Measuring Critical Thinking: Test-Retest Reliability for a New Performance Task

Jackson Schwartz

Department of Psychology, University of Michigan

Dr. Kai S. Cortina

April 1, 2024

Author Note

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Art with Honors in Psychology from the University of Michigan in 2024.

Abstract

Critical thinking (CT) has become a relevant skill that employers look for when hiring college students and recent graduates. The International Performance Assessment of Learning Collaborative (iPAL) constructed a framework for assessments measuring CT in individuals or institutions. However, no systematic reliability analysis of the derived performance tasks is documented. We developed three CT performance tasks (Bryn Bower Series, BBS) based on the categories suggested by Zlatkin-Troitschanskaia et al. (2019) and a coding method that allows retesting necessary or test-retest reliability checks.

This study demonstrates the internal validity and reliability of the most recently developed task of the series piloted in the fall 2023 term. We ran both an immediate retest study and a semester-long longitudinal study. The measurement properties of the new task are commensurate with the properties of the two earlier-developed tasks. Having more equivalent tasks allows for more flexibility in longitudinal studies because every repeated measurement requires a different task to avoid memory effects.

Keywords: educational psychology, developmental psychology, critical thinking, performance task, assessment

Acknowledgments

I would like to thank those who helped me write this honors thesis. I want to express my gratitude to Dr. Cortina for your constant support and meaningful mentorship. You have been essential to my interest in research, making it tangible for me to contribute to the field without much experience. Additionally, to soon-to-be doctor Blake Ebright-Jones, I appreciate you introducing me to the lab and allowing me to join the collaborative team. The internal workings of our research lab are complex, and your teaching made it easy to learn and apply them. I also appreciate the members of the Cortina-Ebright Lab and their effort and work on the multiple projects that were included. I have been lucky to have landed this type of support system, and I greatly appreciate all of their unwavering support throughout this journey.

Measuring Critical Thinking: Test-Retest Reliability for a New Performance Task

Critical thinking (CT) is the process of analyzing, weighing, and synthesizing information to solve a problem or decide on a course of action regarding a given question or decision. In educational psychology, (CT) is often defined more broadly than in cognitive psychology, as it usually includes the skill to communicate the underlying deliberative process (Shavelson et al., 2019).

There is little doubt that college education increases subject-specific critical thinking, for example, epistemological reasoning in advanced natural science classes. But this does not necessarily generalize to authentic, real-world situations. Despite some controversy (Arum & Roksa, 2011; Pascarella et al., 2011), there is consensus that .5 SD is a solid ballpark guess for gains in CT over the course of four years of full-time college education (Pascarella & Terenzini, 1991; 2005). However, in their synopsis, Pascarella and Terenzini (2005) combined research with various operationalizations of the construct. In review, many measures used in the included studies disregard the multifaceted nature of the broader construct of CT as it is outlined in the assessment framework of the International Performance Assessment of Learning Collaborative (iPAL). Braun et al. (2020) highlight several subskills unique to the domain-unspecific and foundational conceptualization of the CT construct. For example, the trustworthiness of an information source is rarely an essential aspect of scientific/disciplinary reasoning. Still, it is a crucial aspect of everyday CT, reflecting potential biases and knowing what and who to trust.

Defining Critical Thinking in Current Times and the iPAL Framework

Shavelson et al. (2019) discuss CT as an important twenty-first-century skill, one of the domain-general skills that students are expected to pick up in a rigorous academic setting like college. In this broader sense, CT is congruent with the skill set that employers refer to as CT when hiring college students and recent graduates. However, traditional measures of CT, in particular those used in personnel recruitment, use tests with questionable validity (Shavelson et al., 2019). This criticism led members of the iPAL consortium to develop PA tasks that are drawn from complex real-world problems characterized by ambiguity and unclear solutions. The goal is to provide information sources with varying relevance and biases, reflecting the key aspects of CT as a twenty-first-century skill.

Bryn Bower Series (BBS)

Based on the iPAL coding framework (Zlatin-Troitschanskaia et al., 2019), we distinguish three aspects of CT (Ebright-Jones, 2024) that can be measured: a) argumentative depth (identifying pro and con arguments), b) bias/critical distance (reflecting relevance and trustworthiness of information), and c) communication skills (succinct integration of argumentative depth and bias/critical distance in a written document).

Those aspects can be assessed using curated tasks (Performance Assessments, PA) based on real-life problems of societal relevance (e.g., immigration, alternative energy, etc.), where students write critical essays to demonstrate critical reflection. Zlatkin-Troitschanskaia et al. (2019) used a multidimensional assessment rubric to assess the CT levels reflected in each student essay. They reported sufficient interrater reliability for their coding protocol. However, the assessment remains contingent on substantial coder training, is time-consuming, and lacks transparency. Therefore, this approach is not easy to scale up for research on a larger scale (e.g., in classroom settings).

In their research lab, Kai Cortina and Blake Ebright-Jones initially developed two PA tasks (called Bryn Bower Series, BBS) that differ from the traditional iPAL concept in one key aspect: the number of curated documents provided and the total number of arguments pertinent to the decision-making task is limited and known, which allows for a straight-forward coding of the student's responses, namely by counting the arguments provided and reflective statements regarding the relevance and trustworthiness of the sources. They were able to show that the interrater reliability of this coding outperforms traditional iPAL task coding that is based on rating scales. However, a balanced test-retest design study with the two BBS tasks revealed a low immediate repeated-measurement reliability, suggesting that the two tasks cannot be considered equivalent measurements in the sense of classical test theory, suggesting a substantial effect of the specific content of each PA.

While the research group was able to solve the interrater reliability shortcoming that was limiting the use of performance assessment tasks in the past, they still cannot present a measure that is reliable enough for individual diagnostics. However, Ebright-Jones (2024) was able to use the BBS to demonstrate CT improvement of college students through various levels of training. Note that on the aggregate level (freshmen vs. senior students), the BBS can already produce reliable CT mean scores if the tasks are randomly assigned to the members of each cohort.

Over the last summer, the lab developed a third BBS task, which is instrumental for estimating the between-task variation in the measurement properties more precisely. The purpose of this study is to demonstrate that the new task has the same measurement qualities as the first two tasks. More precisely, we hypothesize that all three BBS tasks measure the same latent construct, namely CT. Additionally, we hypothesize that the measurement qualities for the three subcomponents are equal across the three tasks.

- H1. The three tasks measure the same latent construct, i.e., critical thinking
- H2. The measurement qualities for the three subcomponents are equally measured across the three tasks.

Method

Participants

Study 1: The sample included 150 undergraduate students from the University of Michigan enrolled in Psych 356 (Educational Psychology) in the fall of 2023. Data were collected as part of the teaching unit "critical thinking." Participation was voluntary; students received extra credit for participating. Data were collected at three time points: in the third week of the semester (end of September), the 8th week (end of October), and finally in week 12 (last week of November). Participants' ages ranged from 18 to 25, consisting of men, women, and other genders, while having a culturally diverse sample. The study procedures were approved and received an exemption from the Institutional Review Board.

Study 2: A sample of 88 undergraduate students from the University of Michigan enrolled in Psych 111 (Introduction to Psychology) in the fall of 2023 participated in a two-hour data collection, working on two BBS tasks back-to-back. Participation was voluntary; students received 2 hours of subject pool credit. Our participants' ages ranged from 18 to 22, and their backgrounds varied, reflecting the cultural diversity of the university's undergraduate population. The study received exemption status from the Institutional Review Board.

Materials

The Bryn Bower Series (BBS) of critical thinking performance tasks were modeled after the iPAL framework. BBS tasks are characterized by limiting the total number of documents to 7-9 one-to-two-page items. This, in turn, limits the number of arguments pertinent to the decision-making to approx. 30 per task. BBS tasks are dilemma-based, which means that there are roughly as many arguments in favor as opposed to the given decision question at hand. The limited number of arguments allows for a straightforward coding of the student's responses, namely by counting the arguments provided and reflective statements regarding the trustworthiness of the sources.

Procedure

Participants had 50 minutes to work on each task (recommended: 25 minutes reading the documents and 25 minutes writing the essay). Participants were randomly assigned one of three BBS tasks in the first hour and then randomly assigned to a different task the second time than they had completed prior. In the second study (retest-study), the random assignment of tasks was restricted so that the subjects worked on the new task ("Cull") either as their first or second task. To limit fatigue, participants also watched a short video in between the two PA tasks in study two (2 hours total).

Coding

There was a substantial amount of pre-processing required for the data analysis of the written essays as "raw data": cleaning of the documents (removing names, adding random code) and eliminating any formatting differences between documents (Standard: Times Roman, 12 pt., double-spaced, 1-inch margins, no title lines). Coders identified and counted the number of distinct arguments (pro and contra a proposal), evaluations of documents (explicit references and trustworthiness assessments), and the use of falsehoods or fabrications. Each PA task has a separate, comprehensive list of viable arguments the participant could possibly have derived from the provided documents. Coding instructions outline when certain codes are given. Writing errors were also marked and counted.

Coder Training

Coder training consists of four 30-minute sessions where sample essays are coded and discussed. Each coder is specialized in one PA task; the team consists of two to three independent coders for each of the three tasks. It is crucial to have high interrater reliability between coders, particularly in essay coding, where arguments are occasionally up for interpretation. Coding for this project is a two-step process. First, coders learn to read and assess participants' responses, how to count arguments, and what defines a trustworthiness statement, falsehoods/fabrications, and writing errors. Secondly, coders learned how to enter the data into a standardized spreadsheet. After this process, the data of all coders were combined and converted into an SPSS file.

Analyses

All statistical analyses were performed using SPSS. For the purpose of the current paper, the total score of critical thinking was calculated according to the arithmetic coding suggested by Ebright-Jones (2024). Since the difference between tasks in absolute numbers on the subscores and total CT score are not relevant to the Research question at hand, the overall CT scores were each standardized within tasks so that scores were comparable across tasks on a z-score metric.

The critical outcomes for our analyses are the intercorrelations of total CT z-scores across tasks and the intercorrelation of raw subscores.

Results

Total Score

Table 1 reports the correlations of the total scores (on z-metric) of the pooled data of study 1 and study 2. Correlations between Cull and Plume, *r*(110) = 0.317, and between Cull and Legacy, *r*(110) = 0.345, indicating that the third BBS task correlates very similarly with the other tasks in the BBS series. The Cull correlates stronger with the Plume and Legacy tasks than those two correlate with each other. However, the differences between the three correlation coefficients are not statistically significant, $\chi^2_{(2)} = 6.13$, $p = 0.421$.

Table 2 looks at results from two studies conducted by the lab team last year, the same design as study 1 and study 2 in this experiment. When analyzing the correlation between Plume and Legacy, $r_{(66)} = 0.262$, $p < .001$, we observe a stronger correlation than the same correlation in this year's study. This demonstrates that all three BBS tasks correlate positively at around the same level. The results suggest consistent patterns of association between the BBS.

The two samples differed with respect to the time between measurements and the fact that study 1 included interventions between measurements, causing a "fan-spread effect" of increasing variance in scores over time which often results in lower correlations. Table 3 shows the correlation matrix for study 2 data only. We intentionally did not differentiate between tasks but simply between the first and second z-scores. This means that we treated the tasks as interchangeable and the correlation as the estimate of the test-retest reliability. As expected, the correlation is higher, $r = .401$, p < .001. Overall, the correlation pattern is consistent with the hypothesis that all three tasks measure CT similarly. It also replicates the finding from last year that the retest reliability on the individual level for BBS tasks is weak.

Subcategory Analysis

The key feature of the BBS approach is the clear link between codes and total scores. In calculating the CT score on a task, coders follow a set of precise instructions of to grade participants' essays. Included in the subcategories of criteria of CT are frequency of relevance and trustworthiness statement, number of arguments, myside bias, and writing errors. Past research on the Plume and Legacy BBS tasks indicated some of these criteria have higher correlations with each other and are strongly associated with overall CT in the BBS tasks.

In order to keep the sample size for the subcategories as large as possible, Tables 4-8 show the results always first for the pooled data set and then separate for the two studies. Note, that even in the pooled analyses, the correlation coefficients between all Plume and Legacy subcategories are exclusively based on study 1 data because none of the subjects in study 2 worked on both tasks since one of the tasks was always the Cull by design.

Table 4 looks at the intercorrelations of three “relevance” counts (relevance is the frequency of explicitly mentioning a source) using the pooled data. All correlations are statistically significant and are similar, which was expected based on hypothesis 2. Table 5 reports the same analysis separates the two samples. The intercorrelations are higher and more consistently above r = .5 for the data in study 2.

Table 6 shows the correlation coefficient for the number of arguments across the three tasks (pooled sample). While they are all statistically significant, they are consistently lower than the correlations for relevance. Similarly, Table 7 reveals lower intercorrelations for the myside bias, the difference between the number of pro and con arguments in favor of a person’s decision, which is even lower and, in one case, not significantly different from zero. This suggests that the myside bias score is a weak indicator of CT since there is little consistency in re-testing despite overall stable retest reliability for the total score.

Table 8 shows the trustworthiness correlations between tasks in Studies 1 and 2. The trustworthiness correlation between the Cull and Plume, r(113) = -0.041, the trustworthiness correlation between the Cull and the Legacy, r(117) = 0.433, and the trustworthiness correlation between the Plume and the Legacy, r(87) = -0.049. The trustworthiness correlations are volatile, which is indicative of a distributional problem for this indicator.

Discussion

The purpose of this study was to establish the measurement equivalence of the newly developed task (Cull) with the established measurement equivalence for the other two BBS tasks

(Plume, Legacy). Results from the pooled and separate analysis of the total score support our hypotheses. The mutual correlation coefficients for the direct test-retest study (sample 2) are very similar and not statistically significantly different. Ebright-Jones (2024) discussed the validity of the Plume and Legacy task. Inasmuch as that is accurate, we can extend this validity claim to the Cull task as there is no indication that the total or the subcategories correlate differently with the corresponding variable of the other two tasks as those tasks' scores correlated with each other.

This study also corroborates Ebright-Jones's (2024) observation that the test-retest reliability is too low to warrant the use of the BBS for individual diagnostics. If the Cull task had better measurement qualities compared to the other two, its CT score would have correlated higher with both other CT scores than the correlation between the CT scores of the two old tasks. It is reasonable to assume that developing additional BBS tasks will produce similar findings, namely that the measurement properties are comparable and hence allowing for repeated measurement studies without memory effects, but not reliable enough to make claims about the critical thinking skill of every single individual. This means, among others, that the BBS tasks should not be used for personnel selection. This also sheds a critical light on all those commercially available CT measures that claim to be useful for this exact purpose. For the BBS tasks, the only possible scenario for a defensible strategy for individual diagnostic would be to give subjects two BBS tasks: If we, somewhat optimistically, estimate the retest reliability for BBS tasks to be $r = .5$, the Spearman-Brown formula for test length (Lord & Novick, 1968) would predict retest reliability of $.5*2/(1+1*.5)=.75$, which would be close to what is considered sufficient for diagnostic purposes ($r > .8$).

While the retest reliability is a weakness of BBS, its strength is the interrater reliability. As was true for the first two tasks, it was easy to establish a coding routine for the Cull task that produced high convergence of the codes. Ebright-Jones (2024) reports an interrater correlation of 0.76 for BBS coding. While not analyzed yet, it is reasonable to expect a similar coefficient for the Cull task, given the comparable correlation pattern in this study. If the interrater reliability were lower, all correlations would have to be lower as well, which is not the case.

Limitations

While the study produced some generalizable findings, it is not without some limitations. One general problem with performance assessment tasks clearly applies here as well: Some participants may be assigned a task related to a topic in which they are interested and already have a knowledge base beyond what is provided in the task document. Most likely, they will end up with a higher CT score because it is easier for them, for example, to identify more arguments that they would have with a BBS task regarding a topic they are less familiar with. The fact that the Legacy task yielded the highest number of arguments on average can be seen as an indicator of this effect because legacy admission to college is a more salient issue to current college students than the problem of controlling the deer population. International students, on the other hand, who are taking the Legacy task, are at a disadvantage since legacy admission is not a concept in most other countries, and they learn about this unfamiliar concept through the task and are likely to be at a disadvantage with respect to their CT performance. We have to assume, however, that those effects are random and hence contribute mainly to the error variance in the measurement.

Another limitation of the current study and former rounds of study with the BBS are the floor effects for some of the indicators, in particular the trustworthiness score: Over half of the coded essays have a trustworthiness score of zero, i.e. not a single statement about the trustworthiness of sources was made. While this arguably reflects the low level of college students' usual reflection of this aspect, it does cause serious problems for the statistical analysis, as Table 8 indicates: The lack of distributional properties causes variance estimates to become volatile, which affects all statistics that rely on them, in particular correlation coefficients where the covariance is divided by both standard deviation of the two corresponding variables. Because the trustworthiness scores show floor effects for all three tasks (resulting in the underestimation of variance), their intercorrelation is particularly affected. The range of correlation coefficient from 0.43 to -0.041 is most likely a statistical artifact. In the future, the research team might consider lowering the threshold for students to produce trustworthiness statements, for example, by making the instructions more explicit.

Future Research

While the research produced a valid and reliable assessment of CT on the aggregate level of college cohorts, the psychometric properties of the developed tasks, like repeated measurement reliability, do not allow for individual diagnostic purposes, which strongly limits its use for personnel assessment in the corporate world where an individual success prognosis is the reason why a test is given. If one is interested in developing a tool for individual diagnostics, it is clear that the variation of topics has to be included in every assessment. If the time frame cannot be changed (50 min), one might be able to design slightly less complex tasks so that two can be worked on within that time frame. However, it would run counter to the idea of critical thinking in the sense of the iPAL framework to resort to multiple-choice items because they can only be used if there is an underlying right/wrong distinction. The latter, however, undermines the BBS structure of dilemma tasks where the decision itself is not an element of the CT measure.

References

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Braun, V., Clarke, V., Boulton, E., Davey, L., & McEvoy, C. (2021). The online survey as a qualitative research tool. International Journal of Social Research Methodology, 24(6), 641-654. https://doi.org/10.1080/13645579.2020.1805550

Braun H. I., Shavelson R.J., Zlatkin-Troitschanskaia O. & Borowiec, K. (2020) Performance assessment of critical thinking: Conceptualization, design, and implementation. *Frontiers in Education, 5:*156. doi: 10.3389/feduc.2020.00156.

Cortina, K. S., & Ebright, B. (2024). Performance Assessment of Critical Thinking without expert rating scale: an arithmetic approach.

Ebright-Jones, B. (2024). Development of Critical Thinking in College – Empirical Studies Using Performance Assessment Tasks. Dissertation, University of Michigan.

Lord, F.M. & Novick, M.R. (1968) Statistical Theories of Mental Test Scores. Addison-Wesley, Menlo Park.

Mayhew, M. J., Rockenbach, A. N., Bowman, N. A., Seifert, T. A., & Wolniak, G. C. (2016). *How college affects students: 21st century evidence that higher education works* (Vol. 1). John Wiley & Sons.

Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How robust are the findings of academically adrift? *Change: The Magazine of Higher Learning*, *43*(3), 20-24.

Pascarella, E. T., and Patrick T. T. (2003). *How College Affects Students: A Third Decade of Research. Volume 2*. Jossey-Bass, An Imprint of Wiley. 10475 Crosspoint Blvd, Indianapolis, IN 46256.

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.

Webb, N. M., Shavelson, R. J., & Haertel, E.H. (2006). Reliability and Generalization Theory, *Handbook of Statistics, 26*, 4-44.

Zlatkin‐Troitschanskaia, O., Shavelson, R. J., Schmidt, S., & Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *British Journal of Educational Psychology*, *89*(3), 468-484.

Tables

Table 1

Total score intercorrelations of the three BBS tasks (pooled samples)

		PCTz Zscore (PCTraw)	LCTz Zscore (LCTraw)	CCTz Zscore (CCTraw)
PCTz Zscore (PCTraw)	Pearson Correlation	1	.166	.317
	Sig. (2-tailed)		.130	<.001
	N	147	84	110
LCTz Zscore (LCTraw)	Pearson Correlation	.166	1	.345
	Sig. (2-tailed)	.130		<.001
	N	84	152	110
CCTz Zscore (CCTraw)	Pearson Correlation	.317	.345	1
	Sig. (2-tailed)	<.001	<.001	
	N	110	110	178

Table 2

Total score intercorrelations Plume and Legacy 2022

		PCTzscore (PCTraw)	LCTzscore (LCTraw)
PCTzscore (PCTraw)	Pearson Correlation	1	.265
	Sig. (2-tailed)		.031
	N	77	66
LCTzscore (LCTraw)	Pearson Correlation	.265	1
	Sig. (2-tailed)	.031	
	N	66	78

Table 3

Test-Retest Reliability study 2 (immediate retest)

		zwave1	zwave2
zwave1	Pearson Correlation	1	.401
	Sig. (2-tailed)		<.001
	N	74	67
zwave2	Pearson Correlation	.401	1
	Sig. (2-tailed)	<.001	
	N	67	77

Table 4

Intercorrelation Subcategory "Relevance" (pooled sample)

		Prelevance	Lrelevance	Crelevance
Prelevance	Pearson Correlation	1	.338	.269
	Sig. (2-tailed)		.001	.004
	N	148	87	113
Lrelevance	Pearson Correlation	.338	1	.369
	Sig. (2-tailed)	.001		<.001
	N	87	157	117
Crelevance	Pearson Correlation	.269	.369	1
	Sig. (2-tailed)	.004	<.001	
	N	113	117	182

Table 5

Intercorrelation Subcategory "Relevance" by sample

			Prelevance	Lrelevance	Crelevance
Study 1 (356F23)	Prelevance	Pearson Correlation	1	.338	.067
		Sig. (2-tailed)		.001	.559
		N	113	87	78
	Lrelevance	Pearson Correlation	.338	1	.268
		Sig. (2-tailed)	.001		.020
		N	87	114	75
	Crelevance	Pearson Correlation	.067	.268	1
		Sig. (2-tailed)	.559	.020	
		N	78	75	97
Study 2 (CAL234)	Prelevance	Pearson Correlation	1	a	.573
		Sig. (2-tailed)			<.001
		N	35	0	35
	Lrelevance	Pearson Correlation	a	1	.504
		Sig. (2-tailed)			<.001
		N	0	43	42

Crelevance	Pearson Correlation	.573	.504	1
	Sig. (2-tailed)	<.001	<.001	
	N	35	42	85

Note. a. Cannot be computed because at least one of the variables is constant.

Table 6

Intercorrelation Subcategory "# of arguments" (pooled sample)

		Parguments	Larguments	Carguments
Parguments	Pearson Correlation	1	.358	.145
	Sig. (2-tailed)		<.001	.126
	N	148	86	113
Larguments	Pearson Correlation	.358	1	.275
	Sig. (2-tailed)	<.001		.003
	N	86	156	116
Carguments	Pearson Correlation	.145	.275	1
	Sig. (2-tailed)	.126	.003	
	N	113	116	182

Table 7

Intercorrelation Subcategory "Myside bias" (pooled sample)

		Pmyside	Lmyside	Cmyside
Pmyside	Pearson Correlation	1	-.064	.197
	Sig. (2-tailed)		.564	.039
	N	147	84	111
Lmyside	Pearson Correlation	-.064	1	.224
	Sig. (2-tailed)	.564		.018
	N	84	152	110
Cmyside	Pearson Correlation	.197	.224	1
	Sig. (2-tailed)	.039	.018	
	N	111	110	179

Table 8

Intercorrelation Subcategory "Trustworthiness" (pooled sample)

		Ptrustworthiness	Ltrustworthiness	Ctrustworthiness
Ptrustworthiness	Pearson Correlation	1	-.049	-.041
	Sig. (2-tailed)		.650	.668
	N	148	87	113
Ltrustworthiness	Pearson Correlation	-.049	1	.433
	Sig. (2-tailed)	.650		<.001
	N	87	157	117
Ctrustworthiness	Pearson Correlation	-.041	.433	1
	Sig. (2-tailed)	.668	<.001	
	N	113	117	182