2023

Improving Critical Thinking in Written Assignments: Human vs. ChatGPT Tutor in Socratic Questioning Intervention

Katia Martha

Improving Critical Thinking in Written Assignments: Human vs. ChatGPT Tutor in Socratic Questioning Intervention

by
Katia A. Martha

Submitted to Scripps College in partial fulfillment of the degree of Bachelor of Arts

Professor Groscup
Professor Chatterjee
Professor Morgan

December 15th, 2023

Abstract

The purpose of the proposed study is to trial a short Socratic Questioning (SQ) intervention in the writing process, facilitated by either a human or ChatGPT tutor, and explore the effects that this may have on students' critical thinking (CT), which will be coded from their written responses. Participants will be undergraduate college first years in the local California area who are fluent in English and have no learning disabilities. This study involves two visits, spaced a week apart, to gather pre- and post- test data for evaluating the effectiveness of the SQ intervention in improving CT. Both visits will follow a similar format of participants taking a written exam followed by a survey, which, in visit 1, will ask about ChatGPT account set-up and participant demographics, and in visit 2, will ask about intrinsic motivation. Researchers expect that post-test CT scores will be higher in conditions with the SQ intervention as opposed to the control, and specifically higher for the SQ intervention with the human tutor versus the ChatGPT tutor. This relationship is expected to be moderated by baseline level of CT, with lower levels of baseline CT predicting greater impacts from the SQ intervention, and mediated by intrinsic motivation, where intrinsic motivation is expected to be higher for participants with the human versus ChatGPT tutor. This study will address many gaps in the literature, including how to develop and assess CT in writing, specific methods of using ChatGPT to improve student learning, comparing humans versus ChatGPT as tutors, and looking at SQ as an intervention in the writing process.

Improving Critical Thinking in Written Assignments: Human vs. ChatGPT Tutor in Socratic Questioning Intervention

Critical thinking (CT) and writing abilities have consistently been held as central goals of a good education. This is especially true of higher education, where many colleges claim these as skills that students will develop over the course of their college career (Ennis, 1993; Karanja, 2021; Mok, 2009; Yanning, 2017). Critical thinking, especially, is more important now than ever. With the internet age presenting a world in which it is increasingly easier to be misinformed, harder to parse out what is true, and increasingly polarized due to the preceding two issues, critical thinking has gained a reputation as a key 21st century tool, needed to navigate the floods of information online (Guo et al., 2023; Lombardi, 2023; Koreshnikova & Avdeeva, 2022; Mok, 2009; Sahamid, 2016; Yanning, 2017). Beyond the individual level, a citizenry without dedication to critical thinking puts our very democracy at risk (Westheimer, 2008). The recent release of ChatGPT in 2022 has posed a threat both to student's development of critical thinking and writing abilities, sparking a panic across the educational sector that students will use the AI bot for easy cheating, especially on written assignments. Many schools were quick to ban ChatGPT for these reasons while others refrained, arguing that ChatGPT will become an indispensable tool to both education and the future workplace, meaning that students who do not get to use it in school now will be at a disadvantage later (Grassini, 2023). Presently, educators are still debating what to do about ChatGPT and many feel confused or at a loss for what the best approach would be. This study seeks to flip the common narrative about ChatGPT being a tool used mostly for cheating on essays to instead ask how ChatGPT can be used to enhance students' critical thinking and writing abilities. In so doing, this study also hopes to address the pressing need for more evidence-based guidance around how to use ChatGPT to improve learning outcomes as well as address the many gaps in the literature and in practice around this topic.

Defining Critical Thinking

Despite its celebrated importance across educational contexts, a concrete definition for critical thinking (CT) remains elusive and many studies on the topic of critical thinking open with this acknowledgement that the concept is vague and hard to pin down (Condon & Kelly-Riley, 2004; Ennis, 1993; Facione, 1990; Karanja, 2021; Lombardi, 2023). Contributing to the haziness of the definition is the generalist vs. specifist debate on critical thinking. Generalists argue that critical thinking should be thought of as a general skill, free of context, while specifists argue that critical thinking must be defined in a discipline-specific way (Davies, 2006). Condon and Kelly-Riley (2004) offer in the specifist view, for example, that the critical thinking undertaken by a student in an Orchard Management course should differ greatly from the type of critical thinking a student of Metaphysics Philosophy is engaging with. Davies (2006) calls out the debate for posing a false dichotomy and instead proposes an ‘infusion approach’ to critical thinking, which recognizes that in fact, either side on its own falls short of a satisfactory definition of critical thinking. An infusion approach demands that core principles of reasoning, which are key to critical thinking in any context, and the flexibility to adapt to specific context must both be accepted in a comprehensive definition of critical thinking.

Paul and colleagues (1997) seem to have followed a similar train of thought in their approach to interviewing faculty about their definitions of critical thinking, where they assessed the extent to which faculty understood the ‘minimalist’ or essential elements of critical thinking. While Paul and colleagues claimed not to take on ‘any particular view’ of critical thinking throughout their study, they did admit that there were four core, interrelated components, which they viewed as essential to any definition of critical thinking. These four components consist of: (1) ability to engage in discourse from a place of reasoning, (2) where intellectual standards mandate the reasoning, (3) reasoning involves analytical inferential skills, and (4) there is a dispositional commitment to this type of reasoning in terms of habit of mind or attitude. It became clear from subsequent faculty interviews that most definitions faculty reported also shared a common core set of elements. Thus, Paul and colleagues stood by their ‘minimalist’ conception – or ‘infusion approach’ as Davies (2006) would call it – of critical thinking, while denying attachment to any singular definition of critical thinking and instead claiming that their ‘minimalist’ conception would be consistent with all different types of critical thinking definitions. This infusion or minimalist approach to conceptualizing critical thinking appears reasonable in that it takes both sides of the generalist-specifist debate into account.

One thing there is no debate over, on the other hand, is that critical thinking consists of two main elements: skills and dispositions (Ennis, 1993; Facione, 1990; Paul et al., 1997; Paul & Elder, 2001, as cited in Yanning, 2017). In a seminal piece of critical thinking literature referred to as the “Delphi Report,” Facione (1990) utilized the Delphi Method to develop a definition of critical thinking, which ultimately found critical thinking to encompass both skills and disposition. The Delphi Method entails the formation of a panel of experts who must meet numerous times to work toward a consensus on a topic, where each meeting builds upon the last. In the “Delphi Report,” forty-six experts in critical thinking from a range of different disciplines met over six rounds, which lasted from February, 1988 to November, 1989. Six skills were identified: interpretation, analysis, evaluation, inference, explanation, and self-regulation, each with further-specified sub-skills. A longer list of dispositional elements was proposed, which included being inquisitive, concerned with being well-informed, trusting in the reasoning process, and open-minded, among others (Facione, 1990).

Perhaps the most widely cited definition of critical thinking comes from Ennis (1993), who defines critical thinking as “reasonable reflective thinking focused on deciding what to do or believe” (180). Another popular conception of critical thinking stems from Bloom’s taxonomy, which visualizes a hierarchy of thinking skills, where critical thinking has been said to be reflected in the top three levels: analysis, synthesis, and evaluation (Ennis, 1993; Koreshnikova & Avdeeva, 2022; Walker, 2003). While it is good to be aware of different definitions of critical thinking, many scholars assert that there can be no one definition of critical thinking because its range of applicability and its history is simply too vast to be confined to one definition (Condon & Kelly-Riley, 2004; Paul et al., 1997).

For the purpose of this study, which will assess critical thinking from argumentative writing, the definition of critical thinking (CT) will follow Paul and Elder’s (2001) three-dimensional CT model (Appendix A), which consists of eight elements of thought, nine intellectual standards, and eight intellectual traits (Yanning, 2017). This definition of CT was chosen for several reasons. First, a rubric for coding CT from argumentative writing has already been designed and implemented in a study based on this model (Yanning, 2017). Second, Richard Paul and Linda Elder, who created the model, are well-known scholars in the critical thinking literature and have both been directors of the Foundation for Critical Thinking, which gives the model credibility. Additionally, their experience researching CT is apparent in the comprehensive and complex view of CT the model presents, which has three detailed dimensions. The model also assumes an infusion approach (rather than a specifist or generalist), by having core elements of critical thinking which can be adapted to many disciplines and in this case, easily applied to assessing critical thinking from writing.

Defining Socratic Questioning

Socratic questioning consists of asking open-ended questions in pursuit of evaluating assumptions, beliefs, and one's own thinking process (Etemadzadeh et al., 2013; Jaeger, 2016). It differs from other types of questioning in its systematicity, depth, and the associated concerted effort to discover what is true; it is a disciplined form of questioning, which encourages students to develop a line of reasoning (Elder & Paul, 1998; Paul & Elder, 2007). Socratic questioning is recognized as facilitating students' development as independent learners, actively involved in constructing their own knowledge (Etemadzadeh et al., 2013; Padesky, 1993).

That students are constructing their own knowledge is especially important to the Socratic questioning process, which should be a practice of guided discovery, rather than one of the teacher changing the student's mind (Padesky, 1993). Guided discovery should entail some direction and structure – it should not be completely aimless – but does not require the guide or teacher to have an exact answer in mind; ideally, the teacher does not have one answer in mind, so as to let the student lead in the discovery process (Jaeger, 2016; Padesky, 1993). A teacher who voices a direct challenge to a student belief instead of asking a guiding question disallows that student from uncovering any inconsistencies for themselves, which would have been a more meaningful and motivating process for the student. Directing students toward pre-determined answers represents a quick fix to the current assignment or activity at hand, rather than allowing students to practice building the long-term skills of finding their own solutions and thinking critically (Padesky, 1993). This process of guided discovery with Socratic questioning is a practice more commonly used in cognitive therapy but can easily be translated to the classroom setting.

Socratic questioning originated in ancient Greece with Socrates, who was known for starting somewhat infuriating conversations with other Athenians, where their dialogues would serve to deconstruct the interlocutor's (Socrates' conversation partner's) beliefs, leaving Socrates to always conclude that "all he knew was that he knew nothing" (Jaeger, 2016). While there are some valuable elements to this original form, such as starting dialogue by 'destabilizing commonplaces' or dismantling certain assumptions and beliefs around the topic at hand, other components of Socrates' original approach would be extremely problematic to employ with students today. Socrates' questioning could be quite aggressive and destructive; often, his main aim seemed to be tearing down the beliefs of the interlocutor without any care for building up new ideas or offering emotional closure afterwards. This often left the interlocutor feeling disheartened, confused, or frustrated at the end of a dialogue. Obviously, these are not feelings teachers would want to inspire in students after guiding them through reasoning on a topic. Amendments to the original version of Socratic questioning to better support students might involve 'charitably rehearsing' the student's argument with them before identifying any issues, involving more sensitivity and tact throughout the questioning process, and periodically summarizing what has been discussed, as there can be valuable things to learn even from the un-learning process (Jaeger, 2016; Padesky, 1993).

This study will follow a definition of Socratic questioning (SQ) as open-ended, systematic questions, which are delivered in a sensitive way to guide discovery around assumptions, beliefs, and one's own thinking process. This definition incorporates Jaeger's (2016) and Padesky's (1993) recommendations on good Socratic questioning practice in terms of supporting the student, which notably departs from how Socrates originally conducted Socratic questioning. It must also be acknowledged that there is no one strict form or manual that Socratic questioning must follow (Padesky, 1993; Sahamid, 2016); in fact, Padesky (1993) claims this is a good thing because Socratic questioning should be a very personalized process and not follow a template. However, certain studies investigating the links between Socratic questioning and critical thinking have elected to use some guidance for developing Socratic questions in the form of Paul and Elder's (2001) model of CT (Appendix A), which is informed by a recognition of the inherent connection between SQ and CT (Anderson & Piro, 2014; Piro & Anderson, 2015; Sahamid, 2016). The Socratic questioning intervention in the proposed study itself will not necessarily be informed by Paul and Elder's model, as this would require uploading the model into ChatGPT before beginning Socratic questioning, and researchers are specifically interested in exploring the more natural interactions students can have with ChatGPT without needing to put anything additional into the chatbot before getting to interact with it. Studying the effects of these more natural interactions will be important for informing practices that students and teachers can easily do in their homes and classrooms. However, the rubric which will inform coding of CT from participants' essays will be based on Paul and Elder's model.

To ensure that ChatGPT, in its natural state, without any additional input, reflects the definition of Socratic questioning as suggested above for the proposed study, researchers asked ChatGPT to define Socratic questioning and engaged in several rounds of Socratic questioning practice sessions with the chatbot to assess if its understanding of SQ aligned with that of the researchers. Indeed, ChatGPT's definition of SQ was found to fit well with the researchers' own definition, identifying SQ as a form of dialogue which helps guide individuals in discovering and clarifying thoughts, promoting critical thinking, and encouraging deeper exploration of ideas, assumptions, and beliefs. Additionally, the types of questions ChatGPT posed in the SQ practice sessions with researchers also reflected alignment with the researchers' definition of SQ.

Introduction to ChatGPT

ChatGPT is currently the most advanced AI (Artificial Intelligence) chatbot available. Released by OpenAI on November 30, 2022, ChatGPT shocked the public with its quick, human-like text responses and vast knowledge, which allows it to respond directly to a wide variety of specific prompts (Grassini, 2023). Users can prompt the chatbot with requests ranging from asking it to write an email with specific expectations, to creating an imaginative short story about a dog in outer space, to getting help detecting errors in code, and of course, as central to the reasoning for this study, students can use ChatGPT to generate an essay in mere seconds. Due to its widespread applicability and its never-before-seen abilities, the platform sparked interest quickly, with over one million subscribers in its first week (Grassini, 2023). Since then, it has only grown its following and now boasts over 100 million weekly active users (Malik, 2023). Furthermore, ChatGPT is not only being adopted by the average curious person to play with but is quickly being integrated into the workplace as well. Sam Altman, the CEO of OpenAI, recently shared that 92% of Fortune 500 companies are using ChatGPT (Malik, 2023).

As ChatGPT's reach and influence grows, so does the importance of understanding how it works. The 'GPT' in ChatGPT can be expanded to reveal its function as a "Generative, Pre-trained, Transformer," which means, at a basic level, that the AI chatbot is trained on large amounts of data so that it can learn how to recognize certain patterns and context (Grassini, 2023). By building this pattern recognition, ChatGPT can essentially function as a prediction machine, which predicts the next most likely word in a certain context (Schade, 2023). In other words, it does not consciously 'understand' anything but merely generates text based off of perceived patterns. 'GPT' also indicates the Natural Language Processing engine (NLP) that ChatGPT runs on, of which there have been several iterations. When it was first released, ChatGPT ran on GPT-3, though the free version has since been updated to GPT-3.5, which simply signifies expanded capabilities due to training on more data and involving more parameters in its construction. For users able and willing to pay, there are paid subscriptions available for 'Plus' and 'Enterprise' versions of ChatGPT, which run on GPT-4 and represent the most powerful model yet. For example, GPT-4 was able to not only pass the US bar exam, but test in the 90th percentile of scores, whereas the previous version of ChatGPT tested in the 10th percentile (Grassini, 2023; Schade, 2023). GPT-4 additionally has the expanded capabilities of analyzing and generating visuals (OpenAI, 2023).

While one can ask ChatGPT nearly anything, there are some safety measures in place, which prevent it from answering certain questions. For example, if users ask ChatGPT who the best Nazi was or how to build a bomb, it will respond with a statement that it cannot answer these questions as well as moral cautioning about why it is wrong to ask those questions. However, these safety measures are imperfect and 'jailbreaks' are known to happen, where the way a user phrases a prompt can sometimes trick ChatGPT into answering a question that it would normally object to. A popular jailbreak at one point involved asking ChatGPT to act as the user's deceased grandmother and tell them a story about [insert off-limits topic] from her life (Zhang, 2023). While ChatGPT will no longer be tricked by this phrasing now, people continue to discover new jailbreaks, which potentially poses great risk to society.

Relatedly, it may be tempting to believe that because ChatGPT is a machine and especially because it has this 'moral' coding, that it contains no biases. However, one would be sorely mistaken to assume such a thing. ChatGPT has displayed blatant biases in different scenarios such as a case where a user asked ChatGPT to generate a python function to determine what makes a good scientist based on race and gender; ChatGPT responded by identifying “white” and “male” as the race and gender which make a good scientist (Singh & Ramakrishnan, 2023). OpenAI encourages users to use the up-vote and down-vote functions which appear beside each ChatGPT response to flag biased answers so that similar responses can be avoided in the future (Singh & Ramakrishnan, 2023). While this may help to some extent, it is important users understand that ChatGPT inherently contains biases because it was trained on publicly available data, which is fraught with human biases (Borji, 2023; Grassini, 2023; Schade, 2023).

An additional limitation to ChatGPT is that despite being able to perform at such a high level in some areas, it still makes mistakes. Sometimes these mistakes are on very basic reasoning tasks and sometimes the bot simply ‘hallucinates’ or makes up false information (Borji, 2023). While a response from ChatGPT may come across as extremely confident and eloquent, it may simply be a very confident hallucination. Users need to be aware of this so that they understand the importance of fact-checking any information ChatGPT imparts. All the above limitations must be considered when thinking about how to engage responsibly with ChatGPT.

While there is so much potential to explore in terms of using ChatGPT to enhance student outcomes, it is important to researchers to narrowly focus on one specific method for this study (Socratic questioning as a writing intervention). Much like the case with the internet, it is unhelpful on a practical level to investigate the effects of general use of ChatGPT (as opposed to specific methods of interaction) on student outcomes because there are so many ways to interact with ChatGPT, each which would likely engage students in different processes with different outcomes. Shen (2018) highlights this point by investigating school-related internet information seeking styles in relationship to academic self-efficacy. Shen found that there were students who turned to the internet for ready answers that they could copy and paste without thought (executive internet information seekers) and others who sought information on the internet only to the extent where they could then complete the rest of the task on their own (instrumental internet information seekers). Unsurprisingly, instrumental internet information seekers were found to have the highest level of academic self-efficacy while the executive internet information seekers had the lowest. These findings emphasize the idea that students will use the internet and relatedly, ChatGPT, in different ways and while some will abuse it, others will engage in adaptive interactions with it, which can work to enhance learning and should therefore be explored in more depth. In this way, the proposed study falls under what Elliot (1985) refers to as ‘action research,’ which is primarily focused on addressing practical problems faced by practitioners and providing evidence-backed strategies which might address these problems (Sahamid, 2016).

Developing and Assessing Critical Thinking in Writing

Despite claims by many colleges that they prioritize developing critical thinking and writing abilities, many are not engaged in the practice of directly developing or assessing students’ critical thinking in writing (Condon & Kelly-Riley, 2004; Karanja, 2021). Often, critical thinking and writing are thought of as intrinsically linked, where it is simply assumed that developing good writing goes hand in hand with developing good thinking (Beazley & Kearney, 1991; Condon & Kelly-Riley, 2004; Karanja, 2021; Paul et al., 1997). In fact, several studies have investigated and disproven this commonly held myth, finding that much depends on how the writing assignment is designed, graded, and how the process is scaffolded (Condon & Kelly-Riley, 2004; Karanja, 2021). While it is irrefutable that some amount of thought is required for writing, critical thinking does not always have to be present for a paper to receive a good grade.

Condon and Kelly-Riley (2004) found that faculty, more often than not, developed writing assignments which asked for and were graded on accurate information recall and summary rather than any critical thinking. Unsurprisingly, this meant that students could write a paper which succeeded in accurately recalling and summarizing information, thus securing a good grade, but failed to display any critical thinking. While teachers may assume that critical thinking will inevitably follow from their course instruction, there is evidence that many students in higher education are not developing critical thinking (Condon & Kelly-Riley, 2004; Karanja, 2021), indicating a gap between teacher beliefs and values on critical thinking in learning and real student outcomes.

Paul and colleagues (1997) affirmed that there is a disconnect between the value most teachers hold for critical thinking in education and the extent to which they are teaching any critical thinking in their courses. They found that despite a majority of teachers assigning high value to critical thinking, most struggled to define it, were confused about how to integrate teaching critical thinking into their courses, and ultimately, most did not teach any critical thinking in their courses at all (Paul et al., 1997). In a survey of over 100 Iranian teachers, Aliakbari and Sadeghdaghighi (2012) similarly found that a majority of teachers lacked confidence in their abilities to teach critical thinking and less than 10% believed other teachers could explain the departmental definition of critical thinking.

Karanja (2021) points to another problem area in developing critical thinking in writing: the grading rubrics. Oftentimes, rubrics are vague in what they are assessing, which translates to unhelpful feedback for students on how they might develop critical thinking (i.e. a rubric listing ‘persuasive writing’). Beyond this, there is a general dearth of available assessments for how to grade CT from written work and of what exists, there are many inconsistencies, perhaps reflecting the vagueness of the term CT itself.

While there are some recommendations of instructional strategies in the literature for developing critical thinking in writing, there remains a disconnect in implementing and assessing these in school practice. Some suggestions for improving critical thinking in writing include offering engaging writing prompts, peer/instructor feedback, explicitly stating CT development as a goal of the writing process, and multiple-step writing (Karanja, 2021; Yanning, 2017). Socratic questioning as an intervention in the writing process has rarely been explored or practiced. The proposed study seeks to fill this gap and demonstrate how SQ can be used as a practical strategy for improving CT in writing.

Systemic Barriers to Developing Critical Thinking in School

Beyond the lack of a concrete definition for critical thinking and the related confusion about how to integrate CT into instructional practice, there exist other barriers to developing CT in schools, namely the traditional Western education system itself. Western educational culture has generally been dictated by valuing the ‘product’ over the ‘process’ in learning, which can be traced to the value placed on grading and standardized testing (Sobo, 2023). This view steers the form of instruction, where to prepare students to produce good ‘products’ of schooling like a perfect test score or a ‘well-written’ essay, for example, the traditional approach for teaching is a didactic one of lecturing and telling students the answers they expect them to later regurgitate onto a test or paper (Koreshnikova & Avdeeva, 2022; Sobo, 2023). Freire (2005) defines this approach as the ‘banking method’ of education, where students act as passive receptacles, which teachers simply fill with knowledge.

Because students are expected to simply take in knowledge and push it out for assessment in this didactic, banking approach to teaching, students often learn to take that knowledge for granted, without questioning it or reflecting upon it. Koreshnikova and Avdeeva (2022) explicitly call out that didactic teaching is not conducive to critical thinking for these reasons and provide evidence for this assertion with their study comparing the effects of didactic and constructivist teaching styles, which found that didactic instruction mainly functioned through extrinsic motivation rather than intrinsic and had no relationship with the development of critical thinking. Nie and Lau (2010) similarly found that didactic teaching only predicted surface processing strategies and not deeper processing strategies, which include critical thinking.

The value placed on preparing students for standardized tests and producing ‘products’ of learning, using didactic teaching as a method of achieving this, informs practical, systemic barriers for teachers to develop CT in students. For example, Mok (2009) details an observational study conducted on two EFL teachers in Hong Kong with the objective to determine if they were implementing the 1999 critical thinking syllabus in class as they were supposed to or not. Mok found that not only were both teachers not adhering to the CT syllabus, but both were engaging in one stark violation of CT development: teachers were not providing enough ‘space of learning’ or processing space for students to reflect and respond to questions or think critically. Upon later discussion with researchers, the teachers shared that they felt they had to get students to finish a certain product (written task) in class and simply did not have enough time to allow for more space for students to process questions or ask questions themselves. This case study highlights how difficult it can be for teachers to incorporate CT development into their classrooms, often due to a perceived lack of time and space, which is informed by external pressures to be product-oriented. Walker (2003) relatedly highlights that providing ‘wait time’ and space for reflection is critical to developing CT.

Other barriers to developing CT are more explained by poverty and lack of resources than product-oriented values. Kati Haycock and her team of researchers from Education Trust observed a variety of low-income classrooms and found that coloring was “the single most predominant activity (…) right up through middle school” (Delpit, 2012, p.124) a phenomenon referred to as the “Crayola Curriculum.” Delpit (2012) reports on her own classroom observations and noted the stark difference in the typical writing assignment prompts between a predominantly white school and a predominantly African American and Latino school. The predominantly white school’s writing prompt asked students to write an essay on Anne Frank and included detailed instructions about what was expected of the structure and content. An example of a typical writing assignment in the predominantly African American and Latino school was a fill-in-the-blank worksheet, titled “About Me,” with prompts like “a car I want” and “my heartthrob.” Delpit (2012) emphasizes the development of CT as a necessity for liberation, especially for low-income students of color, who are oppressed in a variety of ways- if they do not learn to question the structures they operate in and the directions they are given, they will never be able to liberate themselves. This sentiment is also echoed by Freire (2005), who outlines a pedagogy of oppressed, which essentially a call to action particularly for oppressed peoples to develop critical thinking and a praxis so that they can fight against oppressive systems.

Elder and Paul (1998) along with Holmes (2015) argue that teaching should be more focused on posing and developing questions rather than giving answers. Giving students the answers instead of guiding their discovery can signal to students that they do not need to think critically about the material they are being given (Elder & Paul, 1998). Elder and Paul (1998) go so far as to say that traditional methods of teaching have deadened students' minds and urge teachers to inject more questioning into instruction to give students "artificial cogitation" or artificial respiration for their minds. Holmes (2015) arguably goes farther to assert that not only should teachers be involving questioning more, but they should actively be teaching about ignorance. Holmes expands to say that not teaching about the limits of our knowledge can signal to students that we already know everything there is to know and there is nothing more to ask questions about. Furthermore, exploring the edges of our islands of knowledge brings excitement, engagement, and questions.

Sobo (2023) proposes that the rise of ChatGPT may represent a beacon of hope for changing Western educational culture rather than the harbinger of doom many educators believe it signifies. Sobo argues that ChatGPT exposes the faults of the education system's product-oriented values; it reveals that teachers are essentially asking students to fill in answers or produce products in the way that a machine can. To adapt to the integration of ChatGPT calls for a re-evaluation and an evolution of educational values back towards 'process' over 'product,' where teachers will be able to verify and value the humanity of student's learning through a focus on their process (Campbell, 2023; Sobo, 2023).

This study's focus on how Socratic questioning and ChatGPT can be used to develop critical thinking in the writing process is grounded in a recognition of the necessity to shift Western educational culture towards valuing process over product and the importance of developing critical thinking, as this will create a culture that values students' humanity and functions as a liberatory practice for the oppressed.

The Relationship Between Socratic Questioning and Critical Thinking

Socratic questioning (SQ) and critical thinking (CT) are deeply related. In fact, Paul and colleagues (1997) claim that critical thinking can be traced back 2,500 years to start with Socrates and his questioning technique. It is important to note, however, that a focus on Greco-Roman history is a common bias of Western academia and it is likely that there were others from other cultures who practiced critical thinking but whose methods were not as neatly recorded or honored as Socrates' has been.

At a very basic level, asking questions in general is recognized as an important facet of actively engaging students in learning and evolving their thinking around a topic (Elder & Paul, 1998; Erdogan & Campbell, 2008; Etemadzadeh et al., 2013; Mok, 2009). The quality of the questions asked determines the level of thinking, learning, and overall engaging students will do (Etemadzadeh et al., 2013). Socratic questioning, in particular, is known to facilitate critical thinking by employing the components of reasoning, which are central to critical thinking, in order to form deep, disciplined questions, and develop a line of inquiry (Elder & Paul, 1998; Paul & Elder, 2007). Erodgan and Campbell (2008) especially highlight how open-ended questions, which are characteristic of Socratic questions, represent a higher quality question format than close-ended as they invite more thought and often require students to reveal their reasoning and justification, whereas a yes/no question does not necessarily prompt further thought. Walker (2003) relatedly asserts that questions for promoting critical thinking should go beyond mere information recall or questions that have one set answer, which can be thought of as belonging to a class of lower-level cognitive questions.

Socratic Questioning as a Constructivist Learning Methodology

This study adopts the framework of Constructivist Learning Theory or Constructivism, as proposed by Lev Vygotsky, to explain how Socratic questioning (SQ) can promote critical thinking (CT). Constructivism posits that learners construct their own knowledge through active learning and meaning making (Hatmanto & Sari, 2023; Kwan & Wong, 2015; Nie & Lau, 2010). Thus, constructivist teaching is often characterized by creating an interactive learning environment (Koreshnikova & Avdeeva, 2022). Hatmanto and Sari (2023) identify the main components of constructivism as active engagement, learner autonomy, and knowledge construction. Socratic questioning can be seen as a form of constructivist teaching and learning because its practice meets all three of these main components. First, as a personalized, deeply probing dialogue, SQ requires active engagement of its participants. Walker (2003) supports this point by identifying high-level questioning and particularly Socratic questioning as a strategy of active learning to promote CT. Second, as a mutual process, where both parties involved can receive and give questions and thoughts, learner autonomy is also central to SQ. Finally, as Padesky (1993) highlights, SQ should be a process of guided discovery, where the student is not told the answer nor should the teacher or facilitator have a particular answer in mind, but both commit to reasoning through the discovery process, thereby constructing knowledge. The guided discovery and active learning aspects of SQ involve drawing on prior knowledge and experience, which also involves more meaning making as compared to simply being told an answer.

In a slightly different conceptualization of constructivism, Nie and Lau (2010) operationalized constructivist learning strategies as consisting of three different main components: emphasis on deep understanding, substantive and elaborated communication, and making connections with real-world situations. Even with a slightly different conceptualization, SQ can still be seen as a constructivist learning method through aligning with each of these three different elements, which further grounds SQ as a constructivist learning method. As SQ endeavors to challenge underlying assumptions and overall, seek truth and clarification, it aligns with the element of cultivating deep understanding. The element of substantive and elaborated communication is quite straightforwardly embodied by SQ. Finally, as stated above, in matching SQ to Hatmanto and Sari’s (2023) conceptualization, the guided discovery and active learning process of SQ involve drawing on prior knowledge and experience, which is a way of connecting to real-world situations. Additionally, as SQ often involves destabilizing commonplaces (Jaeger, 2016) or deconstructing commonly accepted beliefs and assumptions, this can invoke personal reflection, which also brings a connection to the real world.

To understand the effects SQ can have on CT as a constructivist learning method, studies on how constructivist learning methods in general have affected CT can be reviewed. A relationship between constructivist learning and critical thinking has been well-supported (Koreshnikova & Avdeeva, 2022; Kwan & Wong, 2015; Nie & Lau, 2010), with critical thinking being explicitly stated as a goal of constructivist learning in several conceptions of the theory (Koreshnikova & Avdeeva, 2022; Kwan & Wong, 2015).

Kwan and Wong (2015) designed a cross-sectional survey design, involving 967 grade nine students in Hong Kong, to investigate the relationships between students’ perception of constructivist learning environment and their critical thinking ability, also looking at cognitive and motivational variables as potential mediators or ways to explain the relationship. Of note, critical thinking was measured by students taking the Cornell Critical Thinking Test (Level X), which is a well-regarded test of assessing general CT skills, established by well-known CT scholar, Robert Ennis. Results found constructivist learning environment had a direct effect on

CT skills and that cognitive and motivational variables fully mediated this relationship. In other words, it was found that cognitive and motivational variables can fully explain how a constructivist learning environment effects student CT skills. Surprisingly, constructivist teaching via cognitive abilities had an indirect negative effect on CT skills, which is in contradiction to other literature, though this may be explained by the fact that students may have higher usage of cognitive strategies and simply employ them inappropriately or unskillfully. Goal orientations (motivation) positively mediated the relationship, which is in alignment with other literature.

Nie and Lau (2010) similarly conducted an investigation of the relationship of constructivist vs. didactic teaching style on cognitive (surface and deep learning) and motivational variables. Though CT was not directly measured, it was included in the description of 'higher-order' strategies, which characterized 'deep' cognitive processes. In alignment with the literature and to counter what Kwan and Wong (2015) found in their study, constructivist instruction was found to predict deep processing strategies (including critical thinking). Also in alignment with the literature and Kwan and Wong's results, there was a positive relation found between constructivist instruction and student motivation (task value).

Koreshnikova and Avdeeva (2022) add more nuance to the investigation of constructivist teaching style on critical thinking development with motivation as a mediator by grounding the exploration in Self-Determination theory (SDT). SDT is a theory of motivation which claims that motivation is influenced by the satisfaction or frustration of basic psychological needs and recognizes that motivation does not just vary in quantity but also in type (i.e. intrinsic vs. extrinsic) (Ryan & Deci, 2022). Results found, in line with the literature, that intrinsic motivation indeed mediated the relationship between a constructivist style of teaching and critical thinking.

Overall, SQ can be clearly recognized as a constructivist teaching/learning method as it aligns with the core elements of constructivism and constructivist practice is clearly linked to promotion of critical thinking in students. It follows that by understanding Socratic questioning from a constructivist learning theory lens, SQ will likely predict promotion of CT. The proposed study will investigate this prediction and contribute to the gap in the literature of identifying specific constructivist teaching methods because despite being a well-accepted educational theory, it remains a challenge to translate constructivism theory into practical classroom practice (Nie & Lau, 2010).

Socratic Questioning in the Writing Process

Socratic questioning (SQ) is not a common step or intervention in the writing process, possibly because, as other scholars on the topic of CT in writing have pointed out, most English/writing teachers are unsure of how to teach CT in the context of writing and fall prey to the common assumption that teaching good writing skills will invariably inform the development of better thinking (Condon & Kelly-Riley, 2004; Karanja, 2021). Most work published on SQ as a writing intervention is not empirical, but reflective and simply proposes SQ as a possible strategy rather than investigating its effects.

One empirical study on SQ as an intervention determined that it was a beneficial addition to the writing process. Etemadzadeh and colleagues (2013) conducted a quasi-experimental study on 60 Malaysian secondary students and found that a Socratic questioning intervention in the writing process improved writing quality through comparing pre- and post- writing responses. Classroom observations revealed that students displayed more active participation in the experimental group as opposed to the control group where students adhered to ‘traditional’ writing guidelines. However, the exact questions that were used and the criteria upon which essays were graded was not made available, which detracts from understanding how exactly the questioning impacted writing quality and if it was truly the Socratic questioning style which led to an increase in writing quality.

Though not an empirical study, Jaeger (2016) reflected on how writing center tutors are known to employ Socratic questioning as a technique of helping tutees work through the writing process. Jaeger suggested certain amendments to the original form of Socratic questioning as Socrates practiced it to make it a less aggressive approach and more supportive to the student. Suggestions included starting a session by 'charitably rehearsing' the student's argument with them before making any criticism and focusing on the goal of elevating the best version of the student's argument rather than tearing it down, as Socrates was known to do.

Another proposal for SQ to be used in the writing process focused on how teachers can scaffold writing in legal courses to both enhance writing quality and legal analysis, which can be compared to critical thinking in terms of a dedication to properties of reasoning and logic (Beazley & Kearney, 1991). Beazley and Kearney (1991) suggest the Socratic method be used at multiple points in the writing process, both in the form of teachers leaving Socratic questions as feedback on student drafts and in later student-teacher conference, when discussing the writing.

Chatbots: For Socratic Questioning and Developing CT

Several studies have investigated the use of chatbots, albeit not ChatGPT, but earlier iterations of an AI chatbot, on the development of argument quality and CT. One such study was conducted by Guo and colleagues (2023), investigating the effects of an AI chatbot, Argumate, on assisting students in developing arguments for a debate. This study involved 44 college students from a university in China who were enrolled in an academic speaking and writing class taught by one of the researchers. The study consisted of three 90-minute sessions, one which delivered instruction on argumentation and two which engaged in argumentation practice. Students were split into a chatbot-assisted condition or a conventional learning condition. In the chatbot-assisted condition, the chatbot was used pre-debate, in helping brainstorm ideas, as well as during-debate. The chatbot's assistance appeared to help develop argumentation in multiple areas. In terms of argument content, students who used the chatbot were found to use more claims, data, and warrants in their arguments. In terms of argument flow, students who used the chatbot were found to have arguments with improved organization, sufficiency, and elaboration in their arguments. Students were also found to be more engaged and to have exerted more effort in the chatbot task than the conventional learning condition.

Goda and colleagues (2014) tested the effects of a 10-minute Socratic dialogue intervention with a chatbot on participant's later performance in discussions across multiple measures. Participants consisted of 130 undergraduate sophomore students in Japan and were split between the experimental condition (chatbot intervention) and control group ('conventional' pe-discussion preparation such as reflecting on paper and searching on the internet). The study involved two visits following a similar format of a pre-discussion activity (chatbot or conventional) followed by a group discussion, where researchers measured different variables in the different visits. From visit 1 results, it was found that participants in the chatbot condition displayed a significantly higher frequency of contributions to the discussion, as compared to the control. From visit 2, no significant differences in CT were found between the control and experimental groups but notably, the experimental group displayed higher awareness of critical thinking and a more inquiring mindset.

Though not a study empirically testing the use of ChatGPT to enhance CT, Hatmanto and Sari (2023) identify ChatGPT in general as a tool in alignment with Constructivist Learning

Theory, and specifically highlight under this classification, ChatGPT's potential to be used for Socratic dialogue.

The Importance of Human, Student-Teacher Relationships on Motivation and CT

The importance of positive, close student-teacher relationships on improving student learning in general has been well-documented in the literature (Longobardi et al., 2016; Johnson, 2017; Kilday & Ryan, 2019; Rimm-Kaufman & Sandilos, 2010; Skinner & Belmont, 1993). Longobardi and colleagues (2016) investigated student-teacher relationships (STR) as a potential protective factor during the tumultuous transition from middle to high school in 122 Italian eighth graders. Based on data from self-report questionnaires and reported student grade averages collected once in eighth grade and once in ninth grade, they found that student-teacher closeness was linked with higher academic achievement.

Likewise, Kilday and Ryan (2019) examined the impact of quality relationships with peers and teachers on student engagement in school through self-report questionnaires, which were collected from a sample of 761 fifth and sixth grade students in the Midwestern United States. Quality relationships with teachers and peers were found to positively impact student engagement in school, though quality relationships with teachers were found to be more impactful and were specifically important in determining student's emotional engagement or their enjoyment/interest in the class subject.

A way of understanding this positive impact of the student-teacher relationship on student learning is through the lens of Self-Determination Theory (SDT), specifically the mini-theory on basic psychological needs (Ryan & Deci, 2022). Self-determination theory posits that for people to thrive and grow, or indeed learn and engage most effectively, the three basic psychological needs of competence, autonomy, and relatedness must be fulfilled (Ryan & Deci, 2022). SDT holds that satisfying these three psychological needs will increase intrinsic motivation.

Support that SDT explains the link between student-teacher relationships and student engagement can be found in Skinner and Belmont’s (1993) widely cited study, which followed 14 teachers and 144 elementary-aged students (grades 3-5) over a school year, tracking student engagement and the extent to which teachers fulfilled the three basic psychological needs. They found that teacher involvement (addressing the need for relatedness), autonomy support, and structure (addressing the need for competence), all had an impact on student learning and classroom experiences. In particular, involvement (addressing the need for relatedness), was found to be a highly impactful element; if controlled for, it was found that autonomy support no longer had a statistically significant impact on student learning.

In terms of the impacts of student-teacher relationships on critical thinking specifically, a survey study conducted on a sample of 217 sixth-grade students found that student-parent and student-teacher relationship closeness was correlated with student critical thinking, where student-teacher closeness was the strongest predictor (Etemadizadeh et. al., 2022).

As discussed earlier, through review of studies demonstrating the positive relationship between constructivist teaching style and critical thinking, the fact that student motivation mediated this relationship time and again reveals the role of student motivation in predicting student critical thinking (Koreshnikova & Avdeeva, 2022; Kwan & Wong, 2015; Nie & Lau, 2010). Koreshnikova and Avdeeva (2022) also found that intrinsic motivation, specifically, leads to the development of CT (as opposed to extrinsic motivation). If teachers can cultivate trusting relationships and classroom environments which are able to fulfill the three basic psychological needs of autonomy, competence, and relatedness, students will experience more intrinsic motivation, which might then predict more critical thinking.

Proposed Method

Participants

Participants will consist of at least 198 first year college students. This sample size was calculated through a power analysis assuming α = .05, desired power = .8, a 3x3 ANOVA design, and a medium effect size (Cohen, 1992). The medium effect size was assumed for this study based on the medium effect size found in Guo and colleagues (2023)'s study, which has a similar design to the proposed study. Participants will be recruited through physical flyers around the campuses of the Claremont Colleges as well as through digital outreach via social media posts and emails to psychology and education research listservs. The sample will consist of undergraduate college first years in the local California area who are fluent in English and who have no learning disabilities. First year college students are specified because they should all be at a relatively similar stage in terms of writing development, at least as compared to sophomores, juniors, and seniors in college, who have had more time in college courses to develop their writing abilities beyond a high school level. Fluency in English is listed as a requirement to ensure that critical thinking outcomes are not skewed by confusion around interpretation or expression of thoughts in English. Likewise, the sample is restricted in terms of learning disability so that critical thinking outcomes are not skewed as a result of difficulty with reading or writing comprehension and because this study involves specifically timed segments, it will not be able to account for students who would regularly need more time to complete written exams due to learning disability accommodations. This eligibility criteria will be clarified in the study advertisement. Participants of all races, ethnicities, and gender identities will be included and are expected to be within a range of 18 to 20 years old.

Materials

Critical Thinking

Critical thinking will be coded from argumentative essays after visit 1 and visit 2 using Yanning's (2017) rubric for assessing CT from writing (Appendix B). The rubric includes 9 intellectual standards of CT, as based on Paul and Elder's (2001) three-dimensional CT model, which will each be scored from 1 to 5, with 1 representing 'very poor' and 5 representing 'very good.' Each score from 1-5 has descriptive qualifiers in the rubric for what a score of 1 or 3, for example, look like for each intellectual standard. Final CT scores will be calculated by averaging the scores across all 9 intellectual standards for one overall CT score.

Of great importance for scoring CT from writing is that coders not only have a good understanding of the rubric and Paul and Elder's (2001) CT model, which it is based on, but have experience grappling with what critical thinking is, what strong critical thinking can look like in the essay responses, and how to engage in critical thinking themselves around coding the essays (Facione, 2008). Before research assistants can code official essays for the study, they will undergo a training period of writing and reviewing each other's essays (the same ones the participants will later be taking) themselves, then coming to a consensus on scores, which is crucial to not only establish inter-rater reliability, but also for coders to get a sense of the participant experience and exposure to what a range of responses to the selected prompts can look like. To ensure essays of a range of quality level will be available for coding practice, coders will also code and discuss the sample responses to the AP written exams, which have been made available on College Board’s website and include previous responses which were scored poorly, highly, and in-between.

The training process will be informed by Facione’s (2008) recommendations for how to establish inter-rater reliability and deeper understanding of critical thinking as a construct in coders. Facione states this training period as a precondition for coders before they can use his own CT assessment rubric, the Holistic Critical Thinking Scoring Rubric, which was not chosen to be used in this study due to its more limited range of scores (1-4), with lower potential to accurately reflect participant improvement between pre- and post- tests. Nevertheless, Facione’s recommendations for training coders in coding CT from essays can be applied to any CT assessment rubric. Following Facione’s guidance, coders will first read the rubric as a group, also reviewing and discussing the model of CT this study will be following (Paul and Elder’s (2001) model). There will be an emphasis on simply assessing critical thinking from essays and not getting distracted by other characteristics of the essay like grammatical correctness or other persuasive flourishes in writing quality such as rhetorical style, which do not necessarily constitute critical thinking. While recognizing that these elements of writing quality are important to an overall writing product, the focus for this study must be on simply assessing critical thinking elements from the writing. The inspiration for having coders also respond to the essays themselves and discuss the results amongst each other stemmed from the guidelines around coding the International Critical Thinking Essay Test (ICTET), which was not chosen to be used in this study for various reasons, but which described a valuable training period that researchers thought a reasonable addition to building coder’s understanding of critical thinking and inter-rater reliability.

Intrinsic Motivation

Intrinsic motivation will be measured by the ‘Interest/Enjoyment’ subscale (Appendix C) of the Intrinsic Motivation Inventory (IMI), because this subscale is meant to be used for self-reporting on intrinsic motivation, as participants of this study will do. The subscale consists of 7 items, measured on a 7-point Likert scale, where a selection of 1 represents ‘not at all true,’ a selection of 4 represents ‘somewhat true,’ and a selection of 7 represents ‘very true.’ Items include statements for participants to rate such as “I enjoyed doing this activity very much” and “I would describe this activity as very interesting.” Two of the seven items will need to be reverse-coded to calculate the final scoring. This will be done for each of the two items by subtracting that item response from 8 and using the resulting number as the item score. The final subscale score will then be calculated by averaging the scores across all the subscale items.

The proposed study suggested that intrinsic motivation would be influenced by self-determination theory, specifically the extent to which basic psychological needs were satisfied. Notably, this scale was created by Richard Ryan and Edward Deci, the scholars who developed self-determination theory, and was indeed informed by self-determination theory, just as the proposed study wanted to measure (Ostrow & Heffernan, 2018).

This subscale is further validated to be used in the proposed study based on Ostrow and Heffernan’s (2018) study, which tested the validity and reliability of four subscales from IMI, including the ‘Interest/Enjoyment’ subscale, specifically within the AIED field. Using an iterative factor analysis and item reduction approach, Ostrow and Heffernan established substantial convergent validity as well as high reliability for the subscales of ‘Interest/Enjoyment,’ ‘autonomy,’ and ‘competence.’ They reported that results also pointed toward high discriminant validity and high face validity for all four subscales. The IMI has been applied in many contexts and so developers have encouraged researchers using the scales to test for its validity within their specific discipline. Importantly, not only does Ostrow and Heffernan's study establish high validity and reliability of the 'Interest/Enjoyment' subscale, but it specifically does so in an AIED online learning environment, which makes it translatable to the proposed study, involving an online learning condition with ChatGPT.

Written Exams

This study will administer question 1 from the 2019 AP English Language and Composition exam in visit 1 as the written exam and will administer question 1 from the 2017 AP English Language and Composition exam in visit 2 as the written exam. Question 1 of both exams follows a similar format of briefly posing an argumentative writing topic, asking students to choose a side, and asking them to support their argument by using at least 3 of the available sources. More detailed information on these tests can be found on College Board's website, where previous AP English Language and Composition exams are available for public viewing. However, part of the instructions above the prompt will be modified for this study to suggest that in addition to participants dedicating 15 minutes to reading the questions and the sources, they should also take at least 5 minutes to outline their argument, with the argument they will be making and the sources they will be using to support that stance.

Starter Socratic Questions Packet

This packet will be used in the human tutor condition to start Socratic dialogue by letting the participant view a range of Socratic questions based on their argument. There will be 2 pages, representing the two sides of the argument on the Q1 2017 AP English Language and Composition prompt of if libraries are out-dated or still necessary in the digital age. Each page will contain a different list of potential starter questions based on the specific participant argument. These will be generated from researchers interacting with ChatGPT, from assuming both different argument positions and taking the first list of generated questions that ChatGPT produces, as this will be exactly the list of starter questions that participants in the SQ with ChatGPT tutor condition will see.

Reference Sheet of Possible Follow-Up Socratic Questions

This packet will be used in the human tutor condition and is only for the eyes of the research assistant. A reference list of possible Socratic questions will be created for the research assistant who is acting as the human tutor to reference in their dialogue with the participant. These potential follow-up questions will be pulled from interactions between researchers and ChatGPT before the study's commencement, where researchers will act as participants of both possible argument positions, responding to different starter questions, to get a range of possible follow-up questions. These questions will then be categorized by the potential argument position on the reference sheet so that the research assistant can flip to that section, depending on what the participant's argument is, and have an idea of questions they can ask throughout the dialogue or if a lull in the dialogue comes up.

Procedure

The proposed study consists of two visits, spaced a week apart, which each follow a similar format of taking a written exam (up to 1 hour, 20 minutes) followed by a survey. In this quasi-experimental design, participants will be randomly assigned to one of three conditions: control (no Socratic questioning intervention), Socratic questioning with human tutor, or Socratic questioning with ChatGPT tutor. Participants will come individually to the lab to partake in the visit. This was decided on the fact that some participants will be in a condition with a human tutor and if many participants were in the same room, overhearing other conversations, this could provide distraction or even give them ideas from other participants' conversations.

In the first visit, all participants, no matter what condition they were assigned, will follow the same process, without intervention, to establish baseline CT scores. Upon explaining what the visit will entail and gaining informed consent, participants will sit at a desk and take the written exam, which will be question 1 from the AP English Language and Composition exam. Participants have up to 1 hour and 20 minutes to complete the exam. This time frame is based on the time frame required in visit 2, including the SQ intervention, and was applied across conditions to ensure that CT scores did not vary simply because the groups in the intervention had more time to take the test. If participants are done early, they can call out 'done' and a research assistant will enter from the next room with an iPad so that participants can take the post-exam survey.

Depending on the condition they are assigned, participants may take one or two surveys. Those in the control condition and the condition of SQ with a human tutor will simply fill out their demographic information, which will include indicating gender, race/ethnicity, and age. This demographic data will not be used in any data analysis but simply for providing transparency around who the sample consists of, and thus which populations the results can be generalized to later. Those in the SQ with a ChatGPT tutor condition will fill out their demographic information and then an additional short survey, simply asking about email information for ChatGPT account setup. Participants will be asked if they have an email address that they would be willing to share so that researchers might set up a ChatGPT account for them in preparation for the next visit. If participants indicate that their email is already in use for a ChatGPT account or that they are uncomfortable giving their personal email account for the creation of a ChatGPT account, researchers will make an email and a related ChatGPT account for the participant. This step is included so that ChatGPT can already be logged into with the appropriate settings modified when participants come in for visit 2.

Participants assigned to the SQ with a ChatGPT tutor condition will additionally be told to look out for an email after the visit is completed, where they will have been sent a link to an instructional video, designed by the researchers, briefly explaining what ChatGPT is, including its limitations, and modeling how participants will be expected to engage with it in visit 2.

The second visit will be a week later and follow a similar format and time frame as the first, except the post-test survey will be filling out a short scale on intrinsic motivation, and of course the SQ interventions will take place for the participants assigned to those conditions. The control condition will simply follow the same process as the first visit except for filling out the different survey. Both SQ interventions will be 10 minutes long and will happen at the 20-minute mark, from when participants sat down to read through the test. At this point in the exam, participants should have taken 15 minutes to read through the question and sources and 5 minutes to create a rough outline of their essay, including the stance they are taking in the argument and the sources they will be using to back this stance. After the 10-minute SQ intervention in both SQ intervention conditions, the participant will take 5 minutes to summarize their learning from the SQ intervention and integrate it into their outline. The rest of the remaining time will be for finishing the written exam and taking the post-test survey.

In the SQ with a human tutor condition, a research assistant will enter with a clipboard, which two packets- the starter Socratic questions and the reference sheet. Upon entering for the SQ intervention, researchers will ask the participant to see their outline so that the research assistant can determine what page to flip to in both packets. This will not be an abrupt introduction, as the participant and research assistant should already have met in the first visit and in a greeting at the start of visit 2.

Once the participant's argument is determined and the research assistant flips to the correct page in each packet, they will show the correct page of starter Socratic questions and ask the participant to choose one to start the dialogue. The other packet is the reference sheet of possible follow-up Socratic questions, also specified to the participant's argument, that the researcher can reference themselves if they feel unsure of what question to ask next or if there is a lull in the dialogue.

In the SQ with a ChatGPT tutor condition, a research assistant will enter at the 20-minute mark to turn on the computer for the participant. They will remind the participant briefly of the function and limitations of ChatGPT and how the participant is supposed to interact with ChatGPT in SQ, as they should have previously heard from watching the instructional video after the first visit. They will also remind participants that their chat history will be recorded and reviewed to ensure they stuck to the expected SQ mode. The research assistant will then ask if the participant has any clarifying questions about how to interact and answer any questions the participant may have, if not too extensive or beyond the scope of the activity.

There will be a pre-pasted prompt in the chat box on the screen, which the participants will simply have to fill in with their argument. An example of this looks like: "Please engage in Socratic questioning mode with me to help me develop my reasoning on this argumentative essay topic: In the debate on if libraries are still relevant in the digital age, I believe that _____, based on source __, source __, and source ___." Research assistants will already have uploaded the sources (labeled in alphabetical order) to ChatGPT before the participants have come in. This will allow ChatGPT to understand the references participants make when they type “based on source A, source D, and source F,” for example.

Across conditions, after participants reach the 1 hour, 20-minute mark for the written exam or call out ‘done,’ the research assistant will enter with an iPad for participants to fill out the post-exam survey on intrinsic motivation. Across all conditions, at the end of visit 2, participants will also be debriefed on the study and given $50 in compensation, which is meant to cover both visits.

Ethical Considerations

The benefits of this study would far outweigh the risks, where the benefits are many and the risk is minimal. Benefits to participants might include learning how to use ChatGPT as a study tool, specifically in the Socratic Questioning style, developing argumentative writing skills, and developing critical thinking skills. This study will also make contributions to the broader community and society by addressing many gaps in the literature including how to develop and assess CT in writing, specific methods of using ChatGPT to improve student learning, comparing humans and ChatGPT as tutors, and looking at SQ as an intervention in the writing process.

Overall, this study can be considered of minimal risk because participants will only encounter risks that they might be naturally exposed to in their everyday life. The task centers around asking participants to respond to a short argumentative essay prompt, which students will be familiar with and have experienced many times in school by the time they are a first year in college. The participant data collected will simply consist of critical thinking, as coded from written responses, demographics, and intrinsic motivation, and will not include asking for sensitive information (i.e. immigration status or health information). Therefore, in answering surveys and completing the writing task, participants will not likely encounter any topics which cause them an unusual amount of distress or discomfort. The proposed study will not specifically target a protected or vulnerable population for recruitment into the sample. Debriefing the study will be a straightforward affair as the study will not involve deception of any kind nor will consent be ambiguous. This study can be considered truly voluntary as participants will receive detailed informed consent and be directly asked if they consent to participate in the study.

Participant data will be strongly protected. Data will be stored in special folders on lab desktop computers, which will be protected with a password. These desktop computers will stay in the labs, on campus, behind locked doors. Each registered participant will be assigned a numerical identification number to replace their name and protect the confidentiality of participants. The name on the top of each written response will be blacked out and the participant's identification number will be added so that when research assistants read through written responses to assess critical thinking skills, they will not know the name of the participant who wrote the response. Similarly, the survey data will not be connected to a participant's name but with the participant's identification number.

Anticipated Results

Data Analysis

A 3 (mode of Socratic Questioning: no SQ vs. SQ with human vs. SQ with ChatGPT) by 3 (baseline level of critical thinking: lower third of ratings vs. middle third of ratings vs. higher third of ratings) analysis of variance test will be run to assess the effects of the predictor variables 'mode of SQ' and 'baseline level of critical thinking (CT)' on the dependent variable, 'post-test CT scores.' A main effect of each of the predictor variables, 'mode of SQ' and 'baseline level of CT,' will be interpreted after running the test to ensure there is a difference in post-test CT scores between the three groups within each predictor variable, as hypothesized. Then, to further clarify where the differences lie between groups, within each predictor variable, a post hoc test of least significant difference will be conducted. Post hoc analyses will also be conducted to reveal the interaction effects of ‘mode of SQ’ and ‘baseline level of CT’ (split into thirds).

Third splits in ‘baseline level of CT’ will be calculated by dividing the distribution of ratings across the sample into thirds. This division will translate ‘baseline level of CT’ from a continuous to a categorical variable, thus allowing an ANOVA test to be run. Importantly, this means that the third splits do not indicate standard ‘low,’ ‘middle,’ and ‘high’ ratings of critical thinking, where there would be specific numerical bounds to define these, but ‘low*er*,’ ‘middle,’ and ‘high*er*’ ratings, which will instead depend on whatever the final distribution of ratings is.

The dependent variable, post-test CT scores, will be coded from assessing participant essays with Yanning’s (2017) rubric. For CT scores to have any validity, written exams must only be coded after raters have undergone a rigorous training period to build recognition and consensus around scoring CT from essays. After this training period, inter-rater reliability on coding essays using Yanning’s (2017) rubric will be assessed using Kappa statistics, with a cut-off of .80, as widely recognized to be representative of a good level of inter-rater reliability.

Next, a mediational analysis in line with Baron and Kenny’s (1986) version will be run to test for the mediating effects of intrinsic motivation on the relationship between ‘mode of SQ’ and ‘post-test CT scores.’ The first step in the three-step process involves establishing that there is a relationship between the main predictor variable, ‘mode of SQ,’ and the dependent variable, ‘post-test CT scores.’ This will be accomplished by running the 3x3 ANOVA, as described above. The second step involves two sub-steps: establishing that the proposed mediator, intrinsic motivation, is related to the main predictor, ‘mode of SQ,’ and that it is also related to the dependent variable, ‘post-test CT scores.’ To establish the relationship between ‘mode of SQ’ and intrinsic motivation, an independent samples t-test will be run, due to ‘mode of SQ’ being a categorical variable and intrinsic motivation being a continuous variable. To then establish the relationship between intrinsic motivation and ‘post-test CT scores,’ a correlational test can be run, as both variables are continuous. The final step in Baron and Kenny’s mediational analysis process requires a test that evaluates both ‘mode of SQ’ and intrinsic motivation as predictor variables of the dependent variable, ‘post-test CT scores,’ which will reveal how intrinsic motivation explains the relationship between ‘mode of SQ’ and ‘post-test CT scores’; in other words, meaning that the effects of intrinsic motivation on ‘post-test CT scores’ should stay significant if it is a true mediator and one would then expect ‘mode of SQ’ to weaken in its effects on ‘post-test CT scores’ and possibly become entirely non-significant. To test this third step, a multiple regression can be run. As a multiple regression requires both predictor variables to be continuous, ‘mode of SQ,’ which is categorical, must first be translated to a continuous variable by dummy-coding or contrast.

Anticipated Results

This study proposes three hypotheses. First, it is hypothesized that ‘mode of Socratic questioning’ (three levels: no SQ vs. SQ with human tutor vs. SQ with ChatGPT tutor) will affect ‘post-test CT scores’ such that the two conditions with Socratic questioning will show greater post-test CT scores than the control with no SQ, and that participants with human tutors for the SQ intervention will show higher post-test CT scores than those of participants with ChatGPT tutors for the SQ intervention. This hypothesis is supported by the wide recognition of questioning, especially higher-level questioning as important to student engagement and learning

(Elder & Paul, 1998; Erdogan & Campbell, 2008; Etemadzadeh et al., 2013; Mok, 2009). Additionally, SQ is an example of a constructivist learning method, as it aligns with all the components of constructivism. Because employing a constructivist teaching style has been shown to promote CT (Koreshnikova & Avdeeva, 2022; Kwan & Wong, 2015; Nie & Lau, 2010), it follows that SQ, as a constructivist teaching/learning method, would also promote CT.

Additionally, support for the importance of student-teacher relationships to the learning process in general is well-established in educational and psychological literature (Johnson, 2017; Rimm-Kaufman & Sandilos, 2010; Skinner & Belmont, 1993). One of the explanations for the strong impact student-teacher relationships can have on student outcomes comes from Self-Determination Theory, and specifically the mini-theory of basic psychological needs, which states that for people to grow to their full potential and be intrinsically motivated, the three basic psychological needs of competence, autonomy, and relatedness must be satisfied (Ryan & Deci, 2022). Skinner and Belmont (1993) verified the importance of teachers fulfilling these needs to positively impact student learning experiences. While ChatGPT tutors might do very well at addressing the needs for competence and autonomy support, human tutors are likely better able to fulfill the need of relatedness. It follows from the basic psychological needs mini-theory then that human teachers would be more likely to promote intrinsic motivation in students, which can promote higher student engagement in general and more critical thinking in students. The 3x3 analysis of variance test and subsequent post hoc tests are expected to find results consistent with this first hypothesis.

Second, it is hypothesized that intrinsic motivation will mediate the relationship between ‘mode of SQ’ and ‘post-test CT scores,’ where intrinsic motivation will be higher for participants engaging in SQ with a human tutor than with a ChatGPT tutor and these higher levels of motivation will predict higher post-test CT scores. In other words, Socratic questioning, especially with a human tutor, will promote intrinsic motivation, which in turn, will promote critical thinking. Recognizing SQ as a constructivist teaching/learning method provides direct, substantial support for the idea that such a mediational relationship will exist because many previous studies have demonstrated that motivation mediates the relationship between constructivist teaching/learning methods and CT (Koreshnikova & Avdeeva, 2022; Kwan & Wong, 2015; Nie & Lau, 2010). As was discussed in the justification for hypothesis 1, intrinsic motivation is likely to be promoted not only by engaging in a constructivist teaching/learning method, but to be further promoted by engaging with a human tutor (vs. a ChatGPT tutor). This is because human tutors are more likely able to fulfill all three basic psychological needs of students, which, according to SDT (Ryan & Deci, 2022), makes it more likely for them to promote intrinsic motivation in students. Based on Koreshnikova and Avdeeva's (2022) finding that it is specifically intrinsic motivation, as compared to extrinsic motivation, which can lead to CT development, it follows that if human tutors can promote more intrinsic motivation in students, they will likely also promote more CT development. The mediational analysis will serve to evaluate this second hypothesis, where results are expected to verify the hypothesis.

Third, it is hypothesized that baseline CT scores will moderate the relationship between mode of SQ and post-test CT scores such that participants with lower baseline CT scores will see greater impacts from the SQ intervention (greater differences between the control and SQ intervention post-test CT scores). This relationship is anticipated due to the ceiling effect, where scales may not be able to detect any improvement from participants who enter the study with an already high level of CT ability, as they are already at or close to the maximum scores possible on the scale. In other words, the intervention will not show as strong of an effect for people who already perform at a high level in CT, whereas there is more potential to see improvement from a CT-targeted intervention in people who enter with a lower level of CT ability. This might be understood through acknowledging that the intervention may present an exercise of thinking that is redundant or known already by the people who entered with high levels of CT, and on the other hand, the intervention may present new ways of thinking and opportunities for developing or enhancing new skills and dispositions in CT for those who entered with lower level of CT. The 3x3 analysis of variance test and subsequent post hoc tests will serve to evaluate this third hypothesis along with the first hypothesis, where results are expected to verify the third hypothesis.

Scholarly Merits and Broader Impacts

Scholarly Merits

The proposed study will function to address many gaps in the literature. As ChatGPT was so recently released, there have been hardly any empirical investigations in the psychological or educational literature yet about student engagement with ChatGPT, nor hardly any investigations into specific methods of interaction with ChatGPT which might support student learning. Empirical exploration in these areas is a pressing concern as students and teachers are struggling right now with figuring out how to interact or not interact with the chatbot on their own. Students and teachers will be best served by having evidence-based recommendations to inform classroom expectations and guidelines around the use of ChatGPT, of which there are hardly any available now. This study will contribute to addressing this practical need and gap in the literature.

Naturally, as a result of the lack of studies around ChatGPT in general, there have also yet to be any studies conducted on the effects that human vs. ChatGPT tutors have on student outcomes. The proposed study will be the first known study to contribute to filling this gap.

The proposed study will also work to address gaps in the literature around SQ as a specific constructivist teaching/learning method and as an intervention in the writing process to improve CT, where there have been almost no studies conducted on either of these topics. Additionally, this study will contribute to the literature around developing and assessing CT from writing in general, which is an area of the literature clouded with different definitions, vague assessments, and confusion.

Broader Impacts

At a societal level, results from this study can inform ChatGPT implementation in classrooms and workplaces across the United States and beyond. This study can provide some of the first clear, evidence-backed guidance on how exactly one can make use of ChatGPT in a beneficial way, specifically for improving critical thinking in writing. Additionally, ChatGPT represents an amazing potential to be a free or relatively cheap technological tool, which is easily accessible to anyone with a device and internet connection. In this way, it can also serve to help close the achievement gap in schooling.

A reason the study specifically looked at baseline CT as a moderator was to potentially show how a SQ intervention could be especially impactful to students with lower CT abilities. While a human tutor would likely be best, a ChatGPT tutor is expected to be a close second in terms of improving CT with a SQ intervention. Having the option for this outside-of-school guidance and practice in developing critical thinking can allow low-income students who might not have had the opportunity to develop critical thinking in their regular schooling to catch up with their high-income peers who have likely had more opportunities to develop CT. Low-income students are also less likely to have access to help at home either in the form of a tutor or an adult who can assist them with homework. While this study does not recommend that

ChatGPT replace human tutors, it does point to how ChatGPT can prove to be a beneficial tutor if used in the right ways and certainly a more accessible tutor than most human tutors.

It must be stated clearly that there is no intention in this investigation to encourage more rapid development and adoption of chatbots and AI in general. On the contrary, by highlighting the importance of conducting empirical tests on specific methods of using ChatGPT, this study is intended to be part of a message about paving the way for responsible AI adoption/integration, which tends to be in alignment with slower development/deployment of AI. This is important to state because the exponential rate at which AI is currently developing is beyond the comprehension of researchers and contains potential catastrophic consequences (Harris & Raskin, 2019-present). Governments are still coming to consensus on what regulation for safe use of AI tools looks like but have acknowledged the importance of proceeding in AI development with caution and care, as evidenced by President Biden’s recent executive order on AI development.

References

Aliakbari, M., & Sadeghdaghighi, A. (2012). Teachers' perception of the barriers to critical thinking. *Procedia – Social and Behavioral Sciences, 70,* 1-5. https://doi.org/10.1016/j.sbspro.2013.01.031

Anderson, G., & Piro, J. (2014). Conversations in Socrates café: Scaffolding critical thinking via Socratic questioning and dialogues. New Horizons for Learning (NHFL), 11(1), 1-9.

Beazley, M. B., & Kearney, M. K. (1991). Teaching students how to 'think like lawyers': Integrating Socratic method with the writing process. *Temple Law Review, 64*(4). 885-908.

Borji, A. (2023). *A categorical archive of ChatGPT failures*. arXiv. https://doi.org/10.48550/arXiv.2302.03494

Campbell, S. H. (2023). *What is human about writing?: Writing process theory and ChatGPT*. ResearchSquare. https://doi.org/10.21203/rs.3.rs-3208562/v1

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Condon, W., & Kelly-Riley, D. (2004). Assessing and teaching what we value: The relationship between college-level writing and critical thinking abilities. *Assessing Writing, 9*(1), 56-75. https://doi.org/10.1016/J.ASW.2004.01.003

Davies, W. M. (2006). An 'infusion' approach to critical thinking: Moore on the critical thinking debate. *Higher Education Research & Development, 25*(2), 179-193. https://doi.org/10.1080/07294360600610420

Delpit, L. (2012). Picking up the broom: Demanding critical thinking, '*Multiplication is for white people': Raising expectations for other people's children*. The New Press.

Elder, L., & Paul, R. (1998). The role of Socratic questioning in thinking, teaching, and learning.

The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 71(5), 297-301. https://doi.org/10.1080/00098659809602729

Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice, 32*(3), 179–186.

Erdogan, I., & Campbell, T. (2008). Teacher questioning and interaction patterns in classrooms facilitated with differing levels of constructivist teaching practices. International Journal of Science Education, 30(14), 1891-1914. https://doi.org/10.1080/09500690701587028

Etemadizadeh, H., Mohamadi, H., & Ariapooran, S. (2022). Correlation of components of child-parent and student-teacher relationships with tendency to critical thinking in sixth-grade students. *Quarterly Journal of New Thoughts on Education, 18*(3), 187-202. https://doi.org/10.22051/JONTOE.2022.31404.3046

Etemadzadeh, A., Seifi, S., & Roohbakhsh Far, H. (2013). The role of questioning technique in developing thinking skills: The ongoing effect on writing skill. *Procedia - Social and Behavioral Sciences, 70*, 1024-1031. https://doi.org/10.1016/j.sbspro.2013.01.154

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction – The Delphi report* (ED315423). ERIC. https://eric.ed.gov/?id=ED315423

Facione, P. A. (2008) Using the holistic critical thinking scoring rubric to train the discovery of evidence of critical thinking. In Facione, N. C., & Facione, P. A. (Eds.), *Critical thinking and clinical reasoning in the health sciences: An international multidisciplinary teaching anthology*. California Academic Press.

Freire, P. (2005). *Pedagogy of the oppressed*. The Continuum International Publishing Group Inc.

Fryer, L. K. (2023). *A psychological platform for chatbot and human co-piloting in education*.

PsyArXiv. https://doi.org/10.31234/osf.io/dqxp4

Goda, Y., Yamada, M., Matsukawa, H., Hata, K., Yasunami, S. (2014). Conversation with a chatbot before an online EFL group discussion and he effects on critical thinking. Information and Systems in Education, 13(1), 1-7. https://doi.org/10.12937/EJSISE.13.1

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education sciences, 13*(7). https://doi.org/10.3390/educsci13070692

Guo, K., Zhong, Y., Li, D., Chu, S. K. W. (2023). Effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation. Computers & Education, 203. https://doi.org/10.1016/j.compedu.2023.104862

Harris, T., & Raskin, A. (Hosts). (2019-present). *The AI Dilemma* [Audio podcast]. Your Undivided Attention. https://open.spotify.com/show/4KI3PtZaWJbAWK89vgttoU?si=451ad7d4d7424d6c

Harunasari, S. Y. (2023). Examining the effectiveness of AI-integrated approach in EFL writing: A case of ChatGPT. International Journal of Progressive Sciences and Technologies (IJPSAT), 39(2), 357-368. https://doi.org/10.52155

Hatmanto, E. D. & Sari, M. I. (2023, November). Aligning theory and practice: Leveraging ChatGPT for effective English language teaching and learning. *International Conference on Environment and Smart Society (ICEnSO 2023), 440*. https://doi.org/10.1051/e3sconf/202344005001

Holmes, J. (2015). The case for teaching ignorance. *The New York Times*. http://nyti.ms/1EepO30

Jaeger, G. (2016). (Re)examining the Socratic method: A lesson in tutoring. *Praxis: A Writing*

Center Journal, 13(2), 14-20.

Johnson, D. (2017). The role of teachers in motivating students to learn. *BU Journal of Graduate Studies in Education*, 9(1), 46-49.

Karanja, L. (2021). Teaching critical thinking in a college-level writing course: A critical reflection. *International Online Journal of Education and Teaching (IOJET), 8*(1). 229-249.

Kilday, J. E., & Ryan, A. M. (2019). Personal and collective perceptions of social support: Implications for classroom engagement in early adolescence. *Contemporary Educational Psychology*, *58*, 163–174. https://doi.org/10.1016/j.cedpsych.2019.03.006

Koreshnikova, Y. N., & Avdeeva, E. A. (2022). Interest cannot be forced. The role of academic motivation and teaching styles in the development of students' critical thinking. *Voprosy Obrazovaniya / Educational Studies Moscow*, *3*. https://doi.org/10.17323/1814-9545-2022-3-36-66.

Kwan, Y. W. & Wong, A. F. L. (2015). Effects of the constructivist learning environment on students' critical thinking ability: Cognitive and motivational variables as mediators. *International Journal of Educational Research*, *70*, 68-79. https://doi.org/10.1016/J.IJER.2015.02.006

Lombardi, D. (2023). On the horizon: The promise and power of higher order, critical, and critical analytical thinking. *Educational Psychology Review, 35*(2), 37-40. https://doi.org/10.1007/s10648-023-09763-z

Longobardi, C., Prino, L. E., Marengo, D., & Settanni, M. (2016). Student-teacher relationships as a protective factor for school adjustment during the transition from middle to high school. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01988

Malik, A. (2023, November 6). OpenAI's ChatGPT now has 100 million weekly active users. *Tech Crunch*. https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/

Mok, J. (2009). From policies to realities: Developing students' critical thinking in Hong Kong secondary school English writing classes. RELC Journal, 40(3), 262-279. https://doi.org/10.1177/0033688209343866

Nie, Y. & Lau, S. (2010). Differential relations of constructivist and didactic instruction to students' cognition, motivation, and achievement. *Learning and Instruction, 20*(5), 411-423. https://doi.org/10.1016/j.learninstruc.2009.04.002

OpenAI (2023, October 19). Dall-E is now available in ChatGPT Plus and Enterprise. *OpenAI*. https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise

Ostrow, K. S., & Heffernan, N. T. (2018). Testing the validity and reliability of intrinsic motivation inventory subscales within ASSISTments. In Rosé, C. P., Martínez-Maldonado, R., Hoppe, H. U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., & Boulay, B. D. (Eds.), *Artificial Intelligence in Education* (pp. 381-394). Springer, Cham. https://doi.org/10.1007/978-3-319-93843-1_28

Padesky, C. A. (1993, September 24). *Socratic questioning: Changing minds or guiding discovery?* [Keynote address]. European Congress of Behavioral and Cognitive Therapies, London, England.

Paul, R., Elder, L., & Bartell, T. (1997). *California teacher preparation for instruction in critical thinking: Research findings and policy recommendations*. Sacramento, CA: California Commission on Teacher Credentialing. Retrieved November 20, 2023 from https://www.criticalthinking.org/store/products/california-teacher-preparation-forinstruction-in-critical-thinking/147

Paul, R., & Elder, L. (2007). Critical thinking: The art of Socratic questioning. *Journal of Developmental Education*, *31*(1), 36-37.

Piro, J., & Anderson, G. (2015). Discussions in a Socrates Café: Implications for Critical Thinking in Teacher Education. Action in Teacher Education, 37(3), 265-283. https://doi.org/10.1080/01626620.2015.1048009

Rimm-Kaufman, S., & Sandilos, L. (2010). Improving Students' Relationships with Teachers to Provide Essential Supports for Learning. *APA*. https://www.apa.org/education-career/k12/relationships#:~:text=Teacher%2Dstudent%20relationships%20contribute%20to%20students'%20resiliency.,important%20as%20having%20positive%20relationships

Ryan, R. M., & Deci, E. L. (2022). Self-determination theory. *Encyclopedia of Quality of Life and Well-Being Research*, 1-7. https://doi.org/10.1007/978-3-319-69909-7_2630-2

Santos, R. P. D. (2023). *Enhancing chemistry learning with ChatGPT and Bing chat as agents-to-think-with: A comparative case study*. ArXiv. https://doi.org/10.48550/arXiv.2311.00709

Sahamid, H. (2016). Developing critical thinking through Socratic questioning: An action research study. *International Journal of Education & Literacy Studies*, *4*(3), 62-72. https://doi.org/10.7575/aiac.ijels.v.4n.3p.62

Schade, M. (2023). How ChatGPT and Our Language Models Are Developed. *OpenAI*. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

Shen, C. (2018). Does school-related internet information seeking improve academic self-efficacy? The moderating role of internet information seeking styles. *Computers in Human Behavior, 86,* 91-98. https://doi.org/10.1016/j.chb.2018.04.035

Singh, S., & Ramakrishnan, N. (2023). *Is ChatGPT biased? A review*. https://doi.org/10.31219/osf.io/9xkbu

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571-581.

Sobo, E. J. (2023). Could ChatGPT prompt a new golden age in higher education? *Teaching and Learning Anthropology Journal*, *6*(1), https://doi.org/10.5070/T36160114

Walker, S. E. (2003). Active learning strategies to promote critical thinking. *Journal of Athletic Training, 38*(3), 263-267.

Westheimer, J. (2008). No child left thinking: Democracy at-risk in American schools. Education and Politics, 3(2), 10-15.

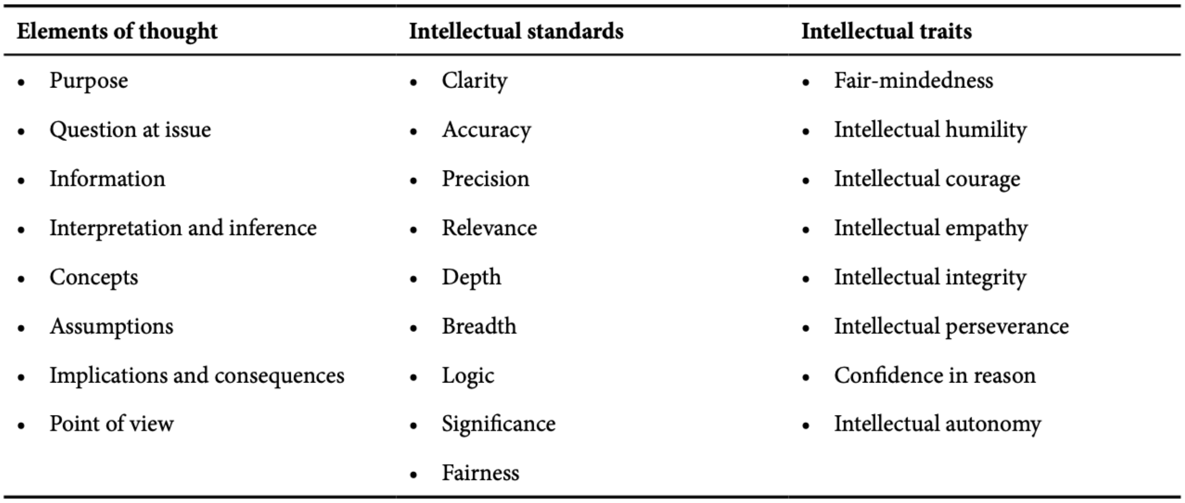
Yanning, D. (2017) Teaching and assessing critical thinking in second language writing: An infusion approach. *Chinese Journal of Applied Linguistics, 40*(4), 431-451. https://doi.org/10.1515/cjal-2017-0025

Zhang, M. (2023, April 21). Users unleash “grandma jailbreak” on ChatGPT. *Artisana*. https://www.artisana.ai/articles/users-unleash-grandma-jailbreak-on-chatgpt

Appendices

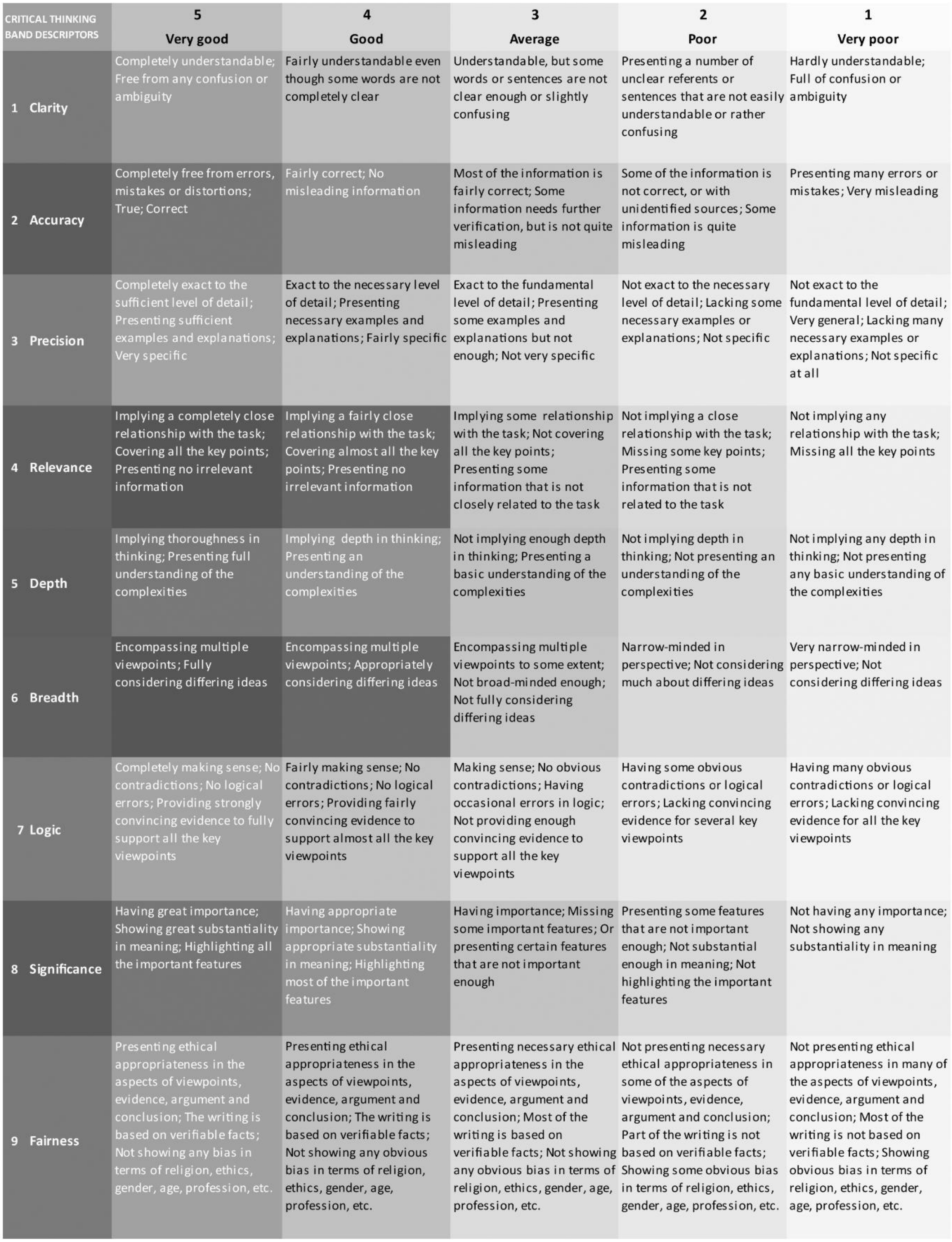
Appendix A

Table 1. Paul and Elder's (2001) CT Model



Elements of thought	Intellectual standards	Intellectual traits
• Purpose	• Clarity	• Fair-mindedness
• Question at issue	• Accuracy	• Intellectual humility
• Information	• Precision	• Intellectual courage
• Interpretation and inference	• Relevance	• Intellectual empathy
• Concepts	• Depth	• Intellectual integrity
• Assumptions	• Breadth	• Intellectual perseverance
• Implications and consequences	• Logic	• Confidence in reason
• Point of view	• Significance	• Intellectual autonomy
	• Fairness	

Appendix B – Yanning's (2017) rubric for assessing CT from writing



CRITICAL THINKING BAND DESCRIPTORS	5 Very good	4 Good	3 Average	2 Poor	1 Very poor
1 Clarity	Completely understandable; Free from any confusion or ambiguity	Fairly understandable even though some words are not completely clear	Understandable, but some words or sentences are not clear enough or slightly confusing	Presenting a number of unclear referents or sentences that are not easily understandable or rather confusing	Hardly understandable; Full of confusion or ambiguity
2 Accuracy	Completely free from errors, mistakes or distortions; True; Correct	Fairly correct; No misleading information	Most of the information is fairly correct; Some information needs further verification, but is not quite misleading	Some of the information is not correct, or with unidentified sources; Some information is quite misleading	Presenting many errors or mistakes; Very misleading
3 Precision	Completely exact to the sufficient level of detail; Presenting sufficient examples and explanations; Very specific	Exact to the necessary level of detail; Presenting necessary examples and explanations; Fairly specific	Exact to the fundamental level of detail; Presenting some examples and explanations but not enough; Not very specific	Not exact to the necessary level of detail; Lacking some necessary examples or explanations; Not specific	Not exact to the fundamental level of detail; Very general; Lacking many necessary examples or explanations; Not specific at all
4 Relevance	Implying a completely close relationship with the task; Covering all the key points; Presenting no irrelevant information	Implying a fairly close relationship with the task; Covering almost all the key points; Presenting no irrelevant information	Implying some relationship with the task; Not covering all the key points; Presenting some information that is not closely related to the task	Not implying a close relationship with the task; Missing some key points; Presenting some information that is not related to the task	Not implying any relationship with the task; Missing all the key points
5 Depth	Implying thoroughness in thinking; Presenting full understanding of the complexities	Implying depth in thinking; Presenting an understanding of the complexities	Not implying enough depth in thinking; Presenting a basic understanding of the complexities	Not implying depth in thinking; Not presenting an understanding of the complexities	Not implying any depth in thinking; Not presenting any basic understanding of the complexities
6 Breadth	Encompassing multiple viewpoints; Fully considering differing ideas	Encompassing multiple viewpoints; Appropriately considering differing ideas	Encompassing multiple viewpoints to some extent; Not broad-minded enough; Not fully considering differing ideas	Narrow-minded in perspective; Not considering much about differing ideas	Very narrow-minded in perspective; Not considering differing ideas
7 Logic	Completely making sense; No contradictions; No logical errors; Providing strongly convincing evidence to fully support all the key viewpoints	Fairly making sense; No contradictions; No logical errors; Providing fairly convincing evidence to support almost all the key viewpoints	Making sense; No obvious contradictions; Having occasional errors in logic; Not providing enough convincing evidence to support all the key viewpoints	Having some obvious contradictions or logical errors; Lacking convincing evidence for several key viewpoints	Having many obvious contradictions or logical errors; Lacking convincing evidence for all the key viewpoints
8 Significance	Having great importance; Showing great substantiality in meaning; Highlighting all the important features	Having appropriate importance; Showing appropriate substantiality in meaning; Highlighting most of the important features	Having importance; Missing some important features; Or presenting certain features that are not important enough	Presenting some features that are not important enough; Not substantial enough in meaning; Not highlighting the important features	Not having any importance; Not showing any substantiality in meaning
9 Fairness	Presenting ethical appropriateness in the aspects of viewpoints, evidence, argument and conclusion; The writing is based on verifiable facts; Not showing any bias in terms of religion, ethics, gender, age, profession, etc.	Presenting ethical appropriateness in the aspects of viewpoints, evidence, argument and conclusion; The writing is based on verifiable facts; Not showing any obvious bias in terms of religion, ethics, gender, age, profession, etc.	Presenting necessary ethical appropriateness in the aspects of viewpoints, evidence, argument and conclusion; Most of the writing is based on verifiable facts; Not showing any obvious bias in terms of religion, ethics, gender, age, profession, etc.	Not presenting necessary ethical appropriateness in some of the aspects of viewpoints, evidence, argument and conclusion; Part of the writing is not based on verifiable facts; Showing some obvious bias in terms of religion, ethics, gender, age, profession, etc.	Not presenting ethical appropriateness in many of the aspects of viewpoints, evidence, argument and conclusion; Most of the writing is not based on verifiable facts; Showing obvious bias in terms of religion, ethics, gender, age, profession, etc.

Appendix C

For each of the following statements, please indicate how true it is for you, using the following scale:

1	2	3	4	5	6	7
Not at all true			somewhat true			very true

Interest/Enjoyment

I enjoyed doing this activity very much
This activity was fun to do.
I thought this was a boring activity. (R)
This activity did not hold my attention at all. (R)
I would describe this activity as very interesting.
I thought this activity was quite enjoyable.
While I was doing this activity, I was thinking about how much I enjoyed it.