practice among domains, and the teachers should explicitly teach them to do so. Lastly, Bailin (2022) considers that there is a need for a field specific definition since the criteria for evaluation of arguments, standards, or pieces of evidence might depend on the domain. The criteria needed for evaluating the quality of a sculpture are different than the ones of an article. This situation producing difficulties in creating a concise, clear and uniform definition of CT.

Nowadays, CT is considered an essential learning goal in psychology by the American Psychological Association (APA Guidelines for the Undergraduate Psychology Major), which considers psychological critical thinking (PCT) as one of the four skill-based goals representing the expectations for undergraduate psychology majors across different educational contexts (APA, 2012). The University of Groningen (UG) also includes psychological critical thinking as a learning goal for undergraduate psychology students (University of Groningen, n.d.). By introducing in their curriculum courses such as Academic Skills where students learn how to effectively apply critical thinking in their professional and personal life.

There is paucity of evidence in the way of assessing CT because of the variability of the criteria depending on the field of interest (Lipman, 1998). There are different recommendations and assessments for it, such as using both multiple-choice and open-ended formats (Ku, 2009). Different measures for CT have been proposed, however no instrument

is ideal for all situations, as they all have shortcomings. For instance, the Watson-Glaser Critical Thinking Appraisal has as an advantage that it is widely used. However, it is multiple-choice format, has validity issues and measures general CT (Watson & Glaser, 1980). Oppositely, the Ennis-Weir Critical Thinking Essay Test is more advantageous, as it has an essay-like format, high validity, reflects real-life commands, and tests the ability to analyze responses, however it measures only general CT (Ennis & Weir, 1985). Additionally, the Psychological Critical Thinking Exam (PCTE) has the drawback of having short written answers format instead of essay-like format, nevertheless it has benefits, as it is domain-specific, and has good content and predictive validity (Lawson, 1999). Haw (2011) and Stark (2012) suggest that by using the PCTE we can determine how students' scores differ when trained at different levels. The APA Guidelines for the Undergraduate Psychology Major (APA, 2012) have recommended it as an adequate measure of critical thinking in psychology. As shown there are many different options for measuring CT but not so many have been validated for measuring domain-specific critical thinking. Therefore, there is a need for a PCT measure that overcomes the shortcomings of the PCTE.

For creating a measure, first, we need a clear definition of PCT. The PCTE defines it as "the ability to evaluate claims in a way that explicitly incorporates basic principles of psychological science" (Lawson, 1999, p. 207). We define PCT as "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 based on the definition from the Valid Assessment of Learning in Undergraduate Education (VALUE) rubric, an assessment approach designed to measure students' abilities, skills, and dispositions (Rhodes, 2010). It contains five aspects that many researchers mention in their definitions (Halpern, 1998; Lamont, 2020; Lewis et al., 1993; Sternberg, 1986; Willingham, 2007) and students can score between 0 and 4 for each aspect.

The first aspect evaluates the clarity and description of the student's consideration of the problem (explanation of issues). Another criterion measures the extent to which selection and usage of information are correctly made to support a point of view (evidence) and an aspect for analyzing how the student is influenced by different contexts and assumptions (influence of context and assumptions). The fourth aspect assesses the student's position, considering the perspective and hypothesis used (student's position). The last one determines if the conclusions of the student's are logical and in line with the evidence (conclusions and related outcomes) (Rhodes, 2010).

Our measure, the Groningen Psychological Critical Thinking Task (GPCTT) will focus on measuring skills recommended by the APA guidelines, what tasks the UG considers that undergraduate psychologists should learn, and on the aspects mentioned in the VALUE rubric, evaluating the students on five criteria. The first point of the VALUE rubric is Methodology is directly related to the learning goals of different courses at UG, such as Introduction to Research Methods, where theory is learned about validity, reliability of constructs (Rijksuniversiteit Groningen, n.d). It evaluates the extent to which the participant took into account the methods used in the articles, such as mentioning internal validity. Secondly, the aspect of Fallacy mentioned in the APA guidelines (2012) is one of the skills needed to interpret psychological phenomena using scientific reasoning correctly. This facet focuses on the participant's ability to identify different fallacies, such as the status-quo bias or the appeal to authority fallacy. Furthermore, the Assumption of Authors, based on the criterion" influence of context and assumptions" mentioned in the VALUE rubric, assesses the ability to spot claims lacking supporting evidence. Additionally, the Bias of Participant and Synthesis aspects, both based on the VALUE rubric and the APA guidelines (2012), focus on whether the participant uses information external to the one provided and whether the participant can combine and weigh contradictory evidence.

To sum up, our goal in this study is to create and validate a domain-specific CT measure for psychology bachelor students from the University of Groningen. For doing so, the assessment measure developed by Lawson (1999), the Psychological Critical Thinking Exam (PCTE), will be used as a comparison to our created measure Groningen Psychological Critical Thinking Task (GPCTT). We hypothesize that the student's scores on GPCTT will significantly correlate with the scores on PCTE.

Method

Participants

A total number of 78 Bachelor 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 data consisted of 67 psychology students (54 first-year, 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% were non-native speakers. 1% came from an Asian country, 87% came from Western countries, and 12% from other countries. 82% indicated that they put in their best effort, 18% did not put their best effort in the task. First-year psychology students were recruited through SONA; any participants from a higher semester or different bachelor's 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 randomized to avoid a possible order effect. Before the survey was distributed to the potential participants, it 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 that had been sent to them. Before the survey started, the participants would see a screen with information about the study and were informed about the amount of SONA credits they would receive and 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 to proceed with the survey. The GPCTT starts with an instruction that states that participants will be presented with three articles about resit exams at the University of Groningen. Participants are instructed to write an essay in which they critically evaluate the articles and come to a conclusion about resits (Appendix A). Subsequently, the PCTE continues with seven research-related scenarios and asks the participants to state whether there is a problem with the researcher's explanation and conclusion.

After finishing both tests, participants were asked about demographic information (age, gender, major, native language, and ethnicity). Lastly, 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

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 RUG-Board 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. 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 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 recognize this Fallacy. The last article we included in our assessment packet is one 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 of Methodology, Fallacy, Assumption of Authors, Bias of Participants, and Synthesis. For the aspects Methodology, Fallacy and Assumption of Authors, the participant can score on a 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 twice of the aspects correctly). For Bias of Participant and Synthesis, each participant can 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 be 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. Lawson and colleagues (2015) have developed and validated this version. For each scenario, a conclusion was reached (Appendix C), 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.

Results

GPCTT Reliability Checks

A pilot study was conducted which served as training for the raters and as an opportunity to gather feedback. The answers for the study were scored independently by two randomly assigned blinded raters. Then, discrepancies in scoring were resolved so that one final score for each question was given. Inter-rater reliability was assessed using Fleiss' Kappa (κ) because it determines the level of agreement between two or more raters. In addition, it does not assume that there is always the same rater for all items. For the items of methodology (κ=.584, 95% CI [.433, .735]), fallacy (κ=.592, 95%CI [.412, .772]) and synthesis (κ=.505, 95%CI [.283, .727]) there is a moderate agreement, between raters and for the items of assumptions (κ=.341, 95%CI [.153, .529]) and bias of participants (κ=.505,

95%CI [.283, .727]) there is a fair agreement (according to an interpretation provided by Hartling et al., 2012). Overall, with an average of κ=.466, there can be concluded that there is a moderate agreement between raters.

Then, we tested the internal consistency to see if the items in the GPCTT are measuring the same thing. For doing so, a Cronbach's Alpha test was done, resulting in a Cronbach's Alpha of .495, showing poor internal consistency according to UCLA: Statistical Consulting Group (n.d.). Additionally, the mean inter-item correlation value was .184; this confirms that the internal consistency is low, meaning that the items as a group are not closely related.

Correlation between GPCTT and PCTE

For assessing normality, the Kolmogorov-Smirnov test was used. This is an important check as many parametric statistical tests rely on the assumption that the variables are normally distributed. This measure tests the null hypothesis that the data set is normally distributed, using an alpha level of 0.05. For the GPCTT, this test gives the value .193 (df=78, p<.001) and for the PCTE a value of .133 (df=78, p=.002). It can be concluded that the null hypothesis is rejected, and the data is not normally distributed.

For testing our hypothesis, the correlation between both tasks needs to be examined. Since our data is not normally distributed and the assumptions of ordinal data and monotonicity were met. We used Kendall's tau (τ) non-parametric correlation between the total scores of the GPCTT and the total scores of the PCTE, obtaining a score of τ=.068, p=.433. This means that there is insuficient evidence to conclude that there is a significant relationship between the scores.

Discussion

The goal of the study was to create and validate a measure for psychological critical thinking for the psychology bachelor students at the UG. For doing so, the participants'

results from the created GPCTT were compared against the validate measure (Lawson, 2015). We hypothesized that results from both measures would significantly positively correlate. We did not find support for it, meaning that the correlation was nonsignificant (τ=.068, p=.433) between scores. A reason for this could be the different aspects of psychological thinking that each measure is grading. On the one hand, the PCTE focuses more on different aspects of methodology inside different research, and on the other hand, that is only one criterion out of the five the GPCTT is measuring. Another reason for not finding a significant correlation could be the different types of tasks each test is asking for. The PCTE demands short written answers and the GPCTT for an essay-based answer. Evidence shows that a better way to encapsulate the features of CT is using open-ended measures (Ku, 2009).

The findings should be interpreted carefully in light of the limitations highlighted by the domain-specificity CT. First, the sample of participants is unbalanced; only 14.1% were non-psychology participants, so we could not generalize the findings to participants from different bachelors. A future direction would be recruiting a larger and more diverse sample. As Prat-Sala & Van Duuren (2020) mention in their study, the level of CT skills differs over different study years, but this aspect was not controlled for in the current research. This is a crucial aspect to consider as there were two types of participants, the first-year students from SONA pool and the ones we sampled, probably from the second or third year or above. Therefore, it might be that the generally low scores and the nonsignificant correlation between tests can be attributed to the high percentage of first-year participants. However, not much research has been done in this area so further studies should explore the impact it has on CT.

Additionally, another limitation that could have influenced the scores of the study was motivation. As Facione (2000) mentioned, there is a tendency to score higher in critical thinking when the motivation is high. In this case, there were two different motivation factors, first-year students were rewarded in credits for their courses, and second and third-year

students were given the chance to enter a lottery to win 15€. This could have affected participant's effort in our study, thus influencing our results.

Additionally, to keep validating the GPCTT, we did a variety of reliability tests. For instance, we checked the test's internal consistency, which showed that the aspects we selected as criteria for measuring psychological critical thinking are not the same characteristics of critical thinking. A reason for this could be that we tried to have as many diverse criteria as possible, thus causing us to measure aspects and skills not highly related to critical thinking, such as synthesis, which is more focused on the ability to combine evidence than on evaluating it. For improving this, some aspects could be redefined as methodology itself could be divided into two different criteria as it is a very broad aspect.

Moreover, as five researchers graded the measures, we assessed the inter-rater reliability. This results in a moderate agreement between raters, meaning that the data collected by researchers is moderately similar (Inter-Rater Reliability, n.d.). For this kind of study, inter-rater reliability must be as high as possible to ensure measure's validity. In this case, the results show that there is some consensus on the rating, but there is still a need for more clarity in the criteria, or better training should be offered to the raters. For instance, for the GPCTT, in the synthesis criterion defining what sufficient ability is. Moreover, for the assumptions of authors criterion, giving more examples of quotes used by the authors that are considered as non-valid arguments.

Our suggestion for future research is to create a longitudinal within-subjects study, where the same students are measured along with their bachelor's and among different bachelors. We would recommend doing the first assignment during the first days of the university to provide us with a baseline of their knowledge and how their critical thinking skills developed during university. This could give more input about the domain specificity CT debate, as we would have more data to compare the scores from different bachelor's

degrees. Additionally, the scores can be compared within each participant from the first day of the university until the last day. Along with these advantages, comes a high risk of drop-out which should be taken into consideration (Caruana et al., 2015).

In conclusion, in this study, we shed light on the importance of critical thinking in education, along with the need for a clear domain-specific definition. While our findings show that our measure was not in line with already validated measures of psychological critical thinking, future research on this is needed, considering the limitations we encountered. An intriguing aspect to explore for the future, would be to further elaborate on the quality of education the University of Groningen provides on critical thinking. Also, it might be beneficial for different institutions that want to attain this goal.

References

Allin, J. L, Booher, C. S., Oliver, R., Williams, R. L., & Winn, B. (2003). Psychological critical thinking as a course predictor and outcome variable. *Teaching of Psychology*, 30, 220–223. https://doi.org/10.1207/S15328023TOP3003_04

American Psychological Association. 2013. *APA Guidelines for the Undergraduate Psychology Major: Version 2.0*. Washintong, DC. http://www.apa.org/ed/precollege/undergrad/index.aspx.

APA Dictionary of Psychology (n.d.) https://dictionary.apa.org/internal-consistency

Ammirati, R., Lilienfeld, S. O., & Michal, D. (2012). "Distinguishing Science from Pseudoscience in School Psychology: Science and Scientific Thinking as Safeguards against Human Error." *Journal of School Psychology* 50: 7–36.

Ansari, M., Dryden, D. M. Hamm, M., Hartling, L., Hempel, S., Milne, A., Santaguida, P. L., Shekelle, P., Tsertsvadze, A., & Vandermeer, B. (2012). *Table 2, interpretation of fleiss' kappa (κ) (from Landis and Koch 1977) - validity and inter-rater reliability testing of Quality Assessment Instruments - NCBI BOOKSHELF*. Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments. https://www.ncbi.nlm.nih.gov/books/NBK92295/table/methods.t2/

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

Bensley, D. A., & Murtagh, M. P. (2012). Guidelines for a scientific approach to critical thinking assessment. *Teaching of Psychology*, 39, 5–16. https://doi.org/10.1177/0098628311430642

Bereiter, C., & Scardamalia, M. (2006). "Knowledge Building: Theory, Pedagogy, and Technology." In *Cambridge Handbook of the Learning Sciences*, edited by R. Keith Sawyer, 97–115. New York: Cambridge University Press.

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

Bonk, C. J., & Smith, G. S. (1998). Alternative instructional strategies for creative and critical thinking in the accounting curriculum. *Journal of Accounting Education, 16*(2), 261–293.

Bransford, J. D., Schwartz, D. L., & Sears, D. (2005). "Efficiency and Innovation in Transfer." In *Transfer of Learning: Research and Perspectives*, edited by Jose P. Mestre, 1– 51. Charlotte, NC: Information Age Publishing.

Burke, B. L., Kraus, S., Roberts-Cady, S., & Sears, S. R. (2014). Critical analysis: A comparison of critical thinking changes in psychology and philosophy classes. *Teaching of Psychology*, 41, 28–36. doi:10.1177/0098628313514175

Caruana, E. J., Roman, M., Hernández-Sánchez J., & Solli, P. (2015). Longitudinal studies. Journal of Thoracic Disease, 7(11), 540. https://doi.org/10.3978/j.issn.2072-1439.2015.10.63

Clifton, R. A., Etcheverry, E., Hasinoff, S., & Roberts, L. W. (1996). Measuring the cognitive domain of the quality of life of university students. Social Indicators Research, 38(1), 29–52. https://link-springer-com.proxy-ub.rug.nl/content/pdf/10.1007%2FBF00293785.pdf

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods and Applications*, 19, 497-515.

Despain, L. H., Gray, M. J., & Penningroth, S. L. (2007). A course designed to improve psychological critical thinking. *Teaching of Psychology*, 34, 153–157. https://doi.org/10.1080/00986280701498509

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

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

Ennis, R. H., Fisher, M. B., & Kennedy, M. (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.

Ennis, R. H., Millman, J., & Tomko, T. N. (2005). *Cornell critical thinking test, level Z* (5th ed.). Seaside, CA: The Critical Thinking Co.

Everitt, B. and Landau, S. (2003). A Handbook of Statistical Analyses using SPSS. Chapman & Hall/CRC Press LLC. https://www.academia.dk/BiologiskAntropologi/Epidemiologi/PDF/SPSS_Statistical_Analyses_using_SPSS.pdf

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

Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relation to critical thinking skill. *Informal Logic, 20*(1), 61–84.

Fischer, S. C., Spiker, V. A., & Riedel, S. L. (2009). *Critical thinking training for army officers, volume 2: A model of critical thinking.* (Technical Report). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Gibbons, J. D., & Kendall, M. G. (1990). Rank Correlation Methods. 5th ed. London: Griffin.

Gibbons, J. D., Kendall, M., & Marden, J. I. (1992). Rank correlation methods (5th ed.). Journal of the American Statistical Association, 87(417), 249–249. https://doi.org/10.2307/2290477

Gibson, James J., and Eleanor J. Gibson. (1995). "Perceptual Learning: Differentiation or Enrichment?" *Psychological Review* 62: 32–51.

Glaser, E. M., & Watson, G. (1994). *Watson–Glaser critical thinking appraisal, form S*. San Antonio, TX: Psychological Corporation.

Halonen, J. S. (1995). Demystifying critical thinking. *Teaching of Psychology, 22*(1), 75–81.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*(4), 449–455.

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

Inter-Rater Reliability. (n.d.). In Alleydog.com's online glossary. https://www.alleydog.com/glossary/definition.php?term=Inter-Rater+Reliability

Introduction to SAS, UCLA: Statistical Consulting Group. *What does Cronbach's alpha mean.* https://stats.oarc.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/

Jaroszewski, K., Johnson, E. J., Howell, G. L., & Tuskenis, A. D. (2011). Development and effects of a writing and thinking course in psychology. Teaching of Psychology, 38, 229–236. https://doi.org/10.1177/ 0098628311421318

Kendall, M. G. (1938). A new measure of rank correlation. Biometrika 30: 81–93.

Koziol, S. M., & Moss, P. A. (1991). Investigating the validity of a locally developed critical thinking test. *Educational Measurement: Issues and Practice*, *10*(3), 17–22.

Ku, K. Y. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity, 4*(2009), 70–76.

Laerd Statistics (2019). Fleiss' kappa using SPSS Statistics. *Statistical tutorials and software guides*. https://statistics.laerd.com/spss-tutorials/fleiss-kappa-in-spss-statistics.php

Lai, E. R. (2011). *Critical Thinking: A Literature Review.* Pearson Research Reports. https://www.researchgate.net/publication/297782058_Critical_thinking_A_literature_review

Lamont, P. (2020). The construction of "critical thinking": between how we think and what we believe. *History of Psychology, 23(3*), 232–251. https://doi.org/10.1037/hop0000145

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

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

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

Liu, A. (2004). *The Laws of Cool: Knowledge Work and the Culture of Information*. Chicago: University of Chicago Press.

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

Newson, R. (2002). Parameters behind "Nonparametric" Statistics: Kendall's tau, Somers' D and Median Differences. The Stata Journal, 2(1), 45–64. https://doi.org/10.1177/1536867X0200200103

Nolet, V., & Tindal, G. (1995). Curriculum-based measurement in middle and high schools: Critical thinking skills in content areas. *Focus on Exceptional Children, 27*(7), 1–22.

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

Pallant, J. (2007). SPSS. Survival Manual. *A Step by Step Guide to Data Analysis using SPSS for Windows.* http://www.muthar-alomar.com/wp-content/uploads/2013/01/SPSS_Survival_Manual.pdf

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

Petress, K. (2004). Critical thinking: an extended definition. *Education -Indianapolis Then Chula Vista-*, *124*, 461–466.

Prat-Sala, M., & van, D. M. (2020). Critical thinking performance increases in psychology undergraduates measured using a workplace-recognized test. *Teaching of Psychology*. https://doi.org/10.1177/0098628320957981

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

Rijksuniversiteit Groningen. (n.d.). Ocasys: Online Course Catalog. https://www.rug.nl/ocasys/gmw/vak/showpos?opleiding=4526

Rijksuniversiteit Groningen. (n.d.). Propaedeutic phase. https://student.portal.rug.nl/infonet/studenten/gmw/bachelors/psychologie-en/year-1/

Schuyler, D. (2003). *Cognitive Therapy*. New York: W. W. Norton & Company.

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

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

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

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

Appendix A

Instructions of the Groningen Psychological Critical Thinking Task

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 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 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 the articles

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

Author: Nelly McTally, 2020

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.

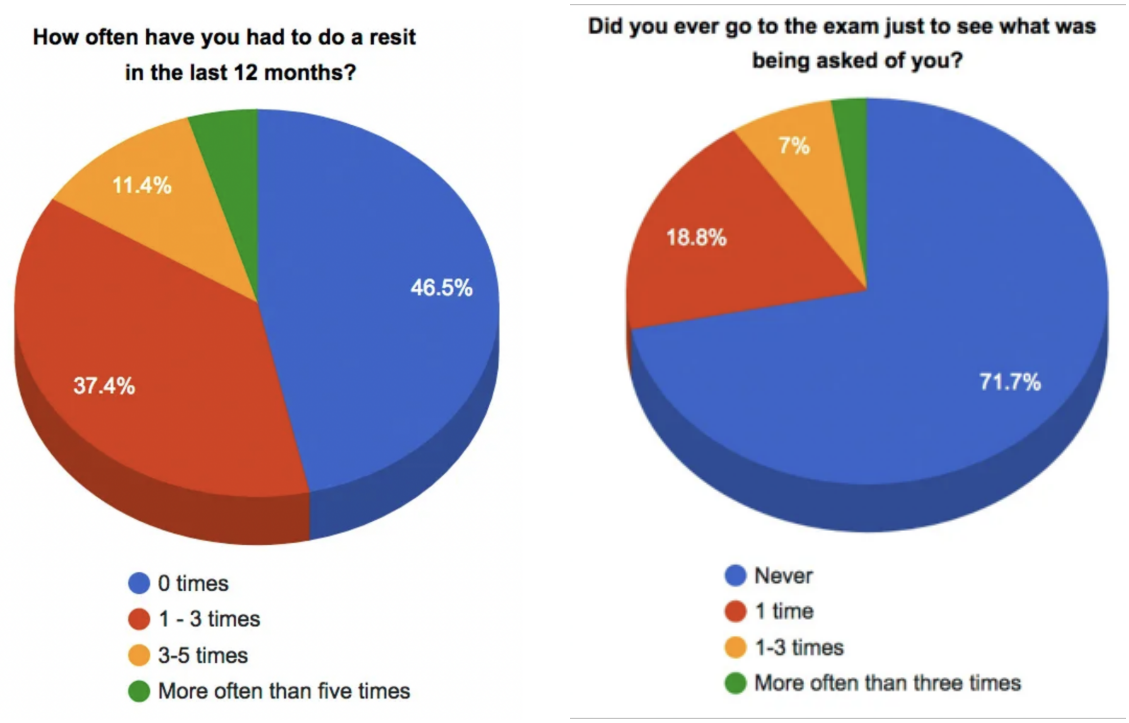
Resits are best planned at the end of the year, which allows students to focus solely on studying for them. It’s annoying for people who’ve planned vacations, but it should be annoying. ‘We have to make passing the norm. Right now, failing is the norm’, says Cohen-Schotanus.

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”)

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.

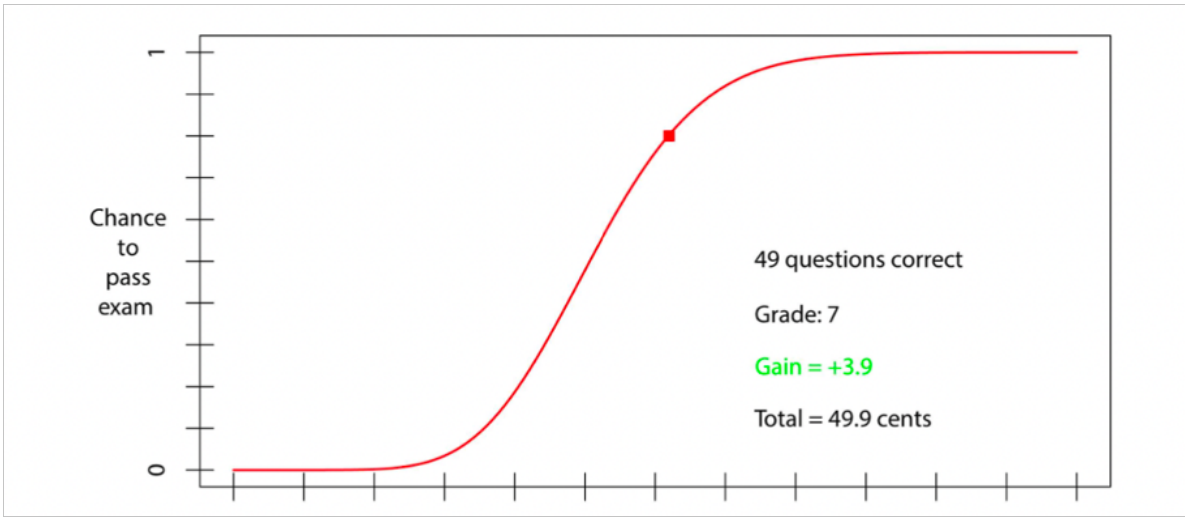
Nevertheless, almost three out of four students are against the measure. 'It would only cause more stress, and the pressure to perform is high enough already', many of them argue. Or: an exam is just a snapshot. Failure happens. Quite a few students argue that they shouldn't be punished for unforeseen 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.

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

Groningen Psychological Critical Thinking Task grading rubric

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

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

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

Psychological Critical Thinking Exam, Mount St. Joseph University, Coder Training Sheet

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.

Sample Answers (with a score in parentheses)

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)**