The final draft

Developing a performance assessment task in the Finnish higher education context: conceptual and empirical insights

Heidi Hyytinen & Auli Toom

1. Introduction

Critical thinking (CT) is a core generic skill for success in higher education and for academic experts working in a range of fields. It includes the ability to elaborate the problem, to evaluate the trustworthiness of information, to reason using trustworthy and relevant information and to avoid judgmental biases. Thus, CT covers a set of cognitive skills and affective dispositions (the American Philosophical Association, 1990). Previous research has shown that there is a huge variety in CT skills among university students in the different phases of their studies (e.g., Arum & Roksa, 2011; Authors, 2015; Authors, 2018). In addition to CT, the complex decision-making situations often require capacity to make moral judgements, i.e. consider moral and ethical attributes that are associated with a situation (Dewey, 1927; Greene, 2013; Haidt, 2012; Toom et al., 2015).

For these reasons, there exists a growing need to investigate the development of higher education students' CT and moral judgement during the studies. This presupposes the development of new research instruments that capture the essential characteristics of CT, not only students' perceptions related to CT. It is fundamental in CT research to utilise authentic open problems, allowing multiple perspectives and making use of various sources possible

when finding solutions to the problems (Shavelson, Zlatkin-Troitschanskaia, & Mariño, 2018). Empirical results on students' CT will allow for the development of pedagogical practices to enhance teaching to support the learning of these skills.

This article presents the development of performance assessment for research on CT and moral judgement to be utilised in Finnish higher education context. We have developed a new performance assessment focusing on CT and moral judgement in a line with general assessment guidelines (American Educational Research Association et al. 2014) and the iPAL framework (Shavelson et al., 2018). The task simulates a real-life decision-making situation and requires students to make and justify their decisions by utilising the available evidence (Shavelson, 2010). It is essential that the tasks trigger respondents' CT processes and their moral judgement, and thus for these to be valid in terms of their cognitive aspects (Brückner & Pellegrino, 2016). Our specific aim is to explore the extent to which respondents' cognitive processes and moral judgement are consistent with the definition of CT and moral dimensions in the context of assessment and how respondent characteristics are related to completion of the task. The instrument, the results of the pilot procedure of the instrument and subsequent phases in the development are presented. Although research on validity evidence based on response processes during task taking is a crucial part of the test development, such research has been seldom reported (Ercikan & Pellegrino, 2017). Thus, there exists a clear need for this kind of methodological research.

2. Theoretical and methodological background

2.1 Conceptualising critical thinking and its measurements

This study uses the definition of CT suggested by the American Philosophical Association's "Delphi report" (1990) which points out that CT is combination of various dimensions of *cognitive skills* (i.e. purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, inference, and found explanation) and *affective dispositions* (i.e. open-minded, flexible in evaluation, willing to reconsider)**.** CT is a process of assessing, evaluating, synthesising and interpreting relevant information that is associated with a situation, and applying that information to solve a problem, to decide on a course of action, to find an answer to a given question or to reach a well-reasoned conclusion (Shavelson et al., 2018). It involves open-minded and self-disciplined thinking about alternative solutions and effective communication (Halpern, 2014; Authors, 2018; Dewey, 1910).

Ethical issues underlying the decision-making situation can be related to CT and reasoning. However, the connection between CT and morality is not straightforward. Several researchers have questioned the role of reason in moral action (e.g., MacIntyre, 1999; Greene, 2013; Haidt, 2012). It can be claimed that one's ability to make moral judgements is not dependent on an ability to think critically. A person who acts rationally may, at the same time, fail in moral judgement. Although people tend to be engaged in moral reasoning, in fact they might rather use morality to justify their own preferences (MacIntyre, 1999). In addition, people do not necessary use reason to make moral judgements, rather, they use intuitions to justify them (Greene, 2013; Haidt, 2012). Nevertheless, moral notions may have some consequences in one's decision-making. For example, in professional contexts, experts act in their professional role, which requires moral reasoning and integrity in professional decisions. Thus, the demanding task of an expert is to fulfil both professional requirements and guiding norms as well as personal aspects in the critical thinking and decision-making, combine and

reason them in a sustainable way (cf. Biddle & Thomas, 1966; Buchmann, 1986; Sockett, 2009; Toom & Husu, 2018).

Researchers have found that CT is extremely difficult to capture with indirect measurements, such as respondents' self-reports of self-perceptions and self-evaluations (Zlatkin-Troitschanskaia et al., 2015). In the literature, a holistic *criterion-sampling approach* has been proposed to overcome the challenges of the measurement of complex thinking and reasoning processes (McClelland, 1973; Shavelson et al., 2018). The idea of the criterion-sampling approach is based on the following notion: if we want to know what students know and can do, students should complete the measurements in which they can demonstrate their actual performance on authentic tasks (Shavelson, 2010; Shavelson et al., 2018). Therefore, for example, if we want to assess students' ability to evaluate information and provide an explanation, we cannot capture it with multiple-choice tests or interviews (McGrew in this special issue; see also McClelland, 1973; Shavelson, 2010). To measure actual abilities and to draw more valid inferences about students' performance, we need to use measurements that assess the students' performance when they evaluate information and give explanations (Shavelson, 2010).

Our aim here was to develop a form of performance assessment that includes elements that invite students to think critically and to make moral judgements. We propose student-centred assessment in which students create their own answers in writing (McClelland, 1973). The task is expected to be authentic in focusing on the real-world societal problem. Next, we provide a sketch of the phases of task development and the task created to assess CT and moral judgement.

2.2 The phases of task development

Task development can be divided into three phases (American Educational Research Association et al. 2014). The first step is to define what is to be measured. After that, the storyline, tasks and scoring criteria must be developed so that they tap into the intended construct (Shavelson et al., 2018). In the third phase, the evidence is gathered by observing and recording performance and ensuring that the task maps directly back to the elements of the construct (Brückner & Pellegrino, 2016; Kane, 2013).

Phase 1: The construct definition

As noted, performance assessment development begins by defining what is to be assessed (Mislevy & Haertel, 2006; Zlatkin-Troitschanskaia et al., 2015; Shavelson et al., 2018). We defined the constructs as the following *cognitive and moral dimensions*:

(1) organising, synthesising, assessing and analysing information (which might be trustworthy/untrustworthy; relevant/irrelevant to the task)
(2) avoiding judgemental errors, such as 'correlation proves causation' embedded in several sources
(3) providing a reasoned explanation for a problem, proposing a solution and recommending a course of action
(4) justifying an explanation by writing arguments and counter-arguments for and against a particular solution using information from the documents
(5) considering moral and ethical implications of decision and action.

The task developed for the performance assessment needs to incorporate all five facets enumerated above. Our previous studies have concentrated on the cognitive dimensions of CT (1-4 above; Authors, 2018; 2015; cf. Halpern, 2014). These dimensions represent crucial CT skills required in various higher education disciplines, through which students construct knowledge and understanding (Authors, 2018). These skills have been found to be essential for university students to progress in their studies (Arum & Roksa, 2011; see also Alexander, this issue).

Furthermore, we have added a new construct into the task being developed—*ethical and moral judgement.* Therefore, the task should be based on a societal problem which challenges students to consider ethical and moral issues as well. Ethical dimensions one way or another underlie most of the professional expert tasks, and exercising integrity and practicing moral judgement is thus necessary. The task should focus on the ethical dimensions relating to the situation and the course of action and invite students to make moral judgements, i.e. judging that solution or action has a moral attribute (Greene, 2013; Haidt, 2012). For example, while completing the task, students need to consider that in a certain position one has to be aware of their duty, and to illustrate that the situation or course of action may cause harm or benefit to many people. Using a task that combines both the constructs of CT and ethical deliberation, it is possible to explore the complex relationship between the two notions among students, how students balance between these two aspects, and even how they might balance between the different premises of ethics in their problem-solving processes.

Phase 2: Generating the storyline, information sources, tasks and scoring rubric

In the next phase, based on the construct definition and considerations of performance assessment we created a storyline and information sources and set forth tasks for students to address (see Figure 1; Mislevy & Haertel, 2006; Pellegrino et al., 2001; Shavelson et al., 2018).

FIGURE 1

The story upon which the assessment is based focused on issues of migration faced by a fictitious country in Europe with given demographic information. Actual information sources (materials or "document library") such as newspaper articles, YouTube videos, statistics, and research abstract were drawn from Finland and UNESCO. In the assessment, we removed and anonymised the sources by using fictitious names. Finally, we included eight documents in the task (Figure 1). At the end of the story, five open-ended questions—tasks—were posed for students to address in their written solution.

The assessment was developed so that the documents contained the necessary information to complete the task and the information is not specialised to the specific field of study (Shavelson et al., 2018). The documents were selected and manipulated so that they included contradictory information, meaning that some was trustworthy and relevant, whereas some was untrustworthy and irrelevant to the problem. In addition, several heuristics, such as representativeness (i.e. refers to situation when one judges probabilities on the basis of resemblance) and baseline (i.e. one prefers information pertaining only to a certain case over general information when the former is available) heuristics (Kahneman, 2011), were added.

In summary, the assessment has been developed so that when combining and elaborating different pieces of information, it invites the students to consider information sources and their trustworthiness and relevance to the problem, and to avoid the heuristics and recognize the biased information. The assessment invites students to slow down, think more thoroughly, and finally to draw a justified conclusion based rational and moral thought.

Phase 3: Evidence

The third phase focuses on a cognitive validity of instrument, that is, whether respondents' CT processes and their moral judgement while performing the tasks are consistent with our construct definition (Brückner & Pellegrino, 2016; Kane, 2013; Karabenic at al. 2007). A variety of methods and assessment protocols can be used to examine evidence according to its cognitive validity (Kane, 2013; Leighton, this issue). The protocol that is used depends on the type of measurement being developed and the type of data that a researcher aims to collect using the measurement (Karabenick at al. 2007). In the present study, cognitive labs with think-aloud protocol and follow-up interviews were conducted with five participants drawn through purposeful sampling (for a description of sampling, see section 2.1; Brückner & Pellegrino, 2016; Leighton, 2017). Through this process, the aim is to gather evidence on the extent to which the task evoked the various dimensions of critical and moral thinking that fit with the construct definition. The aim of cognitive labs with think-aloud protocol is to explore students' interpretations of the performance task, as well as to identify and analyse the cognitive processes and moral judgement occurring during the response process.

Several respondent-related factors, such as students' prior knowledge, academic expertise, expectations, attitudes, self-regulation, and mood are related to and influence on individuals' performance and interpretation (Brückner & Pellegrino, 2016; Mislevy & Haertel, 2006). For example, experts are assumed to be superior to novices in completing the assessment. Participants' performance results from a complex interaction between individual and task characteristics, such as the difficulty of the task and the content area (Brückner & Pellegrino, 2016). Therefore, analyses of cognitive validity should not only focus on task characteristics, but they should also consider individual aspects influencing participants' responses (Brückner & Pellegrino, 2016; Mislevy & Haertel, 2006). By revealing the cognitive processes that occur during the task and how these processes are associated with individual and task characteristics, the present study sheds light on the modifications required to improve the validity of the performance assessment being developed.

3. Pilot Study

The pilot study focuses on claims that the performance assessment tapped into cognitive and moral dimensions as we have defined it. We set out to identify the factors triggering CT processes and moral judgement during task taking. These objectives are approached through the following research questions:

1) What kind of task characteristics triggering critical thinking and moral judgement can be identified?
2) What kind of respondent characteristics emerge when participants complete the assessment?

Here we report the results of think aloud and cognitive labs as well as interviews. More specifically, we describe the participants, the materials and methods used, and the findings.

3.1 Participants

This pilot study was conducted with five participants drawn from purposeful sampling (Table 1). Direct recruitment of potential participants was the chosen strategy in order to gather a representative sample of key informants. Two of the participants were female and three were male. The participants' ages varied from 27 to 59 years. They came from a homogeneous cultural background, and all shared the same first language (Finnish). The sample included beginning and experienced participants in terms of academic expertise from two different domains— educational and social sciences—with various specific experiences related to the assessment. This allowed us to analyse performance and individual characteristics related to solving the performance assessment including the difficulty level of the tasks. We entertained the possibility that the assessment might turn out to be too difficult for the students at the end of their studies or for academic experts. Had that been the case, it would have been necessary to adjust the difficulty level to be suitable for the novice university students. Voluntary participation, informed consent, and anonymity of the participants were ensured in the research process.

TABLE 1

3.2 Procedures and analyses

We collected the data from cognitive labs with think-alouds and interviews (Brückner & Pellegrino, 2016; Leighton, 2017). The data were collected in two phases. In Phase 1, a participant worked 70 to 90 minutes reading and responding to the task in think aloud mode.

They were not interrupted by the interviewer except when they fell silent and the interviewer reminded them to talk aloud. The interviewer took notes on matters such as how much time it required for the participant to complete the task, if the participant struggled with certain terms. In Phase 2, after the assessment was completed, the participant was interviewed about his/her thinking processes in carrying out the task. The aim of the interview was to gather detailed information about individual characteristics as well as participants' interpretations of different parts of the assessment including instructions, documents, and assessment story.

The data collection situations were digitally recorded and transcribed verbatim. The data were analysed using a qualitative abductive approach (Timmermans & Tavory, 2012) aiming to generate novel theoretical insights through dialogue between previously associated and new interpretations The authors firstly systematically coded the data with three qualities in mind: (a) the process by which the respondents approached the task and solved the task, (b) the knowledge and information that the respondents used to carry out the task, and (c) the skills respondents tap in completing the task. Thereafter, the coded extracts were grouped into categories and sub-categories. We identified two main categories triggering CT and moral dimensions, namely (1) task characteristics and (2) respondent characteristics. This phase was theory-driven. In order to identify the categories, we utilised task- and respondent-related factors identified by Brückner and Pellegrino (2016), Mislevy and Haertel (2006) and Leighton (2017). In addition, we utilised the findings from the research, emphasising dimensions of CT (cf. Halpern, 2014; Authors, 2018; 2015; Shavelson et al., 2018) and moral judgement (cf. Sockett, 2009; Greene, 2013; Haidt, 2012; Toom & Husu, 2018). The first category was divided into four sub-categories, which were labelled as the relevance of the task, the response procedure, difficulty of the task, and interpretation of the task items and materials. In addition, we distinguished three sub-categories pertaining to respondent

characteristics: an ability to monitor thinking, motivation and performance, prior experiences and knowledge, and integrity. The final categories and sub-categories are described in Table 2. The analyses were conducted in collaboration with the authors. In the analysis, investigator triangulation (Denzin, 2012) was utilised to confirm the reliability of the findings. The data examples were translated into English.

TABLE 2

4. Preliminary findings

4.1 Task characteristics on triggering CT and moral judgement

The relevance of the task

The problem situation, the "story line" presented in the assessment was perceived as realistic. Due to the high level of authenticity, the assessment theme was perceived as being interesting, motivating and attractive. The participants recognised that the theme dealt with a societal problem that has been a common topic in recent public discussion. They even said that it is ethically multifaceted and politically charged. One participant characterised the assessment while thinking aloud in the following way:

> *"Pretty interesting. So, there are five difficult questions and the answer should be clear, organised, and thorough. So, this is definitely enough work for an hour and a half. But the topic is current and interesting."* (ID2)

The response procedure

Based on think aloud and observations, the response procedure included three phases. In order to respond to the task, all participants first familiarised themselves with the introduction and glanced through the materials provided. Secondly, they read, organised, synthesised and analysed information from eight documents. They moved back and forth between the documents several times. In doing these activities, they assessed their confidence in the information taken from different documents, including the relevance and trustworthiness of the source. They also dealt with conflicting information, as a following extract from think-alouds demonstrates:

> *"This is a pretty brave one, the author* [Document 3] *uses quite brave arguments here, in my opinion. "This is not a singular case. Statistics show that there have been significant changes in the crimes committed by refugees. One of the reasons is definitely the flow of refugees in 2015". It is surely true that this has raised worries among the original population. However, this is contradictory to what is written in Document 1. And then the residents here say in Document 3 that "fear and worry about the future is natural, since disturbances and violence have increased in the refugee centres and the surrounding areas". Quite opinionated, but at least through Document 3 we got a sort of a perspective of the opinions of some of the residents."* (ID1)

Thirdly, the participants used the results of their analysis to decide on a course of action and provided an explanation and justification of their actions, as they were asked to do in the assessment. Three of the five participants also considered ethical issues and consequences related to the alternative courses of action as well as avoiding the judgmental heuristic traps.

Difficulty of the task

The participants perceived the task as being difficult and relatively demanding. The participants emphasised both in think-aloud and interviews that to conclude the task they needed to read and understand several rather challenging documents, combine, evaluate and synthesise the information from the various documents, and to apply that information to reach a conclusion and explanations to the questions presented in the task. They also weighed up moral and ethical consequences relating to the different solutions and the course of action. In particular, they found it difficult to interpret the statistics and the research abstract. In addition, the terminology of the assessment was described as demanding. However, they mentioned in the interview that the information presented in the task was generally recognizable and that such information had been addressed in the different real-life media:

> *"This is in my opinion challenging. This task has many questions. And then these tables* [Document 2] *are difficult. This task is in relation to thousands of aspects. I had to read this research abstract* [Document 4] *several times. Yeah, those terms are really difficult. But the media has also shown these* [in a real-life] *a lot."* (ID1)

Interpretation of the task items and materials

Both the think-alouds and cognitive labs revealed that there was variation in how one question was interpreted. In this question participants were asked to elaborate and recommend a concrete course of action. The question included three alternatives. Three participants thought that they were being forced to select the best option from these three

alternatives. However, two participants thought that these three alternatives were only examples and that their task was to elaborate the alternatives against the materials:

> *"Upon answering this question, you have to remember that this task gives you three choices, but one must base thinking on the materials and think outside the box whether to choose option number four."* (ID2)

In a similar vein, there were different interpretations of the relevance, trustworthiness and fundamentality of the sources. There was variation in interpretations of the relevance and trustworthiness of a letter (Document 1), a newspaper story (Document 3), Youtube (Document 6), and a case report (Document 8). Four participants perceived Document 5 as fundamental and as a baseline for all decisions and recommendations, as the following extract illustrates

> *"Document five is fundamental. Actually, I don't think that any of the principles or recommendations of the expert group should be against the spirit of Document 5."* (ID5)

4.2. Respondent characteristics on performing the task

Ability to monitor thinking, motivation and performance

All participants were able to maintain their motivation during performing the task. However, they differed in terms of their ability to monitor their thinking and performance in order to adapt their thinking processes according to the demands of task as well as their prior experiences and knowledge. Think aloud and cognitive lab data illustrated that the

participants who had prior knowledge of the theme due to their studies or the extent of their expertise or prior experience with the type of assessment, needed 40 minutes to analyse the documents, whereas the others used over 60 minutes to complete the analysis of the documents. The participants who had prior knowledge of the theme or extensive expertise did not find the terminology of the task to be challenging. They smoothly figured out the goals of the task and what they needed to do to accomplish them. They identified all the major ideas presented in the documents with ease. They also weighed up the options and connected related ideas without effort. In contrast, the others struggled with the terminology and combining the range of ideas. The participant with a strong expert identity expressed the view that some aspects of the assessment went beyond his area of expertise, and thus, he would like to check some of the details before making a final decision.

Prior experience and knowledge

Although the assessment evoked both professional and private attitudes related to the theme, the participants who had no prior knowledge of the theme emphasised that the role assigned in the assessment supported their performance. It was also noted in the think-alouds, cognitive labs, and interviews that the role framed, contextualised and gave a neutral perspective to the task as the following extract from an interview illustrates:

> *"The expert group's goal influenced me. So, that in some way brought responsibility to my answers. A certain kind of seriousness was there, as I cannot rely on any individual opinions or experiences in this context but that the answers must somehow more neutral. This could have been more speculative if it had not been linked to such an expert group to draw up some guidelines and make recommendations. So, it could have been more open to playing with ideas."* (ID1)

> *"The role directed me to think about these things more rationally. So, one must be able to say something about the different perspectives present in the documents. Whether you agree with them or not."* (ID3)

On the contrary, the participant with a strong expert identity said that

> *"This is in line with our procedures: This is how we work in a democratic society. This shows how we prepare issues to be clarified further in the democratic society. This resembles the ways in which we do things. I can easily see myself in this kind of role."* (ID5)

Integrity

When working on the task, constructing the answer, and considering their role in solving the task, the participants also pondered the alternatives and consequences of a course of action in the light of their moral principles. They were related to their personal and professional roles and to the ethical attributes underpinning the task. The participants elaborated their ways of thinking and behaving in relation to the individual immigrants and their families, their position in the broader community and in the whole society. Especially, the participants who had prior knowledge of the theme due to their studies or the extent of their expertise expressed their commitments and values when suggesting certain principles and courses of action. The elaboration clearly demonstrated the participants' integrity in principles and course of action.

> *"In this kind of societal questions, it is always necessary to remember that there are human beings, whose life the decisions truly touch and change. As a professional, I*

have to remain truthful to my principles and values – I cannot act against them." (ID5)

5. Discussion

In the piloting process, and based on the participants' experiences, we can say that the performance assessment being developed has relevance to real everyday life. The assessment maintained the participants' interest throughout to completion. The assessment was open-ended, sensitive to prior knowledge and experiences, allowed multiple lines of thinking and did not evoke a standard or fixed response (Shavelson, 2010; Shavelson et al., 2018). The assessment put the participants in the role of expert which was natural for participants who identified their expertise, but challenging for those in the earlier phases of their studies (cf. Brückner & Pellegrino, 2016).

The instructions given for completing the assessment worked well and they triggered participants' CT processes and their moral judgement (cf. Authors, 2018; Toom & Husu, 2018). The assessment was experienced as being relatively challenging, but still possible to complete. It challenged the participants' CT, opened new aspects to the familiar theme they had not thought of before. Perhaps most importantly, the assessment separated participants in terms of their critical and ethical thinking. As assumed on the basis of previous research (Brückner & Pellegrino, 2016), the assessment distinguished between experts and novices. Experts shared a strong adherence to consider alternatives and consequences of a course of action in the light of the ethical attributes (cf. Greene, 2013; Haidt, 2012; Sockett, 2009).

Despite the small number of participants in the pilot testing, we detected both respondent and task characteristics that need to be considered in subsequent phases of the development (cf. Brückner & Pellegrino, 2016; Mislevy & Haertel, 2006; Leighton, 2017). The prior knowledge of the participants is related to the ways of performing the task. For example, due to prior knowledge, the advanced student in the field of social sciences and the expert from the educational sciences did not stick to the concepts in the same way as others; they were able to problematise and further question the data sources. They processed the task smoothly, were able to combine facts and viewpoints from the various sources and perceive the connections underlying various documents. They also perceived the task as a real expert task, and perceived the connections with real expert work, preparation procedures and decision-making conducted in a democratic society. Furthermore, results revealed that the advanced student in the field of social sciences and the expert were aware of ethical issues underlying the situation and the consequences of a course of action.

In subsequent phases, the task will be tested with a bigger group of participants and modified further. For example, the amount and clarity of the statistics as well as the number of documents need to be considered further. The terminology used in the documents also needs to be simplified. In addition, the scoring rubric will be developed that it focuses on the students' use of the different information sources presented in the assessment as well as their elaboration and avoidance of heuristics that lead to errors in judgment, decision-making, argumentation, and elaboration of ethical and moral aspects relating to the situation. Later, the data will be modelled both statistically and qualitatively to bring them to bear on the evidence of the quality or interpretability of the task (Brückner & Pellegrino, 2016; Shavelson et al., 2018).

Although our aim is to develop a test to investigate students' CT and moral judgement, it also has broader connections to the Finnish higher education context and the developments needed there. Once the performance assessment is finalised, students will be able to familiarise themselves with the assessment rubrics and come to learn about the nature of CT and moral judgement as embedded in the assessment. In an ideal situation, students will also receive feedback on their performance, which will help them in learning these issues.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Philosophical Association. (1990). Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction, The Delphi Report: Research findings and recommendations prepared for the committee on pre-college philosophy. P. Facione, (Project Director). ERIC Doc. No. ED 315-423.

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

Authors 2015, 2018

Biddle, B. J. & Thomas, E. J. (Eds.) (1966). *Role theory. Concepts and research*. New York: John Wiley & Sons, Inc.

Buchmann, M. (1986). Role over person: Morality and authenticity in teaching. *Teachers College Record, 87*(4), 529_543.

Brückner, S., & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multilevel models to validate an assessment of higher education students' competency in business and economics. *Journal of Educational Measurement,* 53(3), 293–312.

Denzin, N. K. (2012). Triangulation 2.0. *Journal of Mixed Methods Research*, 6, 80–88.

Dewey, J. (1910). *How we think*. Boston: D.C. Heath & Co.

Dewey, J. (1927). In J. A. Boydston (Ed.), The public and its problems. Reprinted in John Dewey, The later works: 1925_1953 (LW) (Vol. 2, pp. 253_372). Carbondale, IL: Southern Illinois University Press.

Ercikan, K. & Pellegrino, J. W. (2017). *Validation of Score Meaning Using Examinee Response Processes for the Next Generation of Assessments*. NY: Routledge.

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York, NY, US: Penguin Press

Haidt, J. (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion*. New York: Vintage.

Halpern, D. F. (2014) *Thought and Knowledge*. Fifth edition. NY: Psychology Press.

Kahneman, D. (2011) *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., De Groot, E., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, 42 (3), 139–151.

Leighton, J. (2017). Collecting and Analyzing Verbal Response Process Data in the Service of Interpretive and Validity Arguments. In Ercikan, K. & Pellegrino, J. W. (eds.*) Validation of Score Meaning Using Examinee Response Processes for the Next Generation of Assessments*, pp. 25-37. NY: Routledge.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, 42(4), 448-457.

MacIntyre, A. (1985). *After Virtue: A Study in Moral theory*. Second postscript, Duckworth, London.

McClelland, D.C. (1973). Testing for competence rather than testing for "intelligence". *American Psychologist*, 28(1), 1-14.

Mislevy, R. J., & Haertel. G. D. (2006). Implications of Evidence-centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, 25, 6–20.

Shavelson, R. J. (2010). On the measurement of competency. *Empirical Research in Vocational Education and Training*, 2(1), 41-63

Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Mariño, J. (2018). International Performance Assessment of Learning in Higher Education (iPAL): Research and Development. In O. Zlatkin-Troitschanskaia et al., (Eds.). *Assessment of Learning Outcomes in Higher Education – Cross-national Comparisons and Perspectives* (pp. 193-214). Wiesbaden: Springer.

Sockett, H. (2009). Dispositions as virtues: The complexity of the construct. *Journal of Teacher Education, 60*(3), 291-303.

Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: from grounded theory to abductive analysis. *Sociological Theory*, 30, 167-186.

Toom, A. & Husu, J. (2018). Teachers' work in changing educational contexts: balancing between the role and the person. In H. Niemi, A. Toom, A. Kallioniemi & J. Lavonen (Eds.), *The teacher's role in the changing globalizing world* (pp. 1-9). Leiden: Brill.

Toom, A., Husu, J. & Tirri, K. (2015). Cultivating student teachers' moral competencies in teaching during teacher education. In C. J. Craig & L. Orland-Barak (Eds.), *International Teacher Education: Promising Pedagogies (Part C): Advances in Research on Teaching.* Volume 22C. (pp. 13-31). UK: Emerald Publishing. ISSN: 1479-3687/doi:10.1108/S1479-368720150000026001

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education*, 40(3), 393–411. https://doi.org/10.1080/03075079.2015.1004241.

Table 1. Participants of the study.

Participant ID	**Age (years)**	**Gender**	**Educational background, experience**	**Occupation**
1	27	male	MA, Educational sciences, familiar with the general principles of performance tasks	Graduate
2	27	male	BA, Social sciences	Student
3	32	female	PhD candidate, Educational sciences	Doctoral student
4	26	female	MA, Social sciences	Graduate
5	59	male	PhD	Professor of Education

Table 2. The main categories and sub-categories triggering CT and moral judgement

Category	**Sub-categories**	**Definition**
Task characteristics	Relevance of the task	The degree to which the task is realistic and interesting.
	Response procedure	The phases required to accomplish the task.
	Difficulty of the task	The effort required to complete the task.
	Interpretation of the task items and materials	The way of understanding and explaining the meaning of the elements of the task.
Respondent characteristics	Ability to monitor thinking, motivation, and performance	Planning and evaluating one's own thinking, motivation and actions according to the demands of task.
	Prior experiences and knowledge	Using the knowledge the respondent already has before s/he meets new information provided in the task.
	Integrity	Showing an adherence to consider alternatives and consequences of a course of action in the light of the moral and ethical attributes.

Performance task

Story
The Ministry of the Interior has established an expert group to study the challenges the fictional country faces with a rapid inflow of refugees. The Minister has appointed you to an expert group that holds hearings about the situation. You are provided with following documentation:

Document 1: Letter
Document 2: Statistics
Document 3: Newspaper story
Document 4: Research abstract
Document 5: Policy guideline
Document 6: YouTube video
Document 7: Graph
Document 8: Case report

Questions
The chair has asked the expert group to review the documents and answer questions, e.g.:

1. Identify pros and cons, if any, for accepting more refugees.
2. Elaborate and recommend a concrete course of action: stem the flow if refugees are at the border, control the flow of refugees, or take in a quota decided upon agreements. Identify the evidence leading to your recommendation
3. Identify what additional information, if any, should be considered to increase your confidence in the recommendation.

Figure 1. A summary of the performance task