International Journal of Assessment Tools in Education

ISSN-e: 2148-7456 online Journal homepage: http://www.ijate.net/ http://dergipark.gov.tr/ijate

Multi-Trait Multi-Method Matrices for the Validation of Creativity and Critical Thinking Assessments for Secondary School Students in England and Greece

Ourania Maria Ventista

To cite this article: Ventista, O.M. (2018). Multi-Trait Multi-Method Matrices for the Validation of Creativity and Critical Thinking Assessments for Secondary School Students in England and Greece, *International Journal of Assessment Tools in Education*, 5(1), 15-32. DOI: 10.21449/ijate.335167

To link to this article: http://ijate.net/index.php/ijate/issue/archive
http://dergipark.gov.tr/ijate

Int. J. Asst. Tools in Educ., Vol. 5, Issue 1, (2018) pp. 15-32

http://www.ijate.net

e-ISSN: 2148-7456 © IJATE

INTERNATIONAL JOURNAL OF ASSESSMENT TOOLS IN EDUCATION

Research Article

Multi-Trait Multi-Method Matrices for the Validation of Creativity and Critical Thinking Assessments for Secondary School Students in England and Greece

Ourania Maria Ventista *

School of Education, Durham University, Leazes Road, Durham, DH1 1TA, United Kingdom

Abstract: The aim of this paper is the validation of measurement tools which assess critical thinking and creativity as general constructs instead of subject-specific skills. Specifically, this research examined whether there is convergent and discriminant (or divergent) validity between measurement tools of creativity and critical thinking. For this purpose, the multi-trait and multi-method matrix suggested by Campbell and Fiske (1959) was used. This matrix presented the correlation of scores that students obtain in different assessments in order to reveal whether the assessments measure the same or different constructs. Specifically, the two methods used were written and oral exams, and the two traits measured were critical thinking and creativity. For the validation of the assessments, 30 secondary-school students in Greece and 21 in England completed the assessments. The sample in both countries provided similar results. The critical thinking tools demonstrated convergent validity when compared with each other and discriminant validity with the creativity assessments. Furthermore, creativity assessments which measure the same aspect of creativity demonstrated convergent validity. To conclude, this research provided indicators that critical thinking and creativity as general constructs can be measured in a valid way. However, since the sample was small, further investigation of the validation of the assessment tools with a bigger sample is recommended.

ARTICLE HISTORY

Received: 03 April 2017

Revised: 12 August 2017

Accepted: 14 August 2017

KEYWORDS

validation, creativity, critical thinking, assessment, multi-trait multi-method matrix,

1. INTRODUCTION

1.1. Research Purpose

The knowledge demands in the 21st century are not easily predictable. Therefore, the education system of each country should provide the students with skills to adapt in the needs of this changing society. It has been supported that critical thinking and creativity could address these needs (Berliner, 2011). In other words, in the 21st century there is a huge amount of knowledge available to learners. When learners are required to find solutions to their questions, they do not

*Corresponding Author E-mail: o.m.ventista@durham.ac.uk

ISSN: 2148-7456 online /© 2018 **DOI: 10.21449/ijate.335167**

have to simply recall information. Instead, they should be able to identify relevant sources and evaluate them critically. Moreover, economies and societies nowadays change rapidly, so schooling cannot prepare learners to deal with specific problems. By the time learners will finish their schooling, there will be new problems to be solved so they should be able to critically approach these issues and generate solutions creatively.

Consequently, it is not a surprise that the development of critical thinking and creativity are prioritised by school curricula across the world (for example: Australian curriculum, UK curriculum). Similarly, universities expect their students to demonstrate critical and creative thinking and include these skills in their scoring rubrics. Therefore, critical thinking and creativity are judged to be crucially important within educational systems.

Despite their growing importance, the measurement tools of creativity and critical thinking as generic skills are not well established in primary and secondary education. As a result, when primary and secondary school students are assessed, traditional forms of assessment, which focus mainly on attainment, are used.

Hence, this paper investigates to what extent assessments which measure creativity and critical thinking as general constructs can be reliable and valid. To be more precise, concerning reliability, this paper focuses on the internal consistency of the measurement tools. For validity, this paper examines the discriminant (or divergent) and convergent validity. These are important elements to be investigated since there is no sufficient evidence for these psychometric properties. Although there is recent research which examines the relationship of students' performance between sub-sections of Torrance test (Yoon, 2017) or team creativity (Jiang & Zhang, 2014), there is a lack of studies which examine and establish the convergent validity among creativity tests (Plucker & Maker, 2010; Yoon, 2017).

Similarly, for critical thinking there are examples of studies attempting the validation of critical thinking as a subject-specific skill (Tiruneh et al., 2017). However, there is no evidence about the convergent validity between measurement tools of critical thinking.

Even when convergent validity of critical thinking measurement tools is examined, it is not established on comparison of performances in critical thinking assessments. For instance, recently a critical thinking tool for primary school students was developed. The researchers attempted to establish the criterion validity (which is a type of convergent validity) by comparing the performance of students with their grades of students in arts, instead of another critical thinking assessment (Gelerstein et al., 2016). This means that convergent validity was considered, but not in the most rigorous way.

Consequently, there is not sufficient evidence of the validation of creativity and critical thinking measurement tools. Hence, this research contributes to this area and discusses psychometric properties of assessments of creativity and critical thinking. For the purpose of this article, first, the constructs of critical thinking and creativity are defined and operationalised, then, the processes that the validation of measurement tools achieved are discussed. Next, the research methodology is presented, and, finally, the results of this research and its limitations are reported.

1.2. Defining the constructs

Creativity and critical thinking are the focal points of this research. Both terms can be perceived in different ways, but it is fundamental for both constructs to be defined before deciding on their assessments. Critical thinking 'is the intellectually disciplined process of active and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information

gathered from, or generated by, observation, experience, reflection, reasoning, or communication' (The Critical Thinking Community, 2013). According to Ennis (1993), critical thinking involves judging arguments and the credibility or sources, identifying conclusions and assumptions and drawing warranted conclusions. While Ennis (1993) defines "critical thinking as a reasonable reflective thinking that is focused on deciding what to believe and do", Lipman (1987) explains that the use of the word 'reasonable' can lead to circularity and criticised this definition as restrictive. According to Lipman (1987), critical thinking is employed for numerous other aims and does not always lead to a clear-cut conclusion. Lipman (2003) postulates that critical thinking is based on criteria, is self-corrective and sensitive to context. A further definition of critical thinking supports that it involves six basic cognitive aspects: interpretation, analysis, evaluation, inference, explanation and self-regulation (Facione, 1990, 2015). For this research, the working definition of critical thinking consists of observation, analysis, synthesis, evaluation and interpretation of arguments within specific contexts.

Creativity is perceived as a broad term which includes other sub-characteristics such as divergent thinking, convergent thinking, openness to explore new ideas and listening to "inner voice" (Treffinger, Young, Selby, & Shepardson, 2002). According to this paradigm, creativity includes critical thinking. Guilford (1967) supports that problem-solving is the same phenomenon as creative thinking. In order for something to be perceived as creative, it should have two main characteristics: to be original and useful (Rungo & Jaeger, 2012). According to the definition of the National Advisory Committee on Creative and Cultural Education (1999), however, creativity has four - instead of just two - typical characteristics: imagination, purposefulness, originality and a new product with merit. Similarly, Mednick (1962) defines creative thinking as the procedure through which associated components are combined in a new way and this combination is a useful one. In recent years many researchers have accepted the standard definitions of creativity (Weisberg, 2015). By examining studies regarding the definitions of creativity (Kampylis & Valtanen, 2010), it can be concluded that most of the recent definitions involve trivial additions or syntheses of previous ones. Weisberg (2015), however, questions the inclusion of "value" in the definition of creativity, since its evaluation appears to be too subjective and unreliable. As a result, for the purposes of this research creativity is operationalised as a combination of fluency, innovation, novelty and imagination.

1.3. Validation

Having discussed the working definitions of the two main constructs, issues regarding validation of assessment tools are discussed. This paper investigates to what extent critical thinking and creativity assessments can be considered valid. The first issue to be discussed is whether the validity is a psychometric property of a test or a characteristic of the interpretation of the test. On the one hand, it has been supported that a test is valid when it measures what is supposed to measure, so the validity is a psychometric property of the test. On the other hand, it has been supported that the interpretation is the one which can be valid or invalid and a test cannot be itself valid or invalid. This means that a test can be valid for one interpretation, but invalid for another one (Coe, 2012; Newton, 2012).

The second issue concerns the ways that validation can be achieved. Five sources of evidence can support the validation process; test content, response processes, internal structure, relations to other variables and consequences of testing (Sireci, 2009, p. 30). Specifically, about the test content, Kane (2009) states that if the task of a test is close to the performance of interest then there is no need for strong evidence for the content of the test for it to be valid.

With reference to the internal structure as a process of validation, the factors included in a test are considered. This research used Cronbach's Alpha as an indicator of internal structure. Although the relations to other variables is usually called criterion validity, in critical thinking and creativity assessments, there is not a widely accepted gold standard to be considered as criterion. Instead, this research used what Campbell and Fiske discuss (1959) as a validation method: convergent and discriminant validity. Messick (1995) also mentions this method as one aspect of validity, which is related to the external evidence for the quality of an assessment. Convergent validity exists when results from measures that measure the same construct are correlated, while discriminant validaty when the scores of tests which measure different constructs do not correlate. Particularly, convergent validity was sought between the measurement tools which measured the same construct (either creativity or critical thinking) and divergent validity between the measurement tools which measured different constructs (critical thinking and creativity). This implies that this research accepts that critical thinking and creativity are not the same constructs, even though some researchers might have expressed the opinion that they are both part of productive thinking (Facione, 2015; Newton, 2014).

2. METHODOLOGY

2.1. Method

For the selected validation process, collection of data was required. In this case, data was the scores in the assessments. This paper presents the results of research conducted in Greece and its replication in England. As previously mentioned, the validation of the measurement tools attempted to be done with using the multi-trait multi-method matrices (Campbell & Fiske, 1959). This analysis requires the use of at least two traits and two methods. The two traits were creativity and critical thinking and the two methods were written and oral assessments.

As multi-trait multi-method matrices were used, emphasis was put on convergent and discriminant validity. So the hypothesis was that if tests of critical thinking indeed measured critical thinking then the scores that students achieved in both critical thinking tests would be correlated with each other (convergent validity). On the other hand, their critical thinking scores would be less or not correlated with measurements of creativity (discriminant validity), since the assessments measured different constructs. With the exact same logic, there was a similar hypothesis for the creativity measurement tools. If the creativity scores were valid and measured what they supposed to measure, then the scores that the students would achieve in creativity assessments would correlate with each other (convergent validity) and would not correlate with their performance in critical thinking (discriminant validity).

Lastly, because the methodology required correlating scores of the tests, it has to be clarified that there is no lower limit for the sample size when conducting a correlation study. The sample size, however, affects the confidence intervals for the correlation. With small sample sizes, even a slight increase in the number of participants significantly reduces the length of confidence intervals. However, it has been supported that when increasing the number of participants to more than 24 participants, there is a loss of sample size impact on the length of the confidence intervals (Johanson & Brooks, 2010, p. 397). Finally, it has to be mentioned that the recommended number of participants for pilot studies is usually around 30 (Johanson & Brooks, 2010).

2.2. Replication

Seven months later the research was replicated in a secondary school in the North East of England. The purpose of this replication was not the direct comparison of the two countries but to increase the sample size. In Greece, there were only 30 students, so it was judged appropriate to collect some additional data. However, it was interesting to investigate whether the previous results would be also found in a new situation. Moreover, replication was conducted specifically in England in order to exclude the possibility of effects of translation issues, which might have affected the Greek sample.

The results of each study are presented separately because there was one small change in the methodology and because the data collection took place at different times. As I am not a native English speaker, my accent could contribute to a construct irrelevance in the oral assessment of critical thinking. For this reason, students were given three different options than the Greek students. The Greek students had a text read to them, while the English students could choose between the researcher reading the text or them reading it aloud or silently. There is the assumption that they chose wisely in order to maximize their performance in the test and indirectly minimize the potential construct irrelevance.

Even though it would have been preferable to keep the conditions exactly the same as in Greece, it was not possible. Instead of giving them this choice, the alternative of having a recording of the letter read by a native speaker was considered. However, this was too impersonal and could have not taken into consideration the conditions in the room. Hence, it was judged as a bigger change in the methodology compared to allowing the student to choose their preferred method of accessing the text.

2.3. Participants

The initial research took place in a secondary school in Greece with 30 participants aged 13-15 years old. Students of these ages were targeted because there are more available assessment tools for these ages compared to primary school students. The specific school was selected based on the willingness of the headteacher to provide time and space for the research needs. The school was in a suburban area of northern Greece. The students were randomly chosen by the class lists. No student refused to participate and there was no attrition.

In the replication study, the sample was 21 twelve-year old boys who were students in a secondary school. It was not possible to gain access to older students as in the Greek sample. However, the tests were age-appropriate. In this sample 4 participants refused to narrate a fairy tale and this research believes that they felt uncomfortable to do so. British Education Research Association (BERA) guidelines stipulate that participants can withdraw at any point. During the research and during the replication of the research two of the students withdrew (BERA, 2011) .

2.4. Ethics

Before conducting both studies, ethical approval was obtained by the School of Education Ethics Committee at Durham University. Both of the studies followed the BERA guidelines (2011).

3. ASSESSMENT TOOLS

3.1. Critical thinking

The tools used for the critical thinking in the written method were a combination of the deduction items of the Cornell Reasoning Test (Ennis et al., 1964) and items based on the test of appraising observation (Norris & King, 1984). The reasoning test provides "if" statements to the students who should judge whether the last sentence would be a warranted conclusion by deductive reasoning. A choice of "maybe" is also given to the students in this test, as in some cases the data are insufficient for them to decide. The test of appraising observation narrates two stories to the students. Each item of the test provides two statements to the students. The students should judge which of the two statements is more believable. In order to judge effectively, the students should also consider the context of the two stories as a factor.

The time given for these tests was one hour and due to this time limitation only a few items were used. Both tests are quite extensive and, thus, since the aim was not to examine the reliability and validity of the specific existing tools, but to examine whether it was possible to measure critical thinking as a general construct, only a few questions of each test were used. In order to improve the internal consistency of the initial tests, similar questions appear multiple times. In this research, fewer questions were chosen. The questions were judged appropriate and sufficient to operationalise the construct of critical thinking as defined by this research.

Additionally, both of the tests are age appropriate. The Cornell Test Level X (Ennis, Gardiner, Guzzetta, Morrow, Paulus & Ringel, 1964) was deemed appropriate for secondary school students and used in previous studies for evaluating critical thinking in students of this age or even a little older (Iozzi & Cheu, 1978). The last version of appraising observation test is also suitable to assess secondary school students (Norris & King, 1984).

The critical thinking tool used for the oral assessment of critical thinking was based on an established tool (Ennis & Weir, 1985) suitable to test sixth grade to university students. During this assessment, the students were requested to judge presented arguments. The researcher first articulated the main purpose of the letter - the author tried to persuade the listener of the benefits of the prohibition of overnight parking- and then read the letter. The researcher elucidated that students should take a position and either be persuaded or not by the argument in each paragraph to justify their position and share any thought related to the paragraph. The reason why the letter was read by the researcher to the Greek students was to exclude construct irrelevance. It has been supported that the reading ability in tests can play an important role (Hewitt & Homan, 2003). Reading ability is irrelevant to critical thinking and should not be embodied in critical thinking assessments. The oral assessment did not disadvantage students who have reading difficulty. They could also ask for clarification for words that they didn't understand. They had sight of a printed version so as not to disadvantage students who were not used to listening to texts.

3.2. Creativity

For the written assessment of creativity a combination of tests was used (Getzels & Jackson, 1962). Firstly, students had to think as many possible uses for common objects, such as a brick. Secondly, students were given partially complete images and instructed to complete them by drawing around them to illustrate what they imagined the images were. An activity similar to the latter can also be found in the Torrance Test of Creative Thinking (Torrance, Ball & Safter, 2008). The number of responses given by the students and the degree of originality of their responses were assessed.

For the oral assessments of creativity the students were asked to narrate a fairy tale. For the fairy tale a scoring rubric was created. The rubric evaluated the content of the students' stories by combining indicators of imagination. These indicators were the number of mentioned typical elements found in fairy tales, referred to as *functions* (Propp, 1968), the presence of creative characteristics that can be in fairy tales (Rodari, 1996) and the presence of humour and violence in the story. The latter two characteristics are usually connected with creativity (Getzels & Jackson, 1962; Nusbaum, Silvia & Beaty, 2017).

The oral assessment resembled a real-life task with a specific purpose as the communicative language approach would suggest (Richards, 2005). Participants were presented with a real life situation: "*A younger cousin or a sibling of yours has just asked you to narrate a fairy tale. I will give you three minutes to think about the fairy tale you are going to narrate and about this time again to narrate it*". The choice of the activity was grounded in results of prior research investigating gender and ethnicity differences in creativity. Even though males had the self-perception of being more creative on science-analytic and sports tasks and females more on social-communications and visual-artistic tasks, both genders were equally assumed to be creative in verbal-artistic activities (Kaufman, 2006). For this reason a type of verbal activity was set. Nonetheless, it is accepted that for the previous finding, since it is based on self-reported questionnaires there may be a gap between perceived creative strengths and actions, and also that the respondents' opinions and beliefs may not be stable (Foddy, 1993).

3.3. Norm-referenced tests

The two written tests of creativity were norm-referenced measurements because there was a comparison between the performances of the students (Cox & Vargas, 1966). The score of unique answers attributed to the students related to the other participants' responses. Thus, an answer was characterised unique only if no other participant had mentioned this particular answer. Silvia (2015) highlights the significance of this flaw in the creativity tests; the uniqueness grade is sample-dependent. In other words, as the sample increases, the likelihood of a unique answer decreases.

To ameliorate this, the researchers could pre-decide the size of group. For example, the sample for this test could always be 30 students and each reply could be judged unique when it has not been mentioned by the particular number of students. It is accepted that this could not provide a solution for the problem of a student having high performance in a less creative group and be judged to have average performance when compared to a more creative group. Nevertheless, sample-dependence cannot be completely avoided in the norm-referenced tests.

3.4. Matching the assessments to the construct definitions

It is important to discuss the tools used for this research in relation to the aspects of the constructs measured. The appraising observations test assessed the ability of the students to evaluate which statement is more believable. Analyzing and synthesizing can also be assessed by the test (Treffinger et al., 2002). The reasoning test evaluated deductive reasoning. The Ennis & Weir letter (1985) required evaluation of specific arguments. Therefore, these assessments fit the aforementioned definition of critical thinking.

The 'test of different uses for tools' and the 'pattern meanings test' (Getzels & Jackson, 1962) did not have a single correct answer. The only variables measured in this test were originality (how many answers are unique between the answers of all the participants) and fluency (the number of answers mentioned) firstly at the suggestion of the test author (Getzels & Jackson, 1962)

and secondly because these variables can be measured objectively. Concerning the narration of the fairy tale, it mainly attempted to evaluate imagination and innovation, which are characteristics of the creativity (El-murad & West, 2004). Sense of humor as a characteristic of openness was assessed by the oral assessment of creativity. Consequently, creativity assessment also fit the working definition of creativity adopted by this research.

3.5. Translation and adjustment of the Tools in Greek

Measurement instruments were cautiously translated in the Greek language using the back-translation method (Su & Parham, 2002). Furthermore, for the oral assessment of creativity, the content was also slightly adjusted. The town took the name of the town in which the test was administrated, road names were taken from roads in the town and also the name of the authorities 'Director of the National Traffic Safety Council' and the 'National Association of Police Chiefs' were replaced with the respective Greek terms. This aimed to provide the students with a purpose and a motivation to read the test (Richards, 2005).

4. RESULTS AND DISCUSSION (Study in Greece)

The tools are going to be discussed according to their reliability and validity. There are different types of reliability and validity. For the purpose of this research, the reliability is discussed as internal consistency and validity as convergent and discriminant validity.

Table 1. Multi-trait multi-method matrix (Greece)

		WRITTEN TESTS Method 1			ORAL ASSESSMENT Method 2	
		Critical thinking	Creativity: DUO	Creativity: PM	Critical thinking	Creativity
Written tests Method 1	Critical thinking: only reasoning	0.758				
	Creativity: Different Uses of Objects	-0.021	0.817			
	Creativity: Pattern Meanings	-0.376 *	0.719**	0.925		
Oral Assessment Method 2	Critical thinking	0.199	0.139	0.216	0.483	
	Creativity	-0.299	-0.010	0.169	0.257	0.743

* $p < 0.5$ (statistical significance)
** $p < 0.1$ (statistical significance)
Light blue: the cells which show just the internal consistency of the measurement tool
Light green: the cells which show correlation between monomethod and the same trait.
Light pink: the cells which show correlations between heterotrait and monomethod cells (creativity or critical thinking compared with each other and assessed by the same method).
Purple: the cells which show correlations between heterotrait - heteromethod cells.
Orange: the cells which show correlations between monotrait - heteromethod cells.

4.1. Internal Consistency of the Measurement Tools

To consider the reliability of the measurement tools, internal consistency was examined and Cronbach's Alpha was used as an indicator of internal consistency. Cronbach's Alpha should not be used as proof of all types of reliability. It is only related to the correlation of the items and it is the 'mean of all split-half reliabilities for a given test application' (Johnson & Johnson, 2009, p. 14). The internal consistency of the items based on the appraising observation tests was low and it could not be improved even by deleting some items. Thus, these items were excluded by the matrix.

Some of the reasoning items were found to have negative correlation so they were deleted. An item that has negative correlation tends to be answered incorrectly by otherwise high scoring students. One of those items had negative stem. Negative statements in the stem should be avoided (Haladyna, 1994) because it may cause confusion. Two items at the end of the test also had negative correlation, but these items did not seem to differ from the other items. The fact that they were towards the end of the test may be the cause of those items having negative correlation. The students may have been tired or bored by the end of the test.

The results for the reasoning items in the written assessment of creativity had indicated strong internal consistency (a = 0.76). The creativity assessments for the written method also had high reliability (a = 0.81 and a = 0.92), which is comparable with alpha scores required for high-stakes assessment. The oral assessment of creativity had also high internal consistency (a = 0.74). Consequently, even though critical thinking and creativity are multi-facet constructs, when the tests are focused on particular aspects, such as only reasoning or imagination, then high internal consistency can be expected.

The oral assessment of critical thinking was found to have moderate internal consistency (a= 0.48) which could have been a consequence of the test having a few items. With more items, the reliability of the test may have been higher, however, the increase of the number of the items cannot be assumed to substantially increase of the quality of the test even if this is a way to increase internal consistency. For example, by asking similar questions the length of the assessment and Cronabach's alpha increases. However, the quality of assessment remains the same. The low alpha might be explained by the fact that the test was not a multiple-choice test. Multiple choice items are usually preferred in tests because they increase reliability, but this does not mean that they secure the validity of the tests (Burton, Sudweeks, Merrill & Wood, 1991; Lambert & Lines, 2000). Thus, even though the oral assessment had lower internal consistency than the other assessments, it might have been a more valid method of testing critical thinking. Even though there are researchers who support that there cannot be valid inferences without reliability (Koretz, 2006), there are others who advocate that if reliability is perceived merely as consistency among measures then validity may be without reliability (Moss, 1994). Moss (1994) supports that less standardised forms of assessment may be valid without being reliable and 'as assessment becomes less standardised, distinctions between reliability and validity blur' (p.7).

4.2. Convergent and Discriminant Validity

The multi-trait and multi-method matrix presents the convergent and discriminant validity between the measurement tools (Table 1). The written test of critical thinking was validated based on convergent and discriminant validity. Specifically, it was correlated with the oral assessment measuring critical thinking (convergent validity), but not correlated with the creativity assessments (discriminant validity).

The written test of critical thinking had discriminant validity with the three creativity tests (r = - 0.02, r = -0.38 and r = - 0.3). This means that there was not a linear relationship which links the performance in the reasoning items with the performance in the creativity tests of fluency, innovation and imagination. As a result, the reasoning test measured something different from the creativity tests.

The performance of students in the reasoning items had a very weak linear relationship with their performance in the oral assessment of creativity (r = 0.2). This means that the two assessment had, to some extent, convergent validity, but without strong evidence. The low correlation between the scores in the two assessments of critical thinking can be explained because the two tools evaluated different aspects of critical thinking. The written test was focused on deductive reasoning, while the oral assessment on the argument evaluation within a specific context.

The scores of the oral assessment of critical thinking was correlated equally with those of the oral assessment of creativity (r = 0.14 and 0.22) and the written test of evaluating critical thinking (r = 0.2). Similarly, the scores of the oral assessment of creativity was more correlated with the scores of the oral assessment of critical thinking (r = 0.26) rather than those of the creativity assessments (r = - 0.1 and r = 0.17). Thus, the performance of the students in the oral assessments correlated more with each other than with their performance in tests which evaluate the same constructs with different methods. This is not a surprising finding. Paradoxically it is common to identify higher correlation between the scores of heterotrait and homomethod assessments, rather than the homotrait and heteromethod (Coe, 2012).

Furthermore, in this case, slight correlation between the scores that students achieved in critical thinking and creativity assessments is expected, because creativity and critical thinking - as they have already been defined - can be related to each other and be perceived as sub-categories of productive thinking (Newton, 2014).

The scores of the two written assessments of creativity were highly correlated with each other with a strong linear relationship (r = 0.72). In other words, the students who scored highly in the one test also scored highly in the other test, and the students who scored low in one, they also scored low in the other test. This suggests that both tests measured the same thing and that evidence of convergent validity was strong.

This last finding can be considered a positive indicator for future assessment of creativity. For these two tests, it is possible that there is concurrent validity, as they both also have independently high reliability (Lambert & Lines, 2000). Both tests evaluated mainly the same elements of the creativity construct, fluency and innovation by using the same method. The high correlation between their scores demonstrates that as long as the same side of a multifaceted construct is evaluated with the same method using two different assessments, convergent validity between these assessments can be expected.

What requires explanation is the fact that the scores of the two written assessments of creativity were poorly correlated both with those of the critical thinking oral assessment (r = 0.14 and r = 0.22) and with the creativity oral assessment (r = - 0.01 and r = 0.17). More specifically, the low correlation between the written assessments of creativity and the oral assessment of critical thinking can be explained if the two constructs are considered elements of the general construct productive thinking.

The low correlation between the written assessments and the oral assessment of creativity (r = - 0.1 and r = 0.17) can be used as a lucid demonstration that creativity is a multi-faceted concept

and the assessments evaluate different aspects of the same construct. The written test about the use of objects measured fluency and innovation, while the oral assessment measured verbal imagination. Thus, students might have been creative in some aspects, but not in others. In other words, different measurements tools of creativity using different methods were not found to be highly correlated. This finding is line with studies in creativity literature which suggested that people might perform differently in different tasks which require creativity (Hocevar, 1979).

To summarise, convergent and divergent validity were found for the written critical thinking assessment. Similarly, the creativity assessments had high convergent validity only when the same method and the same facets of the construct were assessed. The research in Greece revealed some positive indicators for the evaluation of critical thinking and creativity as general constructs.

5. RESULTS AND DISCUSSION (Replication Study in England)

A few months later the study was replicated in England. The results observed were similar to those derived from the Greek sample.

5.1. Internal Consistency of the Measurement Tools

When the research was replicated, the internal consistency of the measurement tools was also found to be relatively high. The reasoning items in the written assessment were found with similar internal consistency values as in Greece (a= 0.74). All the assessments of creativity had high alpha scores (a = 0.8), similar to the Greek sample data. These values of internal consistency are sufficient to enable the assessments to be used as high-stakes. The high internal consistency values could be explained by the fact that all the three creativity assessments measure a narrow and specific aspect of creativity.

Concerning its internal consistency, the data relating to the questions based on the appraising observation test indicated a low alpha score when implemented in Greece, but with the English sample it was slightly higher (a=0.52). For a multiple-choice test to have such a low alpha score is concerning as it contradicts with the usual expectation of multiple-choice items to be more reliable assessments (Burton et al., 1991).

Finally, the oral assessment of critical thinking had a higher internal consistency (a = 0.57) than the Greek sample. The test was not a multiple-choice test and this might affect its internal consistency.

5.2. Convergent and Discriminant Validity

When replicating the research in England (Table 2) the evidence was similar to the results from the Greek data (Table 1), as the multi-trait multi-method matrices suggested. The written assessment of critical thinking was also validated with convergent and discriminant validity, as with the Greek sample. The evidence for convergent validity in the English sample was stronger than the Greek one, since a moderate linear relationship between the written assessment and oral assessment of critical thinking was found (r=0.44). This relationship suggested that the students who scored highly in one test usually tended to score highly in the other test as well. The relationship between the two tests was much stronger compared to what was found in the Greek sample (r = 0.2). A possible explanation might be an issue of translation or cultural differences in the critical thinking tests in the Greek sample.

Table 2. Multi-trait multi-method matrix (England)

		WRITTEN TESTS Method 1			ORAL ASSESSMENT Method 2	
		Critical thinking	Creativity: DUO	Creativity: PM	Critical thinking	Creativity
Written tests Method 1	Critical thinking: reasoning items	0.741				
	Creativity: Different Uses of Objects	0.251	0.813			
	Creativity: Pattern Meanings	0.208	0.477*	0.879		
Oral Assessment Method 2	Critical thinking	0.437	-0.357	-0.383	0.566	
	Creativity	-0.040	0.159	0.228	-0.332	0.845

* p < 0.5 (statistical significance)
** p < 0.1 (statistical significance)
Light blue: the cells which show just the internal consistency of the measurement tool
Light green: the cells which show correlation between monomethod and the same trait.
Light pink: the cells which show correlations between heterotrait and monomethod cells (creativity or critical thinking compared with each other and assessed by the same method).
Purple: the cells which show correlations between heterotrait - heteromethod cells.
Orange: the cells which show correlations between monotrait - heteromethod cells.

For the written test of critical thinking there was a very weak relationship with the written tests of creativity (r = 0.25 and r = 0.2), but no relationship with the oral assessment of creativity (r = - 0.04). The first two assessments might be slightly correlated because they use the same method (written) as the reasoning items and it has been found that there is correlation between assessments which use the same method independently of the construct (Coe, 2012). However, the lack of relationship between the reasoning items and the oral assessment of creativity established the discriminant validity between the assessments.

Moreover, discriminant validity between the oral assessment of critical thinking and creativity measurement tools was reported (r = -0.36, r = - 0. 38 and r = - 0.33). Therefore, the data from the English sample validated the critical thinking tools with both convergent and discriminant validity.

The scores of the two written creativity tests were found with a sufficient linear relationship to establish convergent validity both in Greece (r= 0.72) and in England (r = 0.48). Thus, as the same side of a multifaceted construct is evaluated and the same method is used, correlation between the tests can be expected.

The results of the two written assessments of creativity were found almost equally correlated with the written assessment of critical thinking (r = 0.25 and r = 0.25) and the oral assessment of creativity (r = 0.16 and r = 0.23). However, as mentioned previously, there are examples of studies

which demonstrate that the method by which students are assessed sometimes plays a more crucial role than the construct on which they are assessed (Coe, 2012).

With reference to the oral assessment of creativity, there was validation of the assessment. Convergent validity was found between the oral assessment of creativity and the two tests of creativity (r = 0.16 and r =0.23). The convergent validity, however, was not supported by high correlation between the creativity assessments. This is expected, because the oral assessment of creativity did not examine the same aspects of creativity concept as the written assessment of creativity. This finding confirmed that creativity characteristics vary within a person and no person can have all the creative characteristics (Treffinger et al., 2002). In multi-faceted constructs like creativity, convergent validity can be sought between assessments which evaluate the same aspects of the construct.

Furthermore, discriminant validity was found since the oral assessment of creativity was not correlated with the two critical thinking assessments (r = - 0.04 and r = - 0.33). The lack of correlation between the performances of the students in the oral assessment of creativity and the critical thinking tests suggested that they measure different concepts. Therefore, there was discriminant validity which also supported the validation of the measurement tools of creativity and critical thinking.

To conclude, the assessments in the multi-trait and multi-method matrix in England were found to be valid concerning their convergent validity and discriminant validity. Consequently, the replication of the study confirmed the findings of the initial study in Greece and supported with even stronger evidence that critical thinking and creativity can be evaluated as general constructs in a valid way.

5.3. Is critical thinking and creativity culture and knowledge dependent?

As it has been previously said, the purpose of collecting data from two different countries was not their comparison. Besides, the sample was too small to enable such a comparison. However, by replicating this study in two different schools in two different countries and by perceiving critical thinking and creativity as general constructs and not subject-specific, it is reasonable to question to what extent the performance of the students was culture and knowledge dependent. For a deeper understanding of potential differences, there was an examination of the recorded material of the oral assessments. This material gave access to the students' thinking process. In the narration of the fairy tale no significant cultural differences were identified. The themes that emerged in the students' stories were similar. Moreover, this task did not demand any knowledge and thus knowledge did not appear to affect the performance of the students.

This was not the case with the relationship between knowledge and the evaluation of arguments in critical thinking assessment. Some students were not critical because of the lack of specific knowledge. Particularly, students were persuaded by an argument presenting results of a one-day experiment. Being students in a secondary school and without research knowledge they could not realise that results of one day experiment could not support generalisation. Therefore, sometimes prior knowledge is required to be critical. This is in agreement with the ideas of some of academics. For example, McPeck (1981, 1990) supports that critical thinking is subject-specific and in order for somebody to be critical they should have knowledge of the topic. This stance opposes Ennis' whose definition and assessments have been broadly accepted by this research. However, it should be recognised that it is valid to evaluate critical thinking as a high-order thinking skill of a subject as the Bloom's taxonomy would espouse (Krathwohl, 2002), when there

are also knowledge requirements in the assessment. Nevertheless, as the findings of this research suggested, critical thinking tests which do not require prior knowledge can be constructed.

No cultural differences were identified when the critical thinking performance of students in England and Greece were compared. However, when one of the arguments in the oral assessment of critical thinking discussed driving to work during rush hour, three students in Greece suggested arriving to work slightly late in order to avoid rush hour traffic. This was not suggested by English students. The sample was too small to lead to generalisation, but this might suggest some cultural differences. Hence, critical thinking assessments could be biased because of cultural differences.

Finally, the arguments used in the oral assessment of critical thinking were adjusted in the Greek language and context by also using a town familiar to the students. This adjustment aimed to make the context more realistic and motivate some students. However, it confused other students who became fixed on the real traffic problems of that specific town. Therefore, if the topic in the critical thinking test is relevant to the daily life of the students, this may affect their judgment. The students might adhere to the specific stimulus provided, which could restrict their judgment. This is in line with what Lipman (2003) supported; critical thinking is -and should be - related to the context.

6. LIMITATIONS

The two matrices in this research can only provide positive indicators for the validation of the tools, because the research design had several limitations. Specifically, the sampling method and the small number of participants do not allow generalisation of the conclusions about the effectiveness of the assessment tools. However, the assessments were conducted by only one researcher and it was infeasible to conduct more oral assessments (each of them lasted approximately 30 minutes). It is suggested that future studies use a bigger sample.

Additionally, the tests had no consequences for the students, and their motive to complete them was not examined. They may have merely guessed several of the questions as there were no aftereffects. What is more, narrating a fairy tale may inadequately motivate teenagers, especially boys. Some teenagers may feel in an inconvenient position when someone asks them to narrate a fairy tale. Moreover, with solely one rater, interrater reliability could not be examined. In the oral assessment halo effects may have been present to some extent which may have influenced marking (Nisbett & Wilson, 1977). Finally, the tests were translated for implementation in Greece. Even though back-translation took place, translation may still affect the results (Su & Parham, 2002).

For future researchers the replication of the research with a bigger sample is recommended. In both matrices, the creativity tool ‘narrating a fairy tale’ used in the oral assessment found highly reliable but not particularly correlated with any other test. This might be either because it evaluates different aspects of creativity or because the gender or the age of the students influenced their motivation and involvement in this task. In future research, it would be useful to pilot this tool with students in primary school and attempt to examine the convergent validity with other established creativity tests which evaluate the same aspect of creativity. Moreover, it is crucial for the convergent validity of this test with linguistic ability tests to be examined. It might be the case that this tool has high construct irrelevance by including general language ability since participants have to express their thoughts and tell a story by not only demonstrating an isolated creativity skill.

7. CONCLUSIONS

Critical thinking and creativity as general constructs can be measured. Most of the assessments had moderate or high internal consistency. Furthermore, internal consistency was found to be independent of the format of the tests, as one of the multiple-choice assessments was found to be the least reliable.

By using convergent and discriminant validity for the tools' validation, there was some evidence that critical thinking and creativity tools which evaluate these constructs as general can be valid. Discriminant validity between critical thinking and creativity tools was identified in almost all of the instances in both countries' data matrices.

The value of convergent validity between the assessments which measure the same constructs in some of the cases has been low. However, this finding is justifiable because in some cases even though both tests measured the same construct, they measured different aspects of the same construct. Hence, if creativity and critical thinking are to be evaluated, the convergent validity of the tests should be sought between tests which assess common sides of the construct. The validation of the tools could not be achieved when the assessment tools measured different sides of the same construct.

In a few cases, assessments using the same method were found highly correlated to each other even though they measured different constructs. This suggests that the assessment method can play a crucial role in the students' performance in the thinking skills assessments.

As a final remark, since critical thinking and creativity are multi-faceted constructs, multi-assessment is recommended, because students might perform well in an assessment which measures one of the facets, but not in another which measures one of the other facets.

Disclosure statement

No potential conflict of interest was reported by the author.

Acknowledgements

I would like to thank Prof Rob Coe for his support in all the aspects of this research (both methodological and administrative). I would also like to thank Prof Stephen Gorard for his insightful comments and the four reviewers of the IJATE for their detailed and constructive feedback on my manuscript. Finally, I would like to thank Miss Sinead Flinders for her invaluable help, which significantly improved my manuscript.

8. REFERENCES

Australian Curriculum (n.d.) *Critical and Creative Thinking*. Available at: https://www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/critical-and-creative-thinking/ (access: 6 August 2017)

BERA (2011). *Ethical guidelines for educational research*. British Educational Research Association. Available at BERA website: https://www.bera.ac.uk/researchers-resources/publications/ethical-guidelines-for-educational-research-2011 (access: 5 August 2017)

Berliner, D. C. (2011). 'The Context for Interpreting PISA Results in the USA: Negativism, Chauvinism, Misunderstanding, and the Potential to Distort the Educational Systems of Nations'. In Pereyra, M.A., Kotthoff, H. & Cowen, R. (ed.) PISA Under Examination:

Changing Knowledge, Changing Tests, and Changing Schools. (pp. 77-96). Rotterdam: Sense Publishers

Burton, S.J., Sudweeks, R.R., Merrill, P.G. & Wood, B. (1991). *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty*. Brigham Young University Testing Services and The Department of Instructional Science.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and Discriminant Validation by the Multitrait-multimethod matrix. *Psychological Bulletin*, 56 (2), 81-105

Coe, R. (2012). 'Conducting Your Research: Inference and Interpretation'. In Arthur, J., Waring, M., Coe, R. & Hedges. L.V. (ed.) Education Research: Methods and Methodologies. (pp. 41-52). London: Sage

Cox, R.C. & Vargas, J.S. (1966). A comparison of Item Selection Techniques for Norm-Referenced and Criterion-Referenced Tests. University of Pittsburgh.

Critical Thinking Society (2013). *Defining Critical Thinking.* Available at: http://www.criticalthinking.org/pages/defining-critical-thinking/766 (Accessed: 28 January 2015).

Department for Education (n.d.) *National Curriculum in England: Framework for key stages 1 to 4.* Available at Gov.UK website: https://www.gov.uk/government/publications/national-curriculum-in-england-framework-for-key-stages-1-to-4/the-national-curriculum-in-england-framework-for-key-stages-1-to-4#the-school-curriculum-in-england (access: 6 August 2017)

El-Murad, J. & West, D. C. (2004). The Definition and Measurement of Creativity: What do we know?. *Journal of Advertising Research*, 44(2), 188-201. doi: 10.1017/S0021849904040097

Ennis, R.H. (1993). Critical thinking assessment, *Theory Into Practice*, 32(3), 179-186. doi: 10.1080/00405849309543594

Ennis, R.H., Gardiner, W., Guzzetta, J., Morrow, R., Paulus, D. & Ringel, L. (1964). *Cornell Critical Thinking Test Series. The Cornell Critical Reasoning Test. Form X*. University of Illinois.

Ennis, R.H. & Weir, E. (1985). *The Ennis-Weir Critical Thinking Test: Test, Manual, Criteria, Scoring Sheet. An instrument for teaching and testing*. Pacific Grove: Midwest Publications

Facione, P. A. (1990). *Critical Thinking: A statement of expert consensus for Purposes of Educational Assessment and Instruction (The Delphi Report).* California State University.

Facione, P.A. (2015). *Critical Thinking: What it is and why it counts.* Revised. Insight Assessment.

Foddy, W. (1993). *Constructing Questions for Interviews and Questionnaires.* Cambridge: Cambridge University Press

Gelerstein, D., del Río, R., Nussbaum, M., Chiuminatto, P., & López, X. (2016). Designing and implementing a test for measuring critical thinking in primary school. *Thinking Skills and Creativity*, 20, 40-49.

Getzels, J.W. & Jackson, P. W. (1962). *Creativity and Intelligence: Explorations with Gifted Students*. London and New York: John Wiley and Sons Inc.

Guilford, J.P. (1967). *The nature of Human Intelligence.* New York: Mc Graw-Hill Book Company

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. UK: Lawrence Erlbaum Associates.

Hewitt, M.A. & Homan, S.P. (2003). Readability level of standardized test items and student performance: The forgotten validity variable, *Reading Research and Instruction*, 43(2), 1-16.

Hocevar, D. (1979). Measurement of Creativity: Review and Critique. *Annual Meeting of the Rocky Mountain Psychological Association*, Colorado, 12-14th April

Iozzi, L.A. & Cheu, J. (1978). Preparing for Tomorrow's World: An Alternative Curriculum Model for the Secondary Schools Paper. *First Annual Conference of the Education Section et the world Future*, Texas, 22nd October

Jiang, H. & Zhang, Q. (2014). Development and Validation of Team Creativity Measures: A Complex System Perspective. *Creativity and Innovation Management*, 23 (3), 264-275.

Johanson, G.A. & Brooks, G. P. (2010). Initial Scale Development: Sample Size for Pilot Studies. *Educational and Psychological Measurement*, 70(3), 394-400.

Johnson, S. & Johnson, R. (2009). *Conceptualising and interpreting reliability*. UK: Ofqual.

Kampylis, P. G., & Valtanen, J. (2010). Redefining creativity-analyzing definitions, collocations, and consequences. *The Journal of Creative Behavior*, 44(3), 191-214.

Kane, M.T. (2009).Validating the Interpretations and Uses of Test Scores. In Lissitz, R.W. (ed.) *The Concept of Validity Revisions, New Directions, and Applications* (pp. 39-64). United States: Information Age Publishing Inc.

Kaufman, J.C. (2006). Self-Reported Differences in Creativity by Ethnicity and Gender. *Applied Cognitive Psychology*, 20, 1065-1082.

Koretz, D. (2006). *Measuring Up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Krathwohl, D.R. (2002). A Revision of Bloom's Taxonomy: an overview. *Theory into Practice*, 41(4), 212-218.

Lambert, D. and Lines, D. (2000) *Understanding assessment: purposes, perceptions, practice*. London: Routledge Falmer

Lipman, M. (1987). Critical thinking: What can it be? *Analytic Teaching*, 8(1), 5-12.

Lipman, M. (2003). *Thinking in Education*. 2nd edn. Cambridge: Cambridge University Press

McPeck, J.E. (1981). *Critical Thinking and Education.* Oxford: Martin Robertson

McPeck, J. E. (1990). Critical Thinking and Subject Specificity: A Reply to Ennis. *Educational Researcher*, 19(4), 10-12.

Mednick, S.A. (1962). The associate basis of the creative process. *Psychological Review*, 69(3), 220-232.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741-749.

Moss, P.A (1994). Can there be Validity without Reliability? *Educational Researcher*, 23(2), 5-12.

Newton, P.E. (2012). Clarifying the Consensus Definition of Validity. *Measurement: Interdisciplinary Research & Perspective*, 10(1-2), 1-29.

Newton, D.P. (2014). *Thinking with Feeling: Fostering productive thought in the classroom*. New York: Routledge

Nisbett, R. E. & Wilson, T. D. (1977). The Halo Effect: Evidence for Unconscious Alteration of Judgements. *Journal of Personality and Social Psychology*, 35(4), 250-256.

Norris, S. P. & King, R. (1984). *The design of a Critical Thinking Test on Appraising Observations. Studies in Critical Thinking. Research Report No1*. Canada: Institute for Educational Research and Development.

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2017). Ha ha? Assessing individual differences in humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, 11(2), 231-241.

Plucker, J A. & Makel., M.C. (2010) Assessment of creativity. In Kaufam, J.C. & Sternberg, R.J. (ed.) *The Cambridge handbook of creativity* (pp.48-73). Cambridge: Cambridge University Press

Propp, V. (1968). *Morphology of the Folk Tale*. Translation by Laurence Scott. The American Folklore Society and Indiana University

Richards, J.C. (2005). *Communicative Language Teaching Today*. SEAMEO Regional Language Center

Rodari, G. (1996). *The Grammar of Fantasy: An Introduction to the Art of Inventing Stories.* Translation and introduction by Jack Zippes. New York: Teachers & Writers Collaborative

Rungo, M. A. & Jaeger, G.J. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, 24(1), 92-96. doi: 10.1080/10400419.2012.650092

Silvia, P.A. (2015). Intelligence and Creativity are Pretty Similar After All. *Educational Psychology Review.* 27 (4), 599-606. doi: 10.1007/s10648-015-9299-1

Sireci, S.G. (2009). Packing and Unpacking Sources of Validity Evidence. In Lissitz, R.W. (ed.) *The Concept of Validity Revisions, New Directions, and Applications* (pp. 19-37). United States: Information Age Publishing Inc.

Su, C.T. & Parham, L.D. (2002). Case Report- Generating a valid questionnaire translation for cross-cultural use. *American Journal of Occupational Therapy*, 56, 581-585.

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, 663-682.

Torrance, E. P., Ball, O. E. & Safter H.T. (2008). *Torrance Tests of Creative Thinking: Streamlined Scoring Guide for Figural Forms A and B.* Bensenville: Scholastic Testing Service Inc.

Treffinger, D.J., Young, G.C., Selby, E.C. & Shepardson, C. (2002). *Assessing Creativity: A guide for educators*. Florida: Center for Creative Learning

Weisberg, R. W. (2015). On the usefulness of "value" in the definition of creativity. *Creativity Research Journal*, 27(2), 111-124.

Yoon, C. H. (2017). A validation study of the Torrance Tests of Creative Thinking with a sample of Korean elementary school students. *Thinking Skills and Creativity*, 26, 38-50.