Università di Foggia

Università degli Studi di Foggia
Dipartimento di Studi Umanistici
PhD course in
"Culture, Communication, and Education"
Curriculum: Educational communication (XXXIII)

PhD thesis in
"Experimental Education"

Title
Defining and Assessing Critical Thinking: toward an automatic analysis of HiEd students' written texts

Supervisor
Professor Antonella Poce

PhD Candidate
Francesca Amenduni

PhD Coordinator
Professor Lorenzo Cantatore

A.A. 2019/2020

Abstract ENG

The main goal of this PhD thesis is to test, through two empirical studies, the reliability of a method aimed at automatically assessing Critical Thinking (CT) manifestations in Higher Education students' written texts. The empirical studies were based on a critical review aimed at proposing a new classification for systematising different CT definitions and their related theoretical approaches. The review also investigates the relationship between the different adopted CT definitions and CT assessment methods. The review highlights the need to focus on open-ended measures for CT assessment and to develop automatic tools based on Natural Language Processing (NLP) technique to overcome current limitations of open-ended measures, such as reliability and costs. Based on a rubric developed and implemented by the Center for Museum Studies – Roma Tre University (CDM) research group for the evaluation and analysis of CT levels within open-ended answers (Poce, 2017), a NLP prototype for the automatic measurement of CT indicators was designed. The first empirical study was carried out on a group of 66 university teachers. The study showed satisfactory reliability levels of the CT evaluation rubric, while the evaluation carried out by the prototype was not yet sufficiently reliable. The results were used to understand how and under what conditions the model works better. The second empirical investigation was aimed at understanding which NLP features are more associated with six CT sub-dimensions as assessed by human raters in essays written in the Italian language. The study used a corpus of 103 students' pre-post essays who attended a Master's Degree module in "Experimental Education and School Assessment" to assess students' CT levels. Within the module, we proposed two activities to stimulate students' CT: Open Educational Resources (OERs) assessment (mandatory and online) and OERs design (optional and blended). The essays were assessed both by expert evaluators, considering six CT sub-dimensions, and by an algorithm that automatically calculates different kinds of NLP features. The study shows a positive internal reliability and a medium to high inter-coder agreement in expert evaluation. Students' CT levels improved significantly in the post-test. Three NLP indicators significantly correlate with CT total score: the Corpus Length, the Syntax Complexity, and an adapted measure of Term Frequency-Inverse Document Frequency. The results collected during this PhD have both theoretical and practical implications for CT research and assessment. From a theoretical perspective, this thesis shows unexplored similarities among different CT traditions, perspectives, and study methods. These similarities could be exploited to open up an interdisciplinary dialogue among experts and build up a shared understanding of CT. Automatic assessment methods can enhance the use of open-ended measures for CT assessment, especially in online teaching. Indeed, they can support teachers and researchers to deal with the growing presence of linguistic data produced within educational

platforms. To this end, it is pivotal to develop automatic methods for the evaluation of large amounts of data which would be impossible to analyse manually, providing teachers and evaluators with support for monitoring and evaluating the skills demonstrated online by students.

Key words: Critical Thinking; assessment; open-ended measures; Natural Language Processing; Higher Education

Abstract ITA

L'obiettivo principale di questa tesi di dottorato è testare, attraverso due studi empirici, l'affidabilità di un metodo volto a valutare automaticamente le manifestazioni del Pensiero Critico (CT) nei testi scritti da studenti universitari. Gli studi empirici si sono basati su una review critica della letteratura volta a proporre una nuova classificazione per sistematizzare le diverse definizioni di CT e i relativi approcci teorici. La review esamina anche la relazione tra le diverse definizioni di CT e i relativi metodi di valutazione. Dai risultati emerge la necessità di concentrarsi su misure aperte per la valutazione del CT e di sviluppare strumenti automatici basati su tecniche di elaborazione del linguaggio naturale (NLP) per superare i limiti attuali delle misure aperte, come l'attendibilità e i costi di scoring.

Sulla base di una rubrica sviluppata e implementata dal gruppo di ricerca del Centro di Didattica Museale – Università di Roma Tre (CDM) per la valutazione e l'analisi dei livelli di CT all'interno di risposte aperte (Poce, 2017), è stato progettato un prototipo per la misurazione automatica di alcuni indicatori di CT. Il primo studio empirico condotto su un gruppo di 66 docenti universitari mostra livelli di affidabilità soddisfacenti della rubrica di valutazione, mentre la valutazione effettuata dal prototipo non era sufficientemente attendibile. I risultati di questa sperimentazione sono stati utilizzati per capire come e in quali condizioni il modello funziona meglio. La seconda indagine empirica era volta a capire quali indicatori del linguaggio naturale sono maggiormente associati a sei sotto-dimensioni del CT, valutate da esperti in saggi scritti in lingua italiana. Lo studio ha utilizzato un corpus di 103 saggi pre-post di studenti universitari di laurea magistrale che hanno frequentato il corso di "Pedagogia sperimentale e valutazione scolastica". All'interno del corso, sono state proposte due attività per stimolare il CT degli studenti: la valutazione delle risorse educative aperte (OER) (obbligatoria e online) e la progettazione delle OER (facoltativa e in modalità blended). I saggi sono stati valutati sia da valutatori esperti, considerando sei sotto-dimensioni del CT, sia da un algoritmo che misura automaticamente diversi tipi di indicatori del linguaggio naturale. Abbiamo riscontrato un'affidabilità interna positiva e un accordo tra valutatori medio-alto. I livelli di CT degli studenti sono migliorati in modo significativo nel post-test. Tre indicatori del linguaggio naturale sono

correlati in modo significativo con il punteggio totale di CT: la lunghezza del corpus, la complessità della sintassi e la funzione di peso tf-idf (term frequency–inverse document frequency). I risultati raccolti durante questo dottorato hanno implicazioni sia teoriche che pratiche per la ricerca e la valutazione del CT. Da un punto di vista teorico, questa tesi mostra sovrapposizioni inesplorate tra diverse tradizioni, prospettive e metodi di studio del CT. Questi punti di contatto potrebbero costituire la base per un approccio interdisciplinare e la costruzione di una comprensione condivisa di CT.

I metodi di valutazione automatica possono supportare l'uso di misure aperte per la valutazione del CT, specialmente nell'insegnamento online. Possono infatti facilitare i docenti e i ricercatori nell'affrontare la crescente presenza di dati linguistici prodotti all'interno di piattaforme educative (es. Learning Management Systems). A tal fine, è fondamentale sviluppare metodi automatici per la valutazione di grandi quantità di dati che sarebbe impossibile analizzare manualmente, fornendo agli insegnanti e ai valutatori un supporto per il monitoraggio e la valutazione delle competenze dimostrate online dagli studenti.

Parole chiave: pensiero critico; valutazione; misure aperta; processamento del linguaggio naturale; università;

Index of the figures

Index of the tables

INTRODUCTION

"Our most unquestioned convictions may be as mistaken as those of Galileo's opponents"

Russell, 1997

In June 2019, I participated in the *39th annual international conference on Critical Thinking* in Leuven, carried out by the *Foundation for Critical Thinking*. I was at the second year of my PhD thesis, stuck in finding the right path for my research on Critical Thinking assessment. At that conference, I met people from all over the world and with a different background. I had the opportunity to speak with many participants about the reasons why they were at that conference. One of the most surprising answers I received was the following: "I decided to participate at the conference when I realised that something was wrong in my way of thinking".

I spent some days reflecting upon that simple answer and its relation with the challenges I was facing in my thesis. Writing about Critical Thinking is an unusual experience: in the best case, you are writing about something that you are using for structuring your writing. In this process, you become aware of the defeats of your thoughts, both in personal and professional life. Acquiring this awareness can be a relief. It disposes you to face a path toward knowledge construction, characterised by unavoidable defeats in your thoughts. As explained by Julia Galef in a TEDTalks[1], negative emotions may bound discovery: "We need to learn how to feel proud instead of ashamed when we notice we might have been wrong about something. We need to learn how to feel intrigued instead of defensive". Changing my mindset has helped me to go on in my research path and to understand the personal relevance of the topic I was dealing with.

Critical Thinking (CT) is an essential driver for progress and knowledge growth in any field and the broad society. The World Economic Forum, in their most recent report "Future of Jobs" (2018), identifies CT as one of the top 5 most important, in-demand job skills for the current and future economy. UNESCO (2017) includes CT as one of the eight key competencies for achieving the Sustainable Development Goals. In a recent review, Cunningham and Villaseñor (2016) found a greater demand for socio-emotional skills and higher-order cognitive skills, included CT, than for basic cognitive or technical skills by employers.

CT is not only relevant for educational and professional reasons, but also to become active and responsible citizens. In the current scenario, the role of CT in orientating behaviours is more evident

---

[1] https://www.youtube.com/watch?v=w4RLfVxTGH4&feature=emb_title

than ever. The Coronavirus disease (COVID-19) is the first pandemic in history in which technology and social media are being used on a massive scale to keep people safe and informed. At the same time, the technology is amplifying *information overload*, also defined as *infodemic*, that undermines the global response and jeopardizes measures to control the pandemic. As reported by the World Health Organization (2020), misinformation could cost lives. Reasoning skills, such as CT, play a pivotal role: our ideas, thoughts, and beliefs affect our behaviours, and our actions could reduce or increase risks related to personal and collective health.

Despite its acknowledged relevance, CT is still a disputed concept with several different definitions and operationalisations that come from many approaches. As a consequence, it is difficult to study and evaluate it through an empirical perspective. Some authors argue that the main limitation in the CT empirical research is the lack of systematic design of instructional interventions (Tiruneh, Verburgh, & Elen, 2014). From a recent review carried out at a European level, authors found that most of the research and practices presented qualitative assessment methods, based mainly on students' and teachers' perceptions, and a few adopted formal CT tests, rubrics, or research designs with an experimental/quasi-experimental nature (Dumitru, Bigu, Elen, Ahern, McNally, & O'Sullivan, 2018). The authors also found that teachers reported several difficulties concerning the assessment of students' CT progression. Those difficulties could have been exacerbated in the current scenario, characterised by an unprecedented shift from traditional face to face teaching and learning to online technology-enhanced learning. The transformation in the educational environment also entails a change in assessment procedures and methods (Khan, & Jawaid, 2020).

Online educational environments offer many opportunities to both support and assess students' CT (Garrison, Anderson, & Archer, 2001a, b; Kuhn & Crowell, 2011). However, teachers can fully take advantage of the opportunities provided by the digital technologies for CT enhancement and assessment only in specific conditions. Firstly, teachers need to work with a clear and valid CT framework to design and assess learning activities. Secondly, systems able to detect students' CT manifestations in online environments (e.g. discussion forums post) would facilitate teachers in monitoring and assessing students learning. In this scenario, research of the last years is trying to develop valid and reliable tools based on Natural Language Processing (NLP), for the automatic assessment of CT in students' written texts, such as constructed response answers, essays, forums posts.

Starting from these assumptions, the main goal of this PhD thesis is to test the reliability of a method aimed at automatically assessing CT manifestations in Higher Education students' written texts. To achieve this goal, I decided to devote a space for the reflection upon what CT is and how to define it. This because "the conceptualization and assessment of CT are interdependent issues that must be discussed together: how CT is defined determines how it is best measured" (Ku, 2009; p.71).

The present thesis is organised into four chapters. The first two chapters aim at reviewing the state of the art of CT definitions and assessment.

In Chapter 1, I present a critical review aimed at proposing a new classification for systematising different CT definitions and their related theoretical approaches. Moreover, the review investigates the relationship between the different adopted CT definitions and CT assessment methods. I carried out a qualitative content analysis of 39 CT definitions to develop a new definitions' classification. I developed a grid for the qualitative content analysis based on the most highlighted features in the literature related to CT definitions. Eleven theory-driven categories have been identified which were clustered in six macro-categories (1) Individual dimension (e.g. Facione, 1990); (2) Inter-individual dimension (e.g. McPeck, 1981; Kuhn, 2019); (3) Normative dimension; (Bailin, 1987; Scriven & Paul, 1987); (4) Focus on Process VS Outcome (Lipman, 1987; Garrison, Anderson, & Archer, 2001a; Liu, Frankel & Roohr, 2014) (5) Transferability (6) Assessment method (Ku, 2009).

The results highlight that the dichotomy between the normative-philosophical and explanatory-psychological definitions could not be the most valid way to classify CT definitions. Differences among CT definitions could be better understood considering the focus of the CT analysis (on the outcome or the process) and the unit of the analysis (the individual thinking or the inter-subjective actions and practices) rather than the experts' field of study. The critical review presented in the first chapter attempts to categorise all the CT definitions developed after the Delphi Report and to quantify the relationship among the theory-based categories, often mentioned in research (e.g. skills, dispositions, process, and outcome) through a quantitative-qualitative approach. This literature review was not aimed to propose a new CT definition but at showing unexplored similarities among different CT traditions, perspectives, and study methods. These similarities could be exploited to open up a dialogue among experts and build up a shared understanding of CT.

In Chapter 2, I explore the strengths and weaknesses of the different methods developed to assess CT, focusing on open-ended measures. I describe and present validity and reliability properties of four standardised open-ended measures for CT assessment: EWCTET (Ennis & Weir, 1985), ICTET

(Paul & Elder, 2006), CLA (Council for Aid to Education, 2000), and HCTAES (Halpern, 2013). Moreover, qualitative approaches to assess CT as a *process* are presented, especially in online discussion forums. The focus on open-ended measures was motivated by the acknowledged importance of this kind of assessment on one side, and their relatively poor adoption, on the other side. Indeed, closed measures present different disadvantages that limit their use, including the difficulty of scoring and their costs.

In the last section of Chapter 2, I explain why and how automatic assessment could be a viable solution to the current limitations of open-ended measures. I present the state of the art of CT automatic assessment, and I conclude by describing current limitations and future research perspectives that guided the empirical part of this thesis. The last two chapters present two empirical studies aimed at testing the reliability of NLP methods to assess CT manifestations in open-ended written texts.

In Chapter 3, I provide a definition of CT that considers the findings of the scientific literature of the last few years (presented in Chapter 1). The definition is used as the theoretical foundation for empirical work. Then, I describe in detail a rubric developed and implemented by the Center for Museum Studies (CDM) research group for the evaluation and analysis of CT levels within open-ended answers (Poce, 2017). This model was used to design an NLP prototype for the automatic measurement of some CT indicators: use of language, argumentation, relevance, importance, critical evaluation, and novelty. Together with my research group, I carried out some preliminary studies to validate the tool on a group of 66 university teachers. The reliability levels of the CT evaluation rubric were satisfactory, while the evaluation carried out by the prototype was not yet sufficiently reliable. We used the results of this validation to understand how and under what conditions the model works better.

In Chapter 4, I present an empirical investigation aimed at understanding which NLP features are more associated with six CT sub-dimensions (Poce, 2017) as assessed by human raters in essays written in the Italian language. Indeed, most of the studies presented in Chapter 2, are based on the English language. NLP analysis applied to the Italian language is preliminary in nature, especially in the context of educational research. Only in a few cases, NLP is applied to assess learning outcomes or cognitive dimensions (Chiriatti et al., 2018). Therefore, the last experimentation aimed at understanding which NLP features are associated with six CT sub-dimensions, as assessed by human evaluators in essays written in Italian. The study used a corpus of 103 students' pre-post essays who attended a Master's Degree module in "Experimental Education and School Assessment" to assess

students' CT levels. Within the module, we proposed two activities to stimulate students' CT: Open Educational Resources (OERs) assessment (mandatory and online) and OERs design (optional and blended). The essays were assessed both by human evaluators by considering six CT sub-dimensions and by an algorithm that automatically calculates different kinds of NLP features. We found positive internal reliability and a medium to high inter-coder agreement of the human evaluators. Students' CT levels improved significantly in the post-test, and there was no difference between 100% online and blended attendance. Three NLP indicators significantly correlate with CT total score: the Corpus Length, the Syntax Complexity, and an adapted measure of Term Frequency-Inverse Document Frequency. I discuss at the end of the chapter limitations and future developments.

The topics of this PhD research are relevant because of different reasons.

Firstly, CT is considered a desirable learning outcome for European HE students, and it should be comparably recognised, according to the Bologna Strategy. Secondly, research is necessary to understand which teaching strategy can foster CT skills in HE. Comparable methods of CT assessment are fundamental to define the effectiveness of instructional strategies.

Multiple-choice measures cannot be proper for the higher-order skills assessment, such as CT; according to some authors, Multiple-Choice items can be answered without reading the respective text passage. These kinds of tests may be answered merely by low-level processing, such as factual recognition and selection (Nicol, 2007). A further concern regarding Multiple-Choice items is that they make test-takers select between pre-determined answers rather than allowing individualised responses as in constructed response tasks. To address the limitations of Multiple-Choice tests, researchers have developed alternative assessment methods, which involve the adoption of open-ended tasks. According to different authors (Ku, 2009; Liu, Frankel & Roohr, 2014), a measurement that elicits both open-ended and MC response formats should be pursued in CT assessment.

However, open-ended measures present some limitations that could be partially overcome through the development of automatic systems for the assessment of CT in written students' texts. Automatic assessment methods can also facilitate online teaching. Indeed, they can support teachers and researchers to deal with the growing presence of linguistic data produced within educational platforms. To this end, it is pivotal to develop automatic methods for the evaluation of large amounts of data which would be impossible to analyse manually, providing teachers and evaluators with support for monitoring and evaluating the skills demonstrated online by students. The development of automatic tools for the evaluation of CT could reduce the costs of manual scoring and improve the reliability of such measures. Moreover, this method can also be used to support the automatic evaluation of open-ended answers in tests administered at school and university level on a large scale.

CHAPTER 1 DEVELOPING A MULTI-DISCIPLINARY PERSPECTIVE ON CRITICAL THINKING

1. Introduction

Nowadays, a debate regarding the role that higher education is supposed to cover in the broader society is present at an international level. The debate refers to a dialectical conflict between two different stances: should university prepare students to fulfil the job market needs? Or is the university supposed to transmit the knowledge without considering the economic pressure and professional skill training?

To which extent is it possible to reconcile these contrasting perspectives? An education system that focuses on developing higher-order skills, especially Critical Thinking (CT), could be a way to overcome this conflict. Enhancing students' CT is not the only necessary skill to enter and fulfil the job market needs (OECD, 2012; Wagenaar, 2018). Also, it provides students with tools to be autonomous thinkers and active citizens (Davies & Barnett, 2015). CT encompasses different educational perspectives and traditions. The development of CT is currently a declared goal in all levels of education included higher education. However, CT operationalization and definition still represents an open challenge, and therefore, it is difficult to study and evaluate it through an empirical perspective.

The first chapter of this thesis faces the problem of CT definition and its theoretical conceptualization.

1.1 Recent history of an ancient concept

The reflection on CT probably started when someone realized that human beings often fail to think critically properly. Socrates, the father of the Western conceptualization of CT, developed his teaching method to force people to examine their own beliefs and the validity of such beliefs, as an antidote to facing human tendency towards thinking mistakes (Leigh, 2007).

More recently, John Dewey (1910), the modern founder of the CT Movement in Education and a precursor of what today we call CT, worked on the concept of reflective thinking. To illustrate his definition of reflective thinking, in the first pages of his book *How We Think*, Dewey first describes what reflective thinking is not by adopting different examples.

Before Dewey, Francis Bacon and John Locke also reflected on the primary sources of our humans' misconceptions and inference mistakes (Dewey, 1910). Bacon described four "idols":

1. *Idols of the Tribe:* standing erroneous methods that have their roots in human nature. An example of that is the universal tendency to notice instances that corroborate a favourite belief more readily than those that contradict it. Similarly, Locke expressed "that which is inconsistent with our principles is so far from passing from probable with us that it will not be allowed possible." (Locke in Dewey, 1910, p. 24) Bacon and Locke identified and described accurately what today cognitive scientists define as *confirmation bias* (Wason, 1960; Dunbar, Fugelsang & Stein, 2007);

2. *Idols of the Market Place:* fallacies and thinking mistakes that come from an ambiguous use of the language;

3. *Idols of the Cave or Den:* mistakes that derive from individual characteristics. These mistakes refer to what Facione (1990) and many other contemporary CT experts (West, Toplak and Stanovich, 2008) define dispositions or personality traits hat could promote or inhibit CT;

4. *Idols of the Theatre*: mistakes that have their sources in fashion or current general ~~current~~ period.

Similarly, Locke described conditions in which people are more likely to make wrong inferences:

1. People who do not want to devote their efforts to thinking and prefer to believe what others, such as parents, neighbours and any "*opinion leaders*" say;

2. People who are more oriented towards affective states than rational thoughts and do not seriously consider people's opinions that contradict personal interests or values; "Men's prejudices and inclinations impose often upon themselves… Inclination suggests and slides into discourse favorable terms, which introduce favorable ideas;" (Locke in Dewey, 1910, p. 23).

3. People who do not have a broad perspective of a specific knowledge domain, even though they try to follow the reason.

Both Bacon and Locke considered social conditions more dangerous than all the individual sources of misbeliefs because they can perpetuate wrong thinking habits. Thus, education should protect

individuals against erroneous tendencies of their minds and undermine the self-perpetuating prejudices of long ages.

Starting from these preliminary considerations, Dewey (1910, p.10) defined reflective thinking as an "active, persistent and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it, and the further conclusions to which it tends." Consequently, through reflective thinking, people make inferences and conclusions, create and challenge beliefs, starting from an analysis of evidence. Mentioning Locke and Bacon's ideas, Dewey explained that thinking does not always take the right direction during this process. He said, indeed, that if, on one side, thinking "frees us from servile subjection to instinct, appetite and routine" (Dewey, 1910, p. 19), on the other, it can also bring to create wrong beliefs. Dewey mentioned three ways through which we can become aware of our misconception:

1. *Direct experience*: a child discovers that fire can burn his finger by touching a candle. Dewey emphasises that direct experience creates stronger imprinting in the child than a lesson on the properties of heat. A few years later, Piaget (1954) worked on a similar idea of knowledge construction, describing the complementary processes of *assimilation* and *accommodation*. Through assimilation, children integrate external elements into evolving or completed structures, whilst accommodation schemes or structures can change according to newly encountered elements. According to Piaget, assimilation without accommodation creates a distorting reality;

2. *Social sanctions or recognition*: parents, teachers, and friends say that people are right or wrong;

3. *Internal control*: despite the usefulness of the first two kinds of controls, direct experience and social conditions have their limits. For example, existing customs could inhibit the use of data and evidence to reach a the right conclusion. Thomas Kuhn (1962) showed that the tendency to confirm well-accepted theories could be detected in social groups with a high education level. Scientists carry out their regular work within a settled paradigm or explanatory framework, without necessarily questioning the underlying assumption of that theory. Besides, our direct experience is limited to a specific time and space, and the limitation of our senses. Thus, according to Dewey, human beings developed more sophisticated strategies to control the observation conditions and to formulate conclusions based on the evidence.

To sum up, Dewey reckons that education should consider both skills and dispositions enhancement. On the one hand, education should cultivate effective behaviours to discriminate tested beliefs from assertions, guesses, and beliefs (skills); on the other hand, education should develop an open-minded preference for properly grounded conclusions (dispositions). No matter how much an individual knows. Dewey (1910) said that if people did not have attitudes and habits, they would fail to apply reflective thinking.

1.2 From Dewey's Reflective Thinking to Critical Thinking in Education

Davies and Barnett (2015) described three historical movements related to CT, following a similar classification to the one that Richard Paul (2011) proposed.

In the 1970s, mainly philosophers tried to introduce formal and informal logic in the schools and universities curricula. The first wave concerned the identification and evaluation of arguments to avoid fallacies in reasoning processes. In that perspective, the emphasis was on argumentation, logic, and reasoning.

That tradition was based on Walters's (1994) asserted idea that CT is "logistic"; thus, a critical thinker becomes someone like Mr Spock in the original Star Trek series: an objective and rational being. Many authors criticized it. Thomson (1998) explained that the traditional CT notion either ignores or rejects the role that emotions should play in CT. One of the most interesting criticisms of the traditional CT view comes from feminist literature. Clinchy and Zimmerman (1985) interviewed female students and asked them to react to a set of statements. Researchers found out that the most common thinking strategy female students adopted, was so-called *connected knowing*. People who adopt this thinking strategy attempt to get into heads of people they want to understand, trying to see an issue through the other's eyes. Contrasted with the traditional CT idea, *connected knowing* embraces empathy as a source for critical thought giving the prototypical example of *the devil's advocate* role. The growing concern towards a reductionist view of CT as a *cognitive and rational machine* brought to an emerging research wave and educational practices. In the 1980s, CT started to be more connected with human beings' inner nature, emphasizing that CT could be inter-related with attitudes, emotions, intuitions, and creativity. Moreover, in the second wave, CT was interpreted as an ideological issue, for example, in German critical theory, phenomenology, and psychoanalysis.

While in the first wave the word *critical* assumes the meaning of *criticism*: identifying weaknesses, correcting a claim or an argument, in the second wave the word assumes the concept of *critique*:

identifying dimensions of meaning that might be missing or concealed behind a claim or an argument. As Davies and Barnett (2015) asserted, both the waves presented weaknesses: the first wave was rigorous, but it neglected many human dimension aspects; on the other hand, the second approach was more abstract and difficult to study from a rigorous perspective. According to Paul (2011), the third wave has been recently emerging by trying to encompass the previous two waves limits and to put emphasis on their strengths: on the one hand, the structure of argumentation and, on the other hand, considering human traits, such as emotion, imagination, and creativity to build a CT theory.

1.3 The Frame of this work

When people talk about CT, they usually refer to many different things, concepts, and traditions. Davies (2015) proposed a complex model to summarise different traditions that refer to CT (Figure 1). The Frame of this work can be understood at the boundary between the CT Movement and the Criticality Movement. The Criticality Movement started to shift the attention from an individual and cognitive perspective on CT to a socio-cultural perspective on CT. The Criticality Movement considers, skills, disposition, critical doing, and critical actions. A specific kind of critical action that will be discussed in this work is the *collaborative knowledge construction* as McPeck (1985) first defined CT as a process where interaction occurs between individuals and the interpretations of knowledge which they create.
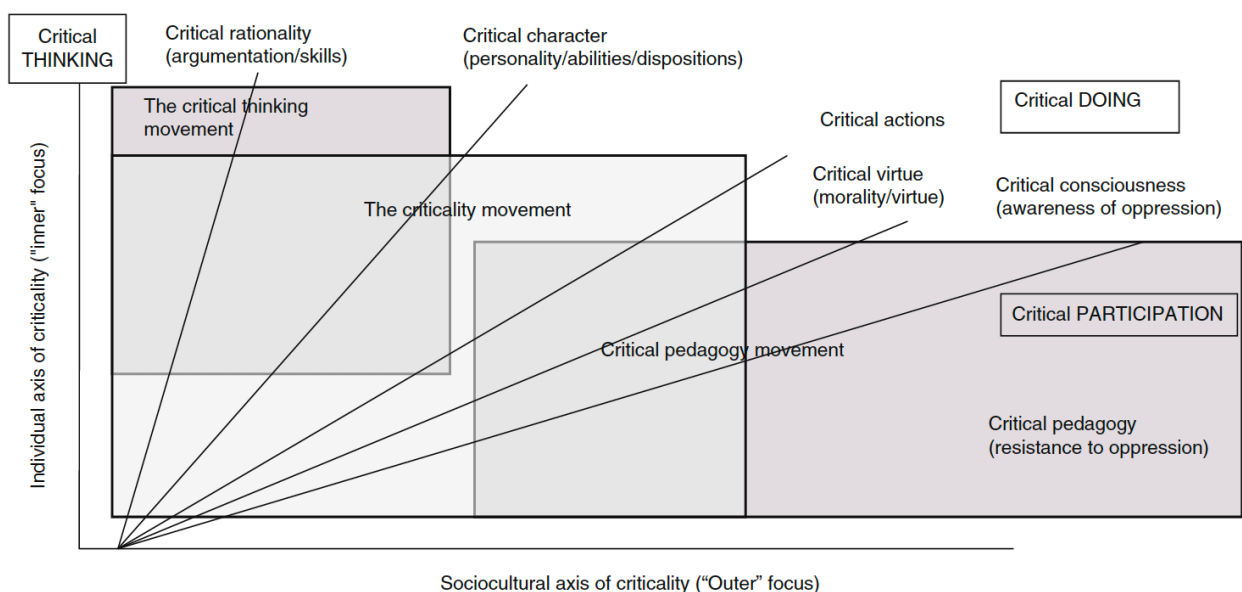
Figure 1 A model for Critical Thinking in Higher education. Retrieved from Davies (2015)

On the other hand, this work will not take into account in-depth the Critical Pedagogy Movement perspectives (Freire, 1972; 1973). Other authors have already analysed them (Giosi, 2009; Cambi,

2009). Furthermore, the Critical Pedagogy Movement does not consider issues related to CT assessment, which is one of the main focuses of this thesis dissertation.

2. The problem of the definition: a critical literature review

CT definition and its related theoretical foundations are affected by at least three interconnected issues:

1. Definitions are abstract and normative rather than being based on any actual reality (Kuhn, 2019; Norris, 1992; Moore, 2013; Atkinson, 1997);
2. Most of the definitions are shaped in a western perspective (Atkinson, 1997; Chen, 2017);
3. They did not consider the network problem (Johnson, 1992; 2000) which means: "the fuzzy relations among certain more or less interchangeable terms, including metacognition, higher-order thinking skills, problem-solving, rationality, and reasoning, that are used when talking about CT." (Johnson, 2000, p. 21 – 22)

According to Atkinson (1997) and Chen (2017), CT definitions are shaped in a Western perspective. The former defines CT as a social practice (1997): even though people cannot define CT clearly, they can still talk about it, understand one another, and even "recognise it... when it occurs", suggesting that CT exists mainly at the level of tacit, common-sense, and social practice. Atkinson, in his work (1997), tried to unpack the CT concept by showing how the implicit assumptions of this concept are strongly related–with to a Western cultural position. Atkinson showed that the relation between language and learning, the role of an individual, and the self-reflection concept could be seen in different ways according to a specific cultural context. Chen (2017) recognised in Confucius philosophy, the Eastern CT conceptualization. According to Confucius, the exemplary thinking focuses more on self-reflectiveness than simple inquisitiveness. Moreover, quietly pondering problems is more valuable than asking a teacher many questions, according to the Confucian philosophy. Thus, Chen affirms that in Eastern Culture, CT can be more easily conceptualised as self-reflexivity; whilst in Western culture, CT is more focused on judgement and decision-making.

Authors also highlighted a conflation risk of related concepts (Byrnes and Dunbar, 2014). Some authors use CT and intelligent thinking as synonyms (e.g. Halpern, 2008) although some research showed a low to a non-significant correlation between CT level and intellectual ability (Stanovich,

West & Toplak, 2013). Problem-solving, logical thinking, and scientific thinking are also commonly overlapped with CT. Another common conflation regards the Critical and Creative Thinking concept.

In addition to the above-mentioned problems, many perspectives proposed by different authors can negatively affect the comparability of empirical research results. According to Barnett (1997) Higher education, which prides itself on critical thought, has done no adequate thinking about CT. Hatcher (2013) has recently disavowed the need for an irreproachable correct definition. Johnson and Hamby (2015) claim that the problem is not the absence of good definitions, but the overabundance of problematic definitions. Although many CT definitions were proposed, a few authors tried to explain what was wrong with the previous definitions before they proposed a brand-new one.

In recent years, different researchers attempted to review CT definitions and their related theoretical foundation. Markle, Brenneman, Jackson, Burrus, and Robbins (2013) synthesised frameworks of higher education student learning outcomes, included CT. The review from Markle and colleagues was one of the most systematic in the field. However, they did not include many theoretical CT definitions. Johnson and Hamby (2015), and Davies (2015), in their respective reviews, included CT definitions as mainly philosophical, excluding other important definitions that came from different approaches. Besides, these reviews tended to be narrative and lack of a systematic approach for classifying the CT definitions and theoretical approaches. Lai (2011), as well as Stendberg (1986), classified the definition in three macro-categories: philosophical, psychological, and educational. Although the classification proposed by Lai (2011) and Stendberg (1986) could be a useful way to look at the *differences* between the CT approaches, they did not consider the complexity to study CT in a multidisciplinary perspective. Cross-fertilization examples in CT disciplines can be detected in many recent works (Mercier & Sperber, 2011; Reznitskaya, 2012; Kuhn, 2019). Classification based on a disciplinary field would make it difficult to identify commonalities and bridges between different disciplines.

Furthermore, no CT definitions' reviews were aimed to comprehend all CT definitions and their related theoretical foundations. Moreover, none of these reviews directly connected the issue of CT definitions with the CT operationalization and assessment. A new way to classify CT definitions to find cross-disciplinary commonalities could be a useful step towards the resolution of the CT definition issue. Thus, the main aim of this critical review was to answer the following research questions:

RQ 1) Can we propose a new classification for systematising different CT definitions and their related theoretical approaches?

RQ2) Which is the relation between the different adopted CT definitions and CT assessment method?

2.1 Methods

2.1.1 Data collection

The first step was trying to select all the CT definitions, mentioned in published research papers after 1990. 1990 is an important year for research on CT. Indeed, in that year, The American Philosophical Association published the Delphi Report led by Peter Facione[2].

This review aims to include as many definitions as possible; therefore, we used a mixed-method which combines a systematic and non-systematic approach. To avoid the studies exclusion relevant to the objective, we carried out a citation analysis included in the published revision studies and a manual search. The main inclusion criteria were the following: definitions had to be connected to a CT theory presented and illustrated in a publication. Thus, we did not consider in the analysis definitions provided in learning outcomes frameworks, for example, Bologna, CAS, Lumina DQP, QAA-FHEQ or USDOL-ETA. We excluded definitions provided in learning outcomes frameworks for two reasons. Firstly, Markle et al. (2013) have already carried out a systematic review concerning CT; secondly, because learning outcomes frameworks usually do not explicitly provide a broader theoretical view on CT. When the same author has provided strongly different CT definitions in a time frame considered, both the definitions were included. Through the method described, we identified 39 relevant definitions (the list of all the definitions are presented below in the Table 8 and the Table 9).

2.1.2 Data analysis

We carried out a qualitative content analysis (Mayring, 2004) to develop a new definitions' classification, through the support of the Software ATLAS.ti[3]. Firstly, we read carefully all the definitions and their related scientific publications.

---

[2] The Delphi report was based on a Delphi method. The Delphi method is an iterative process to collect and distil the anonymous judgments of experts using a series of data collection and analysis techniques interspersed with feedback (Skulmoski, Hartman, & Krahn, 2007). By involving experts in the CT field, the Delphi report expected outcome was to achieve a CT definition agreement.

[3] ATLAS.ti is a workbench for the qualitative analysis of textual, graphical, audio, and video data. It offers a variety of tools for accomplishing the tasks associated with any systematic approach to unstructured data, i. e., data that cannot be meaningfully analyzed by formal, statistical approaches. It offers tools to manage, extract, compare, explore, and reassemble meaningful pieces from large amounts of data in creative, flexible, yet systematic ways.

A grid for the qualitative content analysis (Table 1) was developed based on the most highlighted features in the literature related to CT definitions. Eleven theory-driven categories have been identified: (1) skills and abilities; (2) dispositions; (3) actions; (4) practice; (5) values and standards; (6) process; (7) outcome; (8) generalizable; (9) domain-specific; (10) closed-measures; (11) open-measures. These 11 categories were clustered in six macro-categories (Table 1):

1. *Individual dimension*: CT is mainly conceptualized in terms of an individual working alone on a problem-based task (e.g. Facione, 1990);

2. *Inter-individual dimension*: CT is defined in terms of social and dialogical information exchange (e.g. McPeck, 1981; Kuhn, 2019);

3. *Normative dimension*: CT is defined in terms of an ideal standard to achieve (Bailin, 1987; Scriven & Paul, 1987);

4. *Focus*: CT definitions tend to be focused more on its process or its outcome (Lipman, 1987; Garrison, Anderson, & Archer, 2001a; Liu, Frankel & Roohr, 2014);

5. *Transferability*: some authors emphasised the idea that CT is generalizable to different knowledge domains. Others consider CT as a domain-specific skill. Furthermore, some scholars have intermediate positions (Rear, 2019);

6. *Assessment method*: CT has been commonly assessed through closed-measures (e.g. Multiple-choice), open-ended measure (e.g. essays, short-answers), or mixed methods that combine closed and open-ended measure (Ku, 2009).

Each of 39 definitions was coded through one or more categories following a non-mutually exclusive classification approach (Downe-Wamboldt, 1992). This approach allowed to calculate the categories occurrence and the sub-categories co-occurrence.

Occurrence (O) is the frequency of a specific category in 39 definitions. Co-occurrence (C) indicates the number of times when two categories occur together in the same definitions. Atlas.ti calculates the C-coefficient in co-occurrence analysis. The c-coefficient indicates the strength of the relation between two categories similar to a correlation coefficient (Armborst, 2017). The calculation of the c-coefficient is based on approaches borrowed from the quantitative content analysis. The range of the c-coefficient is between 0 = *codes do not co-occur*, and 1 = *these two codes co-occur wherever they are used.* It is calculated as follows:

$$c = n12/(n1 + n2 - n12)$$

After calculating occurrence and co-occurrence, we have carried out a qualitative interpretation of the co-occurrence results to identify connections and similarities among the definitions, based on the categories previously described.

Table 1 Grid for the qualitative content analysis of the CT definitions

Macro-Category	Category	Explanation
Individual Dimension	Skills / ability	CT as a set, list, taxonomy of internal cognitive competencies and mental operations;
	Dispositions	CT as a set of personal traits and characteristics;
	Actions	CT as behaviours that result from internal mental operations;
Inter-individual Dimension	Practice	CT is defined considering the relation among people and their cultural context;
Normative Dimension	Values and standards	A number of general principles and defined threshold.
Focus	Process	Research focus on what happens during a CT intra and inter-individual activity
	Outcome	Research focus on the outcome of a CT intra and inter-individual activity
Transferability	Generalizable	CT learning outcomes are not specific of a domain knowledge
	Domain specific	CT learning outcomes are not specific of a domain knowledge
Assessment method	Closed-measures	Multiple-choice; self-report;
	Open-measures	Essay; constructed-response tasks; analysis of dialogical exchange;

2.2 Results

2.2.1   Occurrence

Among 39 definitions, 11 of them have been identified before the Delphi Report (Table 2) and 27 definitions after the Delphi Report (Table 3).

It is interesting to see that 3 definitions out of 39 were based on the Delphi Methodology. The first one was formulated by Facione (1990). After that, two Delphi studies were carried out to build a CT definition in the nursing field: (1) Scheffer & Rubenfeld (2000); (2) Paul (2014). These preliminary results could suggest that the Delphi method is not the most effective methodology to achieve an agreement on CT definitions.

The most common categories in 39 definitions are *Skills and ability* (O = 34), *Generalizable* (O = 25), and *Outcome* (O = 22). The focus on the *Process* (O = 16) is slightly lower than the focus on

the *Outcome*. It is possible to detect a greater internal difference in the "Individual" and ~~in~~ the "Transferability" macro-categories. Indeed, *Dispositions* (O = 19) and *Actions* (7) occur less than *Skills and Dispositions*. In the same way, a *Domain-Specific* view on Critical Thinking is less frequent than a *Generalizable* view. *Practice* (O = 3), and *Values and Standards* (O = 8) are among the least frequent categories. Definitions are explicitly connected to assessment methods in less than 50 % of the cases, where *closed-measures* (O = 9) and *open-ended measures* (O = 8) occur with a similar frequency in the definitions.

Table 2 Sub-category Occurrence results

Sub-category	Occurrence
Skills and ability	34
Dispositions	19
Action	7
Practice	3
Values and standards	8
Process	16
Outcome	22
Generalizable	25
Domain specific	11
Closed-measures	9
Open-measures	8

2.2.2   Co-occurrence Analysis and Co-Occurrence Networks

Table 3 contains the analysis of the co-occurrence among the categories included in the *individual dimension, inter-individual dimension, normative dimension,* and *focus and transferability* (the co-occurrence with the categories of the *assessment methods* are presented in the next paragraph). The strongest co-occurrence (0,40 < C < 0,71) can be detected between the following categories: *Skills / Abilities* and *General* (C = 0,71); *Skills / Abilities* and *Outcome* (C = 0,62); *General* and *Outcome* (C = 0,47); *Skills / Abilities* and *Dispositions* (C = 0,44); *Dispositions* and *Outcome* (C = 0,41). Moderate co-occurrence (0,21 < C < 0,39) can be detected between the following categories: *General* and *Dispositions* (C = 0,38); General and *Process* (C = 0,37); *Skills / Abilities* and *Process* (C = 0,36); *Knowledge Specific* and *Outcome* (C = 0,32); *Process* and *Dispositions* (C = 0,30); *Outcome* and *Process* (C = 0,27); *Skills / Abilities* and *Knowledge Specific* (C = 0,26); Practice and *Actions* (C = 0,25); *Actions* and *Disposition* (C = 0,24); *Process* and *Knowledge Specific* (C = 0,23); *Process* and *Action* (C = 0,21); *Skills / Abilities* and *Values / Standards* (C = 0,21).

Table 3 Co-occurrence analysis results

	Actions	Dispositions	General	Knowledge specific	Practice	Process	Skills / Abilities
Actions	0,00	0,24	0,10	0,13	0,25	0,21	0,11
Dispositions	0,24	0,00	0,38	0,20	0,05	0,30	0,44
General	0,10	0,38	0,00	0,16	0,00	0,37	0,71
Knowledge specific	0,13	0,20	0,16	0,00	0,00	0,23	0,26
Outcome	0,16	0,41	0,47	0,32	0,00	0,27	0,62
Practice	0,25	0,05	0,00	0,00	0,00	0,12	0,00
Process	0,21	0,30	0,37	0,23	0,12	0,00	0,36
Skills / Abilities	0,11	0,44	0,71	0,26	0,00	0,36	0,00
Values / Standards	0,15	0,13	0,22	0,06	0,00	0,14	0,21

In this research, we interpreted the co-occurrences to find connections and similarities among the definitions. Thus, we propose two co-occurrence networks. This classification is partially based on Lai (2011) and Stendberg (1989) classification, but it overcomes the idea that definitions could be classified according to the authors' field (philosophy vs psychology vs education).

The first co-occurrence network could be called "Normative - Descriptive network" (Figure 2); it contains most of the strongest identified co-occurrence. According to Stendberg (1986) and Lai (2011), The CT philosophical approach relates to normative – descriptive theories. However, several philosophers were interested in studying CT in terms of *processes* (Lipman, 1981; Van Gelder, 2005; Walton, 1989); as well as psychologists and educators, they developed the normative-descriptive CT definitions (e.g. Halpern, 1998).

The Normative-descriptive approach focuses on what people are capable of doing under the best of circumstances (Lai, 2011). Examples are "perfections of thought", as Paul (1992) described. This appoach also emphasises qualities or standards of thoughts. Critical Thinking *Skills* and *Dispositions* are seen concerning *Outcome* and *Skills,* which are seen as transferable to different domains. Halpern (1998) definition is a good example of the relation among these four categories:

Critical thinking is the use of those cognitive skills or strategies that increase the probability of a desirable outcome. Critical thinking is purposeful, reasoned, and goal-directed. It is the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions. Critical thinkers use these skills appropriately, without prompting, and usually with conscious intent, in a variety of settings. That is, they are   predisposed   to think critically. When we think critically, we are evaluating the outcomes of our thought processes – how good a decision is or how well a problem is solved (p. 450-451).
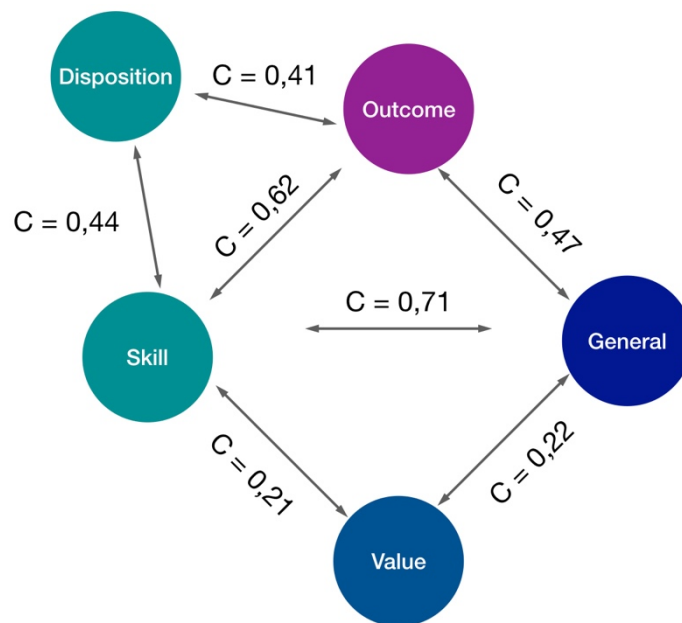
Figure 2 The Co-occurrence Network 1: Normative – Descriptive Network

Some definitions also relate *Skills* to *Value and Standard* to achieve (C = 0,21). For example, according to Scriven & Paul (1987) CT is based on universal intellectual standard that transcend subject matter divisions.

Although also Lipman proposed the idea of standards, his vision is more contextual-based, contrasting the universality of standards idea shown in Scriven & Paul. Indeed, Lipman said that "CT is sensitive to the context" and following this view, he proposed different categories of standards: informal criteria, formal or institutionalised criteria, abstract meta-criteria, and ethical mega-criteria.

To sum up, in Network 1, CT is a synonym of good thinking (Bailin, 1987). Furthermore, in Network 1, CT focuses more on an individual working alone problem-based task; The Delphi Report from 1990 proposed one exemplary definition: "CT is purposeful, self-regulatory judgement that results in

interpretation, analysis, evaluation and inference, as well as explanations of the consideration on which the judgement is based" (Facione, 1990, p. 3).

This definition emphasises the CT concept of CT as an *outcome* – the judgement – and its related features or standards – it is purposeful and self-regulatory. The outcome can be achieved following different cognitive activities, usually defined as CT skills.

The second network could be called "Descriptive – Explanatory network" (Figure 3). The definitions in the second network tend to focus more on how people think rather than how they could or should think under ideal conditions (Sternberg, 1986). Consequently, CT definitions are commonly process-related and personal dispositions to be engaged in a CT process are emphasised, as in the most recent definition: "Critical Thinking is a dialogic practice people commit to and thereby become disposed to exercise, more than an individual ability or skill" (Kuhn, 2019, p. 148).
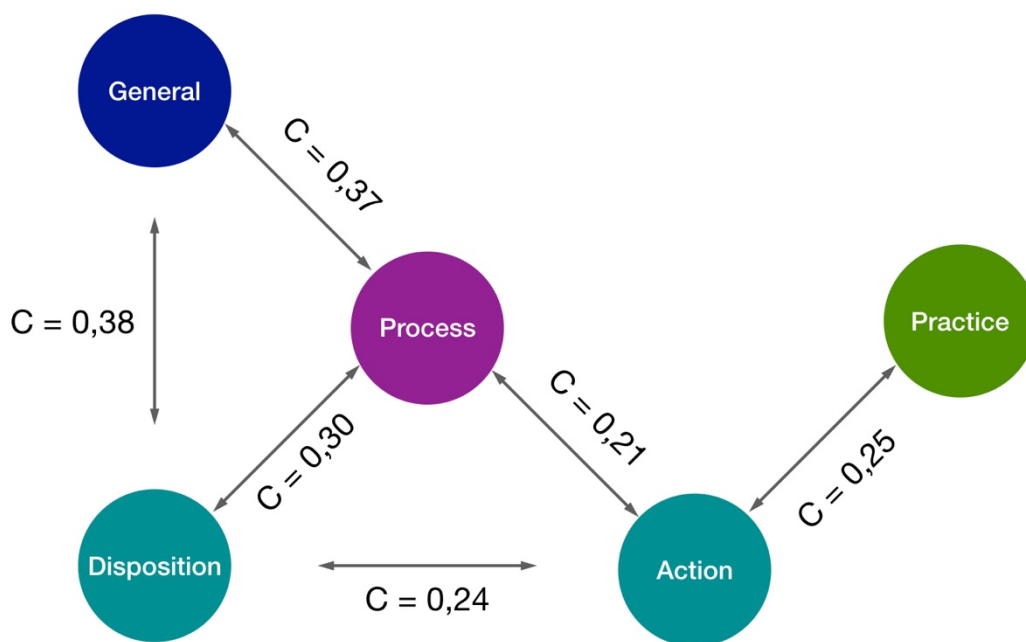
Figure 3 The Co-occurrence Network 2: Descriptive - Explanatory Network

As Kuhn showed, instead of conceptualizing CT as an individual and internal work on a problem-based task, in the second network *Actions* and social *Practices* (C = 0,25) can occur together. The category *Actions* is on the borderline between the individual and the inter-individual dimension. McPeck (1981) proposed an exemplary definition to describe the relation between *Process* and social *Actions:* "CT is a process where interaction occurs between individuals and the interpretation of knowledge which they create." In this case, McPeck includes not only internal mental activities (interpretation) but also a specific CT action: the social construction of knowledge (Scardamalia & Bereiter, 2006).

2.2.3 Co-occurrence with assessment methods

Table 4 describes the co-occurrence between the assessment method categories and the other categories.

Closed measures are more associated with the definitions that emphasised *outcome* (C = 0,35) and *skills* (C = 0,27). On the other hand, open-measures are more commonly associated with the definitions that emphasised *action* (C = 0,25) and *process* (C = 0.26). Both measurement methods are associated with a CT as generalizable.

Table 4 Co-occurrence with assessment method.

	Closed-Measures	Open-measures
Action	0,07	0,25*
Disposition	0,17	0,17
General	0,21*	0,22*
Outcome	0,35*	0,20
Practice	0,00	0,10
Process	0,00	0,26*
Skills / abilities	0,27*	0,21*

* indicates C > 0,20

In Table 5, the assessment methods adopted to assess CT as an outcome and generalizable skill (Co-occurrence Network 1) are presented (for more detailed analysis of the test see Liu, Frankel and Roohr, 2014 and Ku, 2009).

Table 5 assessment methods adopted to assess CT as an outcome and generalizable skill (Co-occurrence Network 1)

Definition	Test	ITEM TYPE
Watson & Glaser (1980)	Critical Thinking Appraisal (Watson & Glaser, 1980)	MC
Norris & King (1983)	Test on appraising observation (Norris & King, 1983; Norris, 1990)	MC
Stendberg (1986)	Sternberg Triarchic Abilities Test (Sternberg and colleagues, 2001)	MC
Ennis (1987)	Cornell Critical Thinking Test (Ennis, 1993)	MC
Facione (1990)	CCTST (Facione & Facione, 1994);	MC
Pascarella & Terenzini (1991)	CAAP Critical Thinking (American College Testing Program, 1989)	MC
Halpern (1998)	The Halpern Critical Thinking Test (2013)	Mixed methods

Most of these tests are directly connected to the CT theoretical definition provided by the author. An exception is for Pascarella and Terenzini (1991). Indeed, they used the CAAP in most of their research although their definition of CT does not directly connect to the test. Most of the tests are multiple-choice (MC) with a correct answer. Halpern (2006) developed the first assessment method that tries to balance MC questions with open-ended questions.

Table 6 presents the assessment methods adopted to assess CT as a *process* and *action* (Co-occurrence Network 2).

Table 6 assessment methods adopted to assess CT as a process and action (Co-occurrence Network 2)

Definition	Test / Tasks	Kind of analysis
Ennis (1987)	Ennis Weir Critical Thinking Essay Test (Ennis & Weir, 1985)	Essay
Riesenmy, Mitchell, Hudgins, & Ebel (1991)	Productive Thinking Program (Covington, Crutchfield, Davies, & Olton, 1972)	Selection and Production tasks (Thinking Aloud);
Garrison (1992)	Online discussion and dialogic interactions (Garrison, Anderson & Archer, 2001a, 2001b)	Content analysis
Kuhn (1992, 2019)	Online discussion and dialogic interactions (Kuhn, Zillmer, Crowell, & Zavala, 2013)	Content analysis
West, Toplak, & Stanovich (2008)	Syllogistic Reasoning Problems with Belief Bias. Heuristics and Biases Tasks (West, Toplak & Stanovic, 2008)	MC

Garrison, Anderson, and Archer (2001a, 2001b); and Kuhn, Zillmer, Crowell, and Zavala (2013) similarly assessed the process of CT in dialogic interactions both online and offline through the content analysis methods. However, they considered different indicators in their respective rubrics. Garrison method is still one of the most used to assess CT in online dialogical interactions, and it inspires many other authors (included Newman, Webb & Cochrane, 1995; Poce, 2017). On the other hand, Kuhn et. al (2013) focus on the dialogical argumentation assessment, providing a hierarchy of argumentation strategies, from the most to the least sophisticated. Examples of strategies are the use of evidence, meta-talk types, and meta-argumentation types).

Riesenmy, Mitchell, Hudgins, and Ebel (1991) assessed CT in children through a thinking-aloud method, considering the following criteria: task-definition, monitoring, strategy and the use of evidence. Also, Ennis (1985) developed a method to assess CT in a constructed-answer task, more specifically through an essay. The essay should be assessed according to the following criteria: getting the point, seeing reasons and assumptions, stating' one's point, giving good reasons, seeing

other possibilities, responding to equivocation; and avoiding equivocation, irrelevance, circularity, a reversal of an if-then relationship, overgeneralisation, credibility, problems, and the use of emotive language to persuade. West, Toplak, and Stanovich (2008) proposed to incorporate the current method adopted to assess *rational thinking* in the CT assessment, specifically by using several syllogistic reasoning problems.

2.3 Conclusive remarks

This work attempts to provide a new way to classify CT definitions, dealing with the complexity and the multidisciplinary nature of the theoretical construct. Although other authors introduced the idea of *normative* and *explanatory* definitions (Sternberg, 1989; Lai, 2011), this research has adopted a more systematic approach for the study of the relationship among different theory-based categories that underpin both the *normative* and *explanatory* CT definitions. Furthermore, this research shows that the dichotomy between the normative-philosophical and explanatory-psychological definitions could not be the most valid way to classify CT definitions. Differences among CT definitions are not simply related to the experts' field of study. Differences could be better understood considering the *focus* of the CT analysis (on the *outcome* or the *process*) and the *unit* of the analysis (the individual thinking or the inter-subjective actions and practices). The *actions* category does not co-occur with most of the sub-categories presented in Network 1. As highlighted by Davies and Barnett (2015), according to many authors, meeting the CT requirements is possible without doing anything. Barnett suggests that by focusing on CT *actions,* pedagogical guidelines could be improved:

> Education can, therefore, potentially do much more than teach students how to demonstrate analytic skills and judgments. It can also prompt students to understand themselves, to have a critical orientation to the world, and to demonstrate an active socio-political stance toward established norms or practices with which they are confronted. (Davies & Barnett, 2015, p. 16)

This work also attempts to highlight unexplored commonalities among CT definitions to reduce the current theoretical fragmentation among the numerous definitions and their related theories.
After the in-depth analysis of different CT perspectives, an interesting and poorly explored concept has emerged: the relationship between CT, *Epistemic Rationality and Epistemological belief/understanding.* Epistemological knowing (Kuhn, 1999) has both a general-philosophical aspect (e.g. "How does anyone know?) and a personal aspect (e.g. "What do I know about my own

knowing?). If we consider CT as "a process where an interaction occurs between individuals and the interpretations of knowledge, which they create" (McPeck, 1981), epistemological belief and understanding could inhibit or support the disposition to adopt CT in the collaborative building-learning environment knowledge (Scardamalia & Bereiter, 2006). Little has been done to explore this relationship both theoretically and practically, despite the attention given by authors from different backgrounds throughout the time frame considered (McPeck, 1981; Kurfiss, 1988; West, Toplak & Stanovich, 2008; Kuhn, 2019).

To sum up, these two networks should not be interpreted as mutually exclusive. On the contrary, they can be seen as different points of view for the study of the same object.

This review presents some limitations. Firstly, I coded the definitions on my own, so showing the results' reliability is impossible. However, in order to partially overcome this limitation, Table 7 shows the transparent way in which each definition was coded. A second limitation concerns the process through which the definitions were identified. With the aim to include as many definitions as possible, a mixed-method has been used which combine a systematic and non-systematic approach. Thus, the process of CT definitions identification is not replicable. Moreover, possibly some CT definitions were not included in the analysis. Despite the illustrated limitations, this work is one of the first that attempts to categorise all the CT definitions developed after the Delphi Report and to quantify the relationship among the theory-based categories, often mentioned in research (e.g. skills, dispositions, process, and outcome). This literature review was not aimed to propose a new CT definition. The main aim was to show unexplored similarities among different CT traditions, perspectives and study methods. These similarities could be exploited to open up a dialogue among experts and build up a shared understanding of CT. The use of comparable research methodology would also be a necessary step to achieve a better understanding of empirical research results on the most debated CT issues.

Table 7 Identified categories for 39 definitions included in the review

Authors	Skills	Disposition	Action	Practice	Values	Process	Outcome	General	Specific
Watson & Glaser, 1980	X	X					X	X	
McPeck, 1981	X	X	X			X			X
Norris & King, 1983	X						X		
Beyer, 1984	X							X	
Stendberg, 1986	X						X	X	
Ennis, 1987	X	X	X				X	X	
Wade & Travis, 1987	X						X		
Lipman, 1987	X				X	X	X	X	
Scriven & Paul, 1987	X		X		X	X	X	X	
Kurfiss, 1988	X					X	X	X	X
Siegel, 1988	X				X				
Facione, 1990	X	X					X	X	
McMurray & Beisenherz, 1991	X						X	X	X
Pascarella & Terenzini, 1991	X							X	
Riesenmy, Mitchell, Hudgins, & Ebel, 1991	X					X	X	X	X
Garrison, 1992	X	X	X			X	X	X	
Kuhn, 1993	X					X	X	X	
Thayer-Bacon, 1993	X	X				X			
Brookfield, 1995	X					X		X	
Atkinson, 1997				X		X			
Fisher & Scriven, 1997	X				X	X		X	
Halpern, 1998	X	X					X		
Bailin, Case, Coombs & Daniels, 1999			X		X		X		X
Scheffer & Rubenfeld, 2000	X	X					X		X
Mingers, 2000	X	X							
Hatcher & Spencer, 2005					X			X	
Willingham, 2007	X	X					X		X
Elder, 2007	X				X			X	
Epstein, 2008	X	X					X	X	
West, Toplak, & Stanovich, 2008		X				X		X	
Moore & Parker, 2009	X						X		
Bailin, Battersby, Clauss 2011	X					X	X	X	X
Mulnix, 2012	X							X	
Johnson, 2014	X	X			X		X	X	
Byrnes & Dunbar, 2014	X	X				X		X	
Paul, A. S. 2014	X					X	X		X
Danvers, 2016			X	X					
Kuhn, 2019		X	X	X		X			
Stendberg & Halpern, 2020	X	X						X	X

Table 8 twelve definitions developed before the Delphi Report and still used in current scientific publications

Authors	Definition
Watson & Glaser, 1980	CT involves an attitude of being disposed to consider thoughtfully the problems and subjects that come within the range of one's experiences; knowledge of methods of logical inquiry and reasoning; some skills in applying those methods. CT calls for a persistent effort to examine any belief or supposed knowledge form in the light of evidence that supports it and the further conclusions to which it tends. As well as the ability to recognize problems; weigh evidence; ~~to~~ comprehend and use language with accuracy and discrimination; interpret data; recognize the existence of logical relationships between propositions; ~~to~~ draw warranted conclusions and generalizations; and ~~to~~ test the conclusions by applying them to new situations to which they seem pertinent.
McPeck, 1981	A process where interaction occurs between individuals and knowledge interpretation which they create; *reflective scepticism*: disposition and skill to do in such a way that E (evidence in any field) is suspended-or temporarily rejected-as sufficient to establish the truth or viability of P (some proposition or action within. CT does not merely refer to the statements' assessment, but it includes thought processes involved in problem-solving and active engagement in certain activities.
Norris & King, 1983	The ability to judge the reports observation credibility.
Beyer, 1984	A set of skills: 1) distinguishing between verifiable facts and value claims; 2) determining the reliability of a claim or source; 3) determining the accuracy of a statement; 4) distinguishing between warranted or unwarranted claims; 5) distinguishing between relevant and irrelevant information, claims, or reasons; 6) detecting bias; 7) identifying stated and unstated assumptions; 8) identifying ambiguous or equivocal claims or arguments; 9) recognizing logical inconsistencies in a line of reasoning; ~~and~~ 10) determining the strength of an argument.
Stendberg, 1986	CT comprises mental processes, strategies and representations that people use to solve problems, make decisions, and learn new concepts.
Ennis, 1987	Reasonable, reflective thinking that is focused on deciding what to believe or do.
Wade & Travis, 1987[4]	The ability and willingness to assess claims and make objective judgments based on well-supported reasons. (a) ask questions and be willing to wonder; (b) define problems clearly; (c) examine evidence; (d) analyse assumptions and biases; (e) avoid emotional reasoning; (f) avoid oversimplification; (g) consider alternative interpretations; (h) tolerate uncertainty.
Lipman, 1987	CT is self-correcting; thinking with criteria; sensitive to the context.
Scriven & Paul, 1987[5]	CT is the intellectually disciplined process of actively and skilfully conceptualizing, applying, analysing, synthesizing, and/or evaluating information gathered from (or generated by) observation, experience, reflection, reasoning, or communication, as a guide to a belief and action. In its exemplary form, it is based on universal intellectual values that transcend subject matter divisions: clarity, accuracy, precision, consistency, relevance, sound evidence, good reasons, depth, breadth, and fairness.
Kurfiss, 1988	An investigation whose purpose is to explore a situation, phenomenon, ~~or~~ question, or problem; and to arrive at a hypothesis or conclusion about it, which integrates all available information and that can, therefore, be convincingly justified. In CT, all assumptions are open to question, divergent views are aggressively sought, and the inquiry is not biased in favour of a particular outcome.
Siegel, 1989	Appropriately thinking is moved by reasons.
Facione, 1990	CT is to be purposeful and have a self-regulatory judgment, which results in interpretation, analysis, evaluation and inference, as well as explanation of the evidential, conceptual, methodological, criteriological or contextual considerations.

---

[4] In Wade, C. (1995). Using writing to develop and assess critical thinking. *Teaching of psychology*, *22*(1), 24-28.

[5] Retrieved from http://www.criticalthinking.org/pages/defining-critical-thinking/766

Table 9 twenty-six definitions developed before the Delphi Report and still used in current scientific publications

Authors	Definition
Mcmurray & Beisenherz, 1991	The ability to reason dialectically or logically in synthesizing multiple frames of reference to resolve new problems. Cognition, evaluation of semantic content, and explanatory components.
Pascarella & Terenzini, 1991	Involves individual's ability to do some or all of the following: identifying central issues and assumptions in an argument; recognising important relationships; making correct inferences from data; deducing conclusions from information or data provided; interpreting whether conclusions are warranted on the basis of the data given; evaluating evidence or authority.
Riesenmy, Mitchell, Hudgins, & Ebel, 1991	CT is a matter of thinkers' problem or conclusion assessing with which they are confronted to determine what it asks or asserts; a matter of organising available evidence into a plan to answer that question, and of evaluating the evidence to determine its acceptability, whether it has been presented by another person or generated by a thinker.
Garrison, 1992	CT is a process of making sense (internal cognitive process) of external experiences through the analysis of issues and information.
Kuhn, 1993	The ability to recognise the possible falsehood of a theory, and the identification of evidence capable of disconfirming it; the ability to justify what one's claim to be true.
Thayer-Bacon, 1993	The ability to be receptive and caring, open to others' ideas and willing to attend to them, to listen and consider their possibilities.
Brookfield, 1995	Identifying and checking the validity of one's assumptions and exploring alternatives for thoughts and actions. As a process, CT involves adults in recognising and researching the assumptions that undergird their thoughts and actions.
Atkinson, 1997	CT is a social practice.
Fisher & Scriven, 1997	Skilled, active interpretation and evaluation of observations, communication, information, and argumentation.
Halpern, 1998	CT is the use of those cognitive skills or strategies that increase the probability of a desirable outcome. CT is purposeful, reasoned, and goal-directed. It is the kind of thinking that is involved in solving problems, formulating inferences, calculating likelihoods, and making decisions. Critical thinkers use these skills appropriately, without prompting, and usually with conscious intent, in a variety of settings. They are predisposed to think critically. When we think critically, we are evaluating the outcomes of our thought processes – how good a decision is or how well a problem is solved.
Bailin, Case, Coombs, & Daniels 1999	CT is seen as analytic. It is a mean to arrive at judgements within a given framework or context. It is done for the purpose of making up one's mind about what to believe or do; a person engaged in thinking is trying to fulfil standards of adequacy and accuracy appropriate to thinking; thinking fulfils the relevant standards to some threshold level.
Scheffer & Rubenfeld, 2000	CT in nursing is an essential component of professional accountability and quality nursing care. CT exhibits these habits of mind: confidence, contextual perspective, creativity, flexibility, inquisitiveness, intellectual integrity, intuition, open-mindedness, perseverance, and reflection. CT in nursing practices cognitive skills of analysing, applying standards, discriminating, information seeking, logical reasoning, predicting and transforming knowledge.
Mingers, 2000	The discipline of being sceptical or questioning about statements, propositions or information.
Hatcher & Spencer, 2006	Thinking that tries to arrive at a judgment only after honestly evaluating alternatives concerning available evidence and arguments.
Willingham, 2007	Seeing both sides of an issue, being open to new evidence that disconfirms people's ideas, reasoning dispassionately, demanding that claims be backed by evidence, deducing and inferring conclusions from available facts, or solving problems.
Elder, 2007[4]	CT is self-guided and self-disciplined thinking which attempts to reason at the highest level of quality in a fair-minded way.
Epstein, 2008	CT is evaluating whether we should be convinced that some claim is true or some argument is good, as well as formulating good arguments.
West, Toplak & Stanovich, 2008	CT is rational thinking, in terms of epistemic and instrumental rationality.
Moore & Parker, 2009	A careful reason application in the determination of whether a claim is true or not.
Bailin, Battersby, & Clauss 2011	A careful reason examination to reach a reasoned judgement.
Mulnix, 2012	CT is an attempt to understand what a rationally justified belief is. As such, CT techniques evaluate some beliefs in light of others.
Johnson, 2014	The articulated judgment of an intellectual product arrived at on the basis of plus-minus considerations of the product in terms of appropriate standards or criteria.
Byrnes & Dunbar, 2014	CT includes the knowledge of the factors that could contribute to claims being inaccurate; it is the ability to recognise flawed reasoning or flawed arguments derived from claims; it also implies also being on guard against being guilty of the same tendencies.
Paul, A. S. 2014	CT is a process that uses a variety of approaches to solve identified problems and requires reflective thinking and the ability to utilize logical problem-solving.
Danvers, 2016	CT is a set of embodied practices that interact with the world and its relations.
Kuhn, 2019	CT is a dialogic practice that people commit to and thereby become disposed to exercise, more than an individual ability or skill.
Stendberg & Halpern, 2020	Attitude + knowledge + thinking skills

2.4 Future directions for empirical research on Critical Thinking

According to different authors (Norris, 1992; van Gelder, 2005; Moore, 2013), there is a lack of empirical basis around CT conceptualisation. Consequently, they argue that more research is necessary.

CT can be studied in contexts of everyday use to examine and begin to understand the factors that contribute to disposition, as opposed to competence to exercise it (Kuhn, 2019). A critical analysis of the empirical research results is also necessary to overcome conflictual ideas regarding CT, such as whether CT is a knowledge domain skill. Kuhn (2019) suggests moving from CT to more measurable concepts, such as argumentation and inquiry. This perspective is especially used in guidelines (U.S. common core standards, 2010; Next Generation Science Standards, 2013). Furthermore, a stronger interconnection with educational practices (e.g. Ennis, 2018) and understanding CT in a developmental perspective (e.g. Byrnes & Dunbar, 2014) could help to overcome the problem of abstract and opposing definitions.

The use of a qualitative, anthropological, and ethnographic method to explore different voices regarding CT is necessary (Chen, 2015; Moore, 2013). This approach could be particularly useful to reflect on the cultural meaning of the word "critical". In Western Culture, the word "critical" has at least three meanings:

> The term 'critical' in ordinary parlance means 'excessively negative'; A second sense of 'critical' is related to the idea of crisis: the patient is in 'critical' condition; negotiations have reached the 'critical' phase (…) The third sense of 'critical' is displaying good judgment about something. We believe that it is this third sense is the sense people have in mind when they discuss 'CT' as an educational ideal. (Johnson & Hamby, 2015, p. 424)

Chen explained that the first translation of the word "*critique*" in Chinese is believed to derive from the English translated version of Kant's essay *Critique of Pure Reason* from 1935. The Chinese word for critical has the most negative connotation of finding fault in something. Consequently, the word *critical* in Chinese is not embedded in pedagogical approaches, and more emphasis is realised on the logical and analytical word. Chen (2015) and Moore (2013) adopted a similar approach: they interviewed students and professors to investigate the implicit and cultural meaning related to CT.

We also need to empirically test situations in which people are required to use two or more inter-related skills to understand better the relation between CT and common overlapping construct (problem-solving, decision-making, creative thinking). The neuroscientific study could help to understand better the relation between CT in its relation with other cognitive functions, such as executive functions (de Acedo Lizarraga, de Acedo Baquedano, & Villanueva, 2012), emotional intelligence (Yao et al., 2018), and problem-solving (Tong et al., 2018). Comparing to other higher-order skills, such as creative thinking, the neuroscientific study related to CT is missing.

In the next sections, some examples of empirical research in CT will be illustrated.

3. Cases of empirical research on Critical Thinking

3.1 Bounded Critical Thinking

In recent years, cognitive scientists have started to test empirically observations that philosophers and educators, such as Lock, Bacon, and Dewey systematically described (see paragraph 1.1). Educational scientists are using this evidence to inform pedagogical practices and develop programs aimed at improving CT:

> People assess probabilities incorrectly; they display confirmation bias; they test hypotheses inefficiently; they violate the axioms of utility theory; they do not properly calibrate degrees of belief; they over project their own opinions onto others; they allow prior belief to become implicated in their evaluation of evidence and arguments. (Stanovich, West & Toplak, 2011, p. 357)

Although numerous evidence indicating that human behaviour can deviate from optimal standards exists, different explanations of reasons why it happens are present.

One possible explanation is the lack of skills and habits (Facione, 1990) to think critically: the tendency to exhaustively examine possibilities; the tendency to avoid my side thinking; knowledge of some rules of formal and informal reasoning; and good argument evaluation skills (Stanovich, West & Toplak, 2011).

However, research showed that also people with analytical thinking skills could fail to apply their skills and procedural knowledge in a specific situation. They could not use their skills because they do not recognise the need to do it like, for example, when a problem is framed through misleading wording, situations, and settings (Kahneman, & Tversky, 2013). According to Stanovich et. al (2011) the *override detection*, which can be defined as the ability to detect the situational cues indicating that

you need to use your CT skills in a specific context, is more related to thinking disposition than a cognitive ability or intelligence.

The human susceptibility to thinking mistakes can be partially explained by the *computational limitations* of the human cognitive system. CT activities, such as selecting information, assessing evidence, and making inferences require cognitive efforts. People constantly decide (with higher or lower awareness) whether a specific problem or task deserves the cognitive and critical effort. According to some authors, a human thinking default processing is not CT (Evans, & Frankish, 2009), but it is a thinking processing with these characteristics: it does not put a heavy load on central processing capacity, it does not require conscious attention, and it is automatically activated.

In an evolutionary perspective, human beings needed to develop a faster-thinking processing in order to solve specific adaptive problems, regulate emotions and apply acquired knowledge in similar situations. These strategies are commonly defined as *heuristics*: "rules-of-thumb that can be applied to guide decision-making based on a more limited subset of the available information. Because they rely on less information, heuristics are assumed to facilitate faster decision-making than strategies that require more information" (APA, 2017, Heuristics). Although heuristics can support decision-making in many cases, they can be improperly adopted in a situation where reflective and analytical strategies are more desirable. Heuristics, such as availability, anchoring, or representativeness can make people more vulnerable to make statistical errors or not recognise fallacies in arguments (Jacowitz & Kahneman, 1995; Kahneman, & Frederick, 2002).

Two commonly recognised fallacies are: formal and informal. Formal fallacies are "those arguments that derive their psychological persuasiveness from their superficial resemblance to valid deductive forms" (Zeidler, Lederman & Taylor, 1992, p. 440).

Philosophers, starting from Aristotle, have been describing formal fallacies for years. Currently, cognitive sciences try to understand under which conditions we are more susceptible not to recognise formal fallacies. An example of formal fallacy is the following:

"All flowers have petals. Roses have petals. Therefore roses are flowers."

This example is a formal fallacy because the structure of the argument is wrong from a logical perspective.

More specifically, this syllogism can be described as following: "X then Y. Z then Y. Z then X".

We experience, in our world, that roses are flowers, and for this reason, it is difficult to recognise that syllogism incorrectness. To recognise the fallacy, people must suppress the tendency to endorse a valid response because of the naturalness of the conclusion, which is that roses are flowers (Kahneman, 2003). On the other hand, informal fallacies can be correct from a logical perspective. However, a problem could be related to the content and the meaning of the elements used in an argument. In informal fallacies, ambiguous or misleading language can be used to deceive. Copi (1986) divided informal fallacies into two further sub-groups: fallacies of relevance and fallacies of ambiguity. Fallacies of relevance are arguments that deceive through the inclusion of at least one statement that is irrelevant to the conclusion. Examples of fallacies of relevance are (1) hominem arguments; (2) appeals to popularity; (3) appeals to authority; (4) circular reasoning.

Ambiguity fallacies can be related to an ambiguous word or term usage. An example is the equivocation case: the repeated use of a term, which implicates that the word is consistently used throughout an argument when the meaning behind each occurrence is not equivalent.

"Sure philosophy helps you *argue* better, but do we really need to encourage people to *argue*? There's enough hostility in this world."

In this sentence, the word "argue" is first used to mean something like "claim, reason and explain" while in the second case, denotes "fight, dispute". To recognise informal fallacies, it is necessary to understand the deep meaning of words according to the context in which the words are used.

It is agreed that one of the skills implied in CT is to analyse argumentation structures and assess their relevance and validity. By looking deeply at the language used in an essay or a public speech, it can be possible to detect clues useful to distinguish among established and demonstrated ideas, hypothesis or predictions.

Although most of the research in this field was focused on the assessment of someone else's argumentation or an external problem to solve, attempts to study people thinking mistakes when they produce their argumentation are also present.

Kuhn (2020) presented research in which she showed that people have difficulties in explaining something with two or more factors. Research showed that the ability to find a cause-effect relationship improves with age. Children can easily confuse a co-occurrence with a causal relationship. Adolescents and adults acquire the ability to distinguish between co-occurrence and

causal relationship. However, they seem to struggle with finding more than one cause to explain a specific kind of phenomenon.

In an experiment, Kuhn (2020) asked a group of people to explain why a middle-aged woman from the Western US voted Donald Trump. Seventeen people out of 24 found only one factor that explained the woman's decision. Usually, when people have to explain the reason why they voted someone, they describe more than one cause.

Suppose it is difficult to find out that different perspectives on the same problem are present. In that case, it could be challenging for people to build argumentation that justifies two different positions, to explain the differences and to reconcile divergent thinking, all necessary skills to be engaged in CT.

Another important constrain for CT is the difficulty to integrate knowledge that contradicts prior knowledge and conceptions. As mentioned before, according to some authors, CT is self-correcting (Lipman; 1987), and it requires being open to new evidence that disconfirms people's ideas (Willingham, 2007). However, scientific evidence shows that self-correcting, in many cases, is something difficult to achieve. The tendency of seeking or interpreting evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand is today defined by cognitive psychologists as "Confirmation Bias" or "My side-bias", identified years ago by Francis Bacon (Nickerson, 1988). Many experimental paradigms have been developed to study this kind of bias (Walton, 1960; van Brussel, Timmermans, Verkoeijen, & Paas, 2020). Dunbar, Fugelsang, and Stein (2007) tried to explain why students have difficulties in changing their misconceptions about a topic in Physics. By using fMRI techniques, they found out that when participants were presented with data that were inconsistent with their preferred theory, the anterior cingulate, precuneus, and dorsolateral prefrontal cortex showed increased levels of activation. They described two principal roles anterior cingulate in cognition, which are *error detection* and *inhibitory control*. On the other hand, when participants had information consistent with their previous theories, brain areas related to memories were activated. The authors conclude that this could explain why generally it could be difficult for people to abandon their preferred theory and also reorganise and absorb the old theory into a new knowledge system.

3.2 Critical Thinking as a meta-cognitive skill

Although the disagreement on CT definitions, most authors, who have addressed the topic, mentioned the role of metacognition. Besides, the meta-cognitive CT nature is one of the most agreed

assumptions both in Western (see Facione 1990) and Eastern conceptualisation (see Chen, 2015) of CT. Also, empirical studies show that meta-cognition levels can predict students score on different CT tests such as WGCTA (Magno, 2010) and CCTST (Sadeghi, Hassani, & Rahmatkhah, 2014). According to Lau (2015), CT implies the ability to think about thinking. Inhibiting automatic answers, recognising the reasons why people can be biased because of their previous beliefs and values are examples of the metacognitive efforts implied in CT. Awareness of how people's mind is functioning regarding strengths and weaknesses (heuristic and bias reasoning) is part of meta-cognition. Meta-cognition includes self-knowledge concerning skills and dispositions, self-regulation, and self-monitoring.

According to scientific literature, described by Lau, four main meta-cognitive aspects that affect CT are present:

- *Meta-conceptions or mindset*: what people think about cognition can affect their performance and learning. Some researchers (Dweck, 2015) show that students who believed their intelligence could be developed (a growth mindset) outperformed those who believed their intelligence was fixed (a fixed mindset). Mindsets are not fixed, and they can change (Dweck, 1986).
- *General knowledge about cognition*: knowledge should include information regarding (1) skills and actions necessarily involved in a CT task; (2) scientific knowledge about processes, such as reasoning and how reasoning performance might be affected by other factors (previous knowledge, motivation, attention, personal values).
- *Meta self-knowledge*: accurate self-understanding is important to know strengths and weaknesses and to identify areas of improvement.
- *Self-regulation*: how to monitor and control cognitive processes and resources effectively and develop cognitive dispositions and personality traits conducive to better thinking and learning, and other positive life outcomes.

Following Kuhn's early work (1999), metacognitive, rather than cognitive competencies, are most relevant to CT. She divided meta-knowing into three broad categories: metastrategic, metacognitive, and epistemological. The third category is poorly explored and more interesting in the adult education field. A belief about the knowledge nature could play a fundamental role in CT (Felton, & Kuhn, 2007).

Level	Assertions	Knowledge	Critical Thinking
Realist	Assertions are COPIES of an external reality.	Knowledge comes from an external source and is certain.	Critical thinking is unnecessary.
Absolutist	Assertions are FACTS that are correct or incorrect in their representation of reality.	Knowledge comes from an external source and is certain but not directly accessible, producing false beliefs.	Critical thinking is a vehicle for comparing assertions to reality and determining their truth or falsehood.
Multiplist	Assertions are OPINIONS freely chosen by and accountable only to their owners.	Knowledge is generated by human minds and therefore uncertain.	Critical thinking is irrelevant or useful only for dismantling absolutist assertions.
Evaluativist	Assertions are JUDGMENTS that can be evaluated and compared according to criteria of argument and evidence.	Knowledge is generated by human minds and is uncertain but susceptible to evaluation.	Critical thinking is a vehicle for evaluating the relative merit of assertions based on an identified set of criteria.

Figure 4 Epistemological understanding and Critical Thinking. Retrieved from (Felton & Kuhn, 2007).

Kuhn described four kinds of epistemological beliefs: *realist, absolutist, multiplist, evaluativists.* Figure 4 describes possible interconnections between these epistemological beliefs and CT. According to Kuhn's hypothesis, the recent empirical research findings (Hyytinen, Holma, Toom, Shavelson, & Lindblom-Ylänne, 2014) showed that students' epistemological beliefs were interwoven into their CT: students used CT as a tool (1) for enhancing understanding and (2) for determining truth or falsehood.

3.3 Critical Thinking: language, dialogue and argumentation

Traditional and mainstream CT perspective casts an individual working alone on a reasoning task. On the other hand, since the ancient Greeks, voices have raised the possibility that intelligence and, more specifically, reasoning can be understood in an interactionist perspective (Mercier & Sperber, 2011). Contemporary research has risen much concern about the idea that cognitive functions can be merely understood considering the inside perspective of an individual actor. Atkinson (1991) proposed the idea of "distributed cognition" rooted in the Vygotskyan idea, presented in the book *Mind in Society*. Vygotsky (1978) argued that every high-level cognitive function appears twice: first on the social level, and then on the individual level. Mercier and Sperber (2011) focused their attention on reasoning skills. According to them, reasoning processes have been developed because through argumentation, social interactions, and communication they are supported:

Humans rely on communication to an unprecedented extent within the Primate order (…) senders usually have incentives to lie, deceive and manipulate receivers. (…) So receivers evolve mechanism of epistemic vigilance that allows them to accept information discriminately. One of the means that can be used is to exchange arguments. (…) Claims that would otherwise have been automatically rejected can now be defended and properly evaluated. (Mercier and Sperber, 2011, p. 6)

Following this perspective, CT, which is a kind of reflective reasoning, is the cognitive ability that evolved in order to help senders find reasons, and receivers to evaluate them. Thus, the reasoning is rooted in argumentative and social practices. According to Kuhn (2019), CT is a dialogic practice that people commit to and thereby become disposed to exercise, more than an individual ability or skill. For example, CT is the activity involved when someone who advances arguments to support a claim, anticipates their defeasibility as a consequence of others' objections. Consequently, the argument quality can be assessed not merely regarding logic coherence but mainly based on its collaborative value as a contribution to social interaction (Grice, 1975). Grice states that the quality of interaction can be assessed through four indicators:

1. The maxim of *quantity*: a person tries to be as informative as they possibly can and gives as much information as is needed, and no more;
2. The maxim of *quality*: a person tries to be truthful and does not give false information or not supported by evidence;
3. The maxim of *relation*: a person tries to be relevant and says things that are pertinent to the discussion.
4. The maxim of *manner*: a person tries to be as clear, as brief, and as orderly as they can in what they say; and avoids obscurity and ambiguity.

Consequently, the argumentation quality can be assessed concerning its knowledge function shared among people. Kuhn (2019) presented her paper studies with the scope to prove the relation between CT and dialogical practices. Many other studies have emphasised the role dialogic and linguistic exchange in the CT development, and, often, dialogic interaction is supported by Communication Mediated Technologies (Newman, Johnson, & Cochrane, 1997; Garrison, Anderson & Archer, 2001b; Guiller, Durndell, & Ross, 2008; Poce & Amenduni, 2019).

In a broader perspective, a dialogical relation should not be considered only when two or more people interact with each other. A dialogical interaction can also be considered an interaction where ~~someone~~

a person tries to anticipate another argument, or an internal dialogue among new information, ~~and~~ evidence, and a person's knowledge system. Mediation tools support internal and interpersonal dialogues. Semantic means, such as "language, various systems of counting, mnemonic technique, works of art, all sorts of connectional signs and son on" (Vygotsky, 1981, p. 137) connect social and individual functioning. Verbal language is one of the most powerful and spread mediation tool used by human:

> If the problem is stated verbally, the meaning which the words convey, is the starting point for the solution. Each word brings up its own trend of association and the process of analysis and selection immediately starts. The interpretation of the language by the individual thus influences in a very significant way the individuals thinking. (Glaser, 1941)

3.4 Soft Critical Thinking: emotions, motivation and dispositions

The relationship between an emotion and a reason is a complex question that philosophers have tried to answer since the beginning of speculation about human beings. Are emotion and thinking separated? Do they compete? Can emotions support thinking? Or do emotions mislead thinking? Historically, three hypothetical descriptions of the relation between emotions and cognition have been discussed throughout the centuries (Thomson, 1998):

1. Emotion is a separate system related to two other systems in an organism, namely cognition and will (Plato, Kant, Mendelsohn, Leibniz).
2. Emotion is a grand system, a coordinator of all developing subsystems in an organism (Freud, Descartes).
3. Emotion is one of many components in a complex organism, which are in constant dynamic interaction with each other (Aristotle, Spinoza).

Recently, Walton (1989) studied negotiation and argumentation using informal logic and CT. Also, Gilbert (1995) pointed out that emotional, intuitive, and physical arguments ought to be considered legitimate and studied just as much as logical arguments. Martinovski and Mao (2009) proposed a model to understand the role of emotion in argumentation and rational thought, according to which emotions contribute to goals and strategies changes during an inter-individual exchange. However, with a few exceptions, the role that emotions play in argumentation and negotiation is generally neglected in CT literature. Walton explains that much of the literature surrounding fallacies warns against reliance on emotions too. Emotions are distrusted and labelled as logical fallacies. However,

empirical research results have often shown the opposite. According to Thomson (1998), CT can be enhanced by a particular kind of emotion: empathy. Empathy, indeed, could have a role in enhancing CT through debates. Nussbaum (1998) reported, for example, that students improved their ability in political argument by adopting a different point of view. Macrae and Milne (1992) found out that empathy intensifies the effects of counterfactual alternatives. Zhang and Zhang (2013) discovered that instructors' positive emotions had positive effects on Chinese and American students' behavioural and cognitive engagement, and CT. However, the effects were largely mediated by students' positive emotions. Moreover, Yao et. al (2018) indicated that people who have higher emotional intelligence exhibit more effective and automatic processing of emotional information and tend to be strong critical thinkers. Despite these preliminary results, more empirical research is needed to understand better the impact of non-cognitive skills on CT.

Other "soft" factors that could have an impact on CT are motivation and dispositions. Since the considerable effort entailed in CT (Alexander, 2014), a disposition to exercise it should not be regarded as a habit but rather as an intention and purpose. The ability to exercise CT counts for little if CT is not exercised. Having the ability does not necessarily mean that people will always adopt it. Many authors argue that people have dispositions to be critical and that they have personal traits to be more or less involved in CT (Ennis, 1987; Facione, 1990; Halpern, 1998). On the other hand, it was shown that people are more motivated to be critical in specific contexts, and they are less motivated in others. For example, a situation in which people commonly more motivated to use CT is when their side is challenged. In an experiment, Klaczinsky (2000) found out that higher-order reasoning was more used by participants to reject theory-incongruent evidence whilst heuristics were used to evaluate theory-congruent evidence.

3.5 Critical Thinking and Knowledge

Commonly, CT is considered as a tool for determining what to believe or do (Ennis, 1987). However, the CT function as a tool for enhancing knowledge understanding, exploration, and construction is less explored. Although different views of the relationship between CT and knowledge can be retrieved from the literature, this relation is acknowledged. The first kind of a relationship between CT and knowledge implies that people need to have some knowledge of a domain to think critically about it (McPeck, 1981; Byrnes and Dunbar, 2014). On the other hand, CT skills could support the deep understanding and evaluation of the knowledge domain. Other kinds of knowledge commonly related to CT are meta-self-knowledge and knowledge about cognition (see paragraph 3.3). Thus, the

concepts of "knowledge" and "knowing" are substantial aspects of conceptualising CT (Hyytinen et al., 2014).

People can have different views of what knowledge is and what knowing something means. Usually, these individual stances are called "epistemological beliefs", and some researchers pointed out that epistemological beliefs are interwoven with CT (Kuhn, 1999; Felton, & Kuhn, 2007; Hyytinen et al., 2014). For example, Hyytinen et al. (2014), by asking 10 university students to think aloud during the resolution of Collegiate Learning Assessment (CLA) tasks, found out two different epistemological beliefs that inhibit the adoption of CT strategies: (1) the trust in authoritative specialists and experts; (2) the trust in scientific method and proof. Students in the first category had difficulties in evaluating information and jumped to conclusions. One of them reported similar difficulties in studying in-depth for university exams. Also, in their view knowledge was uncertain, and they were not capable of evaluating it. On the other hand, although students in the second category described themselves as "critical" and "error seekers", they neither evaluated information from reliable sources nor recognised biased "scientific" sources. They also excluded at priori more than half of the provided documents. Students, who considered both objective and subjective knowledge as useful sources to achieve a conclusion or to improve their understanding, evaluated all the sources and adopted different CT strategies.

To sum up, authors from different backgrounds and time frames have given attention to the relation between CT and epistemological beliefs (McPeck, 1981; Kurfiss, 1988; West, Toplak & Stanovich, 2008; Kuhn, 2019). However, little has been done to explore, both theoretically and practically, the relation between epistemological beliefs and CT. Initial research confirms a relationship between epistemological beliefs and CT. Nevertheless, more research is required to understand this relationship better.

3.6 Conclusive remarks

The previous paragraphs provide examples of empirical research lines that could be explored to enhance CT understanding. Developing a complex theoretical understanding of CT regarding its relation with other constructs (rationality, metacognition, argumentation, language, emotions, and motivation) is necessary to develop pedagogical models and valid assessment methods. Without a strong theory and shared CT understanding, it would be difficult, indeed, to understand: (1) which

pedagogical strategies could be the most effective to promote CT; (2) in which way we should assess CT.

In the next chapter, I will focus on empirical pedagogical research to try to describe the impact of different educational strategies on CT development.

CHAPTER 2 ASSESSING CRITICAL THINKING: CHALLENGES AND OPPORTUNITIES

1. Introduction

In a review on CT assessment methods, Ku (2009) declares that "the conceptualization and assessment of CT are interdependent issues that must be discussed together: how CT is defined determines how it is best measured" (p. 71).

Although Ku's assertion may seem obvious, this is not always the case. Indeed, the theoretical background presented in research does not always justify the CT adopted assessment methods (examples can be retrieved in the following research: Ryser, Beeler, & McKenzie 1995; Saadé, Morin, & Thomas; Gloudemans, Schalk, & Reynaert, 2013). Also, authors who provide CT theoretical definitions do not always propose coherent methods to assess it. In the first part of this thesis, it was noticed that only 15 out of 39 CT definitions were explicitly associated with an assessment method (see Table 7, Chapter 1). According to economic (OECD, 2012), cultural (UNESCO, 2015), and educational research-oriented organizations (IEA, 2018), CT skills are considered a desirable learning outcome in all levels of education (HE included) despite the scepticism towards the possibility to assess and define CT objectively. Concerning the Bologna Declaration 1999, aimed at developing a comparable degree system among European countries, the Tuning Project identified different general and subject-specific skills to develop in HE students. Among general skills, CT related abilities, such as being critical and self-critical, searching, processing, and analysing information from different sources, are included (Gilpin & Wagenaar, 2008). OECD (2012) carried out the AHELO project, which included CT as one of the general skills that should be assessed at an international level. Thus, reflecting upon CT assessment choices is necessary at least for two reasons. Firstly, CT is considered a desirable learning outcome for European HE students, and it should be comparably recognised, according to the Bologna Strategy. Secondly, research is necessary to understand which teaching strategy can foster CT skills in HE.

As Rear (2019) asserted in his recent review, the assessment of CT has become a significant enterprise with several available standardised payable tests.

Assessment tests could be classified in different ways. Hyytinen, Nissinen, Ursin, Toom, and Lindblom-Ylänne (2015) differentiated a self-report from performance-based measurements. Moreover, the performance-based measurements can be classified into multiple-choice (MC) tests, questionnaires, and constructed response tasks (CRT).

Another way to classify CT assessment is to distinguish CT assessment tools focus between an outcome and a process (Garrison, Anderson & Archer, 2001a). In the first part of this research (Table 4, Chapter 1), it was showed that process-oriented CT definitions co-occur more with open-ended measures ($C = 0,26$); and outcome-oriented CT definitions co-occur more with closed-measures ($C = 0,35$).

In this chapter, I will explore the strengths and weaknesses of the different methods developed to assess CT. However, I will focus more on open-ended measures, and I will present reasons for this choice. Lastly, I will explain the need to develop and validate a new set of methodologies for the CT assessment, based on innovative methods, which can face challenges related to the assessment of open-ended answers.

2. Closed Measures – Standardised Assessment

Many standardised CT tests adopt MC format or other kinds of closed measures (e.g. self-report questionnaire).

As shown in the first chapter of this thesis (Table 4), closed-measures are commonly used to assess CT as a general skill ($C = 0,21$) and as an outcome ($C = 0,35$). Examples of these tests are the Watson-Glaser Critical Thinking Appraisal (WGCTA; Watson & Glaser, 1980), the Cornell Critical Thinking Tests (CCTT; Ennis, 1993), the California Critical Thinking Skills Test (CCTST; Facione and Facione, 1994), and the California Critical Thinking Disposition Inventory (CCTDI; Facione, Facione, and Sanchez, 1994). Cases for subject-specific MC tests used in different fields, such as biology (McMurray, Beisenherz, & Thompson, 1991), physics (Tiruneh, De Cock, Weldeslassie, Elen, & Janssen, 2017), and psychology (Bensley, Lilienfeld, & Powell, 2014) are present as well. However, as Lai noted (2011), the well-established assessment tests tend to focus on CT general skills rather than on subject-specific. Various national and international associations and organisations developed the following tests: the Collegiate Learning Assessment (CLA), developed by Council for Aid to Education (CAE)[6]; the Collegiate Assessment of Academic Proficiency (CAAP) Critical Thinking, developed by the American College Testing Program (ACT)[7]; and the ETS HEIghten Critical Thinking Assessment (Liu, Mao, Frankel, and Xu, 2016). Most of these tests focus on CT skills rather than on CT dispositions. Researchers use alternative instruments to measure CT dispositions: the Need for Cognition Scale (Cacioppo, Petty, Feinstein, & Jarvis, 1996) and the

---

[6] The CAE is a non-profit corporation established in 1952 in New York to increase private support to higher education with a view to increase student access. https://cae.org/solutions/
[7] https://www.act.org/content/act/en.html

adapted versions of NEO Five-Factor Inventory (Costa & McCrae, 1992). However, these instruments have not been specifically developed to measure thinking dispositions, and are, therefore, of limited explanatory power (Sosu, 2013).

Table 10 presents some of the most popular CT assessment tests based on closed-measures and their related operationalisation.

It is possible to see the overlapping among these tools regarding some CT sub-skills, such as reasoning, analysis, argumentation, and evaluation. However, the assessment tools also differ along a few dimensions, such as decision-making and problem-solving, included, for example, in the Halpern Critical Thinking Assessment. The use of MC tests could provide different advantages. Firstly, scoring is easier and faster; secondly, assessor's subjectivity and students' language proficiency should not bias the evaluation by contributing to guarantee higher reliability and validity of the measure. For example, Liu, Frankel, and Roohr (2014) reported that the reliability of the CLA 60-minute constructed response task section is only .43. The test-level reliability is .87, primarily driven by the reliability of CLA's 30-minute short MC section. However, in two separate reviews on CT assessment methods (Ku, 2009; Liu, Frankel, & Roohr, 2014), researchers raised concerns about the supposed higher reliability and validity of MC tests. Ku (2009) reported low internal consistency, poor construct validity, unstable reliability, and low comparability of the two CCTST subscales. Similar problems concern the WGCTA, with poor reliability levels and no clear subscale structure. Ku also reported a phenomenon: the studies conducted by researchers not affiliated with the authors of the tests tend to report lower psychometric quality of the tests than the studies conducted by the authors and their affiliates.

According to Liu, Frankel, and Roohr (2014), common problems with existing assessments include insufficient distinct dimensionality evidence, unreliable sub-scores, noncomparable test forms, and unclear evidence of differential validity across test takers groups. The authors reported that only a few studies have looked at the relationship of CT with behaviours, job performance, or life events. Besides these issues, some authors point out that the MC measures use cannot be proper for the higher-order skills assessment, such as CT; according to some authors, MC items can be answered without reading the respective text passage. MC tests may be answered merely by low-level processing, such as factual recognition and selection (Nicol, 2007).

A further concern regarding MC items is that they make test-takers select between pre-determined answers rather than allowing individualised responses as in CRT (Rauch & Hartig, 2010). Another weakness concerns students:

> A student may be able to recognise the correct answer that they would have never been able to generate on their own. In that sense, MC items can present an exaggerated picture of a

students' understanding or competence, which might lead teachers to invalid inferences. (Popham, 2003, p. 81–82)

Table 10 – Tests to Assess CT General Skills and Dispositions Based on Closed Measures.

Test	Format	Developers	Areas	Themes / Scales
California Critical Thinking Disposition Inventory (CCTDI)	Likert scale – extent to which subjects agree or disagree	Facione, Facione, & Sanchez, 1994	Dispositions, general	7 scales: (a) truth seeking; (b) open-mindedness; (c) analyticity; (d) systematicity; (e) confidence in reasoning; (f) inquisitiveness; (g) maturity of judgement
California Critical Thinking Skills Test (CCTST)	Multiple Choice (MC)	Facione & Facione, 1994	General skills	6 scales: (a) analysis; (b) evaluation; (c) inference; (d) deduction; (e) induction; (f) overall reasoning
Collegiate Assessment of Academic Proficiency (CAAP) Critical Thinking	Multiple Choice (MC)	ACT, 1989	General skills	3 areas: (a) analysing elements of an argument; (b) evaluating an argument; (c) extending an argument
Collegiate Learning Assessment (CLA)*	Multiple Choice (MC) and Constructed response task	Council for Aid to Education, 2000	General skills	The MC items assess: (a) scientific and quantitative reasoning; (b) critical reading and evaluation; (c) critiquing an argument
Cornell Critical Thinking Test (CCTT)	Multiple Choice (MC)	Ennis, 1993	General skills	7 dimensions: (a) induction; (b) deduction; (c) credibility; (d) identification of assumptions; (e) semantics; (f) definition; (g) prediction in planning experiments
Critical Thinking Disposition Scale	Likert scale – extent to which subjects agree or disagree	Sosu, 2013	Dispositions, General	2 dispositions: (a) Critical Openness; (b) Reflective Scepticism
HEIghten Critical Thinking Assessment HEICTA	Multiple Choice (MC), multiple selection, multiple choice, select-in-passage, inline choice, and composite items	Educational Testing Service (Liu et al., 2016)	General skills	5 dimensions: (a) evaluate evidence and its use; (b) analyse and evaluate arguments; (c) understand implications and consequences; (d) develop sound and valid arguments; (e) understand causation and explanation
Halpern Critical Thinking Assessment Using Everyday situations (HCTAES)*	Forced choice (MC, ranking or rating of alternatives) and open-ended	Halpern, 2013	General skills	5 skills: (a) verbal reasoning skills; (b) argument and analysis skills; (c) skills in thinking and hypothesis testing; (d) using likelihood and uncertainty; (e) decision making and problem solving
Watson – Glaser Critical Thinking Appraisal tool (WGCTA)	Multiple Choice (MC)	Watson & Glaser, 1980	General skills	5 scales: (a) inference; (b) recognition of assumptions; (c) deduction; (d) interpretation; (e) evaluation of arguments

*indicates that the test includes also CRT items. For more information related to CRT features of these tests, please refer to the Table 11.

Moreover, MC tests can never assess students' skills to synthesise or generate their answers (Popham, 2003). Lastly, all the tests based on MC (presented in Table 10) are chargeable, which limits their accessibility and their use in educational contexts.

To address the limitations of MC tests, researchers have developed alternative assessment methods, which involve the adoption of open-ended tasks. The next paragraph will focus on an in-depth description of standardised open-ended measures.

3. Open and Mixed Measures for Critical Thinking Assessment – Standardised Assessment

Open-ended measures are characterised by the requirements given to the examinees to create their answers to questions. In these measures, students usually need to analyse, evaluate and synthesise complex information and provide a reasoned explanation. It is possible to create more authentic contexts and assess students' ability to generate rather than select responses by using open-ended measures. Research has long established that the ability to recognise is different from the ability to generate (Shepard, 2000). These tasks are sometimes referred to as "authentic assessment" because they elicit the same thinking processes that individuals use when they solve complex problems in their everyday lives (Andrews & Wulfeck, 2014). Indeed, in real-life situations where CT skills need to be exercised, no choices are provided. Instead, people are expected to come up with their own choices and determine which one is preferable based on the question at hand. Thus, according to some authors, open-ended measures could provide a better proxy of real-world scenarios than MC items. Ennis (1993) was one of the first authors who highlighted the need to adopt open-ended measures for CT assessment. According to Ennis (1993), open-ended measures are necessary because MC tests are not comprehensive and miss much important CT elements:

> The MC tests can, to varying degrees, be used for (…) diagnosis, feedback, motivation, impact of teaching, and research. But discriminating judgment is necessary. For example, if a test is to be used for diagnostic purposes, it can legitimately only reveal strengths and weaknesses in aspects for which it tests. The less comprehensive the test, the less comprehensive the diagnosis. For a comprehensive assessment, unless appropriate multiple-choice tests are developed, open-ended assessment techniques are probably needed. Until the published repertoire of open-ended critical thinking tests increases considerably, and unless one uses the published essay test, (…) it is necessary to make your own. (p. 184)

Although it has been almost 30 years since Ennis said these words, the limitations of MC tests for CT assessment have not been completely overcome yet. Contrasting results have been found regarding the comparability of MC and CRT measures for CT assessment.

In a report of 2009, Klein et al. reported high correlation levels between different MC tests and CRT tests for CT (which varies from 0,79 to 0,93). Hyytinen et al. (2015), who found opposite results, compared the two measures used in the OECD's AHELO project for assessing CT skills: the CLA and an MC questionnaire from the Australian Council for Educational Research (ACER). The results showed that the correspondence between the CLA and the MC questionnaire was fully comparable in 45.5% of the students' test performance. Ten percent of the students had opposite test results. These students were further divided into two dissonant groups: (1) students with high MC scores but low CLA scores, and (2) students with low MC scores but high CLA scores. By analysing the CLA responses qualitatively, the authors found out that students' responses in the first group were comprised of isolated and reproduced facts.

In contrast, in the second group, the students' written responses indicated the in-depth material understanding. Based on these features they labelled these groups as (1) Superficial Processing (e.g. students reproduced or slightly modified portions of text sources, without explaining the content of the materials in their own words), and (2) Thorough Processing (e.g. students evaluated the quality of the information and considered its premises, as well as the implications of different conclusions). The authors found out that the reason why the "Thorough Processing" group obtain a low score in the MC questionnaire was not due to the wrong answers, but due to the high number of unanswered questions. The authors concluded that MC tests do not measure students' skills to produce arguments and to give reasoned explanations, which are the essential elements of CT. Although the scoring of the CRT might be challenging, the students' written answers reveal the level of processing and understanding.

Table 11 presents the most known standardised tests to assess CT, which employ open-ended measures. As shown in Table 10, CLA and HCTA presented MC items too; thus, they can be considered the multi-response format assessments. According to different authors (Ku, 2009; Liu, Frankel & Roohr, 2014), a measurement that elicits both open-ended and MC response formats should be pursued in CT assessment.

By looking at general features of these tests, it is possible to retrieve some common characteristics of the open-ended item format:

- They use ill-structured problems. Moss and Koziol (1991) explain that test questions should require students to go beyond the available information in the task to draw inferences or make evaluations. Besides, problems should have more than one plausible or defensible solution, and

sufficient information and evidence should be present within the task to enable students to support multiple views.

- They provide contradictory materials or sources and focus on controversial topics. Fischer, Spiker, and Riedel (2009) argue that CT is a "stimulus-bound phenomenon," meaning that certain external task features may impact whether CT is elicited in a given assessment context. They demonstrated that certain tasks types are more likely to elicit CT than others. The level of consistency, or lack of contradictions, within stimulus materials did have the primary effect; it is more likely to prompt CT while using inconsistent or contradictory materials than consistent and coherent stimulus materials.

Table 11 Validated Tests to Assess CT General Skills and Dispositions Based on Open-Ended Measures

Test	Format of the Open-Ended Measures	Developers	Reported validation evidences	Competences assessed
Ennis Weir Critical Thinking Essay Test (EWCTET)	Given an argumentative passage, the examinees have to evaluate the logic of the passage and defend their own argument.	Ennis & Weir, 1985	Ennis & Weir, 1985; Taube, 1997	Recognising formal and informal fallacies; individuating alternative solutions; assessing quality of the arguments and producing own arguments
International Critical Thinking Essay Test (ICTET)	Given a literary text (e.g. the Art of Loving, Erich Fromm) examinees are required to (1) paraphrase; (2) explicate; (3) analyse; (4) evaluate; (5) role-playing the author.	Paul & Elder, 2012 (first edition 2006)	Hollis, Rachitskiy, Van der Leer, & Elder, 2020	Reflecting, self-monitoring, summarising, exemplifying, synthetising, connecting with daily life experiences, explicating the thesis, analysing the logic, applying standards in writings
Collegiate Learning Assessment (CLA)	Given realistic problems, which include more or less relevant reading materials (e.g. letters, summaries of research reports, articles, graphs), examinees are asked to organise, analyse, synthesise and evaluate these multiple sources of information to arrive at a solution or explanation of a problem.	Council for Aid to Education, 2000	Klein, Benjamin, Shavelson, & Bolus, 2007; Klein et al., 2009; Aloisi, C. & Callaghan, A. 2018; Arum, Cho, Kim, & Roska, 2012; Zahner & James, 2015	Analysis and problem-solving; writing effectiveness; writing mechanics
Halpern Critical Thinking Assessment Using Everyday situations (HCTAES)	Given 20 everyday scenarios, respondents are first asked an opened-ended question (e.g. "Based on this information, would you support this idea? Explain why") which is followed by a forced choice question.	Halpern, 2013	Hau et al., 2006; Butler, 2012;	Verbal reasoning skills; argument and analysis skills; skills in thinking and hypothesis testing; using likelihood and uncertainty; decision-making and problem-solving

Although the importance of using open-ended measures in CT assessment, they are less widespread than closed measures because they present different disadvantages. The most important is the difficulty of scoring (Attali, 2014). The open-answer assessment is characterised as subjective and open to scoring bias because examinees' responses are traditionally scored by using human evaluation. The CRT scoring is also considered time consuming and expensive; a large amount of time and effort is needed to train scorers and to score the responses.

Research is focusing more and more on the CRT automated scoring to solve these issues. Before describing in more details the innovative methodologies, the next paragraphs will focus on an in-depth evidence analysis related to four validated tests to assess CT skills and dispositions, briefly presented in Table 11.

3.1 The Ennis Weir Critical Thinking Essay Test (EWCTET)

The Ennis-Weir Critical Thinking Essay Test (EWCTET) is one of the first tests developed to assess the CT ability in an open-ended format. This test includes a complex argument presented to an examinee who is asked to formulate another complex argument in response to the first one following the essay format. The test authors explained, "although the logical and psychological dimensions of critical thinking are not completely separable, this test with its scoring system emphasizes the logical dimension of critical thinking" (Ennis & Weir, 1985 p. 1). The areas of the CT competence covered by the EWCTET are the following: (a) getting the point; (b) seeing reasons and assumptions; (c) stating one's point; (d) offering good reasons; (e) seeing other possibilities; and (f) responding appropriately to and/or avoiding argument weaknesses. EWCTET includes different scenarios; in one of them (Figure 5), test takers are presented with a letter to the editor in which the writer argues for a ban on parking regulations. The letter includes eight paragraphs, and each paragraph exemplifies one or more formal and informal fallacies. After reading the letter, a test taker is required to write an essay evaluating arguments in each paragraph as a whole. According to the authors, the test can be used both for formative assessment and research purposes, with high school and college students.

In the test manual, the authors describe how to score the essay, providing possible examples of more and less correct answers. The essay should contain an analysis of each eight paragraphs. Each test taker analysis should be scored in the following way: -1 = *judges incorrectly or show bad judgement in justifying;* 0 = *makes no response;* +1 = *judges correctly, but does not justify;* + 2 = *justifies semi-adequately;* + 3 = *justifies adequately.*

The manual includes some taken precautions during the scoring process: (1) focusing on the quality of thinking in the responses, rather than on the expressions mode; (2) focusing on the logical jargon use in the responses.

THE MOORBURG LETTER

230 Sycamore Street
Moorburg
April 10

Dear Editor:

Overnight parking on all streets in Moorburg should be eliminated. To achieve this goal, parking should be prohibited from 2 a.m. to 6 a.m. There are a number of reasons why any intelligent citizen should agree.

1. For one thing, to park overnight is to have a garage in the streets. Now it is illegal for anyone to have a garage in the city streets. Clearly, then, it should be against the law to park overnight in the streets.

2. Three important streets, Lincoln Avenue, Marquand Avenue, and West Main Street, are very narrow. With cars parked on the streets, there really isn't room for the heavy traffic that passes over them in the afternoon rush hour. When driving home in the afternoon after work, it takes me thirty-five minutes to make a trip that takes ten minutes during the uncrowded time. If there were no cars parked on the side of these streets, they could handle considerably more traffic.

3. Traffic on some streets is also bad in the morning when factory workers are on their way to the 6 a.m. shift. If there were no cars parked on these streets between 2 a.m. and 6 a.m., then there would be more room for this traffic.

4. Furthermore, there can be no doubt that, in general, overnight parking on the streets is undesirable. It is definitely bad and should be opposed.

5. If parking is prohibited from 2 a.m. to 6 a.m., then accidents between parked and moving vehicles will be nearly eliminated during this period. All intelligent citizens would regard the near elimination of accidents in any period as highly desirable. So, we should be in favor of prohibiting parking from 2 a.m. to 6 a.m.

6. Last month, the Chief of Police, Burgess Jones, ran an experiment which proves that parking should be prohibited from 2 a.m. to 6 a.m. On one of our busiest streets, Marquand Avenue, he placed experimental signs for one day. The signs prohibited parking from 2 a.m. to 6 a.m. During the four-hour period, there was **not one accident** on Marquand. Everyone knows, of course, that there have been over four hundred accidents on Marquand during the past year.

7. The opponents of my suggestions have said that conditions are safe enough now. These people don't know what "safe" really means. **Conditions are not safe if there's even the slightest possible chance for an accident.** That's what "safe" means. So, conditions are not safe the way they are now.

8. Finally, let me point out that the Director of the National Traffic Safety Council, Kenneth O. Taylor, has strongly recommended that overnight street parking be prevented on busy streets in cities the size of Moorburg. The National Association of Police Chiefs has made the same recommendation. Both suggest that prohibiting parking from 2 a.m. to 6 a.m. is the best way to prevent overnight parking.

I invite those who disagree, as well as those who agree with me, to react to my letter through the editor of this paper. Let's get this issue out in the open.

Sincerely,

Robert R. Raywift

Figure 5 EWCTET letter scenario. Retrieved from p. 13, Ennis & Weir, 1985

Ennis and Weir (1985) provide the following information regarding the test validity and reliability:

- The inter-rater reliability ranged from 0,82 and 0,86. The results were calculated based on the test scoring of 55 students.

- According to the authors, content validity is guaranteed because the test presents a common situation in which CT skills are manifested. However, the authors did not provide any information regarding other kinds of validity, such as predictive, criterion, or concurrent validity.

Although the test authors prescribe to focus on the quality of thinking in the responses rather than on the mode of expressions, according to other authors the subjective scoring process could cause potential biases in favour of test takers who are more proficient in writing (Adams, Whitlow, Stover, & Johnson, 1996). Taube (1997) asserted that the effects of disposition on thinking performance might not be adequately revealed because the highly specific context and the strict structure could restrict test takers' responses. Since only a few studies investigate the EWCTET reliability and validity, more research to assess the test's psychometric properties is needed. Also, the test was thought to consider the logical CT dimension, in agreement with the first wave of CT theories identified by Paul (2011; see Chapter 1, Paragraph 1.2). Thus, EWCTET could not be a valid tool to assess other relevant CT sub-components, such as knowledge interpretation and creation (McPeck, 1985), epistemological belief, (Kuhn, 1999; Felton, & Kuhn, 2007; Hyytinen et al., 2014), or CT dispositions. The following paragraphs will show that CT tests have been further developed and validated to encompass other CT dimensions.

3.2 The International Critical Thinking Essay Test (ICTET)

Richard Paul was one of the first experts who emphasised both the argumentation structure and human traits, such as emotion, imagination, and creativity to build a CT theory. The International Critical Thinking Essay Test (ICTET), developed in its first edition in 2006 and updated in 2012, reflects Paul's emphasis.

The ICTET assesses students' ability to use reading and writing as tools for acquiring knowledge (Paul & Elder, 2012). The test is based on the idea:

> Educated persons routinely read closely and write substantively – to learn new ideas and to correct conceptual misunderstandings. To read closely is to construct accurately the meaning of the texts one reads. It involves constructing the thinking of an author in one's own mind, in such a way that were the author to hear the summary, he or she would say 'Excellent, you understand exactly what I was saying'. (Paul & Elder, 2012, p. 3)

According to Paul and Elder, close reading and substantive writing require the following intellectual abilities to:

1. Clarify purposes;
2. Formulate clear questions;
3. Distinguish accurate and relevant information from inaccurate and irrelevant information;
4. Reach logical inferences or conclusions;
5. Identify significant and profound concepts;
6. Distinguish justifiable from unjustifiable assumptions;
7. Trace logical implications;
8. Identify and think with multiple viewpoints.

Paul and Elder conceptualised that in-depth reading and substantive writing involve specular CT skills (Table 12). For this reason, texts produced by students, provide in-depth information regarding their level of CT skills.

The ICTET operationalisation of CT skills has some common elements with the EWCTET, such as analysis and evaluation of the text logic. However, ICTET also includes new CT components: metacognition (self-monitoring, self-evaluation), connection with the external world, empathy, and the application of what Paul and Elder defined "universal" standards. While doing the test, students are required to:

1. Paraphrase a text sentence by sentence (e.g. "State in your own words the meaning of each sentence you read");
2. Explicate the text thesis (e.g. "State the main points of the paragraph and then elaborate what you have paraphrased");
3. Explicate the text logic (e.g. "Express clearly the author's purpose, the most basic concepts, and conclusion");
4. Evaluate the text logic (e.g. "Assess what you read by applying intellectual standards to it.");
5. Role-play the author (e.g. "Role-play the principal author by constructing a dialogue between him/her and a questioner who asks him/her to explain various text positions").

The test's manual provides three text examples that can be used for the close reading and substantive writing activities: *The United States Declaration of Independence, On The Duty of Civil*

Disobedience[8]*,* and *The Art of Loving*[9]. For each text, the authors also include specimen answers, which should be not considered as right answers, but good examples. Paul and Elder, indeed, explain:

> It is important that both teachers and students understand that there are multiple ways to accurately paraphrase a text, to explicate the thesis of a text, to explicate the reasoning embedded in a text. What you will be assessing in student work is their ability to capture the essence of a sentence, phrase or text, the essence of the authors' reasoning and so forth. (2012, p. 11)

Table 12 Core CT Competences Declined in Close Reading and Substantive Writing Activities (Paul & Elder, 2012)

Core CT competences	Close reading activities	Substantive writing activities
Reflection	Students reflect as they read.	Students reflect as they read.
Self-monitoring	Students monitor how they are reading and distinguishing between what they do and do not understand in the text.	Students monitor how they are writing and distinguishing between what they do and do not understand in the text.
Summarising	Students paraphrase what they read (sentence by sentence); they accurately summarise in their own words texts they read.	Students accurately summarise in their own words texts they read.
Connecting with real world experiences	Students give examples, from their experience; they take the core ideas they obtain through reading and apply them to their lives.	Students give examples from their experience as they write; they write about ideas that apply to their lives.
Integration of concepts	Students connect the core ideas in a text to other core ideas they understand.	Students explicitly connect core ideas to other core ideas in their writing.
Explanation	Students state, elaborate, exemplify, and illustrate (SEEI) in writing their thesis.	Students state, elaborate, exemplify, and illustrate (SEEI) in writing their thesis.
Analysis	Students analyse the text logic of what they read.	Students analyse the text logic in their writing.
Evaluation / Self-evaluation	Students evaluate what they read: clarity, accuracy, precision, relevance, depth, breadth, logic, significance, and fairness.	Students apply universal standards in their writing: clarity, accuracy, precision, relevance, depth, breadth, logic, significance, and fairness.
Empathy	Students accurately role-play the authors' point of view.	

[8] Thoreau, Henry David. 1937. Walden and Other Writings. New York: The Modern Library
https://www.ibiblio.org/ebooks/Thoreau/Civil%20Disobedience.pdf
[9] Fromm, Erich. 1956. The Art of Loving. New York: Harper and Row.

The test authors invite teachers to individuate other kinds of excerpts by creating sample answers to support the assessment process (see examples in Poce, 2017). Students' answers should be assessed on a 1 to 10 point scoring scale, and the scale can be used for every single answer within a form or holistically, for the entire form.

The manual does not contain any information regarding the test reliability. The proper test use should lead to high *consequent validity,* which means that the test use brings teachers to teach how to foster close reading and substantive writing. Like in the EWCTET manual (Ennis & Weir, 1985), Paul and Elder did not provide any information regarding other kinds of validity, such as predictive, criterion, or concurrent validity.

Hollis, Rachitskiy, Van der Leer, and Elder (2020) have recently carried out the first validation study of the ICTET. They assessed the test for inter-rater reliability, internal reliability, and criterion validity. A hundred volunteers completed the ICTET online test, based on the Art of Loving by Erich Fromm, and the EWCTET online Moorburg Letter task. The authors found that the ICTET items inter-rater reliability varied from 0,441 to 0,785; the overall inter-rater consistency of total test scores was excellent. The test had good internal reliability, with Cronbach's alpha values varied from 0,816 to 0,953. Correlation between EWCTET scores was calculated to assess criterion validity. It showed a strong and significant correlation between scores on the ICTET and the EWCTET test, with $r(s) = 0,78$, $p < 0,001$. Factor analysis demonstrated that scores on ICTET items were best explained with one factor, suggesting that the test measures a single construct. In grading responses, the authors found that answers to the last question, which asked for the author's point of view, were often repeating what participants had already written in the previous answers. Similarly, participants' answers to the second question, which asked to identify the key questions addressed in the text, were generally incorrect and participants usually gave answers that were more suitable for question 1, which asked about the purpose of the text answers. Both the second and the last question had the lowest psychometric properties concerning the inter-coder agreement, inter-rater consistency, and internal reliability; and they had a relatively low loading to the single factor that best fit the data. Authors conclude that the ICTET is a valid CT measure and it could be submitted in a short-version by removing the second and the last item, especially when the test is submitted online or in time-constrains circumstances.

Despite the usefulness of these first validation results obtained by Hollis et al. (2020), more studies are necessary to consider them robust. Researchers who are not affiliated with the ICTET authors should carry out future investigations. Indeed, Ku (2009) warned about the risks related to higher

psychometric quality reported by the test authors and their affiliates compared to the non-affiliated researchers.

The test advantage is the adaptability to different stimulus, allowing teachers to personalise the assessment method according to their needs. I will present how the Paul and Elder's model inspired the assessment method design adopted and improved in this PhD thesis.

3.3 The Halpern Critical Thinking Assessment Using Everyday Situations (HCTAES)

The Halpern Critical Thinking Assessment Using Everyday Situations (HTCAES) is a multi-response format test designed to assess CT skills for respondents aged 18 and older, which could be applied in an educational and workplace context (Halpern, 2016). Dianne Halpern developed the test based on the following definition of CT:

> Critical Thinking is the use of those cognitive skills or strategies that increase the probability of a desirable outcome. Critical Thinking is purposeful, reasoned, and goal-directed. It is the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions. Critical thinkers use these skills appropriately, without prompting, and usually with conscious intent, in a variety of settings. That is, they are predisposed to think critically. When we think critically, we are evaluating the outcomes of our thought processes – how good a decision is or how well a problem is solved. (Halpern, 1998, p. 450-451)

Following this definition, Halpern clearly stated that both skills and dispositions compose CT:

1. *Dispositions*: Halpern explained that CT implies an attitude to recognise when a skill is needed and the willingness to apply it. She described the following CT dispositions: (a) willingness to engage and persist at a complex task; (b) habitual use of plans and inhibition of impulsive activities; (c) flexibility or open-mindedness; (d) willingness to abandon non-productive strategies in an attempt to self-correct; and (e) awareness of the social realities that need to be overcome. According to Halpern, dispositions are necessary to turn thoughts into actions.

2. *Skills*: Halpern proposed a taxonomy of 5 CT skills, which is adopted in HCTAES operationalisation: a) verbal reasoning (e.g. recognising the use of pervasive or misleading language); b) argument analysis (e.g. recognising reasons, assumptions, and conclusions in arguments); c) thinking as hypothesis testing (e.g. understanding sample size, generalizations); d) using likelihood and uncertainty (e.g. applying relevant principles of probability such as base rates); and e) decision-making and problem-solving (e.g. identifying the problem goal, generating and selecting solutions among alternatives).

Halpern operationalisation of CT skills has some common elements with the theoretical premises of the two tests analysed in the previous paragraphs, such as *argument analysis*. Like in the ICTET, metacognition is pivotal for Halpern (1998), especially regarding the outcomes self-assessment; people achieve it through CT processes.

However, Halpern did not include the ICTET empathetic element in her theory of CT. On the other hand, in the HCTAES, CT is supported for making better decisions and solving problems. For Halpern, CT is a synonym of *rational thinking*, thinking that fights against irrational thoughts, such as paranormal beliefs.

Although Halpern theories of CT include both skills and dispositions and she claimed the CRT portion of the HCTA attempts to reveal the dispositional component of thinking more; the HCTAES explicitly assessed only five, mentioned above, CT sub-skills. On the HCTAES website[10], it is explained that some CT sub-skills worth more points than others, in their contribution to the total CT score. Different skills' weights are presented below together with the rationale for their contribution to CT:

1. *Decision-making and problem-solving* sub-skills are weighted with more total points (approximately 31%) than the other categories because all CT sub-skills are involved to some extent in decision-making and problem-solving.

2. *Argument analysis* (approximately 23%) implies both the ability to produce reasons and to recognise them and their basic functions.

3. *Thinking as hypothesis testing* (approximately 22%) should not be restricted to formal research, but it needs to be adopted in everyday situations. Faulty thinking involves, for example, hasty generalizations from small samples of behaviour (e.g. if a new friend is late, the new friend must be habitually late).

4. *Likelihood and uncertainty* sub-skills are weighted lower (approximately 13%) than other CT sub-skills despite their importance. They need to be developed through formal instruction programs (e.g. statistics) which are not necessarily included in all universities curricula.

5. *Verbal reasoning* is also relatively low weighted (approximately 11%) to not penalise test takers whose native language is not English and because the connotation of words varies among languages.

---

[10] https://sites.google.com/site/dianehalperncmc/home/research/halpern-critical-thinking-assessment

The test consists of 20 everyday scenarios, four scenarios for each of the five sub-skills. In each scenario, respondents answer to both opened-ended questions and forced-choice questions (see an example in Figure 6).

The scenarios were taken from multiple disciplines, such as medical research, social policy analysis, and other numerous disciplines. These scenarios are examples of situations that might be found in newspapers and everyday conversations (Halpern, 2016).

Figure 6 HTCAES Sample Scenario. Retrieved from Halpern, 2016, p. 33.

The HCTAES offers four test forms in two versions: Version A (S1 and S2) and Version B (S3 and S4). Scenarios used in both versions are analogues, but questions are different. In this way, respondents can take the HCTAES twice without possible memory contamination of the test items. Forms S1 and S3 both use constructed responses and forced-choice alternatives, whilst forms S2 and S4 consist only of forced-choice items.

The HCTAES can be administrated both online or offline. The test administrator, with the use of grading prompts, can do the constructed responses grading. The use of grading prompts is one of the most innovative aspects of HCTAES. Unlike the other three CT tests under scrutiny (EWCTET, ICTET, and CLA/CLA+), the HCTAES does not provide an assessment rubric for grading the respondents' answers. Grading prompts, which are a series of simple questions presented to the rater, support scoring. In the computerised system, each open-ended answer is displayed along with a series of grading prompts.

The scoring module screen is divided into two parts. The upper part contains the scenario, the question posed to the respondents, and the provided answer; the lower part contains a series of simple questions presented to the rater. The rater evaluates to what extent a specific content matter is indicated in the respondent's answer by using the answer alternatives: "Yes" / "No" or "clearly indicated" / "less clearly indicated" / "not indicated".
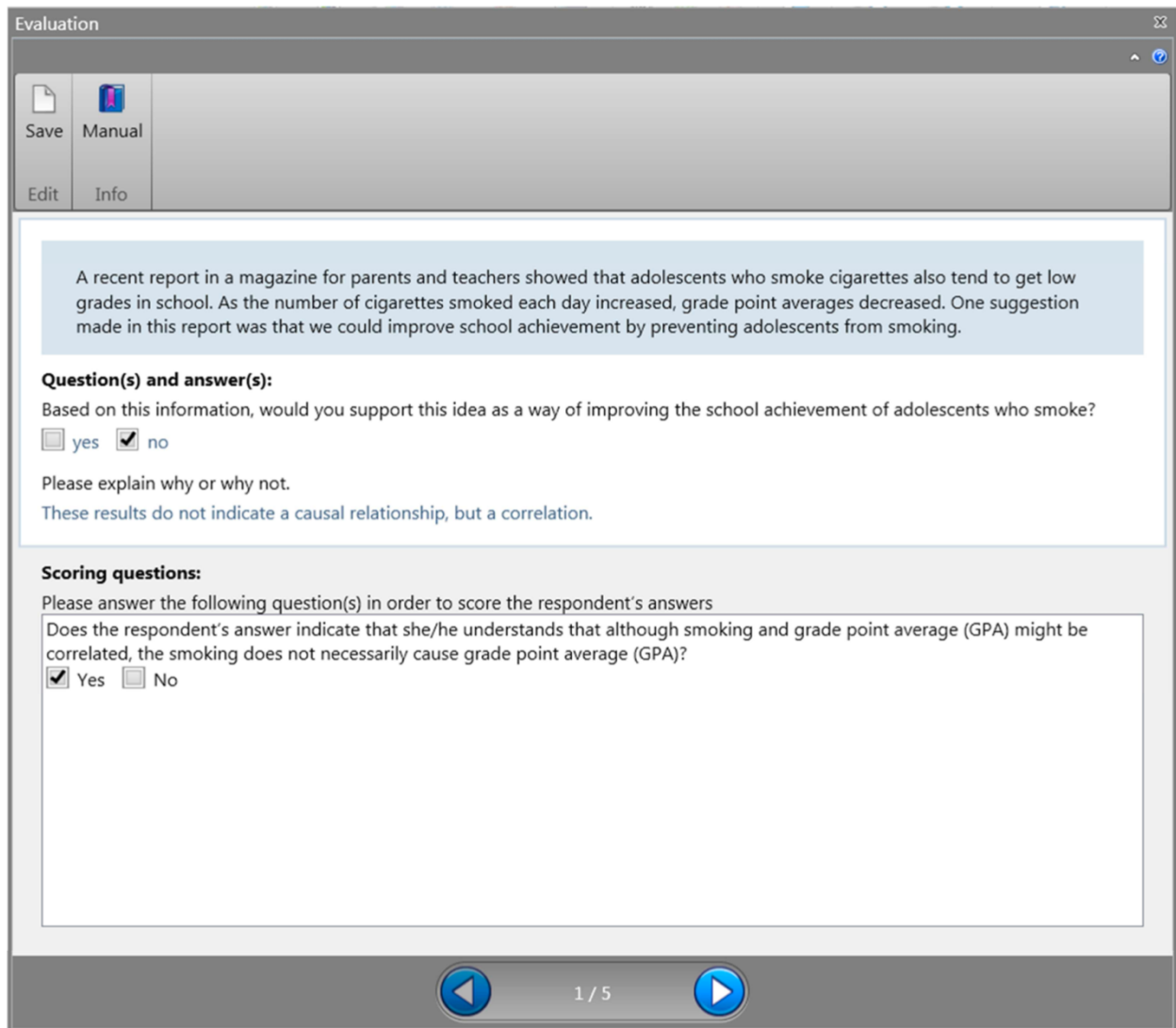
Figure 7 HCTAES Scoring Module. Retrieved from Halpern, 2016, p. 35

Raters are not required to assign general scores to the respondents' answers through the scoring module, and the grading prompts, like in tests that rely on scoring rubrics. The general score is calculated automatically by the scoring module based on the raters' answers to the prompts and questions posed on the lower part of the scoring module screen. The grading prompts are thought to enhance the scoring objectivity of the open-ended and constructed responses.

The manual includes numerous evidence regarding the test's validity and reliability properties.

The inter-rater reliability was calculated with a sample comprising 200 respondents aged between 18 and 72. The results indicate acceptable inter-rater reliabilities for the constructed responses, except *verbal reasoning* (r = 0,60), and *decision-making and problem-solving* (r = 0,53) scales. However, the inter-rater reliability of the main variable *critical thinking* is sufficiently high (r = 0.83). The measurement reliability was calculated with the same sample; Cronbach's coefficient alpha varies from 0,68 to 0,88, depending on the test form and the sample features. Correlation between the constructed responses and the forced-choice items' scores were calculated in different, small case studies (e.g. Verburgh, François, Elen, & Janssen, 2013). Correlation varies between a minimum of 0,39 to a maximum of 0,51. According to the test author (Halpern, 2016), these results suggest the skills required in the constructed response and MC tasks overlapping. However, the moderate level of correlation could suggest that two different constructs are evaluated by each type of question.

According to the test author, content validity is guaranteed by the clear and transparent correspondence between the CT definition provided by Halpern (1998) and the HCTAES dimensions. Construct validity was assessed in different studies through Exploratory and Confirmatory Factor Analysis (Hau et al., 2006; Ku et al., 2006). Results on the factor structure of HCTAES indicated the presence of 10 domains at the lowest level and two separable, though highly correlated, second-order factors representing "CT – free recall" (CRT) and "CT – recognition" (MC tasks).

Criterion validity was assessed in some research through correlational studies in which HCTAES scores were compared with different related-constructs:

- reasoning (measured through the Arlin Test of Formal Reasoning; see Halpern, 2007)
- personality traits (such as Need for Cognition, Conscientiousness, Openness to Experience, and Concern for Truth; see Ku & Ho, 2010)
- academic performance (measured through GRE, SAT, and GPA; see Hau et al., 2006)
- cognitive abilities (measured through Verbal Comprehension Index of the WAIS; See Ku & Ho, 2010)

Butler, in 2012, examined the relationship between HCTAES and the Decision Outcome Inventory (DOI), which measures the quality of respondents' decisions in various everyday life, in a group of 133 adults. Those with higher CT scores reported fewer negative life events than those with lower CT scores, $r(131) = -.38, p < .001$. Results also indicated that the predictive validity of the HCTAES open-ended sections exceed the predictive validity of the HCTAES MC sections and the CT total score. These results indicated that using the constructed response format improves the predictive

validity of the HCTAES considerably. Liu, Frankel, and Roohr (2014) raised some criticism towards Butler's study (2012) because no control for any measures of participants' general cognitive abilities was present. In a more recent study, Butler, Pentoney, and Bong (2017) explored whether CT ability or intelligence was a better predictor of real-life events. Community adults and college students (*n* = 244) completed the HCTAES, an intelligence test (INSBAT), and the real-world outcomes (RWO) event (the adapted version of DOI). Individuals with higher CT scores and higher IQs reported fewer negative life events. CT more strongly predicted life events than intelligence and significantly added to the variance explained by IQ. In a similar study, Franco, Costa, and Almeida (2017) tried to identify if different students profiles that are related to everyday negative outcomes (measured through RWO) result from the lack of CT (measured through HCTAES). Then, they examined whether CT predicted each student profile. The authors found that *thinking as hypotheses testing* and *argument analysis* dimensions of CT are relevant to predicting which students will or will not be "Risk-taking" (i.e. students who incur in risky behaviours, such as drinking alcohol, smoking cigarettes, or hitting something with their car); and which students will tend to feel "Lost in translation" (i.e. students who are experiencing difficulties in the transition or adjustment process to higher education; seeing it as a novel and challenging new phase, a true turning point in their lives).

Despite the amount of evidence produced on HCTAES' validity and reliability properties, researchers have raised some concerns. According to Possin (2013), HCTAES weaknesses concerns: (a) a limited number of informal fallacies addressed, besides the slippery fallacy and the false analogy; (b) accessibility, especially related to costs of the test license; (c) ambiguous and unclear wording in ten scenarios out of twenty. For example, Possin describes that in one item, the test subject is introduced to a proposal in the scenario and then asked to state a position on it, and "Explain your position". According to Possin, this task is ambiguous because it is not clear whether the test taker should provide an *argument*, an *explanation* or a *description*. The scoring module also awards points for restating one's position, thereby penalising those subjects who more efficiently detailed only their reasons. This reflection brings to a further test's limitation: the grading prompts proposed in the scoring module are a virtue to the extent to which the shared alternative answers are correct. Other authors consider the CT definition, used by Halpern, problematic. According to Johnson and Hamby (2015), one common mistake in the CT conceptualization is treating it as if it covered all "good" thinking. They think Halpern's theory is affected by this problem because she defines CT as "thinking that increase the probability of reaching a desirable outcome". Similarly, Byrnes and Dunbar (2014), highlighted a conflation between the CT definition and intelligence in Halpern's theorization (explicated in Halpern, 2007).

To sum up, the HCTAES is a high-structured test for CT, which exploits the benefit provided by the computerised assessment to improve its reliability through an innovative scoring module. The computerised system provides grading prompts to the rater, instead of scoring rubrics, which requires a higher level of assessors' expertise to be properly used. Its psychometric properties were largely explored, included its *predictive validity.* Despite the criticism raised towards Halpern's CT definition, it is explicated and interconnected with the HCTAES scenarios, which allows teachers and researchers to be aware regarding what they measure when they use the Halpern's test.

3.4 The Collegiate Learning Assessment (CLA)

The Collegiate Learning Assessment (CLA) is a computer-administered, open-ended test of analytic reasoning, CT, problem-solving, and written communication skills devoted to HE students. The Council for Aid to Education (CAE) developed the test in 2002. It uses real-world problem-solving tasks to measure students' CT skills. Several HE commissions, such as the U.S. Department of Education (2006); the Association of American Colleges and Universities (AAC&U 2005); and the American Association of State Colleges and Universities (AASCU, 2006), have endorsed the CLA. Also, OECD in the Analysis of Higher Education Learning Outcomes (AHELO) adopted it in 2012. The CLA can be currently considered one of the most widespread and recognised open-ended based CT test. The CLA original purpose was mainly institutional. It was designed to provide a summative assessment of the value-added by the HE institutions curricula respecting certain important learning outcomes. "The CLA's main goal is to provide information that will help HE institutions determine how much their students are improving and whether that improvement is in line with the gains of comparable students at other institutions" (Klein, Benjamin, Shavelson, & Bolus, 2007, p. 418). Between 2013 and 2014, CAE developed an updated version of the test, called CLA+. The CLA+ version is claimed to be sufficiently reliable to be used not only at the institutional level but also at the student and individual level (Aloisi & Callaghan, 2018). The time to complete the CLA+ test is 90 minutes maximum. Three sections compose the test:

1. *Performance tasks*: given realistic problems which include more or less relevant reading materials (e.g. letters, summaries of research reports, articles, graphs), students are asked to organise, analyse, synthesise, and evaluate these multiple sources of information to arrive at a solution or explanation of a problem; tasks derive from a domain of real-world jobs suggested by activities found in education, work, policy, and everyday practice.
2. *Analytical writing*: students are asked either to take a position on a topic or to critique an argument (Shavelson, 2008).

3. *Selected-Response Questions (SRQ)*: it is a 30-minute multiple-choice questionnaire (introduced in the CLA+ version).

The CLA tests are delivered on an interactive Internet platform that produces an online scoring and results report. The system also provides students with their scores on a confidential basis so that they can receive feedback on their performance (Shavelson, 2008). Answers to the essay type tasks are scored using Natural Language Processing (NLP) software whilst human experts score the performance tasks (Klein et al., 2007). Figure 8 illustrates an example of the CLA performance task.

You are the assistant to Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of Dyna Tech's sales force, recommended that Dyna Tech buy a small private plane (a SwiftAir 235) that she and other member of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation.

1: Newspaper articles about the accident
2: Federal Accident Report on in-flight breakups in single engine planes
3: Pat's e-mail to you & Sallys e-mail to Pat
4: Charts on SwiftAir's performance characteristics
5: Amateur Pilot article comparing SwiftAir 235 to similar planes
6: Pictures and description of SwiftAir Models 180 and 235

Please prepare a memo that addresses several questions, including what data support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups, what other factors might have contributed to the accident and should be taken into accont, and your overall recommendation about whether or not DynaTech should purchase the plane.

Figure 8 Collegiate Learning Assessment Performance Task Format

In this task, the president of the DynaTech (a company that produces electronic navigational and communication equipment for small aircraft) asks her collaborator (the test taker) to evaluate the pros and cons of purchasing a plane (called the "SwiftAir 235") for the company. Concerns about the SwiftAir 235 have risen after a crash. Students are required to evaluate the situation by reading and consulting different sources and information. Thus, students have to select which information is relevant; integrate these multiple sources; and provide a solution, decision, and recommendations. Students are invited to answer in a real-life manner by writing a report to their employer; the report includes their analysis, recommendations, and solutions supported by referring to the sources provided. Like in the previously described test (e.g. ICTET), only one possible correct answer and solution does not exist. The test developers provide the evaluators with different possible alternative and correct solutions, and reasoning paths that students could follow.

The CLA performance tasks are assessed through a scoring rubric (Figure 9) divided into three subscales, each situating one aspect of student performance on a proficiency level scale of 1 - 6. The three subscales are: *Analysis and Problem Solving*, *Writing Effectiveness,* and *Writing Mechanics*. Each level on each scale is associated with a performance criterion, reported in the following figure.

	1	2	3	4	5	6
Analysis and Problem Solving Making a logical decision or conclusion (or taking a position) and supporting it by utilising appropriate information (facts, ideas, computed values or salient features) from the Document Library	May state or imply a decision/conclusion/position Provides minimal analysis as support (e.g. briefly addresses only one idea from one document) or analysis is entirely inaccurate, illogical, unreliable or unconnected to the decision/conclusion/position	States or implies a decision/conclusion/position Provides analysis that addresses a few ideas as support, some of which is inaccurate, illogical, unreliable or unconnected to the decision/conclusion/position	States or implies a decision/conclusion/position Provides some valid support, but omits or misrepresents critical information, suggesting only superficial analysis and partial comprehension of the documents May not account for contradictory information (if applicable)	States an explicit decision/conclusion/position Provides valid support that addresses multiple pieces of relevant and credible information in a manner that demonstrates adequate analysis and comprehension of the documents; some information is omitted May attempt to address contradictory information or alternative decisions/conclusions/positions (if applicable)	States an explicit decision/conclusion/position Provides strong support that addresses much of the relevant and credible information, in a manner that demonstrates very good analysis and comprehension of the documents Refutes contradictory information or alternative decisions/conclusions/positions (if applicable)	States an explicit decision/conclusion/position Provides comprehensive support, including nearly all of the relevant and credible information, in a manner that demonstrates outstanding analysis and comprehension of the documents Thoroughly refutes contradictory evidence or alternative decisions/conclusions/positions (if applicable)
Writing Effectiveness Constructing organised and logically cohesive arguments. Strengthening the writer's position by providing elaboration on facts or ideas (e.g. explaining how evidence bears on the problem, providing examples, and emphasising especially convincing evidence)	Does not develop convincing arguments; writing may be disorganised and confusing Does not provide elaboration on facts or ideas	Provides limited, invalid, over-stated, or very unclear arguments; may present information in a disorganised fashion or undermine own points Any elaboration on facts or ideas tends to be vague, irrelevant, inaccurate or unreliable (e.g. based entirely on writer's opinion); sources of information are often unclear	Provides limited or somewhat unclear arguments. Presents relevant information in each response, but that information is not woven into arguments Provides elaboration on facts or ideas a few times, some of which is valid; sources of information are sometimes unclear	Organises response in a way that makes the writer's arguments and logic of those arguments apparent but not obvious Provides valid elaboration on facts or ideas several times and cites sources of information	Organises response in a logically cohesive way that makes it fairly easy to follow the writer's arguments Provides valid elaboration on facts or ideas related to each argument and cites sources of information	Organises response in a logically cohesive way that makes it very easy to follow the writer's arguments Provides valid and comprehensive elaboration on facts or ideas related to each argument and clearly cites sources of information
Writing Mechanics Demonstrating facility with the conventions of standard written English (agreement, tense, capitalisation, punctuation and spelling) and control of the English language, including syntax (sentence structure) and diction (word choice and usage)	Demonstrates minimal control of grammatical conventions with many errors that make the response difficult to read or provides insufficient evidence to judge Writes sentences that are repetitive or incomplete and some are difficult to understand Uses simple vocabulary and some vocabulary is used inaccurately or in a way that makes meaning unclear	Demonstrates poor control of grammatical conventions with frequent minor errors and some severe errors Consistently writes sentences with similar structure and length and some may be difficult to understand Uses simple vocabulary and some vocabulary may be used inaccurately or in a way that makes meaning unclear	Demonstrates fair control of grammatical conventions with frequent minor errors Writes sentences that read naturally but tend to have similar structure and length Uses vocabulary that communicates ideas adequately but lacks variety	Demonstrates good control of grammatical conventions with few errors Writes well-constructed sentences with some varied structure and length Uses vocabulary that clearly communicates ideas but lacks variety	Demonstrates very good control of grammatical conventions Consistently writes well-constructed sentences with varied structure and length Uses varied and sometimes advanced vocabulary that effectively communicates ideas	Demonstrates outstanding control of grammatical conventions Consistently writes well-constructed complex sentences with varied structure and length Displays adept use of vocabulary that is precise, advanced and varied

Figure 9 CLA+ Marked Scheme. Retrieved from Aloisi & Callangham, 2018

The CLA program uses two types of essay questions. The thirty-minute "break-an-argument" task presents an argument and asks students to critique it, including their analysis of the author's arguments validity rather than simply agreeing or disagreeing with the author's position (see Table 13 for an example). The forty-five-minute "make-an-argument" type prompts present an opinion to students about a general interest topic and ask them to respond from any perspective(s) they wish. One of these prompts is "In our time, specialists of all kinds are highly overrated. We need more generalists—people who can provide broad perspectives." Students are instructed to provide relevant reasons and examples to explain and justify their views.

Table 13 Example of a Thirty-Minute Break-an-Argument Prompt. Retrieved from Klein, Benjamin, Shavelson & Bolus (2007).

The University of Claria is generally considered one of the best universities in the world because of its instructors' reputation, which is based primarily on the extensive research and publishing record of certain faculty members. In addition, several faculty members are internationally renowned as leaders in their fields. For example, many of the English Department's faculty members are regularly invited to teach at universities in other countries. Furthermore, two recent graduates of the physics department have gone on to become candidates for the Nobel Prize in Physics. And 75 percent of the students are able to find employment after graduating. Therefore, because of the reputation of its faculty, the University of Claria should be the obvious choice for anyone seeking a quality education.

One of the most innovative CLA aspects is that the answers to the break-an-argument and make-an-argument prompts can be scored automatically through NLP programs. Klein (2007) reported study results in which the NLP reliability was tested. Students' answers to one of the *analytical writing tasks* were assessed both by human experts assessors and by two NLP algorithms developed by Educational Testing Service (ETS): e-rater[11] and c-rater[12]. In Klein's study, human assessors were provided with an assessment guide, which contained 40 separate items (graded 0 or 1) and a 5-point overall communication score. For the latter score, the assessors were asked to consider whether the answer was well organised; whether it communicated clearly; whether arguments and conclusions were supported with a specific reference to the documents provided; and whether the answer used appropriate vocabulary, language, and sentence structure. Readers were instructed to ignore spelling mistakes.

ETS built the e-rater algorithm for the communication score based on grades assigned by human evaluators; it contains modules for identifying the following features relevant to the scoring guide criteria: syntax, discourse, topical content, and lexical complexity. ETS's c-rater, designed for the short-answers assessment, was used to create scores for items 1 through 40.

The correlation between hand and machine assigned mean scores, on the make-an-argument and break-an-argument tasks, was 0.78 (Klein, 2007). This correlation result is close to the 0.80 to 0.85 correlation between two human assessors on these prompts, suggesting a good level of NLP technique reliability.

---

[11] https://www.ets.org/erater/about
[12] https://www.ets.org/accelerate/ai-portfolio/c-rater

Beside the Klein's study, no other studies that tested the reliability of the CLA NLP system were identified. Additionally, Klein did not describe, in the paper, the 40 items assessed through the c-rater system. Consequently, it is difficult to understand on which aspects human assessors and NLP system "agree". A lack of transparency related to the conceptualisation of CT in the CLA has been reported elsewhere. Shavelson (2008) explained that CT was originally conceptualised in the CLA regarding *broad abilities,* which means that complex tasks require integration of abilities that cannot be captured when divided and measured as individual components. However, Aloisi and Callanghan (2018) explained that little remains of this historical and theoretical heritage in more recent CAE documents. Today, CAE simple claimed that the CLA is well aligned with three definitions of CT, developed by Facione (1990), Bok (2006), and Pascarella and Terenzini (1991; 2005). In a recent report, Aloisi and Callangham (2018), examined the CLA theoretical issues together with supporting evidence and threats to the CLA validity and reliability:

1. *Construct Validity*: although authors argue that CLA is well aligned with those three CT definitions, CAE does not explain with sufficient clarity the kind of CT the assessment tries to measure. Indeed, the three CT definitions have some overlapping concepts, but also significant differences. For example, Pascarella and Terenzini, and Bok mention problem-solving in their definitions, but Facione does not. According to Aloisi and Callangham, the absence of an articulated theory of CT undermines the construct validity.

2. *Concurrent Validity*: moderate to high correlation was found between the CLA and other CT tests such as CAAP and MAPP (Klein et al., 2009), which support the CLA concurrent validity. However, Aloisi and Callangham argue that the CLA correlated with CT measures as strongly as it did with skills, sciences, writing, and mathematics measures. These findings should invite to reflect upon how CT, as assessed in the CLA/CLA+, is different from the general academic ability in reading, mathematics, and science literacy.

3. *Predictive Validity*: low scores in the CLA+ test are predictive of unemployment condition, getting into debts, and living at home with parents (Arum, Cho, Kim, & Roska, 2012). In a similar study correlation between CLA+ scores, post-graduate participation, and employment condition was found. However, Aloisi and Challangam warned that these results could be mediated by structural inequalities of the sample (student ethnicity, family income, and segregation level of secondary schools).

4. *Reliability*: the CLA+ has high Cronbach's alpha values (between 0,81 and 0,87) and moderate to strong inter-correlations (from 0,67 to 0,88) whilst year-to-year consistency was low to moderate.

Aloisi and Callangham concluded that a full validation of the CLA+ might reveal that these threats are not that serious to undermine the overall assessment validity. The authors encourage further research on this topic.

Despite reported limitations on validity and reliability, the CLA's psychometric properties are the most explored among all the CT tests based on open-ended tasks. Unlike other tests, such as the EWCTET and the ICTET, poorly tested for validity and reliability; teachers and researchers dispose of a high volume of information related to CLA properties, which should be considered in students' learning outcome analysis at the individual and institutional level.

Moreover, CLA has been trying to overcome one of the most reported limitations of open-ended answers, which is the cost of scoring, by incorporating an NLP system in answers' scoring.

4. Assessing Critical Thinking Processes – Open-ended and Qualitative Methods

All the four standardised assessment methods, described in Paragraph 1.2 of this chapter, assess CT as an outcome. The four methods are based on different CT definitions and conceptualisation that highlighted a disagreement on what CT is (for a more detailed explanation of the CT definition problem, look at Chapter 1 of this thesis). According to Kuhn (2019), to expand and improve a CT theory, it is necessary to "move away from conceptual definitions of critical thinking at an abstract level, in favour of definitions that are tied more closely to specific cognitive behaviours that can be identified and observed" (p. 147). Kuhn states that CT should be studied in contexts of everyday use, such as argumentative exchange or inquiry processes, to examine and begin to understand factors that contribute to a disposition to exercise it. In other words, focusing on CT as a process, rather than as an outcome, could support both researchers' comprehension of CT, and teachers' intervention through formative assessment (Clark, 2012).

In the next paragraphs, I will present some of the most effective methods to assess CT processes in learning contexts characterised by dialogical exchanges. I will also show the CMC role in supporting methodological and theoretical advancements related to CT.

4.1 Studying Critical Community of Inquiry through Content Analysis

Among the first authors who emphasised the importance of studying CT processes rather than outcomes were Garrison, Anderson, and Archer (2001, a). The need to understand how to build a *critical community of inquiry* within Higher Educational contexts, progressively shaped by the

adoption of computer-mediated communication (CMC), moved the authors' attention for CT processes. They stated:

> Critical thinking is both a process and an outcome. As an outcome, it is best understood from an individual perspective—that is, the acquisition of deep and meaningful understanding as well as content-specific critical inquiry abilities, skills, and dispositions. (…) The difficulty of assessing critical thinking as a product is that it is a complex and (only indirectly) accessible cognitive process. However, and most relevant here, from a process perspective, it is assumed that acquiring critical thinking skills would be greatly assisted by an understanding of the process. Moreover, it is assumed that facilitating the process of higher-order learning online could be assisted through the use of a tool to assess critical discourse and reflection. (Garrison, Anderson & Archer, 2001, a, p. 8)

CMC provided unprecedented and unique opportunities for studying and keeping track of CT processes. If CT as an outcome is better understood in an individual perspective, CT as a process can be retrieved in social interaction and dialogical exchange (Byrnes & Dunbar, 2014; Kuhn, 2019). Different online learning environments, especially discussion forums, have been an invaluable source of information for studying CT in its use context (Garrison, Anderson, & Archer, 2001, a, b; Newman, Webb, & Cochrane, 1995; Meyer, 2003; Wang, Woo, & Zhao, 2009)

Garrison et al. (2001, a) were pioneers in the study of CT in discussion forums and have inspired many other authors. The authors developed a model related to different CT phases that could be retrieved both in online and offline contexts of social and collaborative learning. This model starts from an in-depth observation and qualitative analysis of online learning communities in Higher Educational contexts, and it follows four main phases: (1) *triggering event*; (2) *exploration phase*; (3) *integration phase;* and (4) *resolution phase*. According to the model, every CT process starts with a triggering event. Teachers could set a triggering event, but, especially in more informal and democratic learning contexts, any group member may purposively or indirectly add a triggering event to the discourse. The triggering event is a problem-posing event and, therefore, is considered evocative and *inductive* by nature regarding a problem or an issue conceptualization. A triggering event activates an *exploration phase*, in which students start to grasp the problem's nature and move to a fuller and divergent exploration of relevant information. During this exploration, students experience an iterative shift between an internal dimension (the critical thought) and the social dimension, which implies the dialogical exchange. Since exploration consists of searching for relevant information, therefore, it reflects a *divergent* process. After defining what is relevant to the

issue or problem, students are expected to *integrate* the ideas generated in the exploratory phase. At this phase, students begin to assess the applicability of ideas regarding how well they connect and describe the issue under consideration. According to Garrison et al., the integration phase is more challenging to achieve, and students tend to feel more comfortable in a continuous exploration mode; later experimental work confirmed this empirical observation (Kuhn, 2020). The integration phase represents the attempt to achieve a possible solution and, therefore, it implies a *convergent* process. Eventually, students achieve a *resolution* of the starting dilemma. Students develop a solution, which can be interpreted as newly created knowledge that can be tested in a more or less direct way. It represents a commitment to a solution and *deductively* testing its validity.

Table 14 Qualitative Content Analysis Rubric. Adapted from Garrison, Anderson & Archer, 2001(a) p. 15-16

Phases	Indicators	Socio-cognitive processes
Triggering event (inductive)	Recognising the problem	- Presenting background information that culminates in questions
	Sense of puzzlement	- Asking questions - Messages that take discussion into a new direction
Exploration phase (divergent)	Divergence within the online community	- Unsubstantiated contradiction of previous ideas
	Divergence within a single message	- Many different ideas/themes presented in one message
	Information exchange	- Personal narrative, description, facts
	Suggestions for consideration	- Author explicitly characterises messages as exploration
	Brainstorming	- Adds to established points but does not defend/justify/develop addition
	Leaps to conclusions	- Offers unsupported opinions
Integration phase (convergent)	Convergence among group members	- Reference to previous message followed by substantiated agreement - Building on, adding to others' ideas
	Convergence within a single message	- Justified, developed, defensible yet tentative hypotheses
	Connecting ideas, synthesis	- Integrating information from various sources
	Creating solutions	- Explicit characterization of a message as a solution by participants
Resolution phase (deductive)	Vicarious application to real world	
	Testing solutions	
	Defending solutions	

Focusing on the process rather than the outcome has some relevant consequences for the CT assessment. Garrison et al. (2001, a), for example, refuse the idea to assess CT by applying absolute standards to students' written texts (see the universal standards mentioned by Paul & Elder, Paragraph 1.2.2, Chapter 2). The authors share Lipman's view (1987) according to which standards are negotiated and co-constructed within the contexts in which CT processes and practices are realised. The authors' method used to assess CT is a content analysis defined as "a research technique for the objective, systematic and quantitative description of the manifest content of communication" (Borg & Gall, 1989, p. 357). Based on the four model phases, Garrison et al. developed an indicators set defined as concrete examples of how the socio-cognitive processes of each phase manifest themselves in online discussion forums. For example, "motivate the agreement with someone else opinion" is a manifest indicator of a latent CT convergent process typical of the *integration* phase. Table 14 presents the rubric of the qualitative content analysis developed by Garrison et al. to assess and analyse online discussions. Each message in a discussion forum represented an analysis unit (Henri, 1992), which means that each message was coded with one of the indicators presented in Table 14.

In their study of 2001(a) Garrison et al. reported low to moderate level of reliability, with Cohen's kappa value from 0.35 to 0.74. The authors explain these results due to a challenge to assess latent CT processes based on manifest transcripts, and due to a limited number of messages ($N = 95$) codded in their preliminary experimentation.

Garrison, Anderson, and Archer's model is still used for studying CT processes in the online environment. Chou, Wu, and Tsai (2019) found that the Garrison and colleagues' model was the most adopted qualitative method for studying CT in e-learning settings, between January 2006 and November 2017, followed by Gunawardena, Lowe, and Anderson's (1997) interaction analysis model; and Newman and colleagues (1995, 1997) coding framework (see Paragraph 1.3.2, Chapter 2 for an in-depth analysis of this coding framework). Garrison, Anderson, and Archer (2010) argue about the need to conduct multi-methodological studies to augment the shortcomings of quantitative analysis with qualitative content analysis, and other methods, to triangulate results.

An example of a mixed methodological application of the Garrison, et al. original method can be retrieved in a recent study (Oh, Huang, Mehdiabadi, & Ju, 2018) in which the authors combined a qualitative content analysis (based on an adapted version of the rubric presented in Table 14) with Social Network Analysis to study the impact on discussion task design and the specific facilitation strategies on students CT and interaction dynamics.

4.2 Comparing online and offline Critical Thinking Processes through Content Analysis

Newman, Webb, and Cochrane (1995) proposed another very effective approach for the CT processes assessment. Similarly to Garrison and colleagues, they highlighted the need to look for CT signs in social contexts. "Critical Thinking is not just limited to the one-off assessment of a statement for its correctness, but a dynamic activity, in which critical perspectives on a problem develop through both individual analysis and social interactions." (Newman, Webb, & Cochrane, 1995, p. 4). Newman et al., similarly to Garrison and colleagues (2001, a, b), recognised CT as an iterative process between an internal and a social dimension. Their assessment model (1995) was inspired both by the four Garrison's stages (triggering, exploration, integration, and resolution) and Henri's (1992) indicators of cognitive reasoning (e.g. judging the relevance of solutions, making value judgments, and judging inferences). However, Newman et al. (1995) identified some issues in Garrison and Henri's methods, which brought them to develop a new approach for the qualitative content analysis of CT in a dialogic exchange:

1. Individuals in a group discussion are often at different stages in Garrison's CT process; thus, this makes difficult to trace a consistent trajectory through the stages by applying content analysis, especially in an online discussion.
2. Henri's indicators were too broad, although they could be divided into simpler well-defined criteria.
3. Both Henri's indicators and Garrison's stages do not attempt to evaluate the depth of these cognitive skills, distinguishing between critical value judgements and uncritical statements of values.

To face the third problem, Newman et al. (1995) developed a list of paired opposites indicators (Table 15). For each CT macro-indicator, they provided a sub-indicator of a surface/uncritical processing and one of an in-depth/critical processing.

Newman et al. (1995) coded both transcripts of tape-recorded face-to-face discussions and on stored transcripts of online discussion through the indicators presented in Table 15. They considered a unit of analysis phrases, sentences, paragraphs or messages, containing one unit of meaning, and illustrating at least one of the indicators (this approach of meaning definition unit was later criticised by Strijbos, Martens, Prins, and Jochems, 2006).

Macro-indicators	In-depth processing (+) indicators	Surface processing (-) indicators
Relevance (R)	Relevant statements	Irrelevant statements, diversions
Importance (I)	Important points / issues	Unimportant, trivial points / issues
Novelty (N)	(P) New problem-related information	(P) Repeating what has been said
	(I) New ideas for discussion	(I) False or trivial leads
	(S) New solutions to problems	(S) Accepting first offered solutions
	(Q) Squashing, putting down new ideas	(Q) Welcoming new ideas
	(L) New things brought by a learner	(L) Dragged in by tutor
Bringing outside knowledge/experience to bear on problem (O)	(Q) Welcoming outside knowledge	(Q) Squashing attempts to bring in outside knowledge
Ambiguities: clarified or confused (A)	(C) Clear, unambiguous statements	(C) Confused statements
	Discussing ambiguities to clear them up	Continue to ignore ambiguities
Linking ideas, interpretation (L)	Linking facts, ideas and notions; Generation new data from information collected	Repeating information without making inferences or offering an interpretation; Stating that one shares ideas or opinions stated, without taking these further or adding any personal comments
Justification (J)	(P) Providing proof or examples	(P) Irrelevant or obscuring questions or examples
	(S) Justifying solutions and setting out advantages and disadvantages of solutions	(S) Offering judgements and solutions without explanations or justification or offering several solutions without suggesting which is the most appropriate
Critical Assessment (C)	Critical Assessment of own or others' contributions	Uncritical Acceptance or unreasoned rejections
Practical utility (P)	Relate possible solutions to familiar situations	Discussing in a *vacuum* or suggesting impractical solutions
Width of understanding	Widen discussion	Narrow discussion

This formula was done to produce a measure that was independent of the participation quantity, reflecting only the quality of the messages.

For example, if a discussion, online or face-to-face, includes 48 relevant comments (*R*+) and three non-relevant comments (*R*-), the ratio is calculated as follows:

$$R = (48 - 3) / (48 + 3) = 0{,}88$$

A few years later, the authors (Newman Johnson, Webb, & Cochrane, 1997) presented a detailed description of the comparative results between online and face-to-face discussion regarding Critical and In-Depth processing in both. They found more positive ratios for *Bringing in outside information* (O); *Linking ideas and interpretation* (L); and *Important Ideas* (I) in the online discussion transcripts, but slightly lower ratios for *New information, ideas and solutions* (N). Newman et al. (1997) concluded by asserting that face-to-face discussion is better for a creative problem exploration and an idea generation whilst online discussion environments better support later stages of linking ideas, interpretation, and problem integration. These results have been partially confirmed in more recent research (Guiller, Durndell, & Ross, 2008), in which the authors adopted a slightly different system of codification.

More evidence of CT was found in the online discussion than in the face-to-face discussion, as a significantly higher proportion of utterances coded as *Justification with evidence* (J) were noticed in the online discussion. A higher mean was found in the online condition regarding utterances coded as "weighs evidence", although this difference was not significant. According to the authors, the asynchronous discussion groups do promote the use of formal research evidence to support opinions and arguments, comparing to face-to-face discussions.

Newman and colleagues (1995) highlighted some possible issues related to their CT assessment methods:

- A person with the subject knowledge (ideally the class tutor) is needed to mark discussions since indicators are strongly dependent on a specific domain.
- Some teachers found picking out examples of uncritical thinking hard, particularly those who assess work only by looking for positive points. Newman et al. highlighted the importance of training evaluators on a content analysis technique. Asking teachers to mark already scored transcripts before starting on their class could be a viable solution.
- Understanding where statements or points start and end can be challenging in face-to-face transcripts as people interrupt each other or continue across interruptions. It is rarely a problem in CSCL transcripts.

I will present how Newman, Webb, and Cochrane's model inspired the coding framework adopted and improved in this PhD thesis.

4.3 From Dialogic to Individual Argumentation: Assessing CT in Discussions and Essays

Dianne Kuhn (2019) has advanced a CT dialogic concept. In her latest definition, she considered CT a "dialogic practice people engage in and commit to, initially interactively and then in interiorized form with the other only implicit" (Kuhn, 2019, p. 146). Her CT definition is in line with Vygotsky's (1978) developmental theory of higher-order level cognitive functions that appear firstly on the social level and later on the individual level.

Kuhn has directly transferred this idea in her pedagogical approach for both the development and assessment of CT.

Kuhn and Crowell (2011) presented an intervention in which they combined dialogical activities with the assessment of individual essays. In this work, the authors used dialogical activities as a pedagogical tool and individual essays as a way to assess students' argumentative skills (similarly to Ennis & Weir, 1985). Kuhn and Crowell (2011) organised a pre-post essay assessment to test their hypothesis according to which computerised dialogical activities could promote individual students' argumentative skills. In the pre-test, students were asked to write an argumentative essay on the teachers' payoff topic. Students had to choose wheatear all teachers should get the same pay or teachers should be paid according to their years of experience. Moreover, students needed to motivate their choices. In the post-test, students were asked to express their position regarding euthanasia, explaining why doctors should or should not support patients' decision about ending their lives because of incurable illnesses.

Additionally, students were asked to indicate if they have any questions, which answers would help them in writing their essays, and to list all their questions.

Both the quality of the essays and the questions were assessed. More specifically, the students' essays were divided into *idea units,* and each idea unit was classified into one of four categories, from the lowest:

- *No argument:* the student expresses a position without properly motivate it;
- *Own-side argument:* the student include only positives of the favoured position;
- *Dual perspective argument:* the student includes negatives of the opposing position;
- *Integrative perspective argument:* the student includes negative of the favoured position or positives of the opposing position.

Moreover, students' questions were classified into two kinds of questions:

- *General:* questions that have the potential to affect judgement about the issue in general;
- *Case-based:* questions whose answers can have implications for the resolution of the specific case.

Kuhn and Crowell (2011) reported a good level of intercoder agreement, 88% for the teacher-pay essay (Cohen's $k = 0,76$), and 93% for the euthanasia essay (Cohen's $k = 0,91$). The essay's argument quality of the participants involved in computerised dialogical activities exceeded the comparison groups in which students participated in an intervention involving the activity of extensive essay writing practice, along with a whole-class discussion. The intervention group also demonstrated greater awareness of the evidence to argument relevance. In subsequent research, Kuhn et al. (2013) presented an articulated system to assess (1) students dialogical interactions; (2) students evaluation of argumentative passages; and (3) students' argumentative production (Table 16).

In more recent work, Kuhn (2019) highlights the importance to combine the dialogue use, mainly computerised, with essays both for educational and assessment purposes. She reported two significant differences between the electronic dialogues and the essays. In essays, students use a higher number of factual evidence (sources provided by the teacher in the task) and most of their statements are written to support their positions. In dialogues, in contrast, an average of one-third of evidence-based claims served the function of weakening the opposing position. Moreover, writers are more likely to draw on evidence from their prior personal knowledge rather than factual knowledge. Kuhn explains these results by suggesting that a dialogue demands attention to others. The social context of dialogues appears to engage arguers more profoundly and authentically, prompting them to bring what they already know to the exchange. In writing an individual essay, in contrast, the same dialogue participants keep primarily to the information provided to them as the most efficient way to complete their task. It is not because they knew of nothing else to bring to bear, as their quite different dialogue performance confirmed, they instead appeared not to recognise its relevance to the assigned task. Thus, Kuhn concludes that involving students in dialogical written activities could support generalization of argumentative skills from a social to an individual dimension. This transfer process could be tracked through qualitative content analysis of both electronically dialogic interaction and individual written essays. The electronic-dialogue method strength is that it supports students' ability to reflect on what they are saying, not constraining the natural course of these exchanges. Also, the resulting transcripts serve researchers by allowing to examine the evolving norms reflected in peer discourse over a sustained period (Kuhn et al., 2013).

Dimension assessed	Task	Indicators
Online dialogical interaction	Students are asked to participate in electronic discussions on different topics: home-schooling, students' expulsion from schools, animal rights, sale of human organs, and China's one-child policy.	*Meta-talks:* - meta-comprehension - meta-argument - meta-argumentation evaluative - meta-argumentation directive
Evaluation of dialogical arguments. Students were asked to evaluate the following dialogic argument. Students were instructed to evaluate each contribution to the dialogue individually, rather than evaluate the dialogue as a whole and comment on its strength or weakness as an argumentative move.	"*Pat*: Schools should do away with uniforms. They are a bad idea. *Lee*: I think students should have to wear uniforms because then they all look neat and orderly and it's better for learning. *Pat*: Students get tired of wearing the same thing every day. They like to express themselves by looking different from one another. It shows their personality. *Lee*: They have other ways to make themselves look different besides clothing. *Pat*: Some families don't have the money to buy uniforms. *Lee*: Schools usually have a fund for families who need help with school expenses."	*Quality of their evaluation* A respondent's evaluations of each of the six dialogic turns were categorised as: - *Good:* the student recognises the function of an argument and its strengths and weaknesses; - *Weak:* the student recognises an argument's strengths and weaknesses but not its function; - *Poor:* evaluation of the claim alone or no evaluation.
Construction of dialogical arguments Students are asked to construct a discourse between two competitors running for mayors. The arguments advanced needed to be comparative with the candidate's qualifications evaluated concerning the candidate's opponent and vice versa.	Ana Cruz and Maria Diaz are running for mayor of their troubled large city. Among the city's problems are high housing costs, teen crime, traffic, school dropout, and unemployment. Chuck and Doug are TV commentators arguing about who is the better candidate. Write a script of what they might say. They are both experts on the city; they are both expert arguers and evenly matched. So your script should present the most well-argued debate you can construct. Begin your script like this: CHUCK: Cruz should be elected mayor because she'll do better than Diaz. DOUG: I disagree, because xxxxxx Then continue their argument, filling in what each one might say: CHUCK: xxxxxx DOUG: xxxxxxx CHUCK: xxxxx etc. Here is some information about Cruz' positions. She promises to: create job training programs, expand city parks, raise teachers' pay, open walk-in health clinics, reduce rents, impose a teen curfew, employ senior citizens in city schools Here is some information about Diaz' positions. She promises to improve public transportation, open more centers for senior citizens, revise the high school curriculum build a new athletic stadium, improve health care, build more housing.	*Use of evidence:* - No evidence use - Single evidence used to support the position or criticise other position - The integrative use of evidence to support the position or criticise other position; - The integrative use of evidence to compare positions *Kinds of arguments used:* - *Unconnected:* statements that do not have relation to the opponent's proceeding statement - *Counter-alternative:* statements connected to the opponent's argument by proposing an alternative argument - *Counter-critique:* statements that weaken the opponent's claims

5. Road to Critical Thinking Automatic Assessment

In the previous paragraphs, I tried to argue about the importance of combining MC items with open-ended formats for the CT assessment. Open-ended measures are primarily used in teachers daily practice assessing CT related skills and learning outcomes. However, these measures are still poorly used for large-scale studies such as an international survey (the AHELO survey from OECD (2012) is an exception). The scoring difficulties, the enormous amount of time and effort required to train scorers and to score the responses are among the reasons why MC items are preferred for large-scale studies. Liu, Frankel, and Roohr (2014) consider automatic assessment a viable solution to overcome the limitations of open-ended measures.

Automatic assessment of learning outcomes is a "hot topic" in educational research for at least two reasons: firstly, the availability of learning data is growing exponentially due to the spreading of online education. Secondly, researchers in the field of *Big Data*, *Machine Learning,* and *Artificial Intelligence* can provide educators with sophisticated tools for processing an immense amount of linguistic and behavioural data. In the intersection between educational and computer sciences, two primary research approaches are identified: educational data mining and learning analytics.

> Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students and the settings in which they learn. (…) Learning analytics is closely related to the field of Educational data mining and is concerned with the measurement, collection, analysis, and reporting of data about learners and their contexts for purposes of understanding and optimizing learning and the environments in which it occurs. (Kop, Fournier, & Durand, 2017; p. 320)

Siemens and Baker (2012) clarified the similarities and the differences between Educational Data Mining and Learning Analytics Approach. Both approaches analyse similar kinds of data and require comparable researcher skill-sets.

The first difference concerns the goals of these two approaches. Educational Data Mining has a considerably greater focus on automated discovery. Researchers in this field are interested in describing, exploring, and understanding learning. On the other hand, Learning Analytics has a considerably greater focus on supporting human judgment and decision-making.

In line with the greater focus on automated discovery, one of the most crucial applications of Educational Data Mining is the intelligent tutoring system, an application in which computers perform automatically traditional teachers and instructions functions (e.g. providing feedback). By contrast, Learning Analytics models are more often designed to inform and empower instructors and learners. In the next sections, I will present some methods used in both Learning Analytics and Educational Data Mining to assess Learning Outcomes focusing on CT related skills.

5.1 The Use of Natural Language Processing for the Automatic Assessment of Students' Written Texts

Natural Language Processing (NLP) is an analysis of a human language by using computers aimed at automated discourse analysis. The term "natural" was coined to refer to human language in contrast to computer languages. NLP techniques can provide information about multiple levels of text: from the simplest level constituted by the analysis of single words used in a discourse, to the more complex levels which are the semantics as well as the discourse structure (McNamara, Allen, Crossley, Dascalu, & Perret, 2017).

Linguistic Inquiry Word Count (LIWC) is a tool based on the statistical analysis of single words used within a text. The application adopts a quantitative word count approach that aims to reveal the meaning of words taken out of the context from their original setting. Once a text has been processed, the application produces a list of categories and percentages that can be directly read in application programs. The developers carried out an experiment which suggested that the program can detect attention focus, emotional status, and thinking styles by analysing the words used in a text (Tausczik & Pennebaker, 2010) as well as the meaning of words. LIWC software has been recently used to analyse analytical thinking in discussion forums of MOOCs (Moore, Oliver & Wang, 2019).

Among different NLP features, *n-grams* are growingly used in the field of automatic text analysis; they can be defined as groups of characters or words. The letter "n" refers to the number of grams included in the group. For instance, by using the term "bi-grams", we can refer to groups of two words or syllables. N-grams are used among the linguistic features in ETS' c-rater-ML to automatically calculate the short-answers score (Heilman & Madnani, 2015). C-rater-ML specifically calculates words unigrams, words bigrams, and character n-grams (sequences of 2 - 5 characters).

The n-grams and the word count approach allow analysing the explicit content of the text. However, when evaluating the text *relevance* related to a set of concepts, information regarding the latent meaning behind the words is crucial. Latent Semantic Analysis (LSA) is a technique that provides

means to extract semantic meaning from texts and compare text samples for semantic similarities (McNamara et al., 2017). Besides meaning, many other language features can be used to train algorithms in measuring the quality of a given text: parts of speech, syntax, cohesion, and syntax complexity. This information is computed through machine learning techniques to predict learning outcomes.

Among these NLP features, some have been recently used to predict CT related-skills, such as argumentation (Zhu, Liu, & Lee, 2020), reflective writing (Ullman, 2019), discourse coherence quality (Burstein et al., 2013), and the use of evidence (Rahimi et al., 2017). Most of these studies applied NLP to English written texts, and there are only a few attempts to generalise these techniques to other languages.
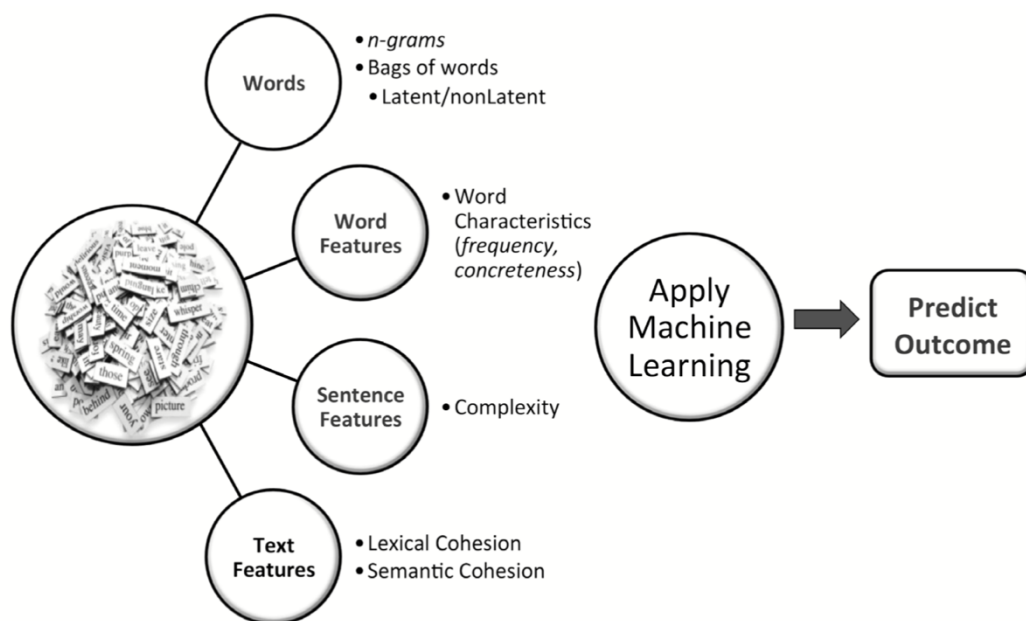
Figure 10 The Work-Flow of a Natural Language Processing System for the Prediction of Learning Outcomes Presented by McNamara et al. (2017)

5.2 Automatic Content Analysis as a Tool to Assess CT

As discussed in previous paragraphs, a tradition of content-analysis-based human interpretation and coding for CT assessment in essays, open-ended answers, and CMC is present. Recently, the application of the information-mining technique to extract semantic meaning from texts has become a prominent trend in the so-called field of Learning Analytics and Educational Data Mining. Computerised content analysis methods are typically based on Latent Dirichlet Allocation (LDA). The LDA assumption is that each document mixes with various topics, and every topic mixes with

different words. In LDA, two layers of aggregations are considered. The first layer is the distribution of categories, and the second layer is the distribution of words within the category (Ma, 2018); the given word order in a text is not considered. On the other hand, methods like the Network Text Analysis (NTA) considers the words positioning in a text by connecting the content analysis with network representations. Links between concepts are established if two words co-occur with a particular frequency. Specific kinds of dictionaries allow connecting words with a related concept category (e.g. subject, domain, place). Thus, networks are formed on the basis of concept-concept relations. These relations can be analysed by using indexes of Network Analysis, such as centrality and cohesion (Hoppe, 2017).

A further automatic content analysis development is called *content analytics*. Whilst traditional content analysis is mainly aimed at assessing latent variables of written texts, content analytics includes different additional analysis forms, such as assessment of student writings, automated student grading, or topic discovery in the document corpora (Kovanovic, Joksimovic, Gaevic, Hatala, & Siemens, 2017). Kovanovic et al. (2017) reviewed the application of content analytics related methods and discovered that one of the earliest application domains was the student essays analysis, also known as *automated essay scoring* (AES).

Based on their analysis, the authors found that the most widely applied technique for automated essay scoring is Latent Semantic Analysis (LSA), used to measure the semantic similarity between two text bodies through the analysis of their word co-occurrence. Regarding AES, LSA can be used to calculate the resemblance of an essay to a predefined set of other essays and the internal document similarity, often considered as a document coherence. Based on those similarities, a numeric measure of the essay quality can be calculated. Another commonly adopted method for AES is the graph-based visualisation, also based on a text's word co-occurrences. Besides approaches based on word co-occurrences, linguistic and rhetorical essays' analysis has been used to assess the quality of *argumentation* (Simsek, Buckingham Shum, Sandor, De Liddo, & Ferguson, 2013). Similar content analytics are used for other types of student-written texts, for example, short answers and online social interactions (e.g., chat, forums). In short-answers cases and AES, a set of "golden-answers" can be used to facilitate the work of automatic scoring systems.

Content analytics feedback systems have specifically been designed for both assessment and instructional purposes. Analytics provides teachers and instructors with visualisation aimed at supporting decision-making and real-time interventions. For example, Lárusson and White (2012) used student essays visualisations to inform instructors about the originality in student writings and

particular points in time when students started developing CT. Similarly, Wegerif et al. (2010) described a computational model to identify moments within e-discussions in which students adopted critical and creative thinking for informing instructors. Other methods to automatically assess CT in online discussions were based on Garrison, Anderson, and Archer's model (2001a, 2001b). McKlin, Harmon, Evans, and Jones (2001) developed a neural network classification system to automate discussion message coding based on the four phases described in Garrison's model: triggering, exploration, integration, and resolutions.

More recently, different studies (Kovanović, Joksimović, Gašević, & Hatala, 2014; Waters, 2015) examined the use of different text-mining techniques for coding messages based on the four stages of the Garrison's model. Kovanović et al. (2014) developed an algorithm that detected Garrison's CT related processes with the accuracy of 58.38% and Cohen's kappa of 0.41. The authors developed their algorithm by computing different linguistic features (i.e. n-grams, part-of-speech n-grams, linguistic dependency triplets, the number of mentioned concepts, and discussion position metrics).

Moreover, Kovanović et al. study (2016) showed that metrics provided by the Coh-Metrix (Graesser, McNamara, & Kulikowich, 2011) and LIWC tools, (in combination with further NLP and discussion-position features) could reliably classify the Garrison's stages almost as accurate as human coders, with the accuracy level of 0,72 and Cohen's kappa of 0,65. Authors described which NLP features were most useful to predict Garrison's CT stages. The most important was the word number in a message: the longer a message was, the higher the message chance was to be in *integration* or *resolution* phase. Also, the number of paragraphs and sentences; and average sentence length showed similar trends, with higher values associated with the later phase of Garrison's model. The standard deviation of the syllable number, which is an indicator of the words different length use, had the strongest association with the *triggering* event phase. In contrast, the *givenness* (i.e. how much in-text information is previously given) had the highest association with the *resolution* phase messages. Finally, the low Flesch Kincaid Grade level readability score and the low overlap between verbs used had the strongest association with messages non-relevant to CT processes. The most important LIWC features were (1) the number of question marks used, which was strongly associated with the *triggering* event phase; (2) the number of first-person pronouns, which was highly associated with messages non-relevant to CT processes; and (3) the use of money-related words, which were mostly associated with the *integration* and *resolution* phases. The authors concluded that while further improvements are needed before educational researchers can widely adopt this system, the progress is promising and has the potential to advance research practices in content analysis.

6. Conclusions

Developing a computational model to identify CT levels in students written comments automatically could provide many advantages. For instance, an automatic program could assist researchers and teachers in finding CT key aspects in immense amounts of data in Learning Management System platforms. In the Learning Analytics field (Siemens & Baker, 2012), a growing number of studies have been focusing on the big corpus of linguistic data automatic analysis (Ezen-Can, Boyer, Kellogg, & Booth, 2015; McNamara et al., 2017). Nevertheless, before adopting these tools to assess CT automatically, the accuracy of automated scores need to be examined. Indeed, it is necessary to be sure they achieve an acceptable level of agreement with valid human scores. However, only a few studies have evaluated the accuracy of the automatic scoring test for CT Assessment (Mao et al., 2018).

Another critical limitation concerns the limited evidence regarding non-English languages. Indeed, all the studies presented in the previous paragraphs (1.4.1 and 1.4.2, Chapter 2) applied NLP features to English written texts and few attempts to generalise these techniques to other languages.

CHAPTER 3 CRITICAL THINKING AUTOMATIC ASSESSMENT IN OPEN-ENDED ANSWER: A PILOT STUDY CARRIED OUT WITH HE TEACHERS

1. Introduction

The first two chapters of this thesis have shown that an evaluative approach to the study of CT represents a challenge in terms of operationalising the construct. In this third chapter, I start to undertake an exploration of the issue of CT assessment through automatic written text analysis, by presenting the results of a pre pilot-experimentation.

I will firstly describe the operative CT definition adopted in this thesis, based on the scientific literature of the last few years. Then, I will present in detail an assessment method developed and implemented by the *Roma Tre University - Centre for Museum Studies* (from here CDM) for the evaluation and analysis of CT levels within constructed response answers (Poce, 2017). This model was used as theoretical framework to design a prototype, based on NLP techniques (from here NLP prototype), able to automatically evaluate CT sub-skills. In collaboration with the CDM research group, led by Professor Antonella Poce (supervisor of this PhD thesis), we conducted a preliminary study to validate the tool on a group of 66 university teachers. The results of this first validation have been used to understand how and in what condition the model works better and how it can be implemented (implementation will be described in the Chapter 4).

1.1 The Critical Thinking definition adopted

Starting from the definitions present in literature, an attempt was made to formulate a definition that would include 1. the elements common to the main definitions 2. the most asserted aspects in the interdisciplinary empirical reference literature.

> Critical thinking is a cognitively expensive thinking process, oriented towards objectives and actions, in which an individual monitors his own mental processes and competently controls the structures that constitutes it through the application of socially co-elaborated standards. Critical thinking implies the interactions between previous knowledge and new information acquired from the context, through which the individual can monitor, correct, and expand his system of knowledge. (Paul & Elder, 2006; Lipman, 1998; Facione, 1990a; Kuhn, 1991; Brynes and Dunbar, 2014; Ennis, 2015)

CT is not a synonym of "quality thinking" or "good thinking", but it represents one of the possible ways of thinking that implies meta-cognitive skills, such as self-monitoring and self-improvement. This way of thinking is cognitively expensive, for this reason it is not and cannot always be active (Kahneman, 2011). Personal *dispositions* (West, Toplak, & Stanovich, 2008) and epistemological beliefs (Felton & Kuhn, 2007), can facilitate or inhibit the activation of this way of thinking. A certain level of agreement has been reached on which skills are at the basis of the CT process: interpretation, analysis, inference, evaluation, argumentation, and self-monitoring (Facione, 1990). At every stage of this process, the critical thinkers are committed to applying some standards to improve the quality of their own thinking (Paul & Elder, 2006) from an inter-subjective perspective, where the individual foresees the interpretation and the meaning that others might ascribe to their inferences and argumentations (Kuhn, 2019). Although such skills and dispositions can be generalized in different contexts, CT can be exercised only if a person is the owner of a domain of knowledge (Brynes and Dunbar, 2014). Therefore, CT needs to be anchored in solid theoretical and cultural foundations, such as the *classical* texts of every discipline (Poce, 2017). Most of the world's knowledge is acquired through direct interaction (for example: dialogues, conversations) or indirect interaction (reading texts or articles) with other individuals rather than through direct experience of the world's phenomena. For this reason, the ability to critically evaluate information has a fundamental evolutionary function for human beings. Therefore, the construct of CT, rather than being understood as a type of individual activity that the subject carries out autonomously whilst engaged in a reasoning task, can be better understood within the theoretical framework of information exchange and symbolic interaction (Brynes and Dunbar, 2014).

1.2 Critical thinking evaluation

The absence of a shared definition of CT has led to the development of multiple methods and tools for the evaluation of this construct (see the Chapter 2 for an in-depth analysis of the issue).

On one side, a high number of tests are available in the standardised testing market (Rear, 2019). On the other side, a recent literature review showed how non standardised instruments created *ad hoc* by the teacher and by the researcher are frequently used too (Tiruneh, Verburgh, & Elen, 2014).

In this work, I tried to take an intermediate position between the need to assess CT validly and ecologically from one side and guarantee measurement validity and reliability on the other side. In accordance with the definition proposed in this thesis, the development of CT can be understood within the theoretical framework of the information exchange and symbolic interaction (Brynes and Dunbar, 2014). Consequently, it is possible to observe CT manifestations or, instead, failures in its

application, in complex communicative acts, mediated by the use of language. For this reason, it is believed that the evaluation of CT within CRT guarantees the highest levels of external and ecological validity. In literature it is possible to identify a wide variety of approaches to the evaluation of the written text, which attempt to improve measurement reliability. More specifically, on the one hand there are approaches that tend to break down students' answers into simpler units to identify reasoning patterns and on the other hand there are rubrics that consider the quality of the reasoning in its complexity. The first approach is less widespread and is particularly used for the analysis of scientific texts. For example, in a recent study Moreira, Marzabal, and Talanquer (2019) modified and adapted a framework for the analysis of speech. This framework had been developed in previous studies by Russ and colleagues (2008) to evaluate the written texts of secondary school students. The scheme of analysis adopted included four elements (entity, property, activity, and organization) and aimed at identifying the existing relationships between these elements. On the other hand, Grimberg and Hand (2009) first observed the mental operations followed by students in the creation of a research report and then built some "cognitive paths" categories, namely a series of mental operations contained in a written text. Some of the operations identified by the authors are observation, measurement, comparison, analogy, clarification, generalization, deduction and argumentation. The authors observed that both those who achieved high or low performances used the same cognitive strategies, but with a different order and organization. In a recent study (Moon, Moeller, Gere, & Shultz, 2019), the authors adopted the 2009 Grimberg and Hand model, identifying ten cognitive operations organized according to level of complexity. The authors propose the adoption of an index called *cognitive complexity*, defined as the density of concepts addressed within the same cognitive operation. A cognitive operation is therefore considered more complex if it links two or more concepts.

The second approach for the evaluation of CT involves the use of rubrics through which it is possible to analyse the quality of the reasoning expressed in a text. This approach is widely implemented through the methods of quantitative and qualitative analysis of the content, in particular for the study of the quality of posts within e-learning discussion forums (Newman, Webb, & Cochrane, 1995; Garrison, Anderson & Archer, 2001, a).

Although the use of open items provides valid information on the actual capacity of adopting CT strategies in real contexts, this method raises some problems such as the reliability of the measurement and the costs of manual evaluation. As described in the chapter 2, automatic evaluation of open-ended questions and essays would represent a solution to these problems (Liu, Frankel, & Roohr, 2014).

Starting from the theoretical conceptualization proposed in this thesis, it is through language that CT manifests itself and develops. For these reasons, we chose to work on tools for the evaluation of CT based on the use of language within different types of CRT. More specifically, we decided to use an evaluation rubric developed by the CDM for the analysis of the levels of CT that occur within the answers to constructed response answers.

1.3 The CDM Critical Thinking Assessment tool

The CDM has been investigating CT assessment for many years (Poce, 2012; Poce, Corcione & Iovine, 2012; Poce, 2015; Poce, 2017). Starting from the insights collected throughout these research, the CDM developed an assessment method aimed at detecting CT manifestations in different kinds of CRT such as short-essays (Poce, 2017), and open-ended questions (Kirsch, Lennon, Yamamoto, & von Davier, 2017).

In a short essay, students have to present their ideas following a linear and logical structure. Students are provided with a stimulus (e.g. a literary text) and a set of questions that students should use as guidelines to organize their argumentation. The idea to assess CT through the use of short essay is based both on both EWCTET (Ennis & Weir, 1985) and ICTET (Paul & Elder, 2012) methodology, presented in details in Chapter 2. The CDM CT assessment uses literary text as a stimulus to activate critical reflection in the short-essay, as in the ICTET. Literary text is used because students apply specific cognitive skills when analysing literary works. If they do it systematically, students learn not only to substantiate their interpretations through well-reasoned arguments but also to become aware of the reasoning process itself (Esplugas & Landwehr, 1996). In an essay-based task, it is important to provide students with questions aimed at stimulating different kinds of CT skills. In the methodological part of this chapter, I will show how the questions proposed in the ICTET (Paul & Elder, 2012) have been adopted and adapted in the context of the pre-pilot experimentation.

Open-ended questions allowed respondents to engage in activities that are similar to those they might perform if they encountered the materials in real life, because they are not constrained by an artificial set of response options (Kirsch et al., 2017). Examples of open-ended questions are copying or paraphrasing information in the stimulus, generating a response, and completing a form.

The CDM assessment tool evaluates students' constructed responses through six macro-indicators, inspired by Newman, Webb and Cochrane (1995). The first macro-indicator, namely *use of the language*, is useful to assess the language form of the text. The macro-indicator called *justification* evaluates students' ability to elaborate on their thesis and support their arguments throughout a

discourse. *Relevance* is a macro-indicator that analyses consistency in the texts produced. For instance, it refers to the correct use of outlines and to the capability to accurately use given stimuli.

Table 17 Rubric for the evaluation of Critical Thinking (Poce, 2017)

Macro-indicators	Indicators	Descriptors	Points
Use of language	Punctuation, spelling, morphosyntax, lexical property.	The expression is:	
		a. Rich and original	5
		b. Appropriate	4
		c. Basically correct	3
		d. Inaccurate	2
		e. Incorrect and inappropriate	1
Explanation/ Argumentation	Formulation of a thesis, arguments and counterarguments.	The argumentation is:	
		a. Rich and comprehensible	5
		b. Clear and well-structured	4
		c. Too concise	3
		d. Not very consistent	2
		e. Inconsistent	1
Relevance	Adherence to the proposed topic.	The topic is developed:	
		a. In a detailed way	5
		b. In a complete and correct way	4
		c. In a general way	3
		d. In a partial way	2
		e. Completely out of topic	1
Importance	Knowledge of the proposed topic (the important aspects related to the topic are mentioned)	The knowledge of the topic is:	
		a. extensive	5
		b. complete	4
		c. appropriate	3
		d. superficial	2
		e. very poor	1
Critical evaluation	Critical reprocessing of documents and sources	The processing is:	
		a. critical and extensive	5
		b. broad and appropriate	4
		c. essential and simple	3
		d. partial	2
		e. contradictory	1
Novelty / Innovation	Additional information, new ideas or solutions are provided.	New information was inserted:	
		a. in a broad, critical, and original way	5
		b. in a detailed way;	4
		c. in a correct way	3
		d. in a simple and partial way	2
		e. no additional information	1

The macro-indicator called *importance* evaluates the knowledge used in discourses. *Critical evaluation* assesses personal and critical elaboration of the sources, data and background knowledge. Finally, *novelty* concerns the development of new ideas and solutions based on the initial hypothesis and personal thesis.

The macro-indicators were organised in form of assessment rubric (Table 17). The assessment rubric was used in this thesis to evaluate constructed response answers.

The assessment method was adopted in several contexts, especially with secondary level school students and university students, and with different knowledge domains. The CDM assessment model is flexible because teachers can individuate different kinds of literary texts and questions to provide to their students. According to the CDM model, teachers should look at the six macro-indicators of CT whatever is the answers' content of students constructed response answers. Teachers and assessors would need to receive the proper training before using the CDM assessment model autonomously.

1.3.1 A NLP Prototype developed by the Center for Museum Studies

The CDM assessment model was used as theoretical basis for the development of a NLP prototype capable of automatically evaluate the CT macro-indicators presented in Table 17 (Poce, De Medio & Amenduni, 2020). In the pre-pilot experimentation presented in this chapter, the NLP prototype was designed to assess four out six sub-skills*: use of language, relevance, importance* and *novelty*.

The NLP prototype is composed of four main modules that allow to perform all the operations necessary to obtain the experimental results (see Figure 11).

In the (1) *Security module*, we implemented an open source Security Framework application to automatically set security processes, such as authentication and authorization. Every operation within the system is logged anonymously. The module allows online registration via email and provides a secure login form to access the services offered.
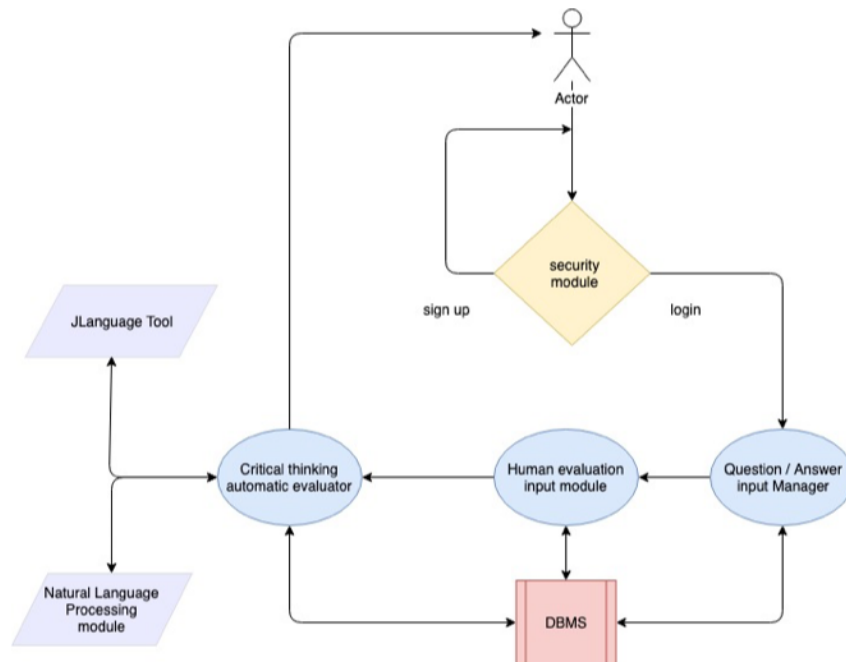
Figure 11 The four modules of the NLP prototype. Retrieved from Poce, Amenduni, De Medio and Re, 2019 p. 66

The (2) *Question / Answer Input* module manages the insertion of the questions and answers to be evaluated. For each question, the assessors have to add the text of the question and a golden answer in the *Question / Answer Input* module (Figure 12). Assessors are also asked to include words representing the *concepts* and the *successors*. *Concepts* could be defined as the topics that should be covered in a correct and exhaustive answer. *Successors* represent, instead, deepening or related topics of the given concepts. *Concepts* and *successors* are be used by the NLP module to evaluate two out of four CT indicators.

Figure 12 Interface of the Question / Answer Input module

The (3) *Human evaluation input* module allows experts assessors to manually evaluate students' constructed response answers (e.g. short-essay). As shown in Figure 13, the expert assessor visualises the text of the question, the text of the answer, and the CT sub-skills to assess. For each student's answer, it is possible to associate one or more anonymous evaluation; these evaluations are compared with the automatic evaluations to verify the reliability of the evaluation.

Question with id = 71

In che modo pensi che l'attività progettata influenzerà il pensiero critico dei partecipanti?

Answer

Attraverso attività volte a sviluppare competenze interculturali e interlinguistiche

Basic Linguistic skills 0

Relevance 0

Importance 0

Novelty 0

Submit Reset

back to the question bank

Figure 13 Interface of the Human evaluation input module

The (4) *Critical thinking automatic evaluator* is the heart of the system which uses two external tools to perform the automatic evaluation.

In order to assess the macro-indicator *use of language*, in the first version of the NLP prototype it was decided to count spelling and punctuation errors of the answers entered by users; the tool chosen for the correction of texts is the JLanguageTool (https://languagetool.org/), an online web-service that allows, giving an input text, to obtain all the errors present and the most likely corrections. The system initially performs an analysis of the type of errors and weighs them according to their severity; a punctuation error or an accent is much less serious than an error within a word (e.g. double letters, words declined in dialect, etc...). In addition, for the other analyses described below, is always used the text with the corrections suggested by JLanguageTool because obviously it would not be functional to the analysis to use the incorrect text, introducing noise in natural language analysis. The value of the indicator is given by normalizing the number of errors considering the number of words that make up the text of the answer.

The macro-indicator *relevance* is assessed using an analysis of the *concepts*. The text is processed by a POS Tagger (specifically https://nlp.stanford.edu/software/tagger.shtml, the most widely used in literature); a POS Tagger is an IT tool that deals with the analysis of the text and tagging it by identifying the various parts of the speech. This tool allows to identify all verbs and their declinations, the subjects of sentences and nouns within a sentence. For our analysis, in this step we extract all the nouns to insert them as input to an algorithm for the calculation of n-grams of length from one to three. Taking a text, the 1-gram set is composed of all the single words taken in order as they appear in the text, while the 2-grams are the set of all the words taken in pairs and so are the 3-grams.

For example, take the sentence:

"all mice love cheese" we can create the three sets in the following way

1-gram: "all", "mice", "love", "cheese"

2-gram: "all mice", "mice love", "love cheese"

3-gram:" all mice love", "mice love chees"

The sets thus generated are compared with the *concepts* defined as necessary for a good answer by the expert who wrote the question. The number of the intersection between n-grams and *concepts* will provide the relevance of the answer to the topic.

The macro-indicator *novelty* is assessed through the same analysis accomplished for the *relevance*, by counting the intersection between n-grams and *successors*.

The macro-indicator *importance* is assessed by performing a knowledge base analysis verified through Wikipedia. Initially, both the question and answer text are sent to an online tagging service through Wikipedia pages (TAGME, https://tagme.d4science.org/tagme/). Given an input text, the service allows to obtain the most important set of notions contained in the text and the links to the related Wikipedia pages. Each defined notion is automatically linked to its Wikipedia page. All the outgoing links of this page are also considered and connected to the related notions. The importance indicator is given by comparing the number of notions extracted from the answer with those extracted from the questions defined by the expert assessor / researcher.

The accuracy of this approach depends on the amount and the accuracy of the notions stored in the knowledge base, in this case Wikipedia. When we analyse texts in English we will be able to get information from Wikipedia.eng which has about 6146000 pages. A Nature investigation reported that Wikipedia comes close to Britannica in terms of accuracy (Giles, 2005). However, the same approach would be less effective in other languages, such as Italian, because of the limited number of Wikipedia pages available in those languages.

2. Objectives and research questions

In this chapter, I will present the preliminary results of validation studies carried out on the CDM NLP prototype aimed at the automatic assessment of CT in constructed response answers.

In this pre pilot experimentation, the data collected for the evaluation of CT consists of English written texts, in the form of short essays and open-ended answers, produced by university teachers. The analysis of the texts was carried out by human evaluators and by the NLP prototype respectively. The performance of the automatic evaluation tool was compared to the scores assigned by expert evaluators. The results of this validation were used to understand how, and under whichs condition the model works best and how it can be implemented.

Therefore, the main research questions that guided the experimentation were the following:

1. What are the levels of CT shown by the participants in the research?
2. What are the levels of reliability of the CT evaluation rubric?
3. What are the levels of reliability shown by the NLP prototype in the automatic evaluation of CT?

3. Methodological and procedural choices

3.1 Participants

In the pre-pilot stage of this research, the data were collected from a group of 66 university teachers from different European nations and from the United States. The first subgroup is composed of 18 teachers participating in a workshop held as part of the Erasmus+ Crithinkedu Summit project (Leuven, Belgium) which took place during the annual event of the *Foundation for Critical Thinking*, the biggest American foundation dedicated to the dissemination of educational practices on CT founded by Richard Paul. The second subgroup is composed of 22 Italian teachers who participated in an assessment session carried out within the *Inclusive Memory* project launched by Roma Tre University. Finally, the remaining 26 teachers were invited to answer an online questionnaire after taking part in the sixth annual conference "*Defining Critical in the 21st Century?*" (New York, United States) on teaching methods for the promotion of CT.

The data collection in English was fundamental to carry out validation studies of the NLP prototype which, at that stage, was designed for the evaluation of texts written in English.

3.2 Design

The three subgroups of university teachers were recruited in national and international meetings that aimed developing the teaching and evaluative methodologies for the promotion of CT at university level (Table 18).

The first subgroup (18 European university teachers) voluntarily participated in the "*How to assess critical thinking skills through writing?*"[13] workshop within the European Project "Crithinkedu" International Summit. Thanks to the great resonance of the event, we managed to involve some of the leading international experts of CT among the participants, such as Ronald Barnett[14]. The workshop aimed to describe the tools for the assessment of CT through written texts analysis. After

---

[13] http://crithinkedu.utad.pt/en/europeansummit-parallel-sessions/
[14] https://www.ronaldbarnett.co.uk/

an introduction to the workshop objectives, the participants were asked to carry out a "paraphrase and comment" task for 30 minutes. Specifically participants were required to read, paraphrase and comment an extract from the text written by Galileo Galilei "*Dialogue Concerning the Two Chief World Systems*" (Figure 14). After this, the participants were invited to reflect upon the relationship between thinking and writing and how to use the written text to assess CT. The discussion also allowed us to collect some feedback on the evaluation tool, that will be presented in the results. The reading and writing activities was an adapted and shortened version of the ICTET (Paul & Elder 2012). The texts written by the teachers during the activity were collected for validation analysis of the CT evaluation rubric and NLP prototype.
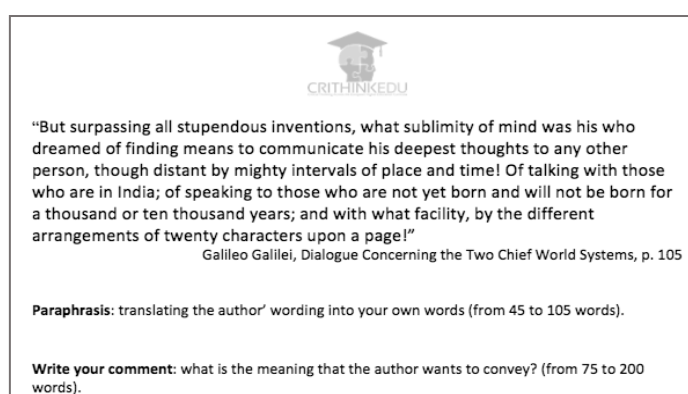
Figure 14 Data collection tool for the evaluation of critical thinking

The second subgroup (22 Italian teachers) participated in an operational meeting within the Inclusive Memory[15] project. One of the main goals of the project is to promote CT among museum visitors to enhance their social inclusion. For this reason, at the end of the meeting, the university teachers were involved in an "open-ended questions" task, by answering to the following three questions:

(Q1) Mention at most three activities that you would adopt for the promotion of CT within your teachings. Explain why you would include these activities in your courses.

(Q2) How do you think the planned activities could influence the development of the participants' CT?

(Q3) How could the development of the participant's CT contribute to other learning objectives?

Again, the texts of open-ended answers were collected for the validation of the rubric of CT evaluation.

Finally, the third subgroup (26 American teachers) was comprised some of the participants in the "*Defining Critical in the 21st Century?*[16]" sixth annual conference held at Berkley College, NYC. At

---

[15] http://host.uniroma3.it/progetti/inclusivememory/
[16] https://ccrwt.weebly.com/uploads/2/2/7/1/22712194/5683_ccrwt_program_onlinedoc_final_pdf.pdf

the end of the conference, an online questionnaire was sent to the conference participants and 26 teachers chose to answer the same three questions asked of the subgroup of Italian teachers.

Table 18 Participants in the pilot study, activities, and types of collected data

Participants	Activities	Description	Collected data
18 European university teachers	"Paraphrase and comment" task	Reading of an extract of the text "*Dialogue Concerning the Two Chief World Systems*", paraphrasing and writing a comment.	36 English texts (18 paraphrases; 18 comments)
22 Italian university teachers	"Open-ended question" task	Participation in a meeting within the *Inclusive Memory* project. Answer three open-ended questions on teaching strategies for the development of CT.	66 Italian texts (3 open-ended answers for each teacher)
26 American university teachers	"Open-ended question" task	Participation in the "*Defining Critical in the 21st Century?*[17]" annual conference, NYC. Answer three open-ended questions on teaching strategies for the development of CT	78 English texts (3 open-ended answers for each teacher)

The entire *corpus* of open-ended answers prototype (Participants = 66; Constructed response answers = 180) was used for the purpose of a preliminary validation of the evaluation of CT rubric developed during previous research by the group coordinated by Antonella Poce (Poce, 2017). Only the English texts (from the first and third sub-groups) were used to carry out preliminary validation studies of the NLP prototype (Participants = 44; Constructed response answers = 114).

3.3 Data analysis

Three evaluators, with previous experience in the assessment of CT, evaluated the open-ended answers, the paraphrasing and the comments using the rubric for the evaluation of CT (Table 17).

The NLP prototype was adopted for the evaluation of answers in English through 4 of the 6 indicators of the model (*use of language, relevance, importance* and *novelty*). It was possible to compare the data collected on the subgroup of Italian and American teachers as for both groups the same data collection tool translated in Italian and English was used (Table 18).

As suggested by Mao and colleagues (2018), the Quadratic-Weighted Kappa (QWK) and Pearson product-moment correlation index can be adopted to assess the degree of agreement between the expert evaluators and between expert's evaluator and NLP prototype. The QWK index is an inter-

---

[17] https://ccrwt.weebly.com/uploads/2/2/7/1/22712194/5683_ccrwt_program_onlinedoc_final_pdf.pdf

rater reliability measure, that quantifies the degree of agreement among raters. The QWK index is a number between 0 and 1, in which 0 indicates the absence of agreement and 1 the perfect agreement (Fleiss & Cohen, 1973). The correlation index of Pearson is another index that allows for the evaluation of the degree of agreement consistency between two evaluators. High levels of inter-rater agreement show that other evaluators, using the same rubric, would reach similar evaluation results, thus proving the evaluation tool reliable.

4. Discussion on collected data

4.1 Results on the group of European teachers – paraphrase and comment task

Figure 15 presents a comparison of the average scores of the six CT macro-indicators respectively in the paraphrasing sections and in the comments. The *novelty* indicator was not calculated in the paraphrasing task as the task does not require the introduction of additional information, ideas and solutions. We observed that the participants obtained slightly higher than average scores in the comment section as compared to that of the paraphrasing task. This result could be explained for two different reasons. The first is that, during the workshop, the European participants stated that they did not have any previous experience with paraphrasing, a practice wide spread in Italy in the teaching of language and literature since primary school[18]. The second possible explanation is that since the paraphrase is an exercise that facilitates the adoption of increasingly sophisticated forms of CT (Paul and Elder, 2006; Poce, 2017), it can elicit a greater use of CT in the comment.

Participants obtained satisfying average scores only for the *use of language* macro-indicator both in the paraphrase and in that of comment (score between 2,9 and 3,4). The average score can be considered sufficient for the macro-indicator's *explanation/argumentation* and *importance* both in paraphrase and in comment, while *critical evaluation* and *relevance* are sufficient only in comment (from 2,3 to 2,8). The average scores are not satisfactory for the *critical evaluation* and *relevance* indicators in paraphrase and for the *innovation* indicator in comment (less than 2,2).

---

[18] http://www.indicazioninazionali.it/wp-content/uploads/2018/08/Indicazioni_Annali_Definitivo.pdf
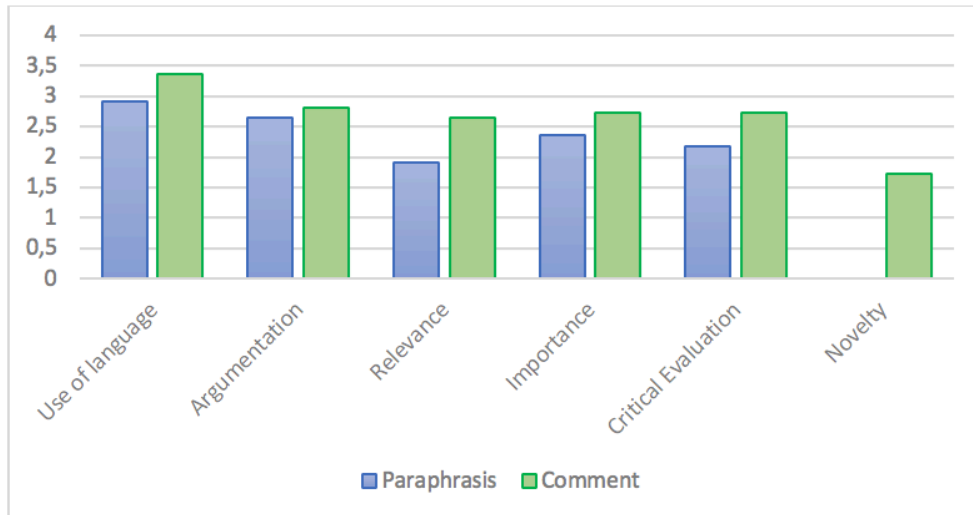
Figure 15 Comparison of critical thinking performance in paraphrase and comment

The average scores assigned by the evaluators were compared with those of the NLP prototype respectively in the paraphrasing task and in the comment section. Figure 16 shows that in paraphrase the NLP prototype provides higher scores than the expert evaluator, except for the *importance* indicator. In comment, instead, we can observe a general tendency of the NLP prototype to assign lower scores than expert evaluators.

Figure 16 Comparison of critical thinking scores calculated by the expert evaluator and by the NLP prototype in paraphrase and comment

As shown in Table 19, the degree of agreement between expert evaluators for the *use of language* indicator is satisfactory both in paraphrase and in comment, with a higher performance in paraphrase (QWK=0,83) than in comment (QWK=0,62). However, there was an absence of correlation between the scores given by expert evaluators and the NLP prototype. This result can be explained by at least three factors: the first is that the text of the answers is quite short (with an average of 35 words per

answer). Previous studies showed that the NLP prototype tends to obtain better scores when the questions are longer (Poce et al., 2019). On the other hand, not all the participants were English native-speakers, and this might have affected their use of language. Finally, the expert evaluators are Italian, and this might have had a further influence on the evaluation of the language used by the English non-native speakers. The degree of agreement between expert evaluators for the *relevance* indicator is satisfactory both in paraphrase and in comment, with a higher performance in comment (QWK = 0,81) than in paraphrase (QWK = 0,68). As for the *relevance* indicator, there was a correlation between the scores given by human evaluators and the NLP prototype (r = 0,47) in the comment that was not significant from a statistical point of view. It is possible to say that the *relevance* indicator is easier to discriminate within the comment both for the expert evaluator and for the NLP prototype. Finally, the degree of agreement between expert evaluators for the *importance* indicator is at the top for the paraphrase task (1), but the agreement drops significantly in the comment (r = 0,64). For this indicator, a non-statistically significant correlation was detected between human evaluators and NLP prototype both in paraphrase (r= 0,45) and in comment (r = 0,43).
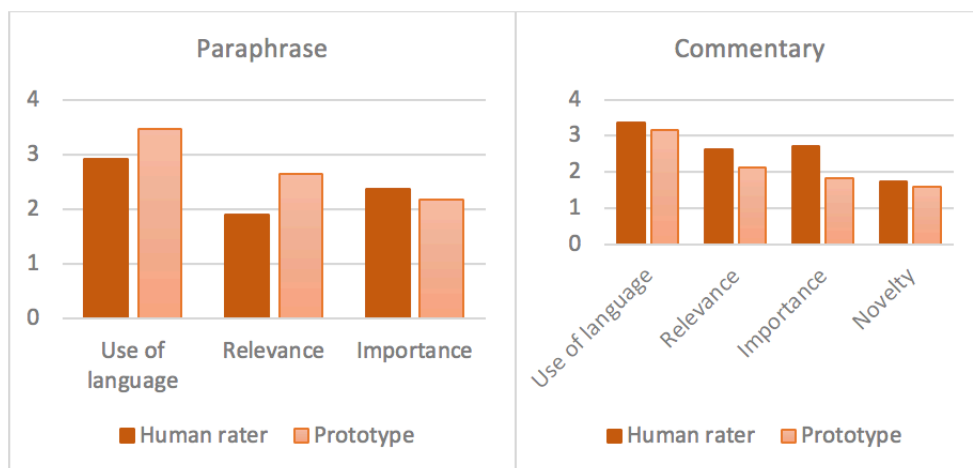
Table 19 Comparison of critical thinking scores calculated by the expert evaluator and by the NLP prototype in paraphrase and comment

Indicators	Correlation between expert evaluators	QWK between expert evaluators
Use of language – paraphrase	0,911*	0,83*
Use of language – comment	0,745*	0,618*
Relevance – paraphrase	0,75*	0,682*
Relevance – comment	0,881**	0,811*
Importance – paraphrase	1,000**	1,000*
Importance – comment	0,642	0,571

The workshop modality allowed for the collection of not only data on the levels of CT of participants, but also some feedback from CT experts too.

A university teacher highlighted as possible critical point in the evaluation system when they pointed out the possibility that basic linguistic skills could create a *bias* in the evaluation of other skills related to CT. Moreover, he added that the evaluation of CT should include an attention towards the willingness to take responsibility for one's positions and he wondered if the method was capable of evaluating this disposition. According to Robert Barnett, the use of writing can contribute to evaluating this disposition because if the written text is designed to be shared with a specific audience (for example a *review*) the assumption of responsibility is encouraged. Another teacher highlighted that through writing alone it is difficult to distinguish the difference between the evaluation of CT

skills and creative thinking skills. The same teacher added that she found the proposed task difficult and therefore assumed that her students would have encountered similar difficulties. Another teacher agreed with her and explained about the students' difficulties in writing. A series of observations emerged from the participants on the reasons why writing represents a vehicle for the development of CT skills. At the end of the debate, Barnett recommended the insertion of a third question, in addition to paraphrase and comment, with a request of critically evaluate the extract.

4.2 Results on the group of Italian and American teachers – three open-ended questions.

As shown in Figure 17, most teachers come from the field of human sciences (43%) and social sciences and education (33%). A lower percentage comes from the field of economic and political sciences (12%), STEM (8%), medicine and healthcare professions (4%).
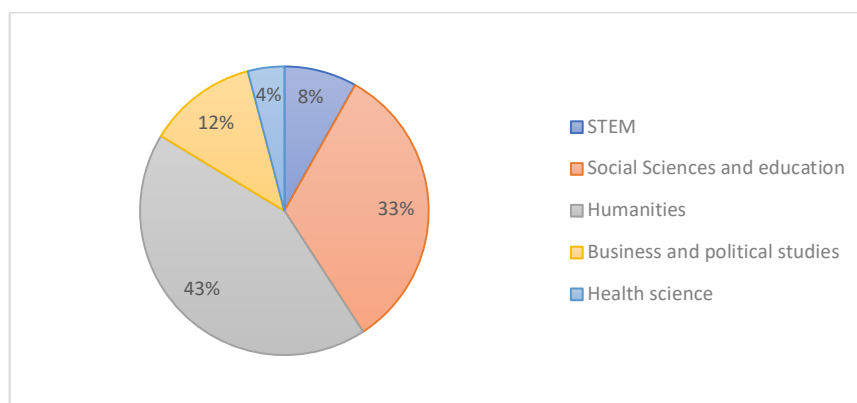
Figure 17 Disciplinary sectors of teachers involved in the analysis

The two groups of Italian and American teachers obtained similar scores in the three questions (Figure 18). For every question, the average score is higher than 17 with a maximum of 30. In addition, the total maximum score is higher than 53 out of a maximum of 90 for both groups. Therefore, the performance obtained on scores of CT can be considered satisfactory both for the Italian group and the American group.
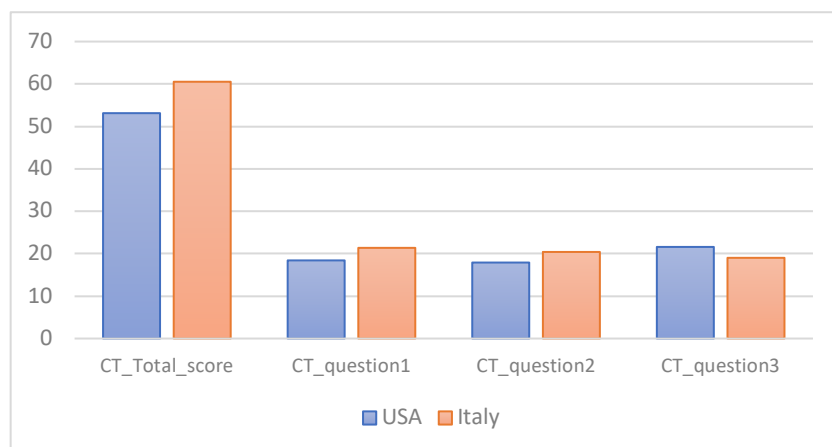
Figure 18 Levels of critical thinking in the Italian and American group

The degree of agreement reached between expert evaluators is shown in Table 20. Almost for every item a satisfactory degree of correlation was observed (i.e., r > 0,69). The results suggest that the evaluation method is reliable when the evaluation is carried out by expert evaluators. On the other hand, there are no significant correlations between the scores given by expert evaluators and the NLP prototype.

Table 20 Pearson correlation index between expert evaluators

Item	Correlation between expert evaluators	Sign.
Question 1	0,785	0,000
Question 2	0,690	0,000
Question 3	0,744	0,000
CT score – total	0,866	0,000
Use of language	0,749	0,000
Relevance	0,873	0,000
Importance	0,807	0,000
Innovation	0,725	0,000

5. Conclusive remarks

This chapter aims to present a theoretical model which was functional to the automatic evaluation of CT within different kinds of CRT (e.g. essays and open-ended questions). This model was adopted to assess English written texts because the first version of the prototype relies on two external applications: JLanguage Tool, a grammar and spell checker for the English language, and Tagme, an online tagging service of Wikipedia pages. We will see in the Chapter 4 how we applied the model to constructed response answers written in Italian language.

Validation studies on the reliability of automated scoring require to collect large amounts of data. The cases presented in the Chapter 3 are therefore not sufficient to draw conclusions. The reliability levels of the CT evaluation rubric calculated on a group of 66 university teachers on 160 constructed response answers are satisfactory but there are areas for improvement. Indeed, the results show higher reliability scores on some indicators and on specific types of stimulus.

In the "paraphrase and comment" task we found that the *relevance* indicator was easier to discriminate within the comment both for the expert evaluator and for the NLP prototype. For the *importance* indicator, a non-statistically significant moderate correlation was detected between human evaluators and NLP prototype both in paraphrase and in comment. The NLP prototype performance was lower for the *use of language* indicator. This lower performance can be possible due to a limitation of the study: the English texts produced by the participants have been evaluated by Italian experts with a good knowledge of the English language. This may have negatively influenced the reliability of the u*se of language* indicator. Moreover, only some participants were English native speakers. In the data collection carried out in Leuven, for privacy reasons, it was not possible to collect additional data on the participants, included participants' mother tongue. Thus, we were not able to see to which extent the participants' mother tongue affected the evaluation of the *use of language* indicator.

In the "open-ended questions" task we found satisfactory reliability level between two human experts but we did not find correlations between the scores given by expert evaluators and the NLP prototype. This could be due to the wider domain of the "open-ended questions" compared to the "paraphrase and comment" task. In other words, the NLP prototype could easier predict topics in the "paraphrase and comment" task rather than in the "open-ended questions" task. Moreover, the prototype can better discriminate relevant and important concepts and notions in the "paraphrase and comment" task because the expected topics are strongly dependent to the provided literary stimulus. For this reason, we decided to use task based on a specific stimulus, such as a literary text, in the following experimentations.

The adoption of a data collection method during the workshop allowed for the possibility of receiving immediate feedback from the participants and to actively involve them in the process of co-construction of the evaluation system. The feedback collected for the most part during the workshop that took place in Leuven within the Summit of Crithinkedu European Project, allowed to implement the evaluation system for the following meetings on data collection. All the collected results were used to improve the evaluation system and to generalise the it to constructed response answers written in Italian.

It is necessary to collect a large amount of data so that it is possible to conduct some *training* in *machine learning* mechanisms for the implementation of the NLP prototype performance (Grimmer,

& Stewart, 2013). At that moment, due to the limited number of cases and the different kinds of stimulus used, it had not been possible to create a *training set* for the application of a *supervised learning model*.

In the future studies it will be necessary to collect data in English produced by mother-tongue participants because the production of English texts by people of different nationalities can introduce a wide variability when it comes to the *use of language*. Alternatively, it would be useful to control the effect of participants' mother tongue, analysing to which extent the variable contribute to the variance on the CT macro-indicators scores.

To guarantee that the NLP prototype evaluation is not biased (Mao et al., 2018), it will be necessary to check the also other variables, such age and gender, collecting some basic information on the participants.

In the next chapter I will present the implementation carried out on the second version of the NLP prototype designed by the CDM research team.

CHAPTER 4 CRITICAL THINKING SUB-DIMENSIONS AND NATURAL LANGUAGE: CORRELATIONS BETWEEN PROCESSING FEATURES IN ITALIAN HIED STUDENTS' PRE-POST TEST ESSAYS

1. Introduction

As showed in Chapter 2, many attempts have been carried out to develop and validate tools for the automatic assessment of CT related-skills. Automated NLP tools have been used to describe linguistic features of writing that predict overall quality and linguistic features that change with development (Crossley, Weston, McLain Sullivan, & McNamara, 2011; McNamara, Crossley, & Roscoe, 2013; MacArthur, Jennings, & Philippakos, 2019). In a recent work, MacArthur et al., (2019) used a corpus of pre-test and post-test argumentative essays to compare changes for the treatment and control groups on linguistic features that affected post-test quality. The authors used Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014), an open-access program that brings together a range of linguistic analysis tools for syntactic parsing, analysis of lexical characteristics and diversity, latent semantic analysis, and other components. Coh-Metrix includes more than over 100 NLP indices organized by linguistic constructs which can be adopted to analyse English written texts. By adopting a structural equation model, MacArthur et al., (2019) found that the NLP constructs *referential cohesion* and *lexical complexity* positively predicted the quality of the argumentation in post-test essays while *syntactic complexity* was negatively related to argumentation quality. Length explained 30% of variance in quality and the full model explained 48.7% of the variance. They also found that the treatment group wrote post-test essays with greater *lexical complexity* and *referential cohesion* and less use of connectives than a control group.

In Chapter 2, the most adopted tools for the automatic analysis of NLP features were presented (e.g. Coh-Metrix and LIWC). All of them have been applied to extract features from English written text. There are a few attempts to generalize these techniques to other languages, included Italian language (Dell'Orletta, Montemagni, & Venturi, 2011). NLP analysis applied on Italian language are preliminary in nature, especially in the context of educational research. Only in a few cases, NLP are applied to assess learning outcomes or cognitive dimensions (Chiriatti et al., 2018). Therefore, in this chapter I present an exploratory work aimed at understanding which NLP features are associated with six CT sub-dimensions, as assessed by human evaluators in essays written in Italian language. The study used a corpus of essays from a quasi-experimental study of an instructional program based on the adoption of Open Educational Resources (OERs) to support students CT.

1.1 OERs and Critical Thinking

Recent research experience has reported the need to adopt a non-formal approach in HE based on the principles of Open Education (Wilson et al., 2011; Tovar & Lesko, 2014; Weller, 2017): (a) reducing or removing access barriers such as financial, geographical, time and entry requirements; (b) modernisation by means of digital technologies; and (c) bridging non-formal and formal education by making it easier to recognise learning achievements (Inamorato dos Santos, Punie, & Castaño-Muñoz, 2016). Open education is understood as a mode of undertaking education using digital technologies and providing alternative, less restrictive access routes to formal and non-formal education. This broad perspective enables a comprehensive view, thus encompassing, for instance, Open Educational Resources (OERs). OERs are digitised materials offered freely and openly for educators, students and self-learners to use and reuse for teaching, learning and research (OECD, 2007). OERs not only comprise individual course components, but also a whole course, a museum collection, an open access journal or a reference work. Over time, the term has also come to cover content management software, content development tools, and implementation resources such as standards and licensing tools for publishing digital resources. Only a few research investigated the impact of OERs on HE students CT skills. For example, Kurubacak (2007) reported that the process of designing OERs proved to be successful when a Project Based Learning methodology is employed to improve students CT levels. However, more research is needed to understand how to employ OERs to support the development of pivotal learning outcomes in HE students.

2. Methods

2.1 Goals of the research

According to the above-mentioned premises, the main goal of this research is understanding which NLP features are best associated with six CT sub-dimensions, as assessed by human evaluators in essays written in Italian: use of language, argumentation, relevance, importance, critical evaluation and novelty (Poce, 2017). We will also try to answer the following Research Questions (RQ):

1. What is the reliability level of human evaluators' assessment?
2. How students CT levels change in a university course designed to support students' CT levels?
3. How students assessed the university course designed to support CT?
4. What is the level of internal coherence of NLP features and how they correlate with CT sub-dimensions?

2.2 Learning activities aimed at stimulating critical thinking skills

An experimentation was carried out within a Master Degree University module in "Experimental Education and School Assessment" at the Department of Educational Sciences (Roma Tre University). The University module lasted 9 months and 202 students (F = 193; M = 7; Prefer not to say = 2; Average age: 23.3) were involved in different kinds of activities designed to foster students CT throughout two semesters. In the first semester students attended a seminar regarding theoretical assumptions of Open Education. After that, they were required to individually search for and assess ten Open Educational Resources (OERs) on topics related to 21st skills and Museum Education for primary school children. Students looked for educational resources in OERs repositories and used a rubric for the OERs assessment developed in the context of the European Erasmus Plus Project "Open Virtual Mobility" (Poce, Amenduni, Re & De Medio, 2019). For each OER identified, students had to assess the following six indicators: (1) quality of the explanation; (2) Support to the lesson (3) Quality of the assessment (4) Quality of the instruction (5) Technological quality (6) Promotion of Higher Cognitive Skills. For each indicator, students provided a score from 0 to 3 or they declared that an indicator was Not Assessable. For example, when a selected OER did not include quizzes or assessment, students inserted "N/A" for the indicator "Quality of the assessment". Students were also invited to insert the link of the OER, a short abstract, the link of the OER repository used and they could also add a facultative comment. Figure 19 shows an example of the spreadsheet filled by a student. This activity was aimed both at stimulating CT *evaluative* skills and preparing students for the second semester not-mandatory assignments where students were given the possibility to design collaboratively their own OERS, following the design principles of the Project-Based Learning (PBL) methodology (Sasson, Yehuda, & Malkinson, 2018).

NAME	LINK	ABSTRACT	Quality of the explanation	Support to the lesson	Quality of the assessment	Quality of the instruction	Technological quality	Promotion of Higher Cognitive skills	OER link	Comment (not mandatory)
21st Century Skills: Empathy	https://badges.newworldofwork.org/quiz/empathy	ASSESSMENT . Vi sono una serie di video didattici e	2	2	3	2	3	2	https://www.oercommons.org/	Bisognerebbe inserire nel curricolo scolastico, le competenze sociali: empatia, collaborazione, condivisione. Sono essenziali per creare rapporti stabili e duraturi nella società.
The Soft Skills Matter	https://www.commonspaces.eu/it/oer/the-soft-skills-matter/	Occorre insegnare a scuola le	3	2	3	2	3	3	https://www.commonspaces.eu/	
iTILT Interactive Technologies in Language Teaching	https://www.commonspaces.eu/it/oer/itilt-interactive-technologies-in-language-teachin/	TECNOLOGIES iTILT è un progetto europeo	2	3	2	3	3	3	https://www.commonspaces.eu/	
Using the Arts to Promote Critical Thinking	https://www.oercommons.org/courses/using-the-arts-to-promote-critical-thinking	THINKING L'arte viene vista come un mezzo	3	2	N/A	3	3	3	https://www.oercommons	La capacità di ragionamento, di elaborazione di idee innovative ed originali sono delle componenti essenziali che vanno valorizzate e trasmesse a tutti gli studenti.
The Importance of Teaching 21st Century Skills	https://www.oercommons.org/courseware/lesson/22117	Nel 21° secolo insegnanti ed educatori devono	3	2	N/A	N/A	3	3	https://www.oercommons.org/	
Valorizzare il patrimonio	https://www.commonspaces.eu/it/oer/valorizzare-il-patrimonio/	CULTURALE Occorre promuovere la	3	2	2	N/A	3	3	https://www.oercommons.org/	
Patrimonio culturale pubblico: "la grande bellezza"	https://www.youtube.com/watch?v=-pojjvz_444	CULTURALE L'Italia offre numerosi museo,	3	3	N/A	N/A	3	2	https://www.youtube.com/	
Museum Week	https://www.commonspaces.eu/it/oer/museumweek/	HERITAGE E DIGITAL TECHNOLOGIES	2	2	2	2	3	2	https://www.commonspaces.eu/	il potere, a mio avviso, di riuscire a coinvolgere gran parte di alunni nello svolgimento di varie attività e nello studio di diversi argomenti piò o meno impegnanti o difficoltosi.
Musei senza barriere	https://www.commonspaces.eu/oer/musei-senza-barriere/view/	HERITAGE AND DIGITAL TECHNOLOGIES	2	3	N/A	N/A	3	3	https://www.commonspaces.eu/	
MUSEO VIRTUALE in "XXI Secolo"	https://www.commonspaces.eu/it/oer/museo-virtuale-in-xxi-secolo/	TECHNOLOGIES SOFT SKILLS IN MUSEUM	2	2	2	2	3	3	https://www.commonspaces.eu/	con la cultura del patrimonio nazionale ed internazionale. I media svolgono un ruolo essenziale nello sviluppo della conoscenza

Figure 19 OERs assessed by one of the students according to the six indicators proposed by the teacher

Students worked in groups in order co-construct their own OERs, by using different kind of technologies. Out of 202 students, 40 students voluntarily participated in the OER design activity by working in 8 groups. OERs, produced by the students, were assessed by the teacher through the same rubric students used in the first semester to assess the OERs retrieved from the repositories. While the OERs individual assessment assignment was mandatory and carried out fully online, the collaborative PBL activity was optional. Students who chose to participate worked in a blended modality, alternating Face to Face meeting at the university with online work. CT level were assessed through a pre-post-test methodology, described in details in the following paragraph.

2.3 Data collection

2.3.1 CT measure

The study used a corpus of pre-post essays written in Italian language by 202 students. Students were asked to read an extract of the "Dialogue concerning two chief world systems" (Galilei, 1632)[19] entitled "Origin of the nerves according to Aristoteles and according to the doctors" (p. 107-8, see Appendix 1). Students completed the same test at the beginning (October 2018) and the end of the course (June 2019). Students were asked to write an essay by including in their arguments the answers to the following six questions:

1. What are the two opposite positions regarding the origin of the nerves described in the text?

2. What are the differences between the methods supported by Simplicio and Sagredo?

3. What does the "principle of authority" consist of? When is it explicitly referred to in the text and when is it implicit?

4. Why do you think the episode was settled in the Republic of Venice?

5. In your opinion, has the principle of authority affected scientific discoveries throughout history? If so, how?

6. Choose one or more elements in the passage that, in your opinion, have played a role in the development of scientific knowledge in the modern and contemporary world. Explain the reasons for your choice.

The choice of the Galilei's stimulus was driven by different reasons, related both with the specific characteristics of the text and the contents. Firstly, the Galilei's text can be classified as literary text. According to some authors (Paul & Elder, 2006, Poce, 2017), when students read literary texts they strive to accurately represent in their own thinking what are they are reading. Reading literary texts

---

[19] Retrieved from: http://actascientiae.org/Galileo_Galilei_Dialogue_Concerning_the_Two_Chief_World_Systems.pdf

requires active engagement, by creating an inner dialog with the text (questioning, summarizing and connecting ideas). Galilei's literary text is characterised by the use of a figurative and allusive language. Since many implicit references are presented in the Galilei's text, students had to go beyond the available information in the task to draw inferences or make evaluations (Moss and Koziol, 1991). A further characteristic of the Galilei's text is the presence of a dialogue on a controversial topic. As shown by Fischer and colleagues (2009) it is more likely to prompt CT while using inconsistent or contradictory materials than consistent and coherent stimulus materials. Last but not least, the Galilei's text concerns relevant topics for the course subject in "Experimental Education and School Assessment" such as the role of empirical research and research methods.

2.3.2   Exam grades

The exam consisted on a MC questionnaire composed by 80 questions aimed at assessing students' knowledge of the course's subject. Results will be presented as percentage of correct answers provided to the MC questionnaire.

2.3.3   Course assessment questionnaire

Data were collected at the end of the course through an online questionnaire developed and adapted in the *Erasmus + Crithinkedu* project. The questionnaire includes both open-ended and multiple choice questions. We received 202 answers from University students. The open questions were the following:

1. What should be continued or kept (related to the development of critical thinking)? Why?
2. What should be changed (stopped and started) (related to the development of critical thinking)? Why?
3. Other considerations?

2.4 Data analysis

In this analysis, to answer to the RQs 1, 2, and 4, 103 students' pre-post tests were included: thus, the corpus is composed by 206 essays. All the essays were assessed by human evaluators and through an algorithm which calculates different kinds of NLP features simultaneously. One human expert assessed all the essays based on a rubric composed by six macro-indicators on a scale from 1 to 5: use of the language, argumentation, relevance, importance, critical evaluation and novelty (based on Poce, 2017). The two remaining human evaluators assessed 80 essays (40 from the pre-test and 40 from the post-test) to perform inter-rater reliability analysis. At the same time, different NLP features were automatically measured: i) corpus length, ii) mean sentence length, iii) readability (Vacca, 1972)

and iv) syntax complexity (Yang, Lu, & Weigle, 2015), since the best essays are more syntactically and semantically complex than others; v) hapax (Poce, 2012) and vi) lexical extension, because the more synonymous and unique words there are in a text, the better the writer (Crossley, Weston, McLain Sullivain, & McNamara, 2011), vii) verbatim copying (Chang & Ku, 2015); viii) TD-IDF (Salton, & McGill, 1983), to evaluate how relevant a word is to a document and to a corpus on the basis of the number of times that word appears in that document and in that corpus, in order to check its relevance. Based on recent research results, TFxIDF is thought to be used to support the assessment of the sub-indicator "Novelty" (Wang, Dong, & Ma, 2019). This because higher is the index, lower is the number of unique concepts introduced in the text compared to all the other students' text. Table 21 describes the assumed correspondence among the six CT sub-skills identified by Poce (2017) and the selected NLP descriptors and features.

Although the algorithm integrates most of NLP features presented in the Table 21, the measurement of some indicators relies on external tools or is not yet fully implemented in the algorithm.

Different methods have been adopted to analyse the data. Descriptive statistics (average, frequencies, SD) were used to describe the sample features and the main variables under investigation. Welch's unequal variance t-test was used when we wanted to test the hypothesis that two populations have equal means. Welch's unequal variance t-test is an adaptation of Student's t-test, and is more reliable when the two samples have unequal variances and/or unequal sample sizes (Ruxton, 2006). Quadratic-Weighted Kappa (QWK) and Pearson product-moment correlation index was adopted to assess the degree of agreement between the expert evaluators. The QWK index is an inter-rater reliability measure, that quantifies the degree of agreement among evaluators. The QWK index is a number between 0 and 1, in which 0 indicates the absence of agreement and 1 the perfect agreement (Fleiss & Cohen, 1973). The correlation index of Pearson is another index that allows for the evaluation of the degree of agreement consistency between two evaluators. High levels of inter-rater agreement show that other evaluators, using the same rubric, would reach similar evaluation results, thus proving the evaluation tool is reliable. Kendall's tau-b (τb) correlation coefficient (Kendall's tau-b, for short) was used to calculate correlation between NLP features and CT indicators as assessed by human evaluators.

A content analysis of the open-ended answers provided to the course assessment questionnaire was performed to answer to the RQ3.

Table 21 Correspondence among six CT sub-skills and the selected NLP descriptors and features

CT Indicators	NLP descriptors	NLP features	State of development
Use of language	Grammar and syntax mistakes	https://scuolaelettrica.it/correttore/correttorea.php https://www.prepostseo.com/grammar-checker	Used as external tools
	Lexicon	- Corpus Length; - Mean Sentence Length (MSL); - Hapax: V1[20]/N x 100; - Lexical extension;	Implemented in the algorithm
Justification / Argumentation	Readability	Flesch reading; F(Reading ease) = 206 – (0,65 x ASW) – ASL[21]	Implemented in the algorithm
	Syntax complexity	Tint (The Italian NLP Tool) was used to count the number of syntactic patterns, typical of persuasive and argumentative texts (e.g. adverb + adjective + conjunction + adjective) included in an essay;	
Relevance	Topics' relevance to a document and to a corpus	TF-IDF (Term Frequency-Inverse Document Frequency) = (sum of all P (t) of: R (p)) / PTotals[22] Higher is TF-IDF, higher will be level of relevance of the essay	Not yet implemeeted in the algorithm
Importance	Coherence, semantic similarity	LSA (Latent Semantic Analysis). Co-occurrence statistics on the content words preceding and following the target word; then weighting of the occurrences and reduction of dimensionality-	Not yet implemeeted in the algorithm
Critical evaluation	Degree of personal elaboration	*Verbatim copying*: number of instances of verbatim copying/four main concepts × the number of students	Implemented in the algorithm
Novelty	Divergent Thinking	TF-IDF. Lower is TF-IDF, higher will be level of relevance of the essay	Implemented in the algorithm

---

[20] V1 is the number of words that only appears once in a work

[21] ASL: Average Sentence Length; ASW: Average Syllables per word

[22] PTotals = all words; T set of texts t; P (t) the set of words p in the text; R(p) the number of repetitions of the word p in all texts of T except t

3. Results

Table 22 shows descriptive features of the group of participants. It is composed by 103 (F = 96; M = 7) Master Degree students enrolled in the course of "Experimental education and School Assessment". 26 out of 103 attended all the course activities in blended modality whilst the remaining 77 students attended only the online activities. Approximately 50% of the students passed the exam with a score higher than 60% at their first try. The lowest score at the exam was 35% of correct answers and the highest was 86,25% (Average = 60,53; SD = 13,72).

Table 22 Descriptive statistics of the group of participants

Variables	Values	Frequency
Gender	Male	7
	Female	96
Attendance	100% online	77
	Blended	26
Exam grades %	Less than 45%	17
	Between 46% and 55%	22
	Between 56% and 65%	17
	Between 66% and 75%	23
	Between 75% and 80%	13
	Higher than 80%	2
	Missing	11
Total		**103**

In the pre-test, students spent in average 58 minutes (SD = 23,36) to complete the CT essay whilst in the post-test students spent in average 30,5 minutes (SD = 29,73).

Regression analysis suggest that time to complete the essay test do not contribute to explain the variability in CT scores, neither in pre-test or post-test. Welch's unequal variance t-test found no significant difference in scores between men (M = 19,85, SD = 4,99) and women (M = 17,23, SD = 4,39) on CT total score (p = .067). Thus, neither gender and time to complete the assessment could explain the variability in CT total score.

3.1 Critical Thinking Human Assessment Reliability

Cronbach's alpha was used to test the internal reliability of items in the CT test. Cronbach's alpha value is 0.894. Utilising Ponterotto and Ruckdeschel's (2007) reliability matrix, an alpha of 0.85 or

above is deemed to be excellent. Thus, Alpha indicates a high level of internal consistency for our scale with this specific group of participants. Table 23 shows the correlation between each CT sub-indicators and CT total.

Table 23 Cronbach's Alpha values for CT sub-indicators assessed by human evaluators

	Correlation between item and total score	Internal reliability with item removed
Use of language	,667	,883
Argumentation	,754	,869
Relevance	,551	**,**898
Importance	,813	,860
Critical Evaluation	,807	,861
Novelty	,706	,877

Three graders marked responses on the CT test scores developed by Poce (2017) in order to test inter expert reliability. Table 24 presents the results. *Use of language* and *Argumentation* obtained the higher level of agreement between evaluators whilst *Critical Evaluation* the lowest. The overall inter-rater reliability is medium to high, which suggest there is still room for improvement in terms of inter expert reliability.

Table 24 Intercoder agreement between experts

	Correlation	QWK
Use of Language	0,815**	0,803**
Argumentation	0,768**	0,742**
Relevance	0,635**	0,488**
Importance	0,599**	0,503**
Critical Evaluation	0,534**	0,430**
Novelty	0,633**	0,549**

3.2 Comparison of Critical Thinking pre-post test scores

The distribution of the CT total score is close to normal distribution both in pre and post-test (see figure 20 for a population pyramid of CT total scores in the pre-tests and in the post-tests). The mean score on the CT test was respectively 15,24 (SD = 2,99) in the pre-test and 19,43 (SD = 4,69) in the post-test. We used Welch's unequal variance t-test to compare the difference between pre-post-test

CT total score. Welch's unequal variance t-test found a statistically significant difference between pre and post CT total score (p < ,000).
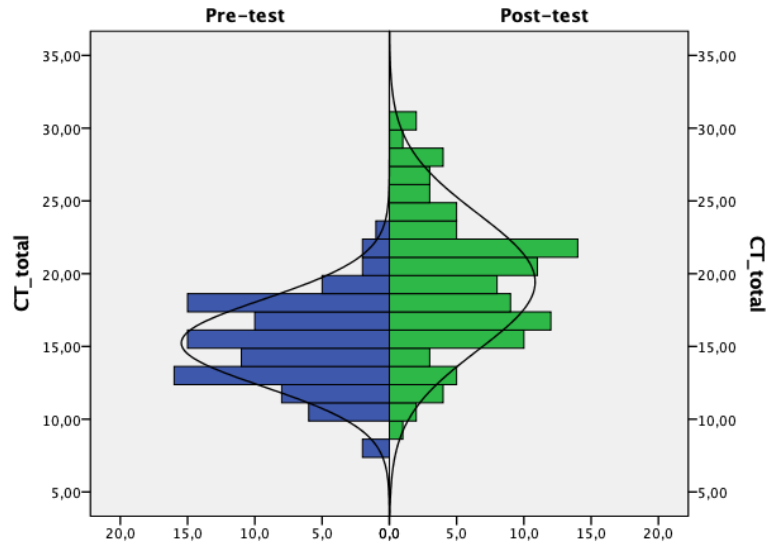

Figure 20 A population pyramid of CT total scores in the pre-tests and in the post-tests

We investigated the difference between CT sub-indicators, as assessed by human experts. Figure 21 shows the comparison of the averages obtained for each sub-indicator. Welch's unequal variance t-test found a statistically significant difference between pre and post CT total score (p < ,000) for all the CT sub-indicators. For almost all the CT sub-indicators, the average was lower than 3 (the median score) in the pre-test, with the exception of the sub-indicator *relevance*. In the post-test, the average was always higher than 3, except for the *novelty* indicator. This suggest that the group of students in average shift from insufficient to sufficient scores. We tried to understand if difference in CT scores could be explained by the attendance of the blended course vs 100% online course.
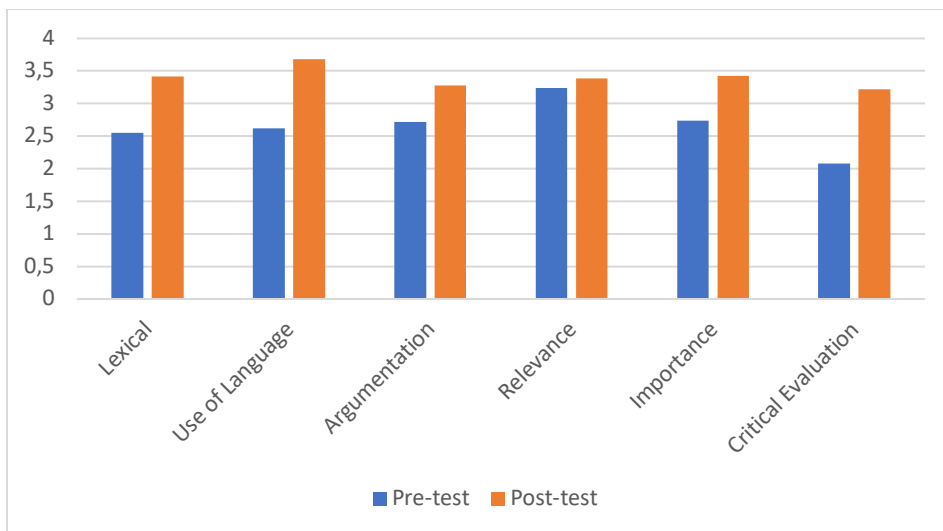

Figure 21 Comparison of the average scores between pre and post-tests as assessed by human evaluators * < 0,05;
** < 0,01 *** <0,001

No statistical difference has been identified in CT level post-test (as assessed by human evaluators) between students who attended the course activities in blended modality compared with students who completed only the online activities. Figure 22 shows the difference between pre-post of blended course attendance vs online 100% students in CT-total average. Both the groups started from similar CT total scores average. Both the groups improved in the post-tests but students who attended only the online activities (blue line) improved a little bit more (Average = 19,92) compared with students who attended the blended activities (green line, Average = 18,00). However, differences in the post-test were not statistically significant between the groups. Thus, the kind of attendance could not explain the difference between pre-post-test.

On the other hand, participants who attended the course had a higher score in the "Exam-grade" (Average = 64,12) compared with students who didn't (Average = 59,23), although the difference between the average is not statistically significant.
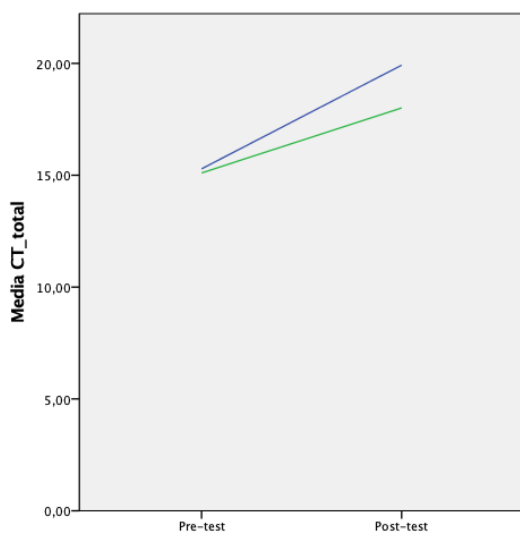
Figure 22 Difference between pre-post of blended course (green line) vs online 100% (blue line) students' attendance in CT-total average
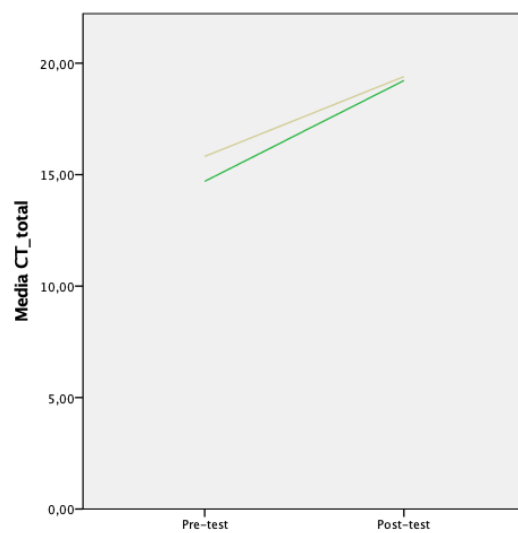
Figure 23 Difference between pre-post of students who obtained a not sufficient (green line) a (yellow line) sufficient exam grade in CT-total average

Figure 23 shows the difference between pre-post of students who obtained an exam grade not sufficient (green line) and a sufficient exam grade (yellow line) in CT-total average. In the pre-test, students who obtained in the final exam an insufficient grade had slightly lower average (Average = 14,69) compared to students who obtained a sufficient grade at the end of the exam (Average = 15,81). However, the difference in CT pre-test was not statistically significant between these groups. Both groups improved in their CT level in the post-tests and the difference between the two groups' averages was reduced, as showed in the figure 5.

Kendall's tau-b (τb) correlation coefficient was used to explore correlation between CT score pre-test and exam grade (Table 25). A low and significant correlation has been identified with two out of six CT sub-indicators: *use of language* and *argumentation*.

Table 25 Correlation between CT sub-indicators and exam grade

		Use of language	Argumentation	Relevance	Importance	Critical Evaluation	Novelty
Exam grade	Tau	,170[*]	,198[*]	,049	-,014	,108	,066
	Sign	,032	,014	,547	,866	,187	,418
	N	88	88	88	88	88	88

3.3 Students' assessment of the course design

202 students assessed the course design activities. 21 students included OERs among course activities that should be continued or kept to support CT. One student reported that the OERs' activity reduced the distance between students who can and cannot attend Face to Face classes (Extract 1).

Extract 1: "Activities regarding digital teaching, OERs and e-learning should be continued because they support exchange with other students, shortening the distance with those who cannot attend Face to Face lessons. Moreover, these activities support open-mindedness".

Indeed, also students who did not attend the Face to Face lessons appreciated the use of OERs in the course, as shown in the Extract 2:

Extract 2: "I only attended two Face to Face classes at the beginning of the course. However, I found interesting and useful researching and evaluating OERs according to the given template".

One of the student connected explicitly the OERs research and assessment activity with CT and information literacy skills stimulation (Extract 3)

Extract 3: "The research and the evaluation of the OERS require students to evaluate materials found online and develop awareness on how to use it in the future".

On the other hand, one student reported that the OERs research and assessment activity was not useful to support CT. She interpreted the assignment as simply filling in a form, without any critical reflection stimulation. Two students reported that, although they recognized the usefulness of the OERs activities, a better preliminary explanation of the OERs would have been useful (Extract 4).

Extract 4: "In my case, only after several lessons and insights I fully understood the concept of OERS. Then, I realised that if I had had that knowledge before, I could have done the OERs research and assessment with much more awareness and I would have grasped more its importance and utility."

All in all, students seems to appreciate the opportunity to structure a Research Project and to understand the strong relationship between research and pedagogical practices. Research is thought as a way to critically reflect upon educational strategies adopted in the classroom.

3.4 NLP features' properties

Table 26 presents descriptive statistics related with the NLP indicators extracted from the students' essays.

Table 26 Descriptive statistics related with the NLP indicators

	Min	Max	Average	SD
Corpus Length	91,00	456,00	245,3795	68,92533
MSL	16,85	73,13	36,9976	9,90324
Hapax	31,36	73,74	52,0249	7,40308
Lexical extension	51,75	85,86	69,7351	5,71921
Reading ease	-376,74	56,38	27,3201	31,27196
Syntax complexity	6,00	61,00	24,2205	10,02870
Verbatim copying	,00	16,00	6,3	,00720
TFxIDF	21,17	51,50	39,3863	5,57181

The essay length was, in average, composed by 245 words and the average length of each sentence in the essay was approximately of 37 words. The Hapax index was in average 37. This indicates that, in average, each essay was composed by 37% of words used only one time. Lexical extension was 69. This indicates that in an essay, the range of words used is of 69%. Reading ease was in average of 27,32 which indicates that a text is written in a complex form, typically adopted by people with higher levels of education. The complexity of syntax is on average 24 which means that students used in average 24 complex argumentative syntax forms in their essays. Students, in average, copied

verbatim 6 times in their essay the words of the test' questions. The TfxIDF was 39,38 in average. This means that each text contains in average 39% of words that are not used in other students' texts. According to the assumptions presented in Table 21, five NLP indicators can be useful to support the assessment of the CT sub-indicator "Use of Language": 1. Hapax 2. Lexical Extension, 3. Corpus Length 4. MSL; 5. Verbatim copying. All these indicators are expected to be related with the Lexicon use. The correlation between hapax and lexical extension is indeed strong: $\tau b = 0,853$ sign. 0,000 (see Figure 24).
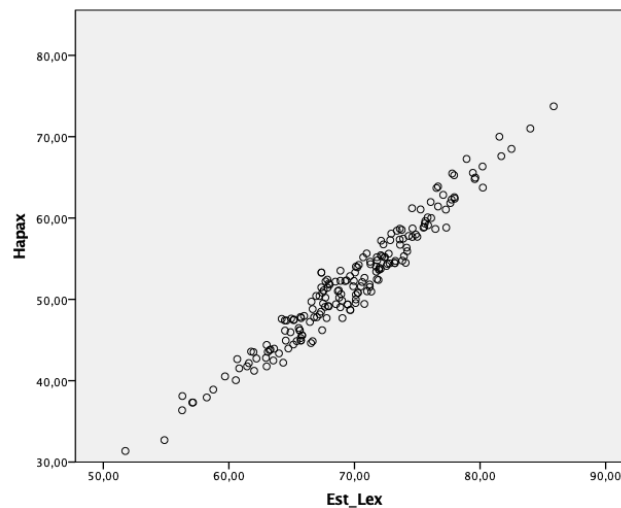
Figure 24 Scatterplot for the correlation between Hapax and Lexical Extension

Corpus Length negatively correlate with both hapax and lexical extension. This means that higher is the number of words used in an essay, lower is the probability that people use unique words in the text (Hapax) and a higher range of words (Lexical Extension). Verbatim copying moderately and negatively correlates with lexical extension. This means that students who copy verbatim the words used in test' questions use a lower range of words. According to the assumptions presented in Table 21, NLP features associated with the sub-indicator "Argumentation" are 1. Reading ease 2. Syntax complexity. A moderate negative correlation has been identified between these two indicators. This means that more difficult is a text too read, higher is the complexity of its syntax. Moreover, syntax complexities strongly correlate with Corpus Length $\tau b = 0,543$ sign ,000. This means that higher is the number of words used in a text, higher is the number of complex structures adopted in that text. TFxIDF was thought to be used to support the assessment of the sub-indicator "Novelty". TFxIDF negatively correlate with Lexical Extension and Hapax and positively correlates with the number of words used, syntax complexity and repetition.

Table 27 NLP features internal coherence

		MSL	Hapax	Lexical extension	Reading Ease	Syntax	Verbatim copying	TFxIDF
Corpus Length	τb	,101*	-,376**	-,421**	-,147**	,543**	,040	,612**
	Sig	,036	,000	,000	,002	,000	,434	,000
MSL	τb	1,00	-,060	-,069	-,694**	,083	,049	,052
	Sig	.	,210	,153	,000	,087	,336	,468
Hapax	τb			,853**	,021	-,216**	-,088	-,315**
	Sig			,000	,667	,000	,082	,000
Lexical extension	τb				,044	-,249**	-,115*	-,321**
	Sig				,359	,000	,022	,000
Reading ease	τb					-,245**	-,021	-,026
	Sig					,000	,673	,723
Syntax complexity	τb						-,019	,445**
	Sig						,709	,000
Verbatim copying	τb							,121
	Sig							,091

Three NLP indicators significantly correlate with CT total score. The Corpus Length, the complexity of the syntax, and the TFxIDF (Table 28).

Table 28 Correlation between NLP features and CT total score

		Corpus Length	Syntax complexity	TFxIDF
CT total score	Tb	,198**	,230**	,228**
	Sign	,000	,000	,001
	N	195	195	195

Table 29 presents the correlation between the 6 CT sub-indicators, as assessed by human experts, and five NLP indicators.

The CT sub-indicator "Use of Language" is moderately and negatively associated with the average sentence length (Figure 7). Average sentence length included between 15 and 45 correspond to score higher than 3 in the evaluation of the sub-indicator "Use of Language". On the other hand, when the average sentence length is higher than 45, the assessment of the sub-indicator "Use of Language" is not sufficient. According to our expectations, the use of complex argumentative syntax forms correlates with the sub-indicator "Argumentation", although the correlation is moderate. This association was graphically explored in the Figure 8, where it is possible to see that an higher numbers

of complex syntax forms correspond to higher level in "argumentation" scores as assessed by human experts. Syntax complexity is also significantly associated with all the others CT sub-indicators. As expected, one of the strongest positive correlation found was between TfxIDF and the sub-indicator "Relevance".

Table 29 Correlation between the 6 CT sub-indicators, as assessed by human experts, and five NLP indicators

		Corpus Length	MSL	Est Lex	Syntax complexity	TFxIDF
UOL_human	τb	,084	-,146**	,056	,144**	,156*
	Sign	,100	,004	,272	,006	,029
Arg_human	τb	,155**	-,071	,014	,175**	,181*
	Sign	,003	,177	,797	,001	,011
Rel_human	τb	,274**	,004	-,126	,235**	,208**
	Sign	,000	,942	,020*	,000	,003
Imp_human	τb	,175**	-,080	,003	,201**	,203**
	Sign	,001	,131	,960	,000	,004
CE_human	τb	,118*	-,099	-,009	,152**	,199**
	Sign	,026	,062	,871	,005	,005
Nov_human	τb	,203**	-,044	-,054	,174**	,141*
	Sign	,000	,412	,312	,001	,050

Higher is the TfxIDF, higher is the coherence with words and concepts used in one essay and the other ones. In other words, higher is the TfxIDF, higher is the relevance of the topics covered in an essay in relation to the others (see Figure 9). Contrary to our expectations, TfxIDF moderately and positively correlates with novelty.

The only indicator that correlates with the exam grade is "syntax complexity" (τb = 0,129 sign < 0,01), which could denote rooms for improvement in terms of criterion validity of NLP towards academic performance.
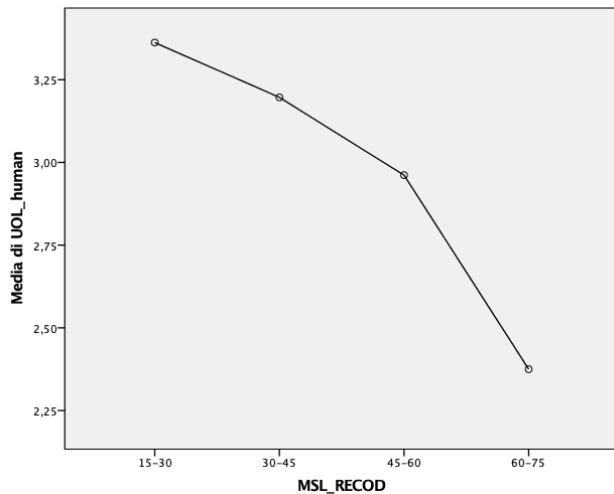
Figure 25 Graphic representation of the correlation between MSL and Use of Language
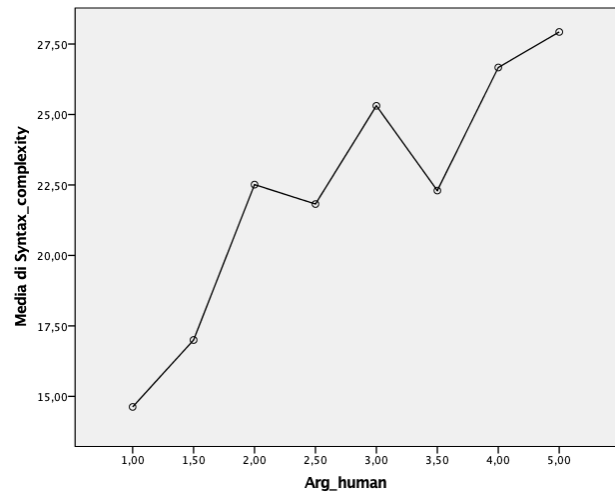
Figure 26 Graphic representation of the correlation between Syntax Complexity and Argumentation
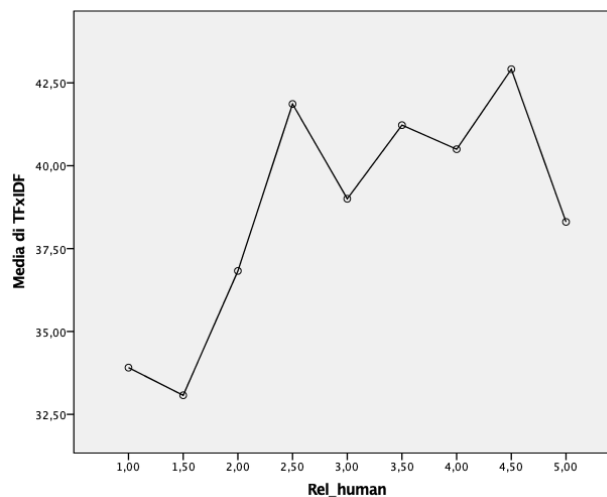
Figure 27 Graphic representation of the correlation between TFxIDF and Relevance

4. Discussion and final remarks

The absence of a shared definition of CT has led to the development of multiple methods and tools for the evaluation of this set of skills, dispositions and behaviours. On one side, a high number of tests are available in the standardised testing market (Rear, 2019). On the other side, a recent literature review showed how non-standardised instruments created *ad hoc* by the teacher and by the researcher are frequently used too (Tiruneh, Verburgh, & Elen, 2014). In this work, we tried to take an intermediate position between the need to assess CT validly and ecologically from one side and the

priority to guarantee measurement validity and reliability on the other side. In our perspective, it is possible to observe CT manifestations or, instead, failures in its application, in complex communicative acts, mediated by the use of language. For this reason, it is believed that the evaluation of CT within CRT guarantees the highest levels of external and ecological validity. Having said that, we acknowledge that CRT scoring can be time consuming and expensive both for teachers and researchers. For this reason, our research team has been working in the direction of "automated scoring" to support and facilitate the scoring of students' written responses. International research shows that the automatic scoring of students written answers has achieved good level of reliability, in specific cases and with English language. However, there are only a few attempts to generalize these techniques to other languages. The present research has started to fill this gap by investigating which NLP features are best associated with six CT sub-dimensions, as assessed by human evaluators, in essays written in Italian: use of language, argumentation, relevance, importance, critical evaluation and novelty (Poce, 2017). The experimentation was carried out with 103 students enrolled in a Master Degree course in "Experimental Education and School Assessment" at Roma Tre University. The course includes different activities aimed at stimulating students CT: OERs individual assessment and OERs collaborative design. The first activity was mandatory and all the students completed it online. The second activity was optional and it was completed by a total of 40 students. Students had to work in groups and alternate Face to Face meeting with online work. In this work, we tried to answer to three RQs. The first RQ concerns the reliability level of human evaluators' assessment. We found an excellent internal reliability and a medium to high inter-coder agreement of the human evaluators. *Use of language* and *Argumentation* obtained the higher level of agreement between evaluators.

Those two indicators also correlate with students' final exam grade and have good internal coherence with CT total scores. Thus, *Use of language* and *Argumentation* can be considered two reliable and valid indicators. On the other hand, *Critical Evaluation* obtained the lowest level of agreement between evaluators. The overall inter-rater reliability is medium to high, which suggest there is still room for improvement in terms of inter expert reliability.

We also wanted to explore how students CT levels change in the university course designed to support students' CT levels. Students CT level improved significantly in the post-test. We compared the CT students' performance of 100% online and blended attendance. Both the groups improved in the post-tests but students who attended only the online activities improved a little bit more (Average = 19,92) compared with students who attended the course in a blended modality (Average = 18,00). However, differences in the post-test were not statistically significant between the groups. Thus, the kind of attendance could not explain the difference between pre-post-test.

In the pre-test, students who obtained in the final exam an insufficient grade had slightly lower average (Average = 14,69) compared to students who obtained a sufficient grade at the end of the exam (Average = 15,81). However, the difference in CT pre-test was not statistically significant between these groups. Both groups improved in their CT level in the post-tests and the difference between the two groups average was reduced. This suggests that CT course entry level could be used to predict students' final exam grade. Moreover, it is possible that the course design had a stronger effect on students' CT level with a lower level of academic preparation. Further research would be necessary to test those hypotheses. In our last research question, we wanted to explore the level of internal coherence of NLP features and how they correlate with CT sub-dimensions.

According to the expectations, we found correlations between 5 NLP features assumed to be associated with the CT sub-indicator "Use of Language": 1. Hapax 2. Lexical Extension, 3. Corpus Length 4. MSL; 5. Verbatim copying. Moreover, a moderate negative correlation has been identified between 1. Reading ease and 2. Syntax complexity, both assumed to be associate with the CT sub-indicator "Argumentation". This means that more difficult is a text too read, higher is the complexity of its syntax. Three NLP indicators significantly correlate with CT total score. The Corpus Length, the complexity of the syntax, and the TFxIDF. As expected, the Medium Sentence Length negatively correlate with the CT sub-indicator *Use of Language*. Complexity of the syntax positively correlate with *Argumentation*. In addition, TFxIDF positively correlates with *Relevance*. On the other hand, some of our expectations were not confirmed.

Although Hapax and Lexical Extension correlates, they did not show any significant correlation with the expected CT sub-indicator *Use of Language*. This result can be explained by an issue retrieved within many of the essays assessed. We found that students often used a not coherent language within their essays, by alternating refined with everyday expressions. In this condition, whilst human evaluators provide low scores to the CT sub-indicator *Use of Language*, the algorithm can find good level of Hapax (number of unique words in the text) and Lexicon Extension.

The second expectations not confirmed concerns the correlation between TFxIDF and the CT sub-indicator *novelty*. We expected a negative correlation between these indicators (based on Wang, Dong & Ma, 2019) but we found a moderate and positive correlation. In previous studies, TFxIDF was used to assess divergent forms of novelty. However, the *Novelty* required in CT should be convergent. Indeed, divergent thought from a single starting point generates varied ideas, whereas convergent thought starting from multiple points seeks one most true or useful conclusion (Brophy, 2001).

In our case, it is likely that algorithm and human evaluators looked for different forms of *Novelty*: the algorithm retrieved divergent new ideas, whilst human evaluators search for convergent new ideas.

This study was exploratory in nature and we acknowledge its limitations: in future studies, we would need to expand the sample size, by including the remaining 200 essays collected. It is necessary to collect a large amount of data so that it is possible to conduct some *training* in *machine learning* mechanisms for the implementation of the NLP prototype performance (Grimmer, & Stewart, 2013). At that moment, due to the limited number of cases, it had not been possible to create a *training set* for the application of a *supervised learning model*. We have also explored a limited number of NLP indicators because of the difficulties to find out Open Tools for Italian Language to be incorporated in our automatic system. In future studies, we are going to use larger corpora and to test new NLP features. These improvements will allow us to carry out more sophisticated statistical analysis such as structural equation modelling and Latent Factor Analysis (MacArthur, Jennings, & Philippakos, 2019).

Sagredo. One day I was at the home of a very famous doctor in Venice, where many persons came on account of their studies, and others occasionally came out of curiosity to see some anatomical dissection performed by a man who was truly no less learned than he was a careful and expert anatomist. It happened on this day that he was investigating the source and origin of the nerves, about which there exists a notorious controversy between the Galenist and Peripatetic doctors. The anatomist showed that the great trunk of nerves, leaving the brain and passing through the nape, extended on down the spine and then branched out through the whole body, and that only a single strand as fine as a thread arrived at the heart. Turning to a gentleman whom he knew to be a Peripatetic philosopher, and on whose account he had been exhibiting and demonstrating everything with unusual care, he asked this man whether he was at last satisfied and convinced that the nerves originated in the brain and not in the heart. The philosopher, after considering for a while, answered: "You have made me see this matter so plainly and palpably that if Aristotle's text were not contrary to it, stating clearly that the nerves originate in the heart, I should be forced to admit it to be true."

Simplicio. Sir, I want you to know that this dispute as to the source of the nerves is by no means as settled and decided as perhaps some people like to think.

Sagredo. Doubtless it never will be, in the minds of such opponents. But what you say does not in the least diminish the absurdity of this Peripatetic's reply; who, as a counter to sensible experience, adduced no experiment or argument of Aristotle's, but just the authority of his bare ipse dixit.

SUMMARY OF THE RESEARCH ACTIVITIES

The results collected during this PhD have both theoretical and practical implications for CT research and assessment. Whoever wants to approach the topic of CT will find difficulties in understanding what CT actually is and, as a consequence, how to teach it and assess it. The problem of the CT definition and operationalisation is far to be solved. However, the PhD work provides an approach to achieve a common understanding of this complex subject.

Firstly, through the method of qualitative content analysis, I was able to identify and quantify commonalities among eleven theory-driven categories in 39 CT definitions. The theory-driven categories are open and controversial issues related to CT definition, such as its transferability, its relation with actions, emotions, and personal dispositions.

Results partially confirmed the CT definitions' classifications proposed by Sternberg (1989) and Lai (2011). I identified two conceptual networks through an analysis of the co-occurrence between the eleven theory-driven categories: the "Normative - Descriptive network" and the "Descriptive – Explanatory network."

Definitions included in the "Normative - descriptive Network" focus on what people are capable of doing under the best of circumstances (Lai, 2011). Examples are "perfections of thought," as Paul (1992) described. This approach also emphasises the qualities or standards of thoughts. CT Skills and Dispositions are seen concerning outcomes, which are considered as transferable to different domains. In Network 1, CT is a synonym for good thinking (Bailin, 1987). Furthermore, in Network 1, CT focuses more on an individual working alone on a problem-based task.

The definitions included in the "Descriptive – Explanatory network" focus more on how people think rather than how they could or should think under ideal conditions (Sternberg, 1986). Consequently, CT definitions are commonly process-related, and personal dispositions to be engaged in a CT process are emphasised. Instead of conceptualizing CT as an individual and internal work on a problem-based task, in the second network, actions and social practices can occur together. The category action is on the borderline between the individual and the inter-individual dimension of CT. The results show that the dichotomy suggested by Sternberg (1989) and Lai (2011) between the *normative-philosophical* and *explanatory-psychological* CT definitions could not be the most suitable way to classify CT definitions. Differences among CT definitions are not merely related to the experts' field of study (philosophy, education, or psychology). Differences could be better understood considering the *focus* of the CT analysis (on the *outcome* or the *process*) and the *unit* of analysis (the individual thinking or the inter-subjective actions and practices). These two networks should not be interpreted as mutually exclusive. On the contrary, they can be seen as different points of view for the study of the same subject.

The results show unexplored similarities among different CT traditions, perspectives, and study methods. These similarities could be exploited to open up an interdisciplinary dialogue among experts and build up a shared understanding of CT. The use of comparable research methodology would also be a necessary step to achieve a better understanding of empirical research results on the most debated CT issues.

The review results also show that closed measures are more associated with the definitions that emphasised outcome and skills (Network 1). On the other hand, open-measures are more commonly associated with CT definitions that emphasised action and process (Network 2).

In this PhD research, I have decided to focus on open-ended measures because, despite the necessity to use it to assess CT, they present several limitations that discourage their adoption. Studies have shown that written responses to constructed-response items provide more information about students' thinking and reasoning processes than the answers to MC items. However, open-ended answer scoring is expensive, and it can be subject to bias. To overcome these limitations, researchers have been working on how to exploit NLP techniques for the automatic assessment of students' written answers. In the second chapter, I presented research that found high levels of agreement in CT assessment of constructed response answers between automatic and human scoring. However, sometimes, research lack transparency probably because researchers develop NLP prototypes for commercial purposes (e.g. CLA, LIWC). We saw, for example, that Klein (2006) reported a high level of reliability correlation between hand and machine assigned mean scores, on the CLA make-an-argument and break-an-argument tasks. However, Klein did not describe the 40 items assessed through the NLP system. Consequently, it is difficult to understand on which aspects human assessors and NLP system "agree". More research is necessary to develop reliable NLP prototypes and to validate the existing ones (Mao et al., 2018).

A further limitation of the current CT automatic assessment tools concerns their transferability to other languages. The most adopted CT automatised assessment tools work only on English written text. There are a few attempts to generalize these techniques to other languages, including the Italian one. NLP analysis, applied to Italian written answers, is preliminary in nature, especially in the context of educational research. Only in a few cases, researchers adopted NLP features to assess learning outcomes or cognitive dimensions.

Two empirical studies have been carried out to overcome the abovementioned limitations, presented respectively in Chapter 3 and 4.

In Chapter 3, I presented a preliminary validation study of an NLP prototype developed by the CDM group. The CDM prototype is based on a CT theoretical framework proposed by Poce (2017). In the first experimentation, the prototype was able to assess four out of six CT sub-indicators (use of language, relevance, importance, and novelty) through different NLP techniques. I used a total of 114 constructed-response answers (18 paraphrases, 18 comments, 78 open-ended answers) written in English by university teachers to test the CDM prototype reliability. Results suggested that the expert evaluator and the NLP prototype can identify the relevance indicator. For the importance indicator, I detected a non-statistically significant moderate correlation between human evaluators and the NLP prototype. The NLP prototype performance was lower for the use of language indicator. Both the NLP prototype and the expert better identified the relevance indicator in the "paraphrase and comment" task. For the importance indicator, I detected a non-statistically significant moderate correlation between human evaluators and the NLP prototype, both in the paraphrase and in the comment. The NLP prototype performance was lower for the use of language indicator.

In the "open-ended questions" task, I found a satisfactory reliability level between two human experts. However, I did not find correlations between the scores given by expert evaluators and the NLP prototype.

The absence of correlation could be due to the wider domain of the "open-ended questions" compared to the "paraphrase and comment" task. In other words, the NLP prototype could easier predict topics in the "paraphrase and comment" rather than in the "open-ended questions" task. Moreover, the prototype can better discriminate relevant and important concepts and notions in the "paraphrase and comment" task because the expected topics are strongly dependent on the provided stimulus. For this reason, we decided to use a task based on a specific stimulus, such as a literary text, in the experimentation presented in Chapter 4.

The research presented in Chapter 4 aimed at seeing how to use NLP features, commonly adopted to assess CT in English written texts, in Italian written texts. The CDM research group identified eight NLP features: i) corpus length, ii) mean sentence length, iii) readability and iv) syntax complexity; v) hapax and vi) lexical extension, vii) verbatim copying; viii) TD-IDF. An algorithm extracted all the NLP features from a total of 206 pre-post-test essays written by 103 university students.

I looked at the correlation between the NLP features extracted from the essays and six CT sub-indicators scores (use of language, argumentation, relevance, importance, critical evaluation, and novelty) assigned by two human experts.

Three NLP features significantly correlate with CT total score, as calculated by human experts. The Corpus Length, the complexity of the syntax, and the TFxIDF. As expected, the Medium Sentence Length negatively correlates with the CT sub-indicator Use of Language. Complexity of the syntax

positively correlates with argumentation. Besides, TFxIDF positively correlates with relevance. On the other hand, we did not find confirmation for some of our expectations.

Although Hapax and Lexical Extension correlate, they did not show any significant correlation with the expected CT sub-indicator Use of Language. A possible explanation of this result is an issue retrieved within many of the assessed essays. We found that students often used a not coherent language within their texts by alternating refined with everyday expressions. In this particular condition, human evaluators provide low scores to the CT sub-indicator Use of Language in contrast with the algorithm that retrieves a high level of Hapax (number of unique words in the text) and Lexicon Extension.

The second expectation we did not confirm concerns the correlation between TFxIDF and the CT sub-indicator novelty. We expected a negative correlation between these indicators (based on Wang, Dong & Ma, 2019). Nevertheless, we found a moderate and positive correlation. In previous studies, TFxIDF was used to assess divergent forms of creativity. However, the introduction of new ideas in CT is a convergent process. Indeed, divergent thought from a single starting point generates varied ideas, whereas convergent thinking starting from multiple points seeks one most true or useful conclusion (Brophy, 2001).

In our case, the algorithm and human evaluators likely looked for different forms of novelty. On one side, the algorithm retrieved divergent new ideas. On the other side, human evaluators search for convergent new ideas.

Limitations and future development

The PhD research presents several limitations to face in future research.

Concerning the critical review (presented in Chapter 1), I am aware of the following limitations. Firstly, I coded the definitions on my own, so I was not able to present the results' reliability. However, to partially overcome this limitation, I inserted Table 7 to show in a transparent way how I coded each definition. In this way, any reader can test the reliability of my codification. A second limitation concerns the selection of CT definitions. I used a mixed-method combining a systematic and non-systematic approach to include as many definitions as possible. Thus, the process of CT definitions identification is not replicable and, possibly, I did not include some CT definitions in the analysis. In future research, it would be necessary to test empirically the theoretical relations I have identified through the qualitative content analysis of CT definitions. Both qualitative and quantitative approaches could provide useful information for a deeper understanding of CT.

For studying CT as a *process*, we should look at contexts of everyday use to examine the factors that contribute to disposition, as opposed to competence to exercise it (Kuhn, 2019).

We can test situations in which people use two or more inter-related skills to improve our understanding of the relation between CT and common overlapping construct (problem-solving, decision-making, creative thinking). The neuroscientific study could help to understand better the relation between CT in its relation with other cognitive functions, such as executive functions (de Acedo Lizarraga, de Acedo Baquedano, & Villanueva, 2012), emotional intelligence (Yao et al., 2018), and problem-solving (Tong et al., 2018). Comparing to other higher-order skills, such as creative thinking, the neuroscientific study related to CT is missing.

Moreover, the use of a qualitative, anthropological, and ethnographic method to explore different voices regarding CT is necessary (Chen, 2015; Moore, 2013). This approach could be particularly useful to reflect on the cultural meaning of the word "critical."

I also acknowledge some limitations concerning the empirical investigations described in Chapters 3 and 4.

Firstly concerning the number of open-ended answers collected. Validation studies on the reliability of automated scoring require large-scale data collection. The cases presented in both Chapter 3 and Chapter 4 are therefore not sufficient to draw conclusions. It would have been necessary to collect a large amount of data to conduct some training in machine learning mechanisms for the implementation of the NLP prototype performance (Grimmer, & Stewart, 2013). Due to the limited number of cases and the different kinds of stimulus used, it had not been possible to create a training set for the application of a supervised learning model.

Regarding the research presented in Chapter 3, a limitation was the difference between the experts' mother tongue and the answers' language. Indeed, Italian native speakers experts assessed English written texts. Although assessors had a high level of proficiency in English, the discrepancy may have negatively influenced the reliability of the use of language indicator by human experts. Moreover, only some participants were English native speakers. In the data collection carried out in Leuven, for privacy reasons, it was not possible to collect additional data on the participants, included their mother tongue. Thus, we were not able to see to which extent the participants' mother tongue affected the evaluation of the use of language indicator.

The experimentation presented in Chapter 4 also includes many limitations. Firstly, we have explored a limited number of NLP indicators because a limited number of Open Tools are accessible for the Italian language. In future studies, it would be necessary to incorporate new NLP features into the algorithm because the most efficient NLP prototypes include more than over 100 NLP indices. Moreover, larger corpora of both English and Italian texts would be necessary. These improvements

will allow to carry out more sophisticated statistical analysis such as structural equation modeling and Latent Factor Analysis.

LIST OF ABBREVIATION

AAC&U = Association of American Colleges and Universities

AASCU = American Association of State Colleges and Universities

ACER = Australian Council for Educational Research

ACT = American College Testing Program

AES = Automated Essay Scoring

AHELO = Assessment of Higher Education Learning Outcome

CAAP = Collegiate Assessment of Academic Proficiency

CAE = Council for Aid to Education

CCTDI = California Critical Thinking Disposition Inventory

CCTST = California Critical Thinking Skills Test

CCTT = Cornell Critical Thinking Tests

CLA = Collegiate Learning Assessment

CMC = Computer Mediated Communication

CRT = Constructed-response tasks

CT = Critical Thinking

DOI = Decision Outcome Inventory

ETS = Educational Testing Service

EWCTET = Ennis-Weir Critical Thinking Essay Test

HCTAES = Halpern Critical Thinking Assessment of the Everyday Situation

HE = Higher Education

ICTET = International Critical Thinking Essay Test

INSBAT = Intelligence Structure Battery

LDA = Latent Dirichlet Allocation

LIWC = Linguistic Inquiry Word Count

LSA = Latent Semantic Analysis

MAPP = Measure of Academic Proficiency and Progress

MC = multiple-choice

MOOC = Massive Open Online Course

OERs = Open Educational Resources

NLP = Natural Language Processing

NTA = Network Text Analysis

OECD = Organization for Economic Co-operation and Development

RWO = Real World Outcome

WGCTA = Watson Glaser Critical Thinking Assessment

LIST OF PUBLICATIONS RELATED TO THE PHD THESIS

o Poce, A., Amenduni, F., De Medio C., Norgini A. (2020) *Assessing critical thinking in open-ended answers: an automatic approach.* In S. K. Softik, D. Andone & A. Szucs (Eds) European Distance and E-Learning Network (EDEN) Proceedings. Human and Artificial Intelligence for the Society of the Future. Inspiring Digital Education for the Next STE(A)M Student Generation. Pp 109-116. DOI: 10.38069/edenconf-2020-ac0008

o Poce, A., Amenduni, F., Re, M. R., & De Medio, C. (2019). Automatic Assessment of University Teachers' Critical Thinking Levels. *International Journal of Advanced Corporate Learning (iJAC)*, *12*(3), 46-58. Retrieved from: https://online-journals.org/index.php/i-jac/article/view/11259/6173

o Re, M. R., Amenduni, F., De Medio, C., & Valente, M. (2019). How to use assessment data collected through writing activities to identify participants' Critical Thinking levels. *Journal of e-Learning and Knowledge Society*, *15*(3), 117-133. https://doi.org/10.20368/1971-8829/1135051

o Poce, A., Amenduni, F., De Medio, C., & Re, M. R. (2019). Road to Critical Thinking automatic assessment: a pilot study. *Form@ re-Open Journal per la formazione in rete*, *19*(3), 60-72.

o Poce, A., Re, M. R., Amenduni, F., De Medio, C., & Valente, M. (2019). Developing a web App to provide personalized feedback for museum visitors: a pilot research project. *Form@ re-Open Journal per la formazione in rete*, *19*(3), 48-59.

o Amenduni, F. (2019). Definire e valutare il pensiero critico attraverso l'analisi del testo scritto. In Giornata della ricerca 2019 del Dipartimento di Scienze della Formazione. V. Carbone, G. Carrus & F. Pompeo (EDS). Pp. 25-28. Roma TrE-Press, Roma, Italia. ISBN 8832136880.

o Poce, A., De Medio C., Amenduni, F. (2020) A Prototype for the Automatic Assessment of Critical Thinking. In: Rehm M., Saldien J., Manca S. (eds) Project and Design Literacy as Cornerstones of Smart Education. Smart Innovation, Systems and Technologies, vol 158. Springer, Singapore;

o Poce, A., De Medio, C., Amenduni, F., & Re, M. R. (2019, September). Critical Thinking assessment: a first approach to the automatic evaluation. In *2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET)* (pp. 1-8). IEEE.

REFERENCES

AAC&U (2005). Liberal education outcomes. Washington, DC: Association of American Colleges and Universities. Retrieved from:
https://www.aacu.org/sites/default/files/files/LEAP/LEAP_Report_2005.pdf

AASCU (2006). Value-added Assessment. Perspectives. Washington, DC: American Association of State Colleges and Universities. Retrieved from:
https://www.aascu.org/uploadedFiles/AASCU/Content/Root/PolicyAndAdvocacy/PolicyPublications/06_perspectives%281%29.pdf

Adams, M. H., Whitlow, J. F., Stover, L. M., & Johnson, K. W. (1996). Critical thinking as an educational outcome: An evaluation of current tools of measurement. Nurse Educator, 21(3), 23-32.

Aloisi, C., & Callaghan, A. (2018). Threats to the validity of the Collegiate Learning Assessment (CLA+) as a measure of critical thinking skills and implications for Learning Gain. Higher Education Pedagogies, 3(1), 57-82

Alexander, P. A. (2014). Thinking critically and analytically about critical-analytic thinking: An introduction. *Educational Psychology Review*, *26*(4), 469-476.

American College Testing Program (1989). Report on the Technical Characteristics of CAAP: Pilot Year 1, 1988-89. Iowa City, IA: Author.

American Psychological Association (2017, 11 November). Heuristics [Blog Post] Retrieved from: https://www.apa.org/pubs/highlights/peeps/issue-105

Andrews, D. H., & Wulfeck, W. H. (2014). Performance assessment: Something old, something new. In Handbook of research on educational communications and technology (pp. 303-310). Springer, New York, NY.

Armborst, A. (2017). Thematic proximity in content analysis. *Sage Open*, *7*(2), 2158244017707797.

Arum, R., Cho, E., Kim, J., & Roksa, J. (2012). Documenting uncertain times: Post-graduate transitions of the Academically Adrift cohort. New York, NY: Social Science Research Council.

Atkinson, D. (1997). A critical approach to critical thinking in TESOL. *TESOL quarterly*, *31*(1), 71-94.

Attali, Y. (2014). A ranking method for evaluating constructed responses. Educational and Psychological Measurement. http://dx.doi.org/10.1177/0013164414527450

Bailin, S., Battersby, M., & Clauss, P. (2011). Reason in the balance: Teaching critical thinking as dialectical.

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of curriculum studies*, *31*(3), 285-302.

Barnett, R. (1997). *Higher education: A critical business*. McGraw-Hill Education (UK).

Bensley, D. A., Lilienfeld, S. O., & Powell, L. A. (2014). A new measure of psychological misconceptions: Relations with academic background, critical thinking, and acceptance of paranormal and pseudoscientific claims. *Learning and Individual Differences*, *36*, 9-18.

Beyer, B. K. (1984). Improving thinking skills: Practical approaches. *The Phi Delta Kappan*, *65*(8), 556-560.

Bok, D.C. (2006). Our underachieving colleges: A candid look at how much students learn and why

Borg, W., and M. Gall. 1989. The methods and tools of observational research. In Educational research: An introduction (5th ed). eds. W. Borg and M. Gall, 473-530. London: Longman.

Brookfield, S. (1995). Adult learning: An overview. *International encyclopedia of education*, *10*, 375-380.

Brophy, D. R. (2001). Comparing the attributes, activities, and performance of divergent, convergent, and combination thinkers. Creativity research journal, 13(3-4), 439-455.

Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (2013). 16 Automated Evaluation of Discourse Coherence Quality in Essay Writing. Handbook of automated essay evaluation: Current applications and new directions, 267.

Butler, H. A. (2012). Halpern Critical Thinking Assessment predicts real-world outcomes of critical thinking. Applied Cognitive Psychology, 26(5), 721-729.

Butler, H. A., Pentoney, C., & Bong, M. P. (2017). Predicting real-world outcomes: Critical thinking ability is a better predictor of life decisions than intelligence. Thinking Skills and Creativity, 25, 38-46.

Byrnes, J. P., & Dunbar, K. N. (2014). The nature and development of critical-analytic thinking. *Educational Psychology Review*, *26*(4), 477-493.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. Psychological bulletin, 119(2), 197.

Cambi, F. (Ed.). (2009). *Pedagogie critiche in Europa: frontiere e modelli*. Carocci.

Chang, W. C., & Ku, Y. M. (2015). The effects of note-taking skills instruction on elementary students' reading. The Journal of Educational Research, 108(4), 278-291.

Chen, L. (2017). Understanding critical thinking in Chinese sociocultural contexts: A case study in a Chinese college. *Thinking Skills and Creativity*, *24*, 140-151.

Chiriatti, G., Della Gala, V., Dell'Orletta, F., Montemagni, S., Pettenati, M. C., Sagri, M. T., & Venturi, G. (2018). A NLP-based Analysis of Reflective Writings by Italian Teachers. In CLiC-it.

Chou, T. L., Wu, J. J., & Tsai, C. C. (2019). Research trends and features of critical thinking studies in e-learning environments: A review. Journal of Educational Computing Research, 57(4), 1038-1077.

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. Educational Psychology Review, 24(2), 205-249.

Clinchy, B., & Zimmerman, C. (1985). *Growing up intellectually: Issues for college women*. Wellesley College, Stone Center for Developmental Services and Studies.

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. Psychological assessment, 4(1), 5.

Covington, M. V., Crutchfield, R. S., Davies, L., & Olton, R. M. (1972). The productive thinking program: A course in learning to think. Columbus, OH: Charles E. Merrill.

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*(3), 282-311.

Cunningham, W. V., & Villaseñor, P. (2016). Employer voices, employer demands, and implications for public skills development policy connecting the labor and education sectors. The World Bank Research Observer, 31(1), 102-134.

Danvers, E. C. (2016). Criticality's affective entanglements: rethinking emotion and critical thinking in higher education. *Gender and Education*, *28*(2), 282-297.

Davies, M. (2015). A model of critical thinking in higher education. In *Higher education: Handbook of theory and research* (pp. 41-92). Springer, Cham.

Davies, M., & Barnett, R. (Eds.). (2015). *The Palgrave handbook of critical thinking in higher education*. Springer.

de Acedo Lizarraga, M. L. S., de Acedo Baquedano, M. T. S., & Villanueva, O. A. (2012). CT, executive functions and their potential relationship. *Thinking Skills and Creativity*, *7*(3), 271-279.

Dell'Orletta, F., Montemagni, S., & Venturi, G. (2011, July). READ–IT: Assessing readability of Italian texts with a view to text simplification. In Proceedings of the second workshop on speech and language processing for assistive technologies (pp. 73-83).

Dewey, J. (1910). *How we think.* D.C. Heath & Co. Boston, New York, Chicago, USA.

Downe-Wamboldt, B. (1992). Content analysis: method, applications, and issues. *Health care for women international*, *13*(3), 313-321.

Dunbar, K. N., Fugelsang, J. A., & Stein, C. (2007). Do Naïve Theories Ever Go Away? Using Brain and Behavior to Understand Changes in Concepts: Kevin N. Dunbar Jonathan A. Fugelsang. In *Thinking with data* (pp. 205-217). Psychology Press.

Dweck, C. (2015). Carol Dweck revisits the growth mindset. *Education Week*, *35*(5), 20-24.

Dumitru, D., Bigu, D., Elen, J., Ahern, A., McNally, C., & O'Sullivan, J. (2018). A European review on critical thinking educational practices in higher education institutions. UTAD.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Series of books in psychology. Teaching thinking skills: Theory and practice* (p. 9–26). W H Freeman/Times Books/ Henry Holt & Co.

Ennis, R. H. (1993). Critical thinking assessment. *Theory into practice*, *32*(3), 179-186.

Ennis, R. H. (2015). Critical thinking: A streamlined conception. In *The Palgrave handbook of critical thinking in higher education* (pp. 31-47). Palgrave Macmillan, New York.

Ennis, R. H. (2018). Critical thinking across the curriculum: A vision. *Topoi*, *37*(1), 165-184.

Ennis, R. H., & Weir, E. E. (1985). *The Ennis-Weir critical thinking essay test: An instrument for teaching and testing*. Midwest Publications.

Epstein, A. S. (2008). An early start on thinking. *Educational Leadership*, *65*(5), 38.

Esplugas, C., & Landwehr, M. (1996). The use of critical thinking skills in literary analysis. *Foreign Language Annals*, *29*(3), 449-461.

Evans, J. S. B. T., & Frankish, K. (Eds.). (2009). In two minds: Dual processes and beyond. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.001.0001

Ezen-Can, A. Boyer, K. E., Kellogg, S. & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (LAK '15). Association for Computing Machinery, New York, NY, USA, 146–150. DOI:https://doi.org/10.1145/2723576.2723589

Facione, N. C., Facione, P. A., & Sanchez, C. A. (1994). Critical thinking disposition as a measure of competent clinical judgment: The development of the California Critical Thinking Disposition Inventory. Journal of Nursing Education, 33(8), 345-350.

Facione, P. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report).

Facione, P. A., & Facione, N. (1994). *The California Critical Thinking Skills Test: CCTST: Test Manual*. California Academic Press.

Felton, M. K., & Kuhn, D. (2007). "How Do I Know?" The Epistemological Roots of Critical Thinking. *Journal of Museum Education*, *32*(2), 101-110.

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

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

Fisher, A., & Scriven, M. (1997). *Critical thinking its definition and assessment*. Centre for research in Critical Thinking.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33, 613–619.

Freire, P. 1972. Pedagogy of the Oppressed. Harmondsworth: Penguin.

Freire, P. 1973. Education for Critical Consciousness. New York: Seabury Press.

Garrison, D. R. (1992). Critical thinking and self-directed learning in adult education: An analysis of responsibility and control issues. *Adult education quarterly*, *42*(3), 136-148.

Garrison, D. R., Anderson, T., & Archer, W. (2001, a). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of distance education*, *15*(1), 7-23.

Garrison, D. R., Anderson, T., & Archer, W. (2001, b). Critical thinking and computer conferencing: A model and tool to assess cognitive presence.

Garrison, D. R., Anderson, T., & Archer, W. (2010). The first decade of the community of inquiry framework: A retrospective. The internet and higher education, 13(1-2), 5-9.

Gilbert, M. A. (1995) Emotional Argumentation, or, Why Do Argumentation Rheorists Argue with their Mates? Analysis and Evaluation: Proceedings of the Third ISSA Conference on Argumentation Vol II.

Giles, J. (2005) Internet encyclopaedias go head to head. *Nature 438*, 900–901 https://doi.org/10.1038/438900a

Gilpin, A., & Wagenaar, R. (2008). Approaches to Teaching, Learning and Assessment in Competence Based Degree Programmes. *Tuning Educational Structures in Europe. Universities' Contribution to the Bologna Process. An Introduction*, 91-118.

Giosi, M. (2009). Prospettive attuali sulla pedagogia critica in area anglo-americana: due modelli. In F. Cambi (Ed.). (2009). *Pedagogie critiche in Europa: frontiere e modelli*. Pp. 53-79. Carrocci.

Glaser, E. M. (1941). *An experiment in the development of critical thinking* (No. 843). Teachers College, Columbia University.

Gloudemans, H. A., Schalk, R. M., & Reynaert, W. (2013). The relationship between critical thinking skills and self-efficacy beliefs in mental health nurses. *Nurse education today*, *33*(3), 275-280.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. Educational researcher, 40(5), 223-234.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41-58). Brill.

Grimberg, B. I. and Hand B., (2009), Cognitive pathways: analysis of students' written texts for science understanding, Int. J. Sci. Educ., 31(4), 503–521.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

Guiller, J., Durndell, A., & Ross, A. (2008). Peer interaction and critical thinking: Face-to-face or online discussion?. *Learning and instruction*, *18*(2), 187-200.

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. Journal of educational computing research, 17(4), 397-431.

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

Halpern, D. F. (2007). Is Intelligence Critical Thinking? Why We Need a New Definition for Intelligence. In P. Kyllonen, I. Stankov, & R. D. Roberts (Eds.), Extending intelligence: enhancement and new constructs (pp. 349-370). Mahwah, NJ: Erlbaum Associates, Inc.

Halpern, D. F. (2008). Is intelligence critical thinking? Why we need a new definition of intelligence. *Extending intelligence. Enhancement and new constructs*, 157-182.

Halpern, D. F. (2013). The Halpern critical thinking assessment: A response to the reviewers. *Inquiry: Critical Thinking Across the Disciplines*, *28*(3), 28-39.

Halpern, D. F. (2016). *Manual Halpern Critical Thinking Assessment*. SCHUHFRIED GmbH, Mödling, Austria. Retrieved from:
https://drive.google.com/file/d/0BzUoP_pmwy1gdEpCR05PeW9qUzA/view

Hatcher, D. L. (2013). Reflections on critical thinking: Theory, practice, and assessment. *INQUIRY: Critical Thinking Across the Disciplines*, *28*(2), 4-24.

Hatcher, D. L., & Spencer, L. A. (2005). *Reasoning and writing: From critical thinking to composition.* American Press, Boston.

Hau, K. T., Halpern, D., Marin-Burkhart, L., Ho, I. T., Ku, K. Y. L., Chan, N. M., & Lun, V. M. C. (2006). *Chinese and United States students' critical thinking: Cross-cultural construct validation of a critical thinking assessment.* American Educational Research Association Annual Conference, San Francisco, 7–11 April.

Heilman, M., & Madnani, N. (2015, June). The impact of training data on automated short answer scoring performance. In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 81-85).

Henri F. (1992) Computer Conferencing and Content Analysis. In: Kaye A.R. (eds) Collaborative Learning Through Computer Conferencing. NATO ASI Series (Series F: Computer and Systems Sciences), vol 90. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-77684-7_8

Hollis, H., Rachitskiy, M., Van der Leer, L., & Elder, L. (2020). Validity and reliability testing of the International Critical Thinking Essay Test form A (ICTET-A).

Hoppe, U. (2017). Computational Methods for the Analysis of Learning and Knowledge Building Communities. In Lang, C., Siemens, G., Wise, A. F., and Gaevic, D., editors, The Handbook of Learning Analytics, pages 23–33. Society for Learning Analytics Research (SoLAR), Alberta, Canada, 1 edition.

Hyytinen, H., Holma, K., Toom, A., Shavelson, R. J., & Lindblom-Ylänne, S. (2014). The Complex Relationship between Students' Critical Thinking and Epistemological Beliefs in the Context of Problem Solving. *Frontline Learning Research*, *2*(5), 1-25.

Hyytinen, H., Nissinen, K., Ursin, J., Toom, A., & Lindblom-Ylänne, S. (2015). Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students. *Studies in Educational Evaluation*, *44*, 1-8.

IEA (2018) *IEA International Computer and Information Literacy Study 2018 assessment framework*. Springer. Switerland.

Inamorato dos Santos, A., Punie, Y., Castaño-Muñoz, J. (2016). Opening up education: A support framework for higher education institutions. JRC Science for Policy Report. doi: 10.2791/293408

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161-1166.

Johnson, R. 1992. The problem of defining critical thinking. In The generalizability of critical thinking, ed. S.P. Norris, 38–53. New York: Teachers College Press.

Johnson, R. 2000. Manifest rationality: A pragmatic theory of argument. Mahwah, NJ: Lawrence Erlbaum.

Johnson, R. H. (2014). *The rise of informal logic: Essays on argumentation, critical thinking, reasoning and politics* (Vol. 2). University of Windsor.

Johnson, R. H., & Hamby, B. (2015). A meta-level approach to the problem of defining 'Critical Thinking'. *Argumentation*, *29*(4), 417-430.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, *58*(9), 697.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, *49*, 81.

Kahneman, D., & Tversky, A. (2013). Choices, values, and frames. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 269-278).

Khan, R. A., & Jawaid, M. (2020). Technology enhanced assessment (TEA) in COVID 19 Pandemic. *Pakistan Journal of Medical Sciences*, *36*(COVID19-S4), S108.

Kirsch, I., Lennon, M. L., Yamamoto, K., & von Davier, M. (2017). Large-scale assessments of adult literacy. In Advancing Human Assessment (pp. 285-310). Springer, Cham.

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. Child Development, 71(5), 1347-1366.

Klein, S. (2007). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. In: D. Nolan & T. Speed (Eds.): Probability and Statistics: Essays in Honor of David A. Freedman. IMS Collections, Vol. 2. Beachwood, OH: Institute for Mathematical Statistics

Klein, S. C., Liu, O. L. E., Sconing, J. A., Bolus, R. C., Bridgeman, B. E., Kugelmass, H. C., ... & Steedle, J. C. (2009). Test Validity Study (TVS) Report.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. Evaluation review, 31(5), 415-439.

Kop, R., Fournier, H., and Durand, G. (2017). A Critical Perspective on Learning Analytics and Educational Data Mining. In Lang, C., Siemens, G., Wise, A. F., and Gaevic, D., editors, The Handbook of Learning Analytics, pages 319–326. Society for Learning Analytics. Research (SoLAR), Alberta, Canada, 1 edition.

Kovanovic, V., Joksimovic, S., Gaevic, D., Hatala, M., and Siemens, G. (2017). Content Analytics: The Definition, Scope, and an Overview of Published Research. In Lang, C., Siemens,

G., Wise, A. F., and Gaevic, D., editors, The Handbook of Learning Analytics, pages 77–92. Society for Learning Analytics Research (SoLAR), Alberta, Canada, 1 edition.

Kovanović, V., Joksimović, S., Gašević, D., & Hatala, M. (2014). Automated content analysis of online discus- sion transcripts. In K. Yacef & H. Drachsler (Eds.), Proceedings of the Workshops at the LAK 2014 Conference (LAK-WS 2014), 24–28 March 2014, IN, Indiana, USA. http://ceur-ws.org/Vol-1137/LA_machinelearning_ submission_1.pdf

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2016). Towards auto- mated content analysis of discussion transcripts: A cognitive presence case. Proceedings of the 6th Interna- tional Conference on Learning Analytics & Knowledge (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 15–24). New York: ACM. doi:10.1145/2883851.2883950

Ku, K. Y. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking skills and creativity*, *4*(1), 70-76.

Ku, K. Y. L., & Ho, I. T. (2010). Dispositional factors redicting Chinese students'critical thinking performance. Personality and Individual Differences, 48, 54-58.

Ku, K. Y. L., Ngai-Man, C., Miu-Chi Lun, V., Halpern, D. F., Marin-Burkhart, L., Hau, K. T., & Ho, I. T. (2006, April). Chinese and US undergraduates' critical thinking skills: Academic and dispositional predictors. In American Educational Research Association Annual Meeting, San Francisco, USA.

Kuhn, D. (1991). *The skills of argument*. Cambridge University Press.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science education*, *77*(3), 319-337.

Kuhn, D. (1999). A developmental model of critical thinking. *Educational researcher*, *28*(2), 16-46.

Kuhn, D. (2019). Critical thinking as discourse. *Human Development*, *62*(3), 146-164.

Kuhn, D. (2020). Why Is Reconciling Divergent Views a Challenge?. *Current Directions in Psychological Science*, *29*(1), 27-32.

Kuhn, D., & Crowell, A. (2011). Dialogic argumentation as a vehicle for developing young adolescents' thinking. Psychological science, 22(4), 545-552 DOI: 10.1177/0956797611402512

Kuhn, D., Zillmer, N., Crowell, A., & Zavala, J. (2013). Developing norms of argumentation: Metacognitive, epistemological, and social dimensions of developing argumentive competence. *Cognition and Instruction*, *31*(4), 456-496 https://doi.org/10.1080/07370008.2013.830618

Kuhn, T. S. (1962). The structure of scientific revolutions. *Chicago and London*.

Kurfiss, J. G. (1988). *Critical Thinking: Theory, Research, Practice, and Possibilities. ASHE-ERIC Higher Education Report No. 2, 1988*. ASHE-ERIC Higher Education Reports, The George Washington University, One Dupont Circle, Suite 630, Dept. RC, Washington, DC 20036-1183.

Kurubacak, G. (2007). Building knowledge networks through project-based online learning: A study of developing critical thinking skills via reusable learning objects. *Computers in human behavior, 23*(6), 2668-2695.

Lai, E. R. (2011). Critical thinking: A literature review. *Pearson's Research Reports*, *6*, 40-41.

Lárusson J., A., & White B. (2012). Monitoring student progress through their written "point of originality". In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12). Association for Computing Machinery, New York, NY, USA, 212–221. DOI:https://doi.org/10.1145/2330601.2330653

Lau, J. Y. (2015). Metacognitive education: Going beyond critical thinking. In *The Palgrave handbook of critical thinking in higher education* (pp. 373-389). Palgrave Macmillan, New York.

Lee, Y. H. (2015). Facilitating critical thinking using the C-QRAC collaboration script: Enhancing science reading literacy in a computer-supported collaborative learning environment. *Computers & Education*, *88*, 182-191.

Leigh, F. (2007). Platonic dialogue, maieutic method and critical thinking. *Journal of Philosophy of Education*, *41*(3), 309-323.

Lipman, M. (1987). Critical thinking: What can it be?. *Analytic Teaching*, *8*(1).

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

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

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

Ma, E. (2018, August). *Two latent methods for dimension reduction and topic modeling*. Towards Data Science A Medium publication sharing concepts, ideas, and codes. Retrieved from: https://towardsdatascience.com/2-latent-methods-for-dimension-reduction-and-topic-modeling-20ff6d7d547

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction?. Reading and Writing, 32(6), 1553-1574.

Macrae, C. N., & Milne, A. B. (1992). A curry for your thoughts: Empathic effects on counterfactual thinking. *Personality and Social Psychology Bulletin*, *18*(5), 625-630.

Magno, C. (2010). The role of metacognitive skills in developing critical thinking. *Metacognition and learning*, *5*(2), 137-156.

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H. S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121-138.

Markle, R., Brenneman, M., Jackson, T., Burrus, J., & Robbins, S. (2013). Synthesizing frameworks of higher education student learning outcomes. *ETS Research Report Series*, *2013*(2), i-37.

Martinovski, B., & Mao, W. (2009). Emotion as an argumentation engine: Modeling the role of emotion in negotiation. *Group decision and negotiation*, *18*(3), 235-259.

Mayring, P. (2004). Qualitative content analysis. *A companion to qualitative research*, *1*, 159-176.

McKlin, T., Harmon, S. W., Evans, W., & Jones, M. G. (2001). Cognitive presence in Web-based learning: A content analysis of students' online discussions. Retrieved from: https://eric.ed.gov/?id=ED470101

McMurray, M. A., Beisenherz, P., & Thompson, B. (1991). Reliability and concurrent validity of a measure of critical thinking skills in biology. *Journal of Research in Science Teaching*, *28*(2), 183-191.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. Behavior Research Methods, 45, 499–515. https://doi.org/10.3758/s13428-012-0258-1.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664.

McNamara, D., Allen, L., Crossley, S., Dascalu, M., and Perret, C. (2017). Natural Language Processing and Learning Analytics. In Lang, C., Siemens, G., Wise, A. F., and Gaevic, D., editors, The Handbook of Learning Analytics, pages 93–104. Society for Learning Analytics Research (SoLAR), Alberta, Canada, 1 edition.

McPeck, J. E. (1981). Critical thinking and education, St. *Mattin's Press, New York.*

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *BEHAVIORAL AND BRAIN SCIENCES*, *34*, 57-111.

Meyer, K. A. (2003). Face-to-face versus threaded discussions: the role of time and higher order thinking. Journal of Asynchronous Learning Networks, 7(3), 55-65.

Mingers, J. (2000). What is it to be critical? Teaching a critical approach to management undergraduates. *Management Learning*, *31*(2), 219-237.

Moon, A., Moeller, R., Gere, A. R., & Shultz, G. V. (2019). Application and testing of a framework for characterizing the quality of scientific reasoning in chemistry students' writing on ocean acidification. *Chemistry Education Research and Practice*. *20*, 484-494.

Moore, B. N., & Parker, R. (2009). *Critical thinking*. Boston, MA: McGraw-Hill.

Moore, R. L., Oliver, K. M., & Wang, C. (2019). Setting the pace: examining cognitive processing in MOOC discussion forums with automatic text analysis. Interactive Learning Environments, 27(5-6), 655-669.

Moore, T. (2013). Critical thinking: Seven definitions in search of a concept. *Studies in Higher Education*, *38*(4), 506-522.

Moreira P., Marzabal A. and Talanquer V., (2019), Using a mechanistic framework to characterise chemistry students' reasoning in written explanations, Chem. Educ. Res. Pract., 20, 120–131.

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

Mulnix, J. W. (2012). Thinking critically about critical thinking. *Educational Philosophy and theory*, *44*(5), 464-479.

Newman, D. R., Johnson, C., Webb, B., & Cochrane, C. (1997). Evaluating the quality of learning in computer supported co-operative learning. *Journal of the American Society for Information science*, *48*(6), 484-495.

Newman, D. R., Webb, B., & Cochrane, C. (1995). A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology*, *3*(2), 56-77.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175-220.

Nicol, D. (2007). E-assessment by design: Using multiple-choice test to good effect. *Journal for Further and Higher Education, 31*, 53–64.

Nicol, D. (2007). E-assessment by design: Using multiple-choice test to good effect. *Journal for Further and Higher Education, 31*, 53–64.

Norris, S. P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement*, *27*(1), 41-58.

Norris, S. P. (Ed.). (1992). *The generalizability of critical thinking: Multiple perspectives on an educational ideal*. Teachers College Press.

Norris, S. P., & King, R. (1983). *Test on appraising observations*. Institute for Educational Research and Development, Memorial University of Newfoundland.

Nussbaum, M. (1998) Transcript of Newshour Inerview by David Gergen, March 5.

OECD (2007). Giving knowledge for free: The emergence of open educational resources. Retrieved from:

http://www.oecd.org/education/ceri/givingknowledgeforfreetheemergenceofopeneducationalresources.htm

OECD (2012) Assessment of Higher Education Learning Outcomes. AHELO Feasibility Study Report. Retrieved from: http://www.oecd.org/education/skills-beyond-school/AHELOFSReportVolume1.pdf

OECD (2012) Assessment of Higher Education Learning Outcomes. AHELO Feasibility Study Report. Retrieved from: http://www.oecd.org/education/skills-beyond-school/AHELOFSReportVolume1.pdf

Oh, E. G., Huang, W. H. D., Mehdiabadi, A. H., & Ju, B. (2018). Facilitating critical thinking in asynchronous online discussion: comparison between peer-and instructor-redirection. Journal of Computing in Higher Education, 30(3), 489-509.

Pascarella, E. T., & Terenzini, P. T. (1991). *How College Affects Students: Findings and Insights from Twenty Years of Research. Vol.[1]*.

Paul, R. (2011). *Critical Thinking Movement: Three Waves*. Retrieved from: http://www.criticalthinking.org/pages/critical-thinking-movement-3-waves/856

Paul, R., & Elder, L. (2006). The miniature guide to critical thinking: Concepts & tools. Dillon Beach, CA: Foundation for Critical Thinking.

Paul, R., & Elder, L. (2012). *The international critical thinking reading and writing test*. Rowman & Littlefield, Tomales (CA). https://www.criticalthinking.org/files/ReadWritingTestOp1.pdf

Paul, S. A. (2014). Assessment of critical thinking: a Delphi study. *Nurse Education Today*, *34*(11), 1357-1360.

Piaget J., (1954). *The construction of reality in the child*. New York: Basic.

Poce, A. (2015). Tecnologia critica, creatività e didattica della scienza. Franco Angeli, Milano.

Poce, A. (2017). *Verba sequentur: pensiero e scrittura per uno sviluppo critico delle competenze nella scuola secondaria*. Franco Angeli.

Poce, A. (Ed.). (2012). *Contributi per la definizione di una tecnologia critica: un'esperienza di valutazione*. Franco Angeli.

Poce, A., & Amenduni, F. (2019, May). Creative writing and Critical Thinking Enhancement at Higher Education. In *Fifth International Conference on Higher Education Advances*.

Poce, A., Amenduni, F., Re, M. R., & De Medio, C. (2019). Establishing a MOOC Quality Assurance Framework--A Case Study. *Open Praxis, 11*(4), 451-460.

Poce, A., Amenduni, F., Re, M., R., & De Medio Carlo (2019). Automatic assessment of university teachers' critical thinking level. In the proceedings of the International Conference on E-Learning in the Workplace 2019. Retrieved from: https://www.icelw.org/proceedings/2019/ICELW2019/Papers/Poce_Amenduni_et_al.pdf

Poce, A., Corcione, L., & Iovine, A. (2012). Content analysis and critical thinking. An assessment study. *Cadmo*.

Poce, A., De Medio, C., & Amenduni, F. (2020). A Prototype for the Automatic Assessment of Critical Thinking. In Project and Design Literacy as Cornerstones of Smart Education (pp. 143-151). Springer, Singapore.

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. Perceptual and motor skills, 105(3), 997-1014.

Popham, W. J. (2003). Test better, teach better. The instructional role of assessment. Alexandria, VA: ASCD.

Possin, K. (2013). Some Problems with the Halpern Critical Thinking Assessment (HCTA) Test. Inquiry: Critical Thinking Across the Disciplines, 28(3), 4-12.

Rahimi, Z., Litman, D., Correnti, R., Wang, E., & Matsumura, L. C. (2017). Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. International Journal of Artificial Intelligence in Education, 27(4), 694-728.

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. Psychological Test and Assessment Modeling, 52(4), 354.

Rear, D. (2019). One size fits all? The limitations of standardised assessment in critical thinking. As-sessment & Evaluation in Higher Education, 44(5), 664-675.

Reznitskaya, A. (2012). Dialogic teaching: Rethinking language use during literature discussions. *The reading teacher*, *65*(7), 446-456.

Riesenmy, M. R., Mitchell, S., Hudgins, B. B., & Ebel, D. (1991). Retention and transfer of children's self-directed critical thinking skills. *The Journal of Educational Research*, *85*(1), 14-25.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology, 17(4), 688-690.

Ryser, G. R., Beeler, J. E., & McKenzie, C. M. (1995). Effects of a Computer-Supported Intentional Learning Environment (CSILE) on students' self-concept, self-regulatory behavior, and critical thinking ability. *Journal of Educational Computing Research*, *13*(4), 375-385.

Saadé, R. G., Morin, D., & Thomas, J. D. (2012). Critical thinking in E-learning environments. *Computers in Human Behavior*, *28*(5), 1608-1617.

Sadeghi, B., Hassani, M. T., & Rahmatkhah, M. (2014). The Relationship between EFL Learners' Metacognitive Strategies, and Their Critical Thinking. *Journal of Language Teaching & Research*, *5*(5).

Salton, G. and McGill, M.J. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill Book Co., New York.

Sasson, I., Yehuda, I., & Malkinson, N. (2018). Fostering the skills of critical thinking and question-posing in a project-based learning environment. Thinking Skills and Creativity, 29, 203-212.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building. *The Cambridge*.

Scheffer, B. K., & Rubenfeld, M. G. (2000). A consensus statement on critical thinking in nursing. *Journal of Nursing Education*, *39*(8), 352-359.

Scriven, M., & Paul, R. (1987, August). Critical thinking as defined by the National Council for Excellence in Critical Thinking. In *8th Annual International Conference on Critical Thinking and Education Reform, Rohnert Park, CA* (pp. 25-30).

Shavelson, R. J. (2008). The collegiate learning assessment. In *Forum for the Future of Higher Education/Ford*. Retrieved from: https://www.researchgate.net/profile/Richard_Shavelson/publication/271429276_The_collegiate_learning_assessment/links/54f5ffc00cf27d8ed71d30d2/The-collegiate-learning-assessment.pdf

Shepard, L. A. (2000). The role of assessment in a learning culture. Educational researcher, 29(7), 4-14.

Siegel, H. (1989). The rationality of science, critical thinking, and science education. *Synthese*, *80*(1), 9-41.

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 252-254).

Simsek, D., Buckingham Shum, S., Sandor, A., De Liddo, A., & Ferguson, R. (2013). XIP Dashboard: Visual analytics from automated rhetorical parsing of scientific metadiscourse. Presented at the 1st Internation- al Workshop on Discourse-Centric Learning Analytics, 8 April 2013, Leuven, Belgium. http://oro.open. ac.uk/37391/1/LAK13-DCLA-Simsek.pdf

Skulmoski, G. J., Hartman, F. T., & Krahn, J. (2007). The Delphi method for graduate research. *Journal of Information Technology Education: Research*, *6*(1), 1-21.

Sosu, E. M. (2013). The development and psychometric validation of a Critical Thinking Disposition Scale. Thinking skills and creativity, 9, 107-119.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2011). Individual differences as essential components of heuristics and biases research. In K. Manktelow, D. Over, & S. Elqayam (Eds.), *The science of reason: A festschrift for Jonathan St B. T. Evans* (p. 355–396). Psychology Press.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259-264.

Sternberg, R. J. (1986). Critical Thinking: Its Nature, Measurement, and Improvement. National Inst. of Education (ED), Washington, DC. Retrieved from:
https://eric.ed.gov/?id=ED272882

Sternberg, R. J., & Halpern, D. F. (Eds.). (2020). *Critical thinking in psychology*. Cambridge University Press.

Sternberg, R. J., Castejón, J. L., Prieto, M. D., Hautamäki, J., & Grigorenko, E. L. (2001). Confirmatory factor analysis of the Sternberg Triarchic Abilities Test in three international samples: An empirical test of the triarchic theory of intelligence. *European Journal of Psychological Assessment, 17*(1), 1–16. https://doi.org/10.1027//1015-5759.17.1.1

Strijbos, J. W., Martens, R. L., Prins, F. J., & Jochems, W. M. (2006). Content analysis: What are they talking about?. Computers & education, 46(1), 29-48.

Taube, K. T. (1997). Critical thinking ability and disposition as factors of performance on a written critical thinking test. The Journal of General Education, 46(2), 129-164.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

Thayer-Bacon, B. J. (1993). Caring and Its Relationship to Critical Thinking. *Educational Theory*, *43*(3), 323-40.

Thomson, P. (1998). A Sociohistorical View of Reason and Emotion in Academic Debate.

Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2017). Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism. *International Journal of Science and Mathematics Education*, *15*(4), 663-682.

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of Critical Thinking Instruction in Higher Education: A Systematic Review of Intervention Studies. *Higher Education Studies*, *4*(1), 1-17.

Tong, D., Lu, P., Li, W., Yang, W., Yang, Y., Yang, D., ... & Zhang, Q. (2019). CT and regional gray matter volume interact to predict representation connection in scientific problem solving. *Experimental brain research*, 1-10.

Tovar Caro, E., & Lesko, I. (2014). Analysis of successful modes for the implementation and use of Open Course Ware (OCW) & Open Educational Resources (OER) in higher education. the virtual mobility case. RIED. Revista Iberoamericana de Educación a Distancia, 17(1), 131-148. doi: 10.5944/ried.17.1.11577

U. S. Department of Education (2006). A test of leadership: Charting the future of U.S. higher education. Washington, DC. Retrieved from:
https://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf

Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning ap-proaches. International Journal of Artificial Intelligence in Education, 29(2), 217-257.

UNESCO (2015). *The future of learning 2: what kind of learning for the 21st century skills?* Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000242996

Vacca, F. (1972). Per una critica quantitativa: romanzi a chilometri, Il Messaggero.

van Brussel, S., Timmermans, M., Verkoeijen, P., & Paas, F. (2020). 'Consider the Opposite'–Effects of Elaborative Feedback and Correct Answer Feedback on Reducing Confirmation Bias–a Pre-registered Study. *Contemporary Educational Psychology*, 101844.

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

Verburgh, A., François, S., Elen, J., & Janssen, R. (2013). The assessment of critical thinking critically assessed in higher education: A validation study of the CCTT and the HCTA. *Education Research International*, *2013*.

Vygotsky, L. S. (1981). The instrumental method in psychology. The concept of activity in Soviet psychology, 135-143.

Wade, C. (1995). Using writing to develop and assess critical thinking. *Teaching of psychology*, *22*(1), 24-28.

Wagenaar, R. (2018). What do we know–What should we know? Measuring and comparing achievements of learning in European Higher Education: initiating the new CALOHEE approach. In *Assessment of Learning Outcomes in Higher Education* (pp. 169-189). Springer, Cham.

Walters, K. S. (Ed.). (1994). *Re-thinking reason: New perspectives in critical thinking*. SUNY Press.

Walton, D. N. (1989) Informal Logic: A Handbook for Critical Argumentation. New York:

Walton, D. N. (1992) The Place of Emotion in Argument. The Pennsylvania State U.P.,

Wang, K., Dong, B., & Ma, J. (2019, May). Towards Computational Assessment of Idea Novelty. In Proceedings of the 52nd Hawaii International Conference on System Sciences.

Wang, Q., Woo, H. L., & Zhao, J. (2009). Investigating critical thinking and knowledge construction in an interactive learning environment. Interactive learning environments, 17(1), 95-104.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, *12*(3), 129-140.

Waters, Z. (2015). Using structural features to improve the automated detection of cognitive presence in online learning discussions (B.Sc. Thesis). Queensland University of Technology.

Watson, G., & Glaser, E. M. (1980). *Watson-Glaser critical thinking appraisal: forms A and B; manual*. Psychological Corporation.

Wegerif, R., McLaren, B. M., Chamrada, M., Scheuer, O., Mansour, N., Mikšátko, J., & Williams, M. (2010). Exploring creative thinking in graphically mediated synchronous dialogues. Computers & Education, 54(3), 613-621.

Weller, M. (2017). The development of new disciplines in Education – the Open Education example. In G. M. dos Santos Ferreira, L. A. da Silva Rosado, & J. de Sà Carvalho (Eds.), Education and technology: critical approaches Sa (pp. 464–486). Rio de Janeiro, BRA: Universidade Estacio de Sà. Retrieved from http://oro.open.ac.uk/49737/

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, *100*(4), 930.

Willingham, D. T. (2007). Critical thinking: Why it is so hard to teach?. *American federation of teachers summer 2007, p. 8-19*.

Wilson, G., Abbott, D., De Kraker, J., Salgado Perez, P., Scheltinga, C., & Willems, P. (2011). The lived experience of climate change: Creating open educational resources and virtual mobility for an innovative, integrative and competence-based track at masters level. International Journal of Technology Enhanced Learning, 3(2), 111-123. doi: 10.1007/978-3-642-16318-0_59

World Economic Forum (2018). *The future of Jobs Report 2018*. Retrieved from: http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf

World Health Organization (2020). Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. Retrieved from: https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. Journal of Second Language Writing, 28, 53-67.

Yao, X., Yuan, S., Yang, W., Chen, Q., Wei, D., Hou, Y., ... & Yang, D. (2018). Emotional intelligence moderates the relationship between regional gray matter volume in the bilateral temporal pole and CT disposition. *Brain imaging and behavior*, *12*(2), 488-498.

Zahner, D., & James, J.K. (2015). *Predictive validity of a critical thinking assessment for post-college outcomes*. New York, NY: Council for Aid to Education. Retrieved from: https://eric.ed.gov/?id=ED582251

Zeidler, D. L., Lederman, N. G., and Taylor, S. C. 1992. "Fallacies and Student Discourse: Conceptualizing the Role of Critical Thinking in Science Education." Science Education 76 (4): 437–450.

Zhang, Q., & Zhang, J. (2013). Instructors' positive emotions: Effects on student engagement and critical thinking in US and Chinese classrooms. *Communication Education*, *62*(4), 395-411.

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. Computers & Education, 143, 103668.