College Students' Critical Thinking: Assessment and Interpretation

Ella Anghel
Boston College

Henry I. Braun
Boston College

Audrey A. Friedman
Boston College

Maria Baez-Cruz
Boston College

Many colleges identify the development of critical thinking (CT) as a key learning outcome. Nonetheless, few studies examined the development of CT during college, and the instruments employed in them are often limited. This article introduces the Critical Reasoning Assessment (CRA), a new instrument based on the Reflective Judgment Model (RJM; King and Kitchener 1994) designed to engage students in analyzing ethical dilemmas while being easy to administer and score. Using the CRA, we measured the CT skills of college students in three studies, both cross-sectionally and longitudinally. The results demonstrated substantial growth in CT skills during the first year and between the first and the fourth years of college; 42% and 60% of the participants advanced to a higher level of CT by the end of their first and fourth year, respectively. This study introduces a comprehensive, theory-based, easy-to-score and interpret instrument measuring CT. Applied to longitudinal data, it adds to limited findings on CT developmental trajectories and quantifies substantively interpretable shifts in the quality of CT.

Keywords: critical thinking, critical thinking assessment, college students development, reflective judgement model, growth modeling

INTRODUCTION

Critical thinking (*CT*) is an important 21st century skill (e.g., Pellegrino and Hilton, 2012) and an important learning objective of higher education (Association of American Colleges and Universities, 2018). Measuring and tracking student growth in CT is challenging, as most existing instruments are limited in scope or have problematic measurement properties. The current study employs the Critical Reasoning Assessment (*CRA*), an open-response instrument to assess college students' CT skills. The CRA improves on existing instruments as it is grounded in a well-established theoretical framework: the Reflective Judgment Model (King & Kitchener, 1994). In contrast to most instruments, it measures CT skills in the

context of moral or ethical dilemmas, instead of more narrow, academic contexts (e.g., Lord et al., 2017; Watson & Glaser, 1980). This paper describes the process of developing and implementing the CRA and presents promising preliminary results. Results show evidence of students' growth in CT throughout the first year of college and between the first and fourth years. This paper not only contributes to the field by introducing this new instrument but also by presenting longitudinal results, providing more direct and rigorous evidence of CT development (Shaw et al., 2020).

We first discuss the literature related to CT development during college and to the Reflective Judgment Model. Second, we describe the development of the CRA, a pilot study, and subsequent instrument revisions. Then, we present three studies using the revised CRA. In Study 1, we administered the CRA to college students from different majors and compared their scores. In Study 2, we used the CRA to measure the growth in CT skills among first-year college students majoring in education and human development (EHD). Finally, Study 3 tracks the growth of students majoring in EHD over four years in college. The final section presents conclusions and suggestions for future research. See Appendix A for a summary of each study.

Literature Review and Theoretical Framework

Despite the importance of CT skills to all citizens in the 21st century, different scholarly traditions disagree on what they entail (Evans, 2020); consequently, several assessment frameworks are linked to the construct. In reviewing those frameworks, Liu et al. (2014) suggest that many include skills such as evaluation of evidence, presentation and analysis of arguments, and making inferences based on the evidence. But CT entails more than logical analysis. In a newer framework, Oser and Biedermann (2019) distinguish between critical analysis involving content knowledge and analytical skills, on the one hand, and critical reflection requiring questioning information and its validity, on the other. They argue that critical analysis is relevant to daily life, and therefore should be assessed using common, real-life ethical dilemmas. They refer to Brookfield's definition (1987):

> Being a critical thinker involves more than cognitive activities such as logical reasoning or scrutinizing arguments for assertions unsupported by empirical evidence. Thinking critically involves recognizing the assumptions underlying beliefs and behaviors...and giv[ing] justifications for ideas and actions. Most important, perhaps, it means we try to judge the rationality of these justifications... by comparing them to a range of varying interpretations and perspectives. We can think through, project, and anticipate the consequences of those actions that rest on these justifications. And we can test the accuracy and rationality of these justifications against some kind of objective analysis of the 'real' world as we understand it. (p. x)

If the goal is to prepare citizens for a global context, it is evident that CT is an important outcome of any educational system. The capacity to conduct systematic analyses of moral-cognitive dilemmas from multiple perspectives across many contexts is essential to developing an informed and engaged citizenry. In fact, college-educated adults are expected to think logically, reason through complex problems, and communicate conclusions effectively (Paul & Elder, 2012).

Relatively little is known about how CT skills develop, particularly during college years. Some findings suggest that students who received greater exposure to clear and organized classroom instruction and deep learning experiences (both associated with liberal education pedagogy) achieved greater gains in CT (Pascarella et al., 2013). Others argue that for many students, a college education contributes little to CT skills (Arum & Roksa, 2011; Caplan, 2019).

Findings that relate background characteristics (e.g., gender, race) and academic factors (e.g., educational achievement and major) to CT and its growth trajectories are also scarce. For example, researchers disagree on the relationship between CT and academic achievement (Halpern, 2010; Harris et al., 2014; Watson & Glaser, 1980) as measured by GPA or SAT scores. Concerning gender, some authors found no differences between males and females in CT scores (Nora et al., 1996; Roska et al., 2017), while

others observed higher scores in CT-related traits in females (Walsh & Hardy, 1999). Findings regarding race and CT are more consistent; not only do African-American students score lower than white students on CT measures when they begin college (Gadzella et al., 1999) but also their skills appear to improve more slowly in comparison (Flowers & Pascarella, 2003). Examining Hispanic students, Kugelmass and Ready (2011) found that although their initial CT skills are lower than their white peers, growth rates are comparable.

Conflicting findings illustrate the need for a valid measure of CT that not only measures students' CT skills, but also these skills' development. Such measures should also explore CT in authentic, everyday contexts making them more relevant to today's increasingly complex, civic climate (Paul & Elder, 2012; Kwak, 2007). Several authors have created such instruments (e.g., Braun et al., 2020; Oser & Biedermann, 2019; Sotiriadou et al., 2020), but many rely on complex performance assessments that are difficult to administer or to score. Other instruments (e.g., Ennis et al., 1985; Facione, 1990) use only multiple-choice or Likert-type response formats, that are limited in their ability to evaluate higher-order thinking skills (Douglas, 2006). Finally, many well-established instruments yield hard-to-interpret scores because they are not explicitly grounded in a CT framework, making score meaning vague or ambiguous (Liu et al., 2014). Thus, the need for an authentic, easy-to-score, theory-base instrument is clear.

The Reflective Judgment Model

King and Kitchener's (1994) Reflective Judgment Model (RJM) is a developmental model of CT easily adapted for assessment in everyday contexts and commonly used in the context of college education (e.g., Franco et al., 2017). The RJM describes the development of complex reasoning and the capacity to justify solutions to ambiguous problems (King & Kitchener, 2004), elements that are integral to CT. The RJM models a range of people's epistemological assumptions, including how they think about knowledge and the role of evidence in decision-making, constituting an ideal framework to track college students' CT development.

The RJM posits three ordered stages of thinking: pre-reflective, quasi-reflective, and reflective. The pre-reflective stage is characterized by relying on direct sources of information, rather than on a careful examination of evidence, and believing that all problems are well-defined and solvable. At the quasi-reflective stage, some uncertainty and limited evaluation of evidence are used to reach an unambiguous conclusion. Reflective reasoning entails both acceptance of uncertainty and clear decision-making; although a conclusion is open to new evidence as reflective thinkers make judgments that are "reasonably certain" based on evidence (King & Kitchener, 2004).

The RJM is a well-established model used in research and practice, and validated in several longitudinal studies, particularly among college students (see a review in King and Kitchener, 2004). Mines et al. (1990) found that students at higher stages of reflective judgment demonstrate better CT skills than do those who use assumptions consistent with lower stages, reflecting true differences in CT. They also argue that acquiring CT skills is developmental, a claim supported by findings showing that specific interventions can raise levels on the RJM (Chen et al., 2020; Lord et al., 2017).

An outgrowth of the RJM framework is the Reflective Judgment Interview (*RJI*), designed to assess people's reasoning about ill-defined dilemmas. This semi-structured interview allows researchers to examine participants' CT in an ecologically valid manner. The instrument asks respondents to reason about a controversial issue such as the use of nuclear energy or creationism vs. evolution, thus allowing for the measurement of CT in the context of a dilemma or a moral issue.

Kitchener and King's (1985) original RJI was scored using five dimensions. They defined three general dimensions: Cognitive Complexity, Reasoning Style, and Openness, and two specific dimensions: Nature of Knowledge and Nature of Justification. Cognitive Complexity addresses the degree to which one acknowledges multiple alternatives or points of view when analyzing a dilemma. Reasoning Style examines the ability to use evidence to move systematically from hypothesis to conclusion that manifests a justifiable solution or worldview. Openness addresses the degree to which the individual acknowledges and understands other points of view and their rationales. Such reasoning is also open to new evidence for evaluation. The Nature of Knowledge dimension addresses the degree to which the individual uses

systematic inquiry to develop a view of knowledge, assigns a moral or logical value to a point of view, and understands how different points of view are plausible across different interpretive frameworks. Finally, the Nature of Justification dimension considers how the individual uses concepts, evidence, and experts' opinions to justify a worldview or point of view.

Originally, the five dimensions jointly determined a person's stage of reflective judgment on a scale of 1-7. King and Kitchener labeled scores of 1-3 scores pre-reflective; 4-5 quasi-reflective; and 6-7 reflective. Using this approach, several studies established the RJI's validity, thus supporting the validity of the entire model (see Wood, 1997 for a review).

Similar to other models of CT, the RJM addresses the analysis of evidence to draw conclusions (Liu et al., 2014). Given its developmental nature, the RJM seems particularly suitable for constructing an instrument measuring CT in a moral/ethical context. However, as an hour-long interview, the RJI is complex and time-consuming to administer. Wood et al. (2002) introduced a Likert-style version of the RJI, the Reasoning about Current Issues Test (*RCI*). Though easier to score, the RCI is harder to interpret and not fully representative of CT because it does not directly ask respondents to reason critically. Essay-based assessments also exist but focus on a specific professional context rather than more generic CT (Van Tyne & McNair, 2019; Wolcott & Lynch, 1997). This article presents the Critical Reasoning Assessment (*CRA*), a constructed response measure based on the RJM that captures respondents' ability to analyze and justify their opinions in the context of ambiguous, moral issues.

Instrument Development

The structure of the CRA follows that of the Reflective Judgment Interview (*RJI*; Kitchener & King, 1985). It employs a set of dilemmas whose formats are similar to those employed in the original RJI (King & Kitchener, 2004). For the CRA, the original RJI dilemmas were modified to address issues that students might encounter in their personal lives. The following is an example of one of the CRA dilemmas (the others are contained in Appendix B):

> 'Some people believe that success is a function of an individual's ability and determination. Regardless of socioeconomic status (class or privilege), race, and/or environmental factors, everyone can succeed if he/she works hard enough and takes advantage of each and every available opportunity. Others believe that such factors as socioeconomic status (class or privilege), race, and/or environmental factors impact an individual's ability to succeed regardless of determination and effort, as each of these factors impacts opportunities that are available to the individual.'

Respondents are then asked to answer a set of questions related to the dilemma (see Appendix B). These questions were modeled on those posed in the RJI, and prompt respondents to consider the reasoning behind their position and arguments that may support other points of view.

Instead of generating a holistic score as does the RJI, we chose an analytic scoring rubric; that is, each of the five dimensions is scored separately (Complexity, Reasoning, Openness, Nature of Knowledge, and Nature of Justification). Consequently, each student received a score profile that allowed finer distinctions among students. We maintained King and Kitchener's 1-7 evaluation scale to rate each of the five dimensions.

We conducted a pilot study to test the CRA. At the first stage of the pilot study (see below), two trained raters scored students' responses, and discrepant ratings were negotiated to achieve consensus. Inter-rater agreement approached 90% prior to consensus, with scores differing by no more than one point. Once a score for each of the five dimensions was agreed upon, the average of the first three dimensions yielded a general dimension score, the average of the other two dimensions yielded the specific dimension score, and their unweighted average yielded a total score for the response. Based on Kitchener and King (1985), we considered scores of 1-3 to correspond to the pre-reflective stage, scores of 4-5 to the quasi-reflective stage, and scores of 6-7 to the reflective stage.

The results of the first phase of the pilot study raised some issues regarding the placement of students in categories. For example, for Cognitive Complexity, a student wrote 'I think that there is truth in the first statement and you cannot have success without ability and determination. However, I also believe that other factors such as race and socioeconomic status greatly affect opportunities that can lead to success.' This response suggests that, on the one hand, the student does not believe the issue is so complex as not to allow for certainty but, on the other, is unable to make a choice. This position does not fit perfectly well with a score of 2 ("knowledge is right or wrong") or 3 ("knowledge is true, false, or uncertain").

Accordingly, we augmented the 7-point scale by introducing half-point scores; the above response was scored as a 2.5 – "knowledge is right or wrong but uncertainty prevents choosing right or wrong." Appendix C contains the revised scoring rubric, with the integer scores derived from Kitchener and King (1985), and the half points based on an analysis of students' responses. This version of the rubric was employed in the spring of the same year in the second phase of the pilot study. For this administration, we found inter-rater agreement approached 96% prior to consensus, with scores differing by no more than 0.5 points.

PILOT STUDY

Procedure

The pilot study consisted of two phases. In phase one data were collected during the fall semester from students participating in a first-year class designed to introduce students to college and, among other goals, foster CT. Participants completed one version of the CRA (see below) as a class assignment and emailed their responses to the researchers. Students could opt out of the study without penalty. The only constraint respondents had was that they rely only on their own experience rather than outside sources (e.g., online searches). One author and a trained graduate student familiar with the RJM scored the responses. Data were collected and matched to demographic data provided by the university's registrar's records, in compliance with the university's IRB.

After this first administration, we expanded the scoring rubric by adding half-point scores (as noted above). To test this new rubric, additional data were collected in the following spring semester from a subset of the original participants, using a similar procedure and the same prompt.

Participants

In the first phase, data were gathered from 85 students majoring in EHD at a private research university in the U.S. Northeast. Students were matched with demographic data provided by the university. Participants also self-identified gender and race. Where administrative data and self-reported data disagreed or were missing, we consulted one of the authors who was personally familiar with all of the students. Based on those data, 85% of the participants were female and 69% were white. Their mean SAT score (either taken by the student or converted from their ACT score) was 1340 (sd = 110).

In the second administration, we gathered data from a smaller group of students (n = 16). The purpose of this administration was to test the utility of the expanded scoring rubric. There were no differences between these participants and those who participated in only the fall with respect to their overall fall CRA score ($t(24.94) = -0.81$, $d = 0.21$). Note that throughout this paper, we are providing the t statistic as a way to support the results' interpretation but not as means for population-level inferences, as the samples were not randomly selected and, therefore, such inferences are inappropriate.

Results and Discussion

Descriptive statistics of the five CRA dimensions in the fall administration and by group are presented in Table 1. Based on the total scores, 62% of participants were in the pre-reflective stage, 35% were in the quasi-reflective stage, and 2% were in the reflective stage. The findings also indicate that males and non-whites scored higher than females and white students, respectively. The differences were particularly large in Cognitive Complexity (d = 0.16 for gender; d = 0.43 for race), Reasoning (d = 0.42 for gender; d = 0.42 for race), and Nature of Justification (d = 0.53 for gender; d = 0.34 for race). However, all five dimension scores correlated very weakly with the students' SAT scores (all $r < |.07|$).

TABLE 1
MEANS (STANDARD DEVIATIONS) OF THE CRA IN THE FIRST PILOT SAMPLE

	Total sample (n = 85)	Males (n = 13)	Females (n = 72)	White (n = 59)	Non-white (n = 26)
Complexity	2.91 (1.19)	3.08 (1.26)	2.88 (1.18)	2.75 (1.10)	3.27 (1.31)
Reasoning	2.73 (1.04)	3.08 (0.86)	2.67 (1.06)	2.60 (1.07)	3.02 (0.93)
Openness	3.14 (1.25)	3.23 (1.09)	3.12 (1.28)	3.06 (1.24)	3.33 (1.26)
Nature of Knowledge	3.20 (1.33)	3.23 (1.54)	3.19 (1.30)	3.15 (1.34)	3.31 (1.35)
Nature of Justification	2.58 (1.28)	3.19 (1.47)	2.47 (1.22)	2.45 (1.24)	2.87 (1.26)

We also examined the instrument's internal structure. Treating each dimension as an item, the Cronbach α reliability of the CRA was .94, with high correlations among the scores (lowest r = .64 between Reasoning Style and Nature of Knowledge). An exploratory factor analysis resulted in a single factor solution accounting for 75% of the variance in the data, and item loadings of .82 or higher, all suggesting a strong single-factor solution rather than a two-factor solution. Therefore, for brevity, we report only the total CRA scores in the following analyses.

The mean in the spring administration was 3.34 (sd = 0.75). The same group's mean in the fall was 3.11 (sd = 0.52), corresponding to a modest increase between administrations ($t(13) = 1.16$, d = 0.31). Small sample sizes prevented comparing scores by demographic groups. Since the new rubric seemed to have functioned as expected, we continued using it in the three studies and examined individual differences in CRA scores and growth trajectories.

RESULTS

Study 1

Procedure

The focus of this study was to examine CRA performance among students in different majors. Data were collected from students in three different courses: one targeted EHD students and the other two enrolled students in different majors. Although all three courses were open to all school years, they primarily enrolled first- and second-year students. In all three courses, data were collected at a single administration during the fall semester. Instructors invited students to participate in the study and received a link to an online version of the CRA (effort vs. privilege). Participants also reported gender, age, race, and major.

Participants

Participants were 405 students from different intended majors at the same university. Of the participants 68% were female (One person did not report gender.), 65% were white, and their mean age was 18.5 (sd = 0.93; Five did not report age.). The students in the sample represent all of the university's schools (see Table 2).

TABLE 2
CRA MEAN SCORES BY SCHOOL IN STUDY 1

	CRA (sd)	n	%
Arts and Sciences	2.51 (1.01)	165	41
Management and Finance	2.18 (0.74)	60	15
Education and Human Development	3.27 (0.78)	103	25
Nursing	2.50 (0.92)	46	11
Undeclared	3.17 (0.87)	26	6

Note: five students did not report their major

Results and Discussion

The Cronbach α (calculated as above) for this sample was .86. Table 3 presents the mean CRA scores of the full sample, as well as by gender and race. Table 2 presents mean CRA scores by school. The mean score for the full sample corresponds to pre-reflective thinking; 63% of the sample were in the pre-reflective stage, 36% were in the quasi reflective stage, and 1% were in the reflective stage. There was no correlation between age and CRA score (r = -.03).

TABLE 3
CRA MEAN SCORES BY GENDER AND RACE IN STUDY 1

	CRA (*sd*)	*n*
Male	2.48 (0.97)	129
Female	2.81 (0.97)	275
White	2.74 (0.99)	265
Non-white	2.63 (0.97)	140
Overall	2.70 (0.98)	405

We conducted a regression analysis predicting CRA scores using the program of study and the demographic variables. Gender and race were treated as binary variables with male and non-white as the reference groups. Age was treated as a continuous variable. The different majors were each coded as dummy variables, with 'undeclared' as the reference group. Major was the only significant predictor of CRA score (adjusted R^2 = .16). Students majoring in EHD and with undeclared majors had higher CRA scores than students in other majors (see Table 4). Adding interaction terms among all variables did not improve the model's fit (adjusted R^2 = .14).

TABLE 4
ESTIMATES OF A REGRESSION MODEL PREDICTING CRA SCORES FROM DEMOGRAPHIC CHARACTERISTICS (STUDY 1)

	Estimate (*SE*)	Standardized estimate	*t*-value
Intercept	4.79 (0.95)	-	5.04
Female	0.02 (0.11)	0.01	0.17
White	0.03 (0.10)	0.01	0.32
Age	-0.09 (0.05)	-0.08	-1.78
Arts and Sciences	-0.67 (0.19)	-0.34	-3.51
Management and Finance	-1.02 (0.22)	-0.37	-4.74
Education and Human Development	0.09 (0.20)	0.04	0.45
Nursing	-0.73 (0.22)	-0.24	-3.22

Note: the comparison group was undeclared major. *SE* = standard error

These results suggest that for this convenience sample, demographic variables were not associated with CRA scores. By contrast, the program of study was associated with CRA scores, with students majoring in EHD and students who did not declare a major having higher mean CRA scores than students in other majors. At this juncture, we were interested in the CRA's ability to track students' CT development. In study 2, we assessed the CT skills of students longitudinally through their first year.

Study 2

Procedure

Data were collected from first-year students taking the same required course employed for the pilot study but in a different year. We collected CRA data from two cohorts, labeled A and B, in three waves through their first year: at the beginning of the fall semester (September), at the end of the fall semester

(December), and at the beginning of May. The first two administrations used a different prompt ('Effort vs. Privilege' in September and 'Fairness' in December; see Appendix B). The Effort vs. Privilege scenario was also used as an instructional tool during the fall semester; the dilemma was discussed without referring explicitly to the instrument's dimensions and scoring criteria. The May administration of the CRA again used the Effort vs. Privilege prompt. Participants' race and gender data were collected and the instructor addressed missing data or inconsistencies.

Participants

Participants were 235 students from the two first-year cohorts (120 in cohort A and 115 in cohort B): 221 students participated in the first, 216 in the second, and 210 in the third administrations. Overall, 199 students participated in all three administrations; 88% of the participants were female, and 71% were white. There were no differences between the cohorts in percentages of females or white students ($\chi^2(1) = 0.79$; $\chi^2(1) = 0.12$, respectively). Cohort A had higher CRA scores at the beginning of the year than cohort B ($t(170.81) = 3.72$, $d = 0.50$), but no differences were detected by the end of the year ($t(198.64) = 0.16$, $d = 0.02$). This is perhaps due to some attrition in class A (11% fewer students between T_1 and T_3), which was not observed in class B. We decided to pool the results over cohorts.

Results and Discussion

Descriptive Statistics. Table 5 presents the CRA scores in all three administrations, disaggregated by gender and race. The differences by race were small at all time points (all $d < 0.05$). Although females had higher scores at the start of the year ($d = -0.16$), males had higher scores by the end of the year ($d = 0.33$). Table 6 displays a transition matrix of CT stages from T_1 to T_3. The table shows that about 42% of participants transitioned to a higher stage, while 53% remained in the same stage. Of the latter, the mean score change was 0.56 ($sd = 0.63$), with 82% showing some improvement without transitioning to a more advanced stage. Looking at the transition matrices disaggregated by gender reveals a similar pattern to that for mean CRA scores: 64% of males and only 39% of females transitioned to a more advanced stage (see Appendix D).

TABLE 5
MEANS (STANDARD DEVIATIONS) OF THE CRA IN STUDY 2

	T_1		T_2		T_3	
	CRA (*sd*)	*n*	CRA (*sd*)	*n*	CRA (*sd*)	*n*
Male	3.35 (1.16)	27	3.86 (1.45)	24	4.83 (1.32)	23
Female	3.52 (0.91)	194	3.87 (1.16)	192	4.45 (0.95)	187
White	3.49 (0.95)	158	3.88 (1.23)	156	4.50 (0.94)	151
Non-white	3.53 (0.93)	63	3.83 (1.09)	60	4.48 (1.15)	59
Overall	3.50 (0.95)	221	3.87 (1.19)	216	4.49 (1.00)	210

TABLE 6
TRANSITION MATRIX FOR STUDY 2 (*N* = 200)

		T_3		
		Pre-reflective	Quasi-reflective	Reflective
T_1	Pre-reflective	10	35	7
	Quasi-reflective	7	83	41
	Reflective	-	5	12

Note: 200 students participated in both the first and the last administrations.

The correlations between administrations were r_{12} = .69, r_{13} = .40, and r_{23} = .51, providing some evidence for the equivalence of the different forms used at T_1 and T_2. Finally, The Cronbach α reliabilities of the five CRA dimension scores were .94, .95, and .94 in T_1, T_2, and T_3, respectively, again strongly supporting a unidimensional structure for the instrument.

Growth Model. We used a longitudinal, multilevel model to track students' growth through the year, with time measured in months of elapsed time in the course as the basic predictor. The time variable's values are 0, 3, and 7 months, corresponding to the September (T_1), December (T_2), and May (T_3) administrations. Note that we did not include the one month of winter break in this count. Multilevel models were estimated using the lme4 R package (Bates et al., 2015). The initial model included only time as a predictor of CRA scores of person *i* at time *t*. We then added gender and race as predictors to examine their associations with initial scores (intercept) and growth (slope). After exploring their effects, the final model was:

$$\begin{aligned} &\text{Level 1: } CRA_{ti} = \pi_{0i} + \pi_{1i}*(TIME_{ti}) + e_{ti} \\ &\text{Level 2: } \pi_{0i} = \beta_{00} + \beta_{01} * \text{gender} + r_{0i} \\ &\qquad\quad \pi_{1i} = \beta_{10} + \beta_{11} * \text{gender} \end{aligned} \tag{1}$$

On average, male and female participants started the course with a CRA score of 3.26 and 3.48, respectively, both corresponding to a high pre-reflective stage. On average, males gained 0.20 points each month (1.40 points over the year), and females gained 0.13 points every month (0.94 points overall). That is, on average both genders reached a quasi-reflective stage by the end of the year. In conclusion, this study demonstrated modest growth in CRA scores through the course of the first year in college. However, we were also interested in tracking students' growth throughout their entire college careers. Study 3 presents the results obtained by following students longitudinally, from their first to their fourth year of college.

Study 3

Procedure

The students in the cohort who participated in the pilot study were asked to participate again in the spring semester of their 4th year. They were contacted through email and were given the same CRA dilemma they completed as freshmen (Effort vs. Privilege). We collected race, gender, and SAT scores from participants and the university's registrar's records. The instructor filled in missing data, resolving any data disagreements between administrations.

Participants

Participants included 165 students in EHD for whom we were able to collect first-year data, fourth-year data, or both. Of the 85 students described in the pilot study, 62 also participated in their fourth year. The attrition was due to transferring out of the school of education or refusing to participate. The 80 students who participated only in their fourth year were mostly unavailable at the time of the initial data collection as they had transferred into the school of EHD at a later time. In the full sample, 85% of the participants were female, and 71% were white. The mean SAT score for all 165 students was 1317 (*sd* = 137). There were no differences between those who participated in both data collections and those who participated once in gender ($\chi^2(1) = 0.37$) and race ($\chi^2(1) = 0.48$). There were only modest differences in mean SAT scores between those who participated in both data collections and those who participated once ($t(153.72) = 1.46$, $d = 0.23$), such that those who participated in both administrations had somewhat higher SAT scores.

Results and Discussion

Descriptive Statistics. Table 7 presents the CRA scores at the end of the fourth year disaggregated by gender and race (results from the first year are presented in Table 1). Differences by gender and race were not significant though males seemed to have a substantially higher score ($t(29.00) = 1.83$, $d = 0.42$). The correlation between the change in CRA scores between administrations and SAT scores was $r = -.01$. Table 8 displays the transition matrix of CT stages from T_1 (first year) to T_2 (fourth year). The table shows that about 60% of participants transitioned to a higher stage by the end of the fourth year, while 34% remained

in the same stage. Of the latter, the mean change was about zero (-0.13; *sd* = 0.63). The overall scores indicate that, on average, students progressed to the quasi-reflective stage.

TABLE 7
MEANS (STANDARD DEVIATIONS) OF THE CRA AT THE END OF THE FOURTH YEAR (STUDY 3)

	CRA (*sd*)	*n*
Male	4.40 (0.93)	21
Female	3.99 (1.02)	121
White	4.10 (0.97)	103
Non-white	3.92 (1.14)	39
Overall	4.05 (1.02)	142

TABLE 8
TRANSITION MATRIX FOR STUDY 3 (*N* = 62)

	T_3			
		Pre-reflective	Quasi-reflective	Reflective
T_1	Pre-reflective	7	25	6
	Quasi-reflective	3	14	6
	Reflective	-	1	-

Growth Model. We used a longitudinal, multilevel model to track students' growth. Since there were only two points of data collection, the time variable values were 0 for the first year and 1 for the fourth year. We estimated a multilevel model with time, gender, race, and standardized SAT scores as the predictors, but the model did not converge, suggesting that the random effects were zero. Consequently, we used a linear regression model to predict the CRA scores, with time, gender, race, and standardized SAT scores as predictors. The results are presented in Table 9 (adjusted R^2 = .22). Only time was a predictor of CRA scores, with an average increase of 1.16 points between the administrations. Adding the interaction terms between time and the other variables did not improve the model's fit (adjusted R^2 = .22), but changed the model's interpretation; time was still an important predictor, but the time and race interaction was also meaningful, meaning that the growth in CT skills was more substantial among white students.

TABLE 9
ESTIMATES OF A REGRESSION MODEL PREDICTING CRA SCORES FROM TIME AND DEMOGRAPHIC CHARACTERISTICS (STUDY 3)

	Model 1		Model 2	
	Standardized estimate	*t*-value	Standardized estimate	*t*-value
Intercept	-	14.11	-	9.96
Time	0.47	8.05	0.33	1.83
Female	-0.11	-1.89	-0.10	-1.08
White	-0.01	-0.19	-0.14	-1.53
Standardized SAT	0.06	0.95	-0.06	-0.54
Time * Female			-0.02	-0.15
Time * White			0.23	1.73
Time * Standardized SAT			0.12	1.08

DISCUSSION

Critical thinking is a crucial component in successfully navigating the complexities of modern life (Dwyer et al., 2014; Kirsch et al., 2016). Universities need to evaluate success in enhancing students' skills such as CT (Shavelson, 2010). This calls for a high-quality, theory-based instrument that is easy to administer and score, but also allows students to critically explore in-depth, real-life ethical issues. The purpose of this study was two-fold: (i) to introduce the Critical Reasoning Assessment (*CRA*), a short, open-response instrument based on King and Kitchener's (2004) Reflective Judgement Model and (ii) to present preliminary empirical results.

The CRA enhances existing measures in several important ways. First, the CRA is relatively quick and easy to score but still allows students to demonstrate complex thinking abilities. This makes the CRA ideal for low-stakes academic contexts such as end-of-course evaluations. Second, the dilemmas presented in the CRA are more grounded in issues that are relevant to students' lives than are the RJI's dilemmas. Thus, it is expected that students will be more engaged and more likely to demonstrate their CT skills. Finally, the CRA's rubric allows for a more nuanced scoring that emerges directly from students' responses.

As our questions and procedure are similar to the one presented in the well-established RJI, we believe that to a large extent, the evidence for the RJI's validity applies to the CRA, as well. This study's findings also provide evidence of the structural and known-group validity of the CRA itself; first-year students were mostly in the pre-reflective stage, and more advanced students had higher CRA scores (King, 2000).

Looking at the developmental trajectories in CRA scores, we found substantial growth in the first year, with most students' scores increasing and transitioning, on average, from the pre-reflective to the quasi-reflective stage. A similar rate of growth was identified between the first and the fourth year. This suggests, perhaps, that most of the growth in CT during college occurs in the first year, though such a conclusion may be unwarranted because different participants were tracked in both studies. Future studies should examine this issue more closely.

Some of the group differences we found were surprising. CT scores were not correlated with SAT scores or age, even though such differences were expected based on the RJI and other existing findings (e.g., Halpern, 2010). It is possible that since data were collected from relatively young and high achieving students, the restriction of range did not allow us to identify differences in CT skills by SAT scores and age. Future studies could address this issue by administering the CRA to a more diverse group of participants. Additionally, inconsistencies in our results in terms of gender and race differences may not be directly related to demographic factors but to initial scores; overall, groups with lower initial scores grew faster.

Finally, we found some substantial differences in CRA scores by major. In particular, students with undeclared majors and students majoring in EHD had higher CRA scores than other students. It is unlikely that these differences reflect differences in training, as the students were assessed in their first semester. They may reflect self-selection; differences in CT skills may be associated with decisions regarding preferred majors. Alternatively, the differences might reflect different levels of engagement with the assessment. Future studies could gather data on students' perceptions of the CRA prompts.

Notwithstanding its strengths, this study is limited in several ways. First, the study's sample is constrained in that most of the participants were first-year students and/or students majoring in EHD at a single institution. Therefore, developmental trends identified here are not generalizable. Related, this study only provides circumstantial evidence as to the role of the college experience in the development of CT, so it is not possible to determine to what extent the growth observed was due to the college experience rather than natural maturation.

In conclusion, this study presents a new tool for higher education administrators and instructors to track their students' CT skills. It also introduces some results regarding students' growth in CT; students improve, but not as much as one might expect. These findings can therefore serve as a starting point for enhancing curricula or designing interventions to support students' CT growth. Nurturing the development of students who reason critically about ill-defined, moral-cognitive dilemmas from multiple perspectives across multiple contexts supports perspective-taking which contributes to empathy, compassion, and altruism

(Klimecki, 2019). Critically reflective thinking can inform more just action toward achieving a more just society for all.

ACKNOWLEDGEMENTS

We would like to thank Gulsah Gurkan and Dan DeCelles for their help in preparing data for this study.

REFERENCES

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

Association of American Colleges and Universities. (2018). *Fulfilling the American dream: Liberal education and the future of work*. Retrieved from https://www.aacu.org/sites/default/files/files/LEAP/2018EmployerResearchReport.pdf

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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*. https://doi.org/10.3389/feduc.2020.00156

Brookfield, S.D. (1987). *Developing critical thinkers: Challenging adults to explore alternative ways of thinking and acting* (pp. xvi, 293). Jossey-Bass.

Caplan, B. (2019). *The case against education: Why the education system is a waste of time and money*. Princeton University Press.

Chen, M-P., Lord, A.Y.Z., Cheng, Y-Y., Tai, K-C., & Pan, W-H. (2020). Collective reflection strategy for moderating conformity tendency and promoting reflective judgment performance. *Journal of Computer Assisted Learning*, *36*(3), 383–396. https://doi.org/10.1111/jcal.12419

Douglas, E. (2006). Critical thinking skills of engineering students: Undergraduate vs. graduate students. *2006 American Society for Engineering Education Annual Conference & Exposition Proceedings*, 11.374.1–11.374.10. https://doi.org/10.18260/1-2--101

Dwyer, C.P., Hogan, M.J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity*, *12*, 43–52. https://doi.org/10.1016/j.tsc.2013.12.004

Ennis, R.H., Millman, J., & Tomko, T.N. (1985). *Cornell critical thinking tests level X & level Z: Manual*. Midwest Publications.

Evans, C. (2020). *Measuring student success skills: A review of the literature on critical thinking* (p.18). National Center for the Improvement of Educational Assessment.

Facione, P.A. (1990). *Technical Report #2 Factors Predictive of CT Skills*. *23*.

Flowers, L.A., & Pascarella, E.T. (2003). Cognitive effects of college: Differences between African American and Caucasian students. *Research in Higher Education*, *44*(1), 21–49. https://doi.org/0361-0365/03/0200-0021/0

Franco, A.R., Costa, P.S., Butler, H.A., & Almeida, L.S. (2017). Assessment of undergraduates' real-world outcomes of critical thinking in everyday situations. *Psychological Reports*, *120*(4), 707–720. https://doi.org/10.1177/0033294117701906

Gadzella, B.M., Masten, W.G., & Huang, J. (1999). Differences between African American and Caucasian students on critical thinking and learning style. *College Student Journal*, *33*(4), 538–538.

Halpern, D.F. (2010). *Halpern Critical Thinking Assessment.* SCHUHFRIED (Vienna Test System).

Harris, K., Stein, B., Haynes, A., Lisic, E., & Leming, K. (2014). Identifying courses that improve students' critical thinking skills using the CAT instrument: A case study. *Proceedings of the 10th Annual International Joint Conferences on Computer, Information, System Sciences, and Engineering*, *10*, 1–4.

King, P.M. (2000). Learning to make reflective judgments. *New Directions for Teaching and Learning*, *2000*(82), 15–26. https://doi.org/10.1002/tl.8202

King, P.M., & Kitchener, K.S. (1994). *Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults.* Jossey-Bass.
King, P.M., & Kitchener, K.S. (2004). Reflective judgment: Theory and research on the development of epistemic assumptions through adulthood. *Educational Psychologist*, *39*(1), 5–18. https://doi.org/10.1207/s15326985ep3901_2
Kirsch, I., Braun, H., Lennon, M.L., & Sands, A. (2016). *Choosing our future: A story of opportunity in America*. ETS Center for Research on Human Capital and Education.
Kitchener, K.S., & King, P.M. (1985). *Reflective judgement scoring manual*.
Klimecki, O.M. (2019). The Role of Empathy and Compassion in Conflict Resolution. *Emotion Review*, *11*(4), 310–325. https://doi.org/10.1177/1754073919838609
Kugelmass, H., & Ready, D.D. (2011). Racial/ethnic disparities in collegiate cognitive gains: A multilevel analysis of institutional influences on learning and its equitable distribution. *Research in Higher Education*, *52*(4), 323–348.
Kwak, D. (2007). Re-conceptualizing critical thinking for moral education in culturally plural societies. *Educational Philosophy and Theory*, *39*(4), 460–470. https://doi.org/10.1111/j.1469-5812.2007.00353.x
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
Lord, A.Y.Z., Chen, M-P., Cheng, Y-Y., Tai, K-C., & Pan, W-H. (2017). Enhancing nutrition-majored students' reflective judgment through online collective reflection. *Computers & Education*, *114*, 298–308. https://doi.org/10.1016/j.compedu.2017.07.010
Mines, R.A., King, P.M., Hood, A.B., & Wood, P.K. (1990). Stages of intellectual development and associated critical thinking skills in college students. *Journal of College Student Development*, *31*, 538–547.
Nora, A., Cabrera, A., Hagedorn, L.S., & Pascarella, E. (1996). Differential impacts of academic and social experiences on college-related behavioral outcomes across different ethnic and gender groups at four-year institutions. *Research in Higher Education*, *37*(4), 427–451.
Oser, F.K., & Biedermann, H. (2019). A three-level model for critical thinking: Critical alertness, critical reflection, and critical analysis. In O. Zlatkin-Troitschanskaia (Ed.), *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)* (pp. 89–106). Springer International Publishing. https://doi.org/10.1007/978-3-030-26578-6_7
Pascarella, E.T., Wang, J-S., Trolian, T.L., & Blaich, C. (2013). How the instructional and learning environments of liberal arts colleges enhance cognitive development. *Higher Education*, *66*, 569–583. https://doi.org/10.1007/s10734-013-9622-z
Paul, R., & Elder, L. (2012). *Critical thinking: Tools for taking charge of your learning and your life* (third). Rowan & Littlefield.
Pellegrino, J.W., & Hilton, M.L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press. https://doi.org/10.17226/13398
Roska, J., Kilgo, C.A., Trolian, T.L., Pascarella, E.T., Blaich, C., & Wise, K.S. (2017). Engaging with diversity: How positive and negative diversity interactions influence students' cognitive outcomes. *The Journal of Higher Education*, *88*(3), 297–322. https://doi.org/10.1080/00221546.2016.1271690
Shavelson, R.J. (2010). *Measuring college learning responsibly: Accountability in a new era*. Stanford University Press.
Shaw, A., Liu, O.L., Gu, L., Kardonova, E., Chirikov, I., Li, G., . . . Loyalka, P. (2020). Thinking critically about critical thinking: Validating the Russian HEIghten® critical thinking assessment. *Studies in Higher Education*, *45*(9), 1933–1948. https://doi.org/10.1080/03075079.2019.1672640

Sotiriadou, P., Logan, D., Daly, A., & Guest, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Studies in Higher Education*, *45*(11), 2132–2148. https://doi.org/10.1080/03075079.2019.1582015

Van Tyne, N.C.T., & McNair, L.D. (2019). *Testing a reflective judgement scale for suitability with first year student reflective responses*. American Society for Engineering Education annual conference, Tampa, FL.

Walsh, C.M., & Hardy, R.C. (1999). Dispositional differences in critical thinking related to gender and academic major. *Journal of Nursing Education*, *38*(4), 149–155.

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

Wolcott, S.K., & Lynch, C.L. (1997). Critical thinking in the accounting classroom: A reflective judgment developmental process perspective. *Counting Education: A Journal of Theory, Practice and Research*, *2*(1), 59–78.

Wood, P.K. (1997). A secondary analysis of claims regarding the Reflective Judgment Interview: Internal consistency, sequentiality and intra-individual differences in ill-structured problem solving. In *Higher Education: Handbook of Theory and Research: Vol. XII* (pp. 243–312). Agathon Press.

Wood, P., Kitchener, K., & Jensen, L. (2002). Considerations in the design and evaluation of a paper-and-pencil measure of epistemic cognition. In *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 277–294). Lawrence Erlbaum Associates Publishers.

APPENDIX A: SUMMARY OF STUDIES

	Purpose	Participants	Comments
First pilot	Testing the first version of the CRA	85 first year EHD students	Fall semester
Second pilot	Testing the revised scoring rubric	16 first year EHD students	Spring semester, a subset of the first pilot
Study 1	Testing the CRA on a larger sample, comparing majors	405 students (various majors)	Fall semester
Study 2	Track growth in CT skills during the first year of college	235 first year EHD students	Students were assessed three times throughout the year
Study 3	Track growth in CT during the four years of college	85 first year EHD students (from pilot) 80 fourth year EHD students	62 students from the pilot sample participated twice (first and fourth years)

Note: CRA = Critical Reasoning Assessment, EHD = School of Education and Human Development, CT = Critical reasoning.

APPENDIX B: CRA PROMPTS

Items

- What is your opinion about this issue?
- How have you come to hold this point of view?
- On what do you base that point of view?
- Can you ever know for sure that your position on this issue is correct? Why or why not?
- When two people differ about matters such as this, is it the case that one opinion is right and one is wrong? If yes, what do you mean by "right"? If no, can you say that one opinion is in some way better than the other? What do you mean by better?
- How is it possible that people have such different points of view about this subject?
- How is it possible that experts in the field disagree about this subject?

Dilemmas

(1) Genetics vs. choice

Some researchers contend that prescription and non-prescription drug abuse (substance use disorder) is due, at least in part, to genetic factors. They often refer to results from numerous studies that support this assertion. Other researchers, however, believe that prescription and non-prescription drug abuse is a choice, claiming that the reason several family members often suffer from prescription and non-prescription drug abuse is because they share common family experiences, socioeconomic status or employment.

(2) Fairness

Some people define *fairness* as treating everyone equally regardless of other factors that may contribute to one's requiring additional economic, academic, emotional or social support. Others, however define fairness as giving each person what he/she needs as other factors may contribute to one's requiring additional economic, academic, emotional or social support.

(3) Compassion

Some people believe that as long as we extend genuine love and care to our family and social group, we are living a life of goodness and compassion. Tacit participation in the dehumanization of others—even though we do not physically, socially, verbally, intellectually or emotionally participate in this lack of inclusivity—is acceptable as long as we are caring for our own in the best possible way. Others believe, however, that living a life of goodness and compassion demands that we extend love and care to those who are not in our familial or social group: the "other." Thus, tacit participation is not inclusive and is discriminatory.

APPENDIX C: THE CRA REVISED SCORING RUBRIC

	GD Cognitive Complexity	GD Reasoning Style	GD Openness	SD Nature of Knowledge	SD Nature of Justification
1	Knowledge is concrete, tangible, singular, & observed; clichéd.	Opinion = fact; no reasoning	Uninformed & naïve; deliberately ignores points of view that are unlike their own	Knowledge needs no justification; right or wrong is known; no other options	Alternatives do not exist; evidence is not evaluated; knowing is egocentric
1.5	Knowledge is concrete; alt. views emerging; not qualified as right or wrong	Reasoning is attempted, but facts are inaccurate	Still naïve & uninformed, recognizes but dismisses other points of view	Knowledge is certain, right or wrong; authorities justify knowledge	Justification combines egocentrism and with what subject has been told by authorities
2	Knowledge is right or wrong; solutions are simple & easy	Reasoning is illogical; opinion & evidence are blurred	Other points of view are possible but wrong.	Knowledge is absolutely certain or un-certain; different views are just wrong	Justifies beliefs via authorities; not based on evidence but what subject has been told
2.5	Knowledge is right or wrong but uncertainty prevents choosing right or wrong	Reasoning is illogical; starting to see difference between opinion & evidence	Open to other points of view but they exist in limbo without right or wrong designation	Knowledge may be uncertain; moving away from labeling views as right or wrong.	Acknowledges good & bad authorities; uses personal experience in combination with good authority to justify decision
3	Knowledge is true, false or uncertain; ambiguity is troubling	Some logic, that is personal & subjective; what feels right	Some openness; rejects beliefs rather than be uncertain	Knowledge will eventually be known; one answer is just as good as another	Decisions are tentative; fact & opinion are different; authority is questioned
3.5	Knowledge is questioned; emerging toleration for ambiguity	Reasoning appears more logical but does not pose any point of view	Open to other points of view; beginning to accept uncertainty.	Knowledge may not be eventually known; beginning to consider that knowledge is uncertain	Moving toward having a strong opinion but still subjective based on unexamined facts & opinions
4	Knowledge is uncertain; issues are complex but without sub-issues	Beginning to realize role of evidence; but does not	Open to other points of view; stubborn or	Knowledge is idiosyncratic, abstract, uncertain,	Expresses strong point of view without objectivity;

		consistently argue with evidence	wishy-washy; contradictory	relativistic; discrete differences; no gray	evidence is incomplete; authority is dogmatic
4.5	Knowledge is complex; certainty & realization of sub-issues are emerging but limited	Reasoning is logical; argument is somewhat explicit with limited evaluation of evidence	Open to other points of view; objective about some/subjective about others.	Knowledge is idiosyncratic and relativistic but not black or white; some ambiguity	Point of view is expressed with some objectivity, but evidence is incomplete
5	Knowledge is complex; experience is limiting; evidence has many sides	Reasoning is logical; with explicit & consistent evaluation of evidence	Sees diverse points of view; objective about all points of view	Knowledge is domain-specific, uncertain, & complex; difference relates to different worldview	Justifies point of view based on evidence that is judged qualitatively using simple/learned rules of inquiry; offers balanced "big picture"
5.5	Knowledge is complex; emerging exploration & analysis	Reasoning is logical; recognizes that some evidence is more compelling; beginning to articulate such claims	Examines many points of view; dismisses some unreasonable claims	Knowledge is complex & contextual; limited comparison of points of view across domains.	Justifies point of view based on evidence but starting to inquire & compare perspectives
6	Knowledge is complex & analyzed across points of view	Reasoning is logical based on evaluation of evidence; articulates more compelling claims	Examines many points of view; dismisses unreasonable claims, but offers no final personal view	Knowledge is judged qualitatively & although valid, not totally defensible; considers credibility of claims	Assumes but does not construct point of view; evaluates strength of evidence & experts' claims
6.5	Knowledge is analyzed & synthesized, but not con-structed to yield a perspective	Reasoning is logical; emerging strategy; with some abstraction	Dismisses unreasonable claims; close to offering personal view	Knowledge is judged qualitatively; some perspectives could be more defensible than other; more inquiry is needed	Beginning to construct point of view using strength of evidence and claims; beginning to synthesize across domains

7	Knowledge is complex, analyzed, synthesized, constructed to form a coherent perspective	Reasoning is logical, strategized, generalized into abstractions supported by evidence	Sees why others hold points of view but owns personal view; is open to new information	Knowledge results from rigorous inquiry across multiple perspectives, across multiple contexts & is more or less reasonable & defensible	States opinions firmly based on evaluated evidence; abstracts across & within domains; constructs higher-order thinking

APPENDIX D: TRANSITION MATRICES FOR STUDY 2 BY GENDER

Males (n = 22)

	T_3			
		Pre-reflective	Quasi-reflective	Reflective
T_1	Pre-reflective	3	3	2
	Quasi-reflective	-	2	9
	Reflective	-	-	3

Female (n =178)

	T_3			
		Pre-reflective	Quasi-reflective	Reflective
T_1	Pre-reflective	7	32	5
	Quasi-reflective	7	81	32
	Reflective	-	5	9