HOW TO USE ASSESSMENT DATA COLLECTED THROUGH WRITING ACTIVITIES TO IDENTIFY PARTICIPANTS' CRITICAL THINKING LEVELS

Maria Rosaria Re, Francesca Amenduni, Carlo De Medio, Mara Valente

University of Roma Tre, Italy
{mariarosaria.re; francesca.amenduni; carlo.demedio}@uniroma3.it;
mar.valente19@stud.uniroma3.it

The present paper aims at presenting the Critical Thinking (CT) Skills assessment results in teachers participating in the Erasmus+ KA203 CRITHINKEDU summit (Critical Thinking Across the European Higher Education Curricula), organised in Leuven in June 2019. Within the summit, a workshop was organized to promote in participants' CT skills knowledge, especially in terms of CT assessment methods through open-ended questions. Based on our theoretical assumptions, description and interpretation activities of written text promote skills such as Analysis, Argumentation, Inference and Critical evaluation, which can also be defined in terms of improvement of language skills. Teachers participating in the workshop were assessed through a test composed by literary text paraphrase and commentary exercises; a prototype for the automatic assessment of CT in open-ended answers was used to evaluate the open-answers. Also three human raters evaluated the answers' texts. The goal of the present research

was to verify the assessment method reliability and to collect some data useful for the implementation of the automatic prototype.

1 Introduction

The definition of Critical Thinking (CT) in education has been representing a crucial issue of scholarly debate for the last century and still is today. It is a central topic of discussion not only in the field of education, given its significant implications in many areas of knowledge, ranging from philosophy to science and from technological innovation to economics. CT skills are more and more defined by educational policy as pivotal for human and social progress in terms of innovation, economic and knowledge growth (World Economic Forum, 2016; Scott, 2015). The promotion of CT learning and teaching methods and assessment tools should be considered as an urgent need in all the formal educational context, taking into consideration the different dispositions and cognitive skills to be promoted at school and university level. According to Paul and Elder (2006), there is a significant relationship between literature and CT development; moreover, Bloom (2000) highlights that reading literature is fundamental in order to know ourselves: close and individual reading allows for memorization, without which we are not able to think (Poce, 2017). According to Esplugas and colleagues (1996), thanks to an in-depth analysis of literary text, many meaningful actions may be encouraged to develop CT, for example: the identification of multiple meanings in the literary text, the use of background knowledge and the recreation of those processes leading the author to conceive the text in the form we read it. Once the great value of the literary text for the purposes of CT development has been shown, it is necessary to reflect upon the best tools suitable to achieve our teaching objectives and for assessing them.

2 Assessing Critical Thinking

Writing is widely considered to be one of the most effective practices for interpretation, elaboration and argumentation purposes. Moreover, writing activities present positive aspects for collecting data useful in terms of CT monitoring and evaluation (Poce, 2017). However, a general lack of agreement on the definition of CT led to the production of different assessment methods. Indeed, the conceptualization and the assessment of CT are interdependent issues that must be discussed together: the definition of CT determines how to best measure it. The most common measurements fall into four categories (Ku, 2009; Liu, Frankel, & Roohr, 2014): 1. multiple choices (e.g. Watson & Glaser, 1980; Facione, 1990b); 2. open-ended answers (e.g. Ennis & Weir, 1985); 3. Self-report measures (e.g. Facione, Facione & Sanchez, 1994); 4.

mixed methods (e.g. Halpern, 2007).

Although multiple choice tests could guarantee a higher reliability, they present problems in terms of validity (Poce, 2017). Ennis (1993) recommends the adoption of the short essay because it allows to assess the CT underlying dimensions and personalize the assessment tool based on the teachers' educational objectives. Open-ended questions offer the benefit of evaluating CT on the basis of all dimensions (skills and dispositions, defined by Facione, 1990a). Ennis (1993, p.185) suggests the adoption of the short essay for assessment purposes and distinguishes three structure levels: high, medium and low. There are numerous examples for each of the three levels. For instance, Ennis Weir Critical Thinking Test (1985) was created for the most advanced structure, while the Illinois Critical Thinking Essay Contest (Powers, 1989) was created for the lowest level. Despite these positive aspects, essays and open-ended measures could present problems related with inter-rater reliability and high-cost of scoring. Automated scoring could be a viable solution to these concerns (Liu, Frankel, & Roohr, 2014).

Starting from these assumptions, the Center for Museum Studies – CDM research group autonomously developed a prototype for CT assessment on the basis of the studies carried out by Ennis and Newman, Webb and Cochrane (1995) which aims at meeting validity and reliability criteria to gain relevant information for future data collection.

The prototype is based on a rubric developed in previous research by Poce (2017) aimed at evaluating CT through short essays or open-ended answers and overcoming the problems of reliability related to CT assessment in open-ended questions. The rubric is composed by six different indicators: *Use of Language*, *Justification*, *Relevance*, *Importance*, *Critical Evaluation* and *Novelty* (Poce, 2017). The prototype has been adopted to automatically assess four of the six CT macro-indicators: *Use of Language*, *Relevance*, *Importance*, and *Novelty*.

In the present paper we will present the results of CT skills in professor participating in the workshop *How to assess critical thinking skills through writing?* organised in June 2019 within the CRITHINKEDU project. The assessment data has been analysed by involving expert human evaluators together with the automatic assessment method in order to collect preliminary validity evidence regarding the use of our CT assessment method. More specifically, the research here presented is aimed at answering to the following research questions:

Which level of CT are shown by participants in the sample analysed?

Which level of reliability are shown respectively by the manual and the automatic assessment methods?

2.1 Context of the research: the CRITHINKEDU Project

The CRITHINKEDU project (Critical Thinking Across the European Higher Education Curricula) is an Erasmus+ KA203 Strategic Partnership project started in September 2017 and lasted 36 months. The universities participating in the project are 10 from 9 different countries: Universidade de Trás-Os-Montes e Alto Douro (coordinator, Portugal), Universidad de Santiago de Compostela (Spain), University of Roma TRE (Italy), University of Wester Macedonia (Greece), University of Thessaly (Greece), National University of Ireland (Ireland), UC Leuven (Belgium), Siuolaikiniu Didaktiku Centras (Lithuania), Vysoka Skola Ekonomicka V Praze (Czech Republic) and Academia de Studii Economice din Bucuresti (Romania).

The project arises from the background and the experience of European Higher Education Institutions, business corporations and Non-Governmental Organizations, and their ongoing concern to improve the quality of learning in universities and across different sectors, which converge in a common need on how to better support the development of CT according to labour market needs and social challenges.

The main objective of the project is to design a model of CT university teaching and learning activities to be adopted at transnational level and in the various partners' courses, promoting CT education around Europe and providing an academic environment that supports the diverse cultural learning needs of international students.

After a first analysis of CT disposition and skills needed in different fields of work and an analysis of the university learning and teaching context in terms of CT promotion, the CRITHINKEDU course was designed in order to promote and support quality teaching on CT (Dominguez, 2018). It provides educational resources and practical training activities within different key topics, such as learning design, teaching methods and CT assessment. By engaging teachers with effective instructional design principles, teaching strategies, and assessment criteria for CT, they were encouraged to integrate them in the daily teaching practice. The CRITHINKEDU project realized and published an educational Protocol on CT development (Elen *et al.*, 2019) which reflects a historically situated, operational understanding of the theoretical and empirical research on CT on one hand, and actual experiences with developing CT on the other.

2.2 Methodology

As part of the CRITHINKEDU research and dissemination activities, the *First European Summit of Critical Thinking* was organized in Leuven in June

3rd, 2019 at KU Leuven in Belgium. The Summit involved higher education researchers and educators, deans, student support agencies, policymakers and employers eager to invest in CT education. During the Summit, different workshops were organized in order to support a deeper analysis on CT learning and teaching methods at university level: teachers from different fields of study had the possibility to enhance their knowledge on the topic.

In particular, the workshop "*How to assess critical thinking skills through writing?"* was aimed at presenting different tools for assessing CT and analysing them from a pedagogical point of view, promoting participants' knowledge acquisition in CT assessment methods context and their critical reflection on the topic. The workshop was composed by the following sections:

1. CT assessment tools presentation: different tools for assessing CT were proposed and analysed from a pedagogical point of view, highlighting the relationship between learning objectives, tools and university teaching methodologies.

2. Text paraphrase and commentary to promote and assess CT skills: the *Verba sequentur* model. The model designed by the research group author of the present paper, within the *Verba sequentur* project, was presented to workshop participants and discussed. The model was designed taking into consideration the research hypothesis by which text description and interpretation through writing led to the development of student CT skills. It was also analysed as an assessment model in different fields of study, from the social sciences to the humanities and STEM. All the indicators of the prototype for CT assessment were in-depth analysed by participants.

3. CT assessment tool design. The prototype for CT assessment was used in order to create new CT tests in different fields of study and teaching. Participants were divided in group taking into consideration their fields of study: Social Sciences, STEM, Humanities, Health, Business and political studies. Each group had to design a teaching activity, addressed to university students and aimed at CT skills promotion, and elaborate the related CT assessment test, taking into consideration the model proposed in the previous section.

4. A final plenary session allowed participants to present the evaluation tools realised and to discuss them together with the workshop presenters.

At the beginning of the workshop, the participants' CT skills level was evaluated through a particular kind of text composed by literary text paraphrase and commentary exercises, elaborated taking into consideration the *Verba sequentur* model.

2.3 How to automatically assess Critical Thinking

In recent years, the idea to support Critical Thinking assessment through automatic scoring has been growing. In a review from Liu, Frankel and Roohr (2014) the authors presented different tools to assess automatically CT both for short-answer and essay questions. Answers' contents (e.g., knowledge accuracy) are mainly assessed in short-answer items. C-rater and c-rater-ML are two tools commonly used to automatically evaluate open answers, both developed by Educational Testing Service (ETS). These two tools utilize natural language processing techniques to score knowledge accuracy (Mao *et al.*, 2018). On the other hand, the writing quality of the responses (e.g., grammar, coherence and argumentation) are usually assessed in short essays. For instance, a functional model to evaluate automatically arguments in dialogical and argumentative contexts was proposed by Gordon, Prakken and Walton (2007). In addition, it was also developed a computational model to identify moments within e-discussion in which students adopted critical and creative thinking (Wegerif *et al.*, 2010). Developing a computational model to identify Critical Thinking levels in students' written comments could provide many advantages. For instance, an automatic program could assist researchers and teachers in finding key aspects of Critical Thinking in big amounts of data in Learning Management System platforms. Results could be used to implement the digital learning environment (Miranda, Marzano, & Lytras, 2017) and students learning engagement (Gaeta *et al.*, 2017). In the field of Learning Analytics (Siemens & Baker, 2012), a growing number of studies have been focusing on the automatic analysis of big corpus of linguistic data (Ezen-Can *et al.*, 2015; McNamara *et al.*, 2017). Nevertheless, before adopting these kinds of tools to automatically assess Critical Thinking, 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 few studies have evaluated the accuracy of automatic scoring test for Critical Thinking Assessment (Mao *et al.*, 2018). From our perspective, more research is needed in terms of development and validation of automatic tools for Critical Thinking assessment. Within the research group, the idea to develop an automatic tool for Critical Thinking assessment has been recently started. The tool is organized in four main modules that allow to perform all the operations necessary to obtain the experimental results. The four modules are described below:

1. *Authentication Manager*: the module allows online registration via email and provides a secure login form to access the services offered. Every operation within the system is logged anonymously.
2. *Input module*: this module manages the insertion of the questions and

answers to be evaluated. A title, the text of the question and a *golden answer* are required for each question. Users are also asked to include words representing the *concepts* and the *successors* respectively for the evaluation of importance and novelty. 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.

3. *Manual evaluator*: through this module, experts can manually evaluate the answers.

4. *Automatic evaluator*: this module is the heart of the system which uses two external tools to perform the automatic evaluation for the four indicators presented.

Use of language: the system uses an external tool that provides a value calculated by normalizing the number of errors considering the number of words contained in the answer.

Relevance: the indicator is assessed carrying out an analysis of the concepts. The text is processed by a Part of Speech Tagger, a software that extracts entities such as nouns and verbs from any kinds of text. After a stemming process that reduce the words to their root, an algorithm is applied on this set of nouns by generating n-grams with a length from one to three. The number of the intersection between the n-grams and the concepts will give the relevance of the answer.

Importance: the system exploits an open source knowledge base. Initially, the text of the answer is sent to an online tagging service through entities pages. The service returns a set of entities pages associated with a given text, in our case the text of the answer. Afterwards, each defined concept is automatically linked to its page. All the outgoing links of this page are considered. The importance indicator is given by the number of known pages that the tagging service system detects respectively from the answers given by the participants and from the concepts defined by the assessor/researcher.

Novelty: the indicator is assessed carrying out an analysis of the successors. As for the relevance indicator, all the nouns and n-grams are extracted from the answers' texts. The frequency of intersections between n-grams and successors results in the novelty dimension of the answer

2.4 Data collection: CT assessment test

18 participants took part in the workshop. The participants were mainly European university teachers involved in the field of CT promotion and evaluation in HEI context. For privacy reasons, data were collected

anonymously. The participants were asked to write in 20 minutes a paraphrase and a comment starting from an extract of Galileo Galilei "*Dialogue Concerning the Two Chief World Systems*".

The participants were provided with a template which included the following instructions (Figure 1):

1. Paraphrase: translating the author's wording into your own words (from 45 to 105 words).
2. Write your comment: what is the meaning that the author wants to convey? (from 75 to 200 words).
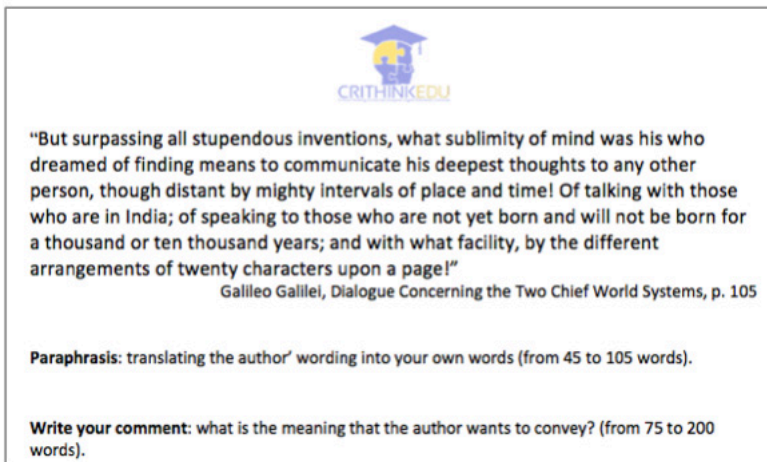

Fig. 1 - CT assessment test used during the workshop

After 20 minutes of the writing activities, participants were invited to reflect upon the assessment of CT through the written analysis of literary texts and providing feedback. The use of paraphrase and commentary exercise depended on the workshop objectives: literary text paraphrase and commentary require the simultaneous use of textual, linguistic and expression skills and they also set up and mobilise CT, analysis and argumentation skills. Paraphrase requires participants to rewrite the literary text by reproducing the original meaning and smoothing out the semantic, lexical, syntactic and content difficulties (Serianni *et al.*, 2003). Paraphrase is based on a thorough understanding of the meaning of the source text and favours the skill in making a comprehensible text in a form that differs from the original one chosen by the author. The commentary of the literary text requires workshop participants to provide a single and deep interpretation of the whole text created by the author, stating, elaborating and exemplifying the thesis of the extract, the author' purpose, the most significant

information and concepts. Accordingly, commentary "actively" involves workshop participants who, while defining the main text elements, must explain and assign the meaning(s) which characterize(s) the text by discussing their interpretation in a critical manner.

At the end of the workshop, participants' written answers were collected and subsequently transcribed in an electronic format in order to be assessed by our prototype for CT Assessment.

2.5 Data analysis

Three human raters with prior experience in CT evaluation, assessed both paraphrase and comment by using a rubric developed by Poce (2017). Although on the comment all the six macro-indicators were applied, the macro-indicator "novelty" was not applied to assess paraphrase since the task does not require the emergence of new ideas. The prototype assessed the answers by applying three macro-indicators on the paraphrase (*Use of Language*, *Relevance*, *Importance*) and four macro-indicators on the comment (*Use of Language*, *Relevance*, *Importance*, and *Novelty*).

The prototype used *concepts* and *successors* provided by the experts and a *golden text* collected during the workshop. As suggested by Mao and colleagues (2018), this study used the quadratic-weighted kappa (QWK) and Pearson product-moment correlation to evaluate the agreement between the three raters' scores and between human raters and the prototype. QWK is a measure of score agreement between raters beyond that expected by chance (Fleiss & Cohen, 1973). The coefficient is a number between 0 and 1, with 0 indicating agreement no better than that expected by chance and 1 indicating perfect agreement. QWK is statistically equivalent to an interrater reliability coefficient (Fleiss & Cohen, 1973). Pearson correlation is another criterion to evaluate consistency between two raters.

2.6 Results

In figure 2, we compared the participants' performance average scores on the six macro-indicators of CT, respectively in paraphrase and commentary.

It is possible to see that participants achieved higher scores in commentary than in paraphrase and this could be explained by two different reasons. Firstly, international participants during the workshop declared they were not familiar with the paraphrase exercise, that is instead commonly used to teach language and literature in Italy from primary schools[1]. On the other hand, according to the

[1] Italian National Guidelines for Primary and Middle School Education, 2012.
http://www.indicazioninazionali.it/wp-content/uploads/2018/08/Indicazioni_Annali_Definitivo.pdf

Verba Sequentur hypothesis (Poce, 2017) supported by Paul and Elder (2006), paraphrase is an exercise that facilitates the adoption of more sophisticated level of CT. Moreover, participants obtained a good average score only for the macro-indicator *Use of Language*, both in paraphrase and commentary (from 2,9 to 3,4). The average score could be considered sufficient for *Argumentation/Justification* and *Importance* both in paraphrase and commentary and also for *Critical Evaluation* and *Relevance* but only in the commentary (from 2,3 to 2,8). The average score could be not considered satisfactory for the indicators *Critical Evaluation* and *Relevance* in the paraphrase and for the indicator *Novelty* in the commentary (less than 2,2).
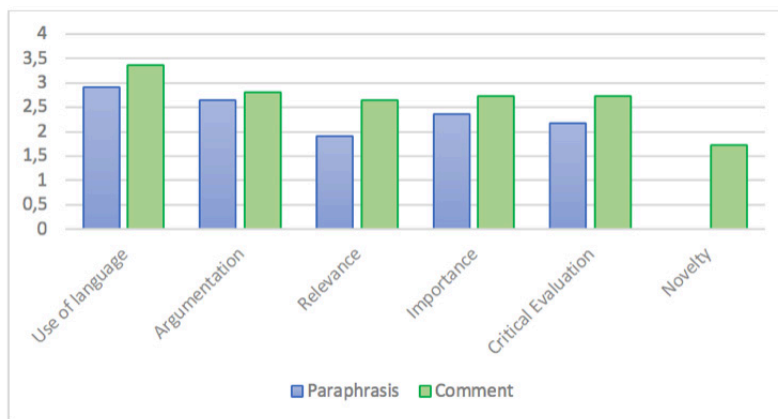

Fig. 2 - A comparison of Critical Thinking performance in paraphrase and commentary.

In order to see whether the prototype could assess CT in a reliable way, we compared the average scores obtained by human raters and prototype respectively in paraphrase and commentary. In figure 3, it is shown that in paraphrase the prototype provides higher score than human raters for the macro-indicators *Use of Language* and *Relevance*. On the other hand, the average score for the indicator *Importance* is slightly higher for human raters than in the prototype. In the commentary, there is a general trend of the prototype to provide lower scores comparing to the human raters. However, it is possible to see that the differences between the average scores for the *Use of Language* scores and *Novelty* in the commentary is quite low.
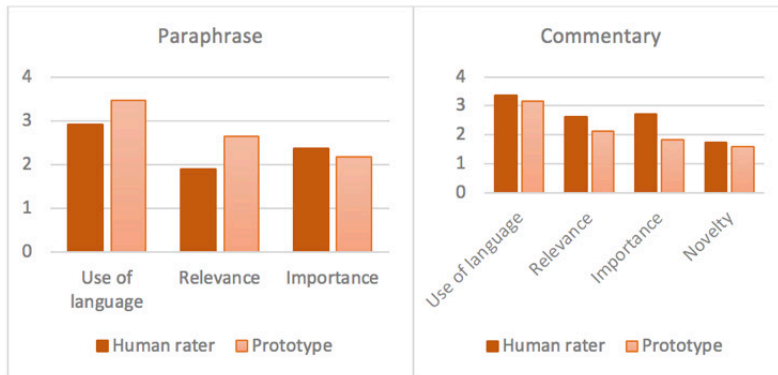
Fig. 3 - A comparison of CT scores calculated by a human rater and the prototype in paraphrase and commentary.

As shown in table 1, the agreement among human raters regarding the indicator *Use of Language* is satisfactory both in the paraphrase and in the commentary, with a higher performance in paraphrase (83% of agreement) comparing to the commentary (approximately 62% of agreement).

Table 1
THE AGREEMENT AMONG HUMAN RATERS REGARDING THE INDICATOR "USE OF LANGUAGE" IN THE PARAPHRASE AND IN THE COMMENTARY. *SIGN. $< 0,05$ **SIGN$< 0,001$

Macro-indicator	H-H Correlation	H-H Quadratic Weighted Kappa
Paraphrase_Use of Language	0,911*	0,83*
Commentary_Use of Language	0,745*	0,618*
Paraphrase_Relevance	0,75*	0,682*
Commentary_Relevance	0,881**	0,811*
Paraphrase_Importance	1,000**	1,000*
Commentary_Importance	0,642	0,571

However, there is no correlation among human raters and prototype. These could be explained by at least three factors: firstly, the texts of the answers are quite short (35 words per sentence) and we saw in previous experiences that the prototype achieved better performance with more elaborated texts (Poce *et al.*, 2019). Secondly, participants were not English native speakers and this might have had an impact on their use of language. Thirdly, human raters are Italian and this could affect their assessment of the use of English language by not native speakers. The agreement among human raters regarding the indicator *Relevance* is satisfactory both in the paraphrase and in the commentary (Table

1), with a higher performance in the commentary (81% of agreement) comparing to the paraphrase (68% of agreement). In the commentary, it is possible to see a tendency to a correlation among the prototype and human raters (r = 0,47) for the indicator *Relevance*, but this correlation is not statistically significant. All in all, we can say that the indicator *Relevance* is easier to detect in the commentary than in paraphrase both for human raters and prototype.

The agreement among human raters regarding the indicator *Importance* is 100% in the paraphrase, but the agreement is lower for the commentary (r = 0,64). There is a tendency of the prototype to correlate with human raters both in the paraphrase (r = 0,45) and commentary (r = 0,43) but correlation is not statistically significant in any case.

2.7 Discussion and conclusive remarks

The present contribute aims to present some preliminary results of validity and reliability regarding a prototype for CT assessment developed by the CDM research group. Data collected and presented in this paper are limited to a pilot activity with a small number of participants (18 in total), so any generalization is not possible. In the sample analysed, mainly composed by European university teachers involved in the field of CT, participants achieved generally good results on CT assessment based on their written answers to two kinds of exercise: a paraphrase and a commentary starting from an extract of the work of Galileo Galilei "*Dialogue Concerning the Two Chief World Systems*". Generally, participants achieved higher scores in comments than in the paraphrase exercise. This result could be explained by a low familiarity with the paraphrase exercise in the European sample or by the fact that writing a paraphrase before the commentary could facilitate the adoption of more sophisticated level of CT (Poce, 2017; Paul & Elder, 2006).

The rubric for CT assessment shows good properties, with satisfactory correlation and inter-rater agreement between human raters. However, the results of the prototype validation are not satisfactory yet and the the accuracy of automated scores still has room for improvement. Interviews were organized with human evaluators in order to understand the reasons for the low correlation values between prototype and human. For the macro-indicator *Use of Language* human evaluators did not give the same weight to spelling errors as the prototype, since human evaluators are not English native speakers. In addition, the human raters rewarded the use of a sophisticated language in terms of words and analyzed the diaphasic and diastratic variation present in open answers. Furthermore, human raters consider the coherence of verbal forms within the text whilst the prototype does not. In the future, we will try to reproduce the

human decision-making process following the instructions of a human-expert evaluator.

The best correlation among human raters and prototype were obtained for the macro-indicators *Relevance* and *Importance* with correlation higher than 0,43. However, correlation could be not considered statistically significant. As shown in other researches (Liu *et al.*, 2014), human raters tended to assign higher scores than our automatic assessment tool in the commentary. On the other hand, in the paraphrase the prototype assigned higher scores than human raters on the macro-indicators *Relevance* and *Importance*. This result could be explained because the prototype is designed to infer concepts from the questions and answers texts. In the paraphrase, the participants are required to report all the text's topics. In this condition, the prototype easily identifies all the concepts, without the need of further analysis. For these reasons, in paraphrase exercise the macro-indicators *Relevance* and *Importance* could obtain higher scores than the other macro-indicators and, more in general, than commentary or argumentation texts. This data leads us to think that it may be necessary to apply changes to the evaluation of the macro-indicators based on the type of stimulus given to the participants (paraphrase, argumentation, commentary, poetry).

Moreover, in recent years, many researchers rely on open data to give a semantic connotation to their analysis (Bovi, Telesca, & Navigli, 2015; Benedetti, Beneventano, & Bergamaschi, 2016). A study of the relationships existing between entities can help in identifying the concepts associated with *Relevance*, *Importance* and *Novelty* and increase the correlation levels associated with the indicators.

The attempt to automatize CT assessment through open-ended questions is at its beginning but it proves to be a useful support to human evaluation. The use of Natural Language Process techniques seems to be a possible direction according to the first results collected in the study herewith presented (McNamara *et al.*, 2017). The research group feels therefore encouraged to follow up the research described above, through further experimentation, working also on different macro-indicators from the Newman, Webb and Cochrane adapted model used so far. A reliable prototype for CT assessment could support researchers and teachers' understanding regarding learning processes related to CT and the environment in which it occurs (Siemens & Baker, 2012).

In future studies, we are going to expand the textual corpus because our prototype achieved slightly better performances with longer and more elaborated open-answers. We will conduct further validation studies with a larger sample and with different kinds of questions.

Acknowledgements

REFERENCES

Benedetti, F., Beneventano, D., & Bergamaschi, S. (2016, October). Context semantic analysis: a knowledge-based technique for computing inter-document similarity. In *International Conference on Similarity Search and Applications* (pp. 164-178). Springer, Cham.

Bloom H. (2000). How to Read and Why. New York: Touchstone

Bovi, C. D., Telesca, L., & Navigli, R. (2015). Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, *3*, 529-543.

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

Dominguez C. (Ed.) (2018). A European review on Critical Thinking educational practices in Higher Education Institutions. Vila Real: UTAD.

Elen, J., Jiang L., Huyghe S., Evers M., Verburgh A., & Palaigeorgiou, G. (2019). Promoting Critical Thinking in European Higher Education Institutions: towards an educational protocol. Vila Real: UTAD.

Ennis R.H. (1993). Critical Thinking Assessment. *Theory into Practice*, *33*(3), pp. 179-186.

Ennis R. H., & Weir E. (1985). The Ennis–Weir critical thinking essay test. Pacific Grove (CA): Midwest Publications.

Esplugas C., & Landwehr M. (1996). The Use of Critical Thinking Skills in Literary Analysis. *Foreign Language Annals*, *29*(3), pp. 449-461.

Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015, March). Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 146-150). ACM.

Facione P. A. (1990a). Executive summary of 'The Delphi Report'. Millbrae (CA): The California Academic Press.

Facione P. A. (1990b). The California Critical Thinking Skills Test. Millbrae (CA): California Academic Press.

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), pp. 345-350.

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

Gaeta, M., Marzano, A., Miranda, S., & Sandkuhl, K. (2017). The competence management to improve the learning engagement. *Journal of Ambient Intelligence and Humanized Computing*, *8*(3), 405-417.

Gordon, T. F., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence*, *171*(10-15), 875-896.

Halpern D. F. (2007). Halpern critical thinking assessment using everyday situations. Background and scoring standards. Claremont (CA): Claremont McKenna College.

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

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

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), pp. 1-23.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice, 33,* 19–28. doi:10.1111/emip.12028

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.

McNamara, D. S., Allen, L., Crossley, S., Dascalu, M., & Perret, C. A. (2017). Natural language processing and learning analytics. *Handbook of learning analytics*, 93-104.

Miranda, S., Marzano, A., & Lytras, M. D. (2017). A research initiative on the construction of innovative environments for teaching and learning. Montessori and Munari based psycho-pedagogical insights in computers and human behavior for the "new school". *Computers in Human Behavior*, *66*, 282-290.

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

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), pp. 56-77.

Paul R.W., & Elder L. (2006). Critical Thinking Reading & Writing Test. Tomales (CA): The Foundation for Critical Thinking.

Poce A. (2017). Verba Sequentur. Pensiero e scrittura per uno sviluppo critico delle competenze nella scuola secondaria. Milano: Franco Angeli.

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

Powers B. (1989). Illinois Critical Chinking Annual. Champaign (IL): University of Illinois College of Education.

Scott C. L. (2015). The futures of learning 3: What kind of pedagogies for the 21st century. *Education Research and Foresight*, *15*, pp. 1-21.

Serianni L., Della Valle V., & Patota G. (2003). L'italiano parlato e scritto. Agenda salvascrittura. Milano: Bruno Mondadori Editori.

Siemens, G., & d Baker, R. S. (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). ACM.

Watson G., & Glaser E. M. (1980). Watson–Glaser critical thinking appraisal. Cleveland (OH): 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.

World Economic Forum (2016). The 10 skills you need to thrive in the Fourth Industrial Revolution. Retrieved from: https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/