Turun yliopisto
University of Turku

CRITICAL THINKING EFFICACY AND TRANSFER AT AN INTERNATIONAL BACCALAUREATE DIPLOMA PROGRAMME SCHOOL IN FINLAND

Shane Horn
Faculty of Education
Department of Teacher Education
Master's Thesis
Learning, Learning Environments, and Educational Systems
May 2018

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

THE UNIVERSITY OF TURKU
Faculty of Education
Department of Teacher Education
HORN, SHANE: Critical thinking efficacy and transfer at an International Baccalaureate Diploma Programme school in Finland
Master's Thesis, 65 pages, 20 appendix pages
May 2018

With the advent of the 'fake news' phenomenon providing a backdrop for relevance, this study is concerned with critical thinking (CT) efficacy and transfer, the former referring to the extent to which facilitation of CT skills meet their intended goals and the latter referring to the extent that CT may extend to external contexts such as the media literacy skills required in daily online interactions. These topics were explored by replicating tasks developed by the Stanford History Education Group (SHEG) at an International Baccalaureate Diploma Programme (IBDP) school in Finland. To further consider the extent that the IBDP may foster growth in CT skills, the tasks were administered to a 'pre-IB' cohort (n=42) preparing to enter the IBDP and an 'IB2' cohort (n=25) preparing to graduate from the IBDP.

While results indicate stronger outcomes on the tasks by the IB2 comparative to the pre-IB, both cohorts in Finland outperformed the U.S. cohorts from the SHEG study (n=170-176) to a far more statistically significant degree. Utilising an existing framework of curricular approaches to facilitating CT skills development as a variable of interest for causal-comparison, it was determined that the Finnish education system and the IBDP develop CT skills by explicitly teaching CT as a separate course, whereas the curriculum guiding the participants in the SHEG study implicitly embeds CT into subject areas. Implications for increasing explicit facilitation in CT course design across curricula, professionalising CT across the field, and the benefits of replicating existing studies in differing socioeducational environments are discussed.

Keywords: critical thinking, media literacy, fake news, Finnish education, International Baccalaureate

Table of Contents

SUMMARY …… ii

LIST OF APPENDICES …… v

LIST OF CHARTS …… v

LIST OF FIGURES …… vi

LIST OF TABLES …… vi

INTRODUCTION …… 1

LITERATURE REVIEW…… 4

2.1 Background and context …… 4

2.2 Defining critical thinking …… 4

2.2.1 Ennis' instructional approaches for critical thinking …… 7

2.2.2 Critical thinking transfer …… 9

2.2.3 Critical thinking efficacy …… 10

2.3 Information and media literacy …… 10

2.4 Curricula under consideration …… 12

2.4.1 International Baccalaureate …… 12

2.4.2 Finnish education system …… 18

2.4.3 U.S. education system …… 22

2.5 Applying Ennis' classification for CT approaches to curricula …… 24

RESEARCH QUESTIONS …… 26

METHODOLOGY …… 27

4.1 Replication …… 27

4.1.1 Stanford History Education Group task development …… 28

4.1.2 Validity …… 33

4.1.3 Reliability …… 33

4.2 Task threats …… 36

4.3 Qualitative data …… 42

RESULTS ……. 43
5.1 Administration of tasks ……. 43
5.2 Demographic data …… 44
5.3 Post-survey data …… 47
5.4 Results on tasks …… 48
5.4.1 Descriptive data …… 48
5.4.2 Comparing outcomes of the pre-IB to the IB2 …… 49
5.4.3 Regression analysis …… 52
5.4.4 Comparing outcomes of the U.S. to the pre-IB and IB2 …… 53
DISCUSSION …… 57
6.1 Limitations …… 61
6.2 Recommendations …… 62
6.3 Further research …… 63
CONCLUSION …… 65

LIST OF APPENDICES

APPENDIX A: Stanford History Education Group assessment tasks for high school ….. 73
Task 1: Argument analysis …… 73
Task 1: Argument analysis rubric …… 74
Task 2: News on Facebook …… 76
Task 2: News on Facebook rubric …… 77
Task 3: Facebook argument …… 79
Task 3: Facebook argument rubric …… 80
Task 4: Evaluating evidence …… 81
Task 4: Evaluating evidence rubric …… 82
Task 5: Comparing articles …… 85
Task 5: Comparing articles rubric …… 87
APPENDIX B: Study script …… 89
APPENDIX C: Survey …… 90
APPENDIX D: Consent form …… 92

LIST OF CHARTS

Chart 1: Participant gender distribution …… 44
Chart 2: Country-of-origin for the pre-IB …… 44
Chart 3: Country-of-origin for the IB2 …… 45
Chart 4: First language for the pre-IB …… 45
Chart 5: First language for the IB2 …… 46
Chart 6: Years studied in the Finnish education system for the IB2 and the pre-IB …… 46
Chart 7: Years studied in the IB programme for the IB2 and pre-IB …… 47
Chart 8: Results of the pre-IB on the five SHEG high school tasks …… 49
Chart 9: Results of the IB2 on the five SHEG high school tasks …… 49
Chart 10: Results of the pre-IB and IB2 on Task 1 …… 50
Chart 11: Results of the pre-IB and IB2 on Task 2 …… 50
Chart 12: Results of the pre-IB and IB2 on Task 3 …… 51
Chart 13: Results of the pre-IB and IB2 on Task 4 …… 51

Chart 14: Results of the pre-IB and IB2 on Task 5 …… 52
Chart 15: Results of the U.S., the pre-IB, and the IB2 on Task 1 …… 54
Chart 16: Results of the U.S., the pre-IB, and the IB2 on Task 2 …… 54
Chart 17: Results of the U.S., the pre-IB, and the IB2 on Task 3 …… 55
Chart 18: Results of the U.S., the pre-IB, and the IB2 on Task 4 …… 55
Chart 19: Results of the U.S., the pre-IB, and the IB2 on Task 5 …… 56

LIST OF FIGURES

Figure 1: IB Diploma Programme design …… 15
Figure 2: IB Theory of knowledge design …… 16

LIST OF TABLES

Table 1: Ennis' (1989) classification for CT with explicit and/or implicit conditions ……. 8
Table 2: Mean differences in students' perceived outcomes of TOK study ……... 17
Table 3: Ennis' (1989) classification for CT applied to curricula …… 24
Table 4: Ennis' (1989) classification with CT taught separately or embedded ……. 25
Table 5: SHEG high school assessment task descriptions with sample size ……. 29
Table 6: Conceptual questions for measurement framework aligned with tasks …… 34
Table 7: Principle component analysis of SHEG tasks matrix …… 34
Table 8: Component plot in rotated space for the SHEG tasks …… 34
Table 9: Mann-Whitney U Test per task, all tasks, and all tasks except Task 2 …… 43

Introduction

While serving as Finland's Director of Communications, Markku Mantila appeared late last year in a U.S. comedy news programme for a segment titled 'Finnish Fake News'. During his interview, Mantila—whose duties included monitoring attempts to influence the Finnish public through campaigns of disinformation—claimed that Finnish citizens have been notably successful in their ability to discern credible information online. "We have trolls. We have fake news," he said. "But the Finns do not buy false news" (Bee, et al., Allen & Pennolino, 2017). Stirred to action by political unrest in Ukraine, frustrated from the inability to manage public perceptions regarding Russo-Finnish child custody disputes, and concerned about disinformation spread via social media on the issue of asylum seekers in Finland, Mantila and the Council of State invited foreign experts to train over 100 members of the Finnish government on the matter ('US experts gird Finnish officials for information war,' 2016). "An important step was our president spoke against disinformation, and he warned us that, whatever you find on the web, be critical," Mantila said, adding this by way of explanation for Finland's success: "We give lessons to students, how to avoid fake news… how to be critical" (Bee, et al., Allen & Pennolino, 2017).

Similar conclusions were echoed in an article published in *Foreign Policy*. Considering the ways in which false news stories and conspiracy theories designed to aggravate existing social problems spread on social media, the author contends that in addition to its extended political history with Russia and a concentrated governmental strategy to deal with the issue, Finland's "strong public education system" has effectively equipped the population to "deflect coordinated propaganda and disinformation" (Standish, 2017, para. 5). Noting that President Niinistö publicly acknowledged information warfare as a homeland reality and declared it the duty of every Finnish citizen to take active measures to individually defend against false news, the article credits the education system as responsible for producing "widespread critical thinking skills among the Finnish population" which in turn provides "a strong defense against concerted outside efforts to skew reality" (para. 10).

A reporting project by the International Press Institute (IPI)—a global organisation which promotes and protects press freedom and journalistic practices in consultation with the UN, UNESCO, and the Council of Europe—also highlighted Finland in its series *Contending with*

'Fake News'. In an article titled 'Finland remains resistant to "fake news", disinformation,' the fake news under consideration again included Russian-led disinformation campaigns in conjunction with anti-immigration and political hate speech. The Finnish journalists and media experts interviewed concluded that the spread of false news and disinformation has had a much smaller impact in their country than it has internationally (Koponen, 2018a, para.3). While the IPI report focused in accordance with its mission statement on the role of Finland's journalistic integrity, it also noted that the country's professional media outlets "benefit from a well-educated populace" (Resistance section, para. 7) when it comes to defending against false information spread online.

Within each of these reports is essentially a three-claim argument: That (a) there are campaigns of disinformation being spread in Finland, but (b) they have been largely unsuccessful, because (c) Finland's outstanding education system has properly equipped its citizenry with the necessary thinking skills to distinguish reliable information from the untrustworthy. The initial claim that there exist campaigns of disinformation within Finland appears factually true based on credible investigation (Aro, 2016; Peters, 2017; Rosendahl & Forsell, 2016). The available evidence to support the follow-up assertion that these attempts at disinformation have been comparatively unsuccessful in Finland, however, remains less clear; while Mantila repeats this directly in various media—and while the media specialists themselves appear similarly convinced of its truth value—there is no evidential support provided to offer substantiation to the claim.

Indeed, the conclusion by Mantila and the media experts does not appear to be generally shared by the Finnish public, as 67% (+/-3.1%) indicated in a poll conducted by the Finnish Media Federation that they believe false news affects citizens' perceptions on issues either 'a lot' or to an 'extreme' degree ('Survey: Majority of Finns feel effects of fake news', 2017). While both the media experts and the public are reporting on what they believe to be true and not necessarily what is measurably factual (as indicated by the divide in opinion), if one takes the position of assuming Finnish citizens are indeed more skilled at determining the truth value of online sources of information relative to other societies, there remains still a larger assumptive leap toward reaching the causal conclusion that Finland's education system must therefore be attributable.

While this research does not deal directly with claims of intergovernmental propaganda and campaigns of disinformation, it does consider the thinking skills involved in evaluating the reliability of evidence found in 'real world' online contexts and the potential role of educational curricula in helping develop those skill sets. The study is broadly concerned with critical thinking (CT) efficacy, or the extent to which the facilitation of CT skills may be producing its intended result (see 2.2.3). More specifically, it considers the ways in which CT skills developed in the classroom may transfer to external contexts such as those found in daily interactions with social media and online news outside the classroom.

To this end, materials developed by the Stanford History Education Group (SHEG) to measure media literacy skills in online contexts were administered to consider the extent to which students at an International Baccalaureate (IB) Diploma Programme (DP) school in Finland effectively applied the questions 'Who is behind the information?' and 'What is the evidence?' to a series of standardised paper-and-pencil tasks. While the SHEG study, which received widespread media attention[1] (Domonoske, 2016; Hunt, E., 2016; Ingram, 2016; Schulten & Brown, 2017; Shellenbarger, 2016; Wineburg & McGrew, 2016) due to its completion coinciding with the advent of the 'fake news' phenomenon, was descriptive in nature and covered much by way of breadth, this study provides some depth by measuring outcomes between groups within curricula specifically designed to develop CT skills, and offers further comparative value with the performance outcomes collected by the SHEG of U.S. high school students on the same tasks.

As a result of the findings, this study may provide empirical evidence supportive of the claim that educational curricula such as the Finnish curriculum and the IBDP which explicitly facilitate CT skills reveal stronger outcomes in students' abilities to evaluate the reliability of information found online than those which implicitly facilitate CT skills such as the U.S. curriculum guiding the participants of the SHEG study. The study overall provides early data gathering toward broader research questions related to CT efficacy and presents a case for the benefits of replicating studies across additional socioeducational environments.

[1] Coverage by prominent outlets included *The Guardian, National Public Radio*, *The New York Times, The Public Broadcasting Service*, and *The Wall Street Journal*, with international coverage across China, England, France, Germany, Indonesia, Israel, Italy, Japan, Norway, Philippines, Poland, Russia, Sweden and Turkey.

Literature Review

The first sections of this literature review explore the background and context of the study (2.1); define the established understandings of CT (2.2); establish a classification system for approaches to teaching CT (2.2.1); define and explain CT skills transfer and efficacy (2.2.2 and 2.2.3); and consider skills development related to information and media literacy (2.3).

It then provides information on the following curricula with an analysis of their approach to CT pedagogy: the IBDP (2.4.1); the Finnish education system (2.4.2); and the U.S. education system (2.4.3). Lastly, it determines which curriculum follows what approach in consideration of existent research.

2.1 Background and context

While the desire for increased CT skills development is "widely cited by national education groups, teacher unions, higher education organizations, and workforce development groups as an imperative for today's students" (Silva, 2009, p. 630) and reflective of a global push across various national curricula toward this end—most visibly found in Australia ("Australian curriculum, assessment and reporting authority," n.d.), Canada (Fillion & Martelli, 2017; "Ontario Ministry of Education," 2017, p.8), England (Glevey, 2008), Singapore (Leen, Hong, Kwan, & Ying, 2014) and some states in the U.S. (Ennis, 2018, p.165; Gewertz, 2008; Silva, 2009, p. 630) as an explicit component of their educational objectives—little work appears to have been conducted in measuring the outcomes of such efforts (Tiruneh, Verburgh, & Elen, 2014, pp. 1-2) to consider the extent to which they are fulfilling their intended purpose. This study strives to fill some of these gaps.

2.2 Defining critical thinking

There emerges a pattern within the existing academic literature on CT, which nearly always contain at minimum a disclaimer to often a full discussion that the concepts, qualities, or processes considered within the defining parameters of CT are difficult to determine with some level of denotative specificity (Ab Kadir, 2007; Ennis, 1991; Larsson, 2017; McPeck,

1990; Orszag, 2015; Paul, 1995; Saiz & Nieto, 2010; Tiruneh et al., 2014), a problem which naturally results in CT being defined variously; or, as Saiz and Nieto (2010) express it: "the number of [CT] definitions is almost as large as the number of researchers engaged in its study" (p. 19). Despite this common concession and problematised viewpoint, there emerges further patterns toward shared defining parameters, which nearly every study then proceeds to list with some semblance of the historicity involved in determining CT's contemporarily accepted meanings.

While that historicity can be traced back at least as far as Socrates, the more contemporary meanings connected to pedagogic practice follow a first wave of CT-oriented educationalists from the early 20th century to an "explosion of interest" (Ennis, 2018, p. 165) beginning in the 1980s and growing exponentially through the ensuing decades. During this time several commonly referenced descriptions for CT emerged:

- "Active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it, and the further conclusions to which it tends… upon firm basis of evidence and rationality" (Dewey, 1910, p.6).
- A process guided by: (1) an attitude of being disposed to consider in a thoughtful way the problems and subjects which come within the range of one's experiences; (2) knowledge of the methods of logical enquiry and reasoning; and (3) some skill in applying those methods[2] (Glaser, 1941).
- "The skill and propensity to engage in an activity with reflective skepticism" (McPeck, 1981, p. 7).
- "Reasonable, reflective thinking that is focused on deciding what to believe or do" (Ennis, 1987, p.10).
- "Skillful, responsible thinking that facilitates good judgment because 1. it relies upon criteria, 2. it is self-correcting, and 3. it is sensitive to context" (Lipman, 1988, p.39).
- "Purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual,

[2] Glaser is amongst the first to popularise the term 'critical thinking', as previously Dewey was defining the term 'reflective thought'.

methodological, criteriological, or contextual considerations upon which that judgment is based" (Facione, 1990, p. 2).

- Halpern (1998) identified components such as determining causal relationships, including isolated data into larger frameworks, recognising and criticising assumptions, assessing components of likelihood and uncertainty, supporting conclusions, and solving problems by using analogies.
- Pascarella and Terenzini (2005) took on the task of summarising the most prevalent definitions and determined that CT skills refer to one's ability to:

> identify central issues and assumptions in an argument, recognize important relationships, make correct inferences from data, deduce conclusions from information or data provided, interpret whether conclusions are warranted based on given data, evaluate evidence or authority, make self-corrections, and solve problems. (as cited in Tiruneh et al., 2014, p. 2)

Ab Kadir (2007) has argued that "while the various theories of critical thinking certainly differ with regard to some key points, they also reveal some common emphases" (p.5) and contends that merely because a precise definition for CT may remain elusive does not mean that the concepts, dispositions and skills associated with the multiple definitions cannot be understood and employed in a coherently meaningful way. Indeed, Ab Kadir asserts that restricting ourselves to a singular conception of CT which discounts other understandings largely overlooks the fluid nature of CT wherein meaning largely depends upon specified context and purpose. In this way, the "resemblance inherent in the competing definitions of critical thinking… are indeed generalizable at the level of context and purpose" (p. 6).

The current examination necessitates establishing that a working definition and conceptual understanding of CT fits within the qualities being measured by the context and purpose of the study. While the task instruments will be explored in more detail in the methodology section (4), the title of the original SHEG study, 'Can students evaluate online sources? Learning from assessments of civic online reasoning' reveals alignment to conceptions of CT with two keywords: evaluate and reasoning.

As an example, the first task developed by the SHEG for the high school level asks students to determine which of two competing arguments posted in a discussion section of an article

on a major news website is stronger and to explain their position. Those who perform at the highest level would identify the stronger argument by thoroughly and accurately explaining how it used evidence from the article while the weaker argument used irrelevant evidence in support of its position. The assessment goals align with understandings of CT from Dewey's (1910) first contemporary conception as providing evidential and reasoned judgments to Glaser (1941) interpreting and evaluating arguments to the expert consensus definition generated for the Delphi Report (Facione, 1990) and nearly every proceeding definition and conceptual understanding to the present day.

All the SHEG tasks are designed to assess students' abilities to evaluate the reliability of sources of information and consider the role of evidence in arguments within online contexts, with the driving categorical questions of 'Who is behind the information?' and 'What is the evidence?' well within the conceptual and defining properties of CT.

2.2.1 Ennis' instructional approaches for critical thinking

A categorisation scheme for approaches to teaching CT which emerged from the literature review serves as a framework for this study. Developed by Ennis (1989) and explored, defined and utilised recently by Tiruneh et al. (2014) in their meta-analysis of CT intervention studies at the higher education level, the *general* approach is defined as one which teaches CT separately from the content of other subjects. The driving assumption for this approach is that the skills developed through explicit CT instruction will naturally transfer for use in other disciplines. The *infusion* approach integrates CT instruction with standard subject matter instruction while making general principles of CT explicit to learners. In this approach, students are encouraged to "acquire and explicitly practice CT skills through deep and well-structured subject matter instruction" (p. 2). The *immersion* approach also integrates CT skills within standard subject matter instruction, but general CT principles and procedures are not made explicit. The assumption with this approach is that learners will naturally acquire the thinking skills from engaging in CT-oriented subject instruction without studying it separately. The *mixed* approach consists of a combination of the general approach with either infusion or immersion. In a mixed approach "there is a separate thread or course aimed at teaching general principles of CT, but students are also involved in subject-specific CT instruction where either the objectives of CT are explicit or implicit" (p. 3).

A natural question which occurs is to consider which approaches are more measurably effective than others when compared, an enquiry which does not yet yield a concrete consensus conclusion. The more answerable quantification is determining which approaches are most commonly utilised. Systematic reviews of CT instructional approaches (Abrami et al., 2008; Behar-Horenstein & Niu, 2011; Tiruneh et al., 2014) reveal that over three-fourths of intervention studies utilise either the immersion or infusion approach. Despite their popularity, Tiruneh et al. (2014) found students' abilities to transfer acquired CT skills from subject-matter instruction to new tasks equally ineffective via the immersion or infusion approaches. The general or mixed approaches were found to more effectively enhance CT skills development, though the researchers caution that the smaller number of studies utilising these approaches may limit their generalisability.

Abrami et al. (2008) reached a different conclusion by finding the immersion approach less effective than the infusion approach, and indeed the variable upon which there is more agreement is whether the CT skills being developed are made explicit (Ennis, 1989; Mayer, 1992, McPeck, 1990; Nickerson, 1988; Resnick, 1987), such as would be found in the general, infusion, and mixed approaches. The available evidence suggests that "direct teaching strategies, which are based on explicit and detailed explanation of CT principles, are more effective than the implicit teaching strategies" (Tiruneh et al., 2014, p. 8). The evidence on the overall effectiveness of implicitly embedding CT skills such as asking higher-order questions, concept mapping, and facilitation of small group discussion within subject matter instruction without any explicit instruction in CT development remains inconsistent and inconclusive (pp. 5-6).

The classification for approaches to teaching CT can be graphed as follows:

CT Approach	**Description**	**Condition**
General	CT skills facilited directly in separate course	Explicit
Infusion	CT skills facilitated directly in subject courses	Explicit
Immersion	CT skills facilitated indirectly in subject courses	Implicit

Mixed	General + Infusion General + Immersion	Explicit + Explicit Explicit + Implicit

Table 1. Ennis' (1989) classification for CT with explicit and/or implicit conditions.

2.2.2 Critical thinking transfer

The central role of transfer with CT skills facilitation appears throughout its varying definitions and is particularly well highlighted in Tiruneh et al. (2014), where they state:

> CT instruction is mainly based on the assumption that there are identifiable and definable thinking skills which are domain-independent and can be taught to students to recognize and apply them appropriately in daily life situations and future careers. The goal of CT instruction is, therefore, to help students acquire and transfer those domain-independent thinking skills to solve problems faced in everyday life (p. 3)

The precedents for this determination can be found in Glaser (1984), Ennis (1989), Halpern (1988)—whose research into teaching CT for transfer across domains provided focus on "training in the structural aspects of problems and arguments to promote transcontextual transfer of critical-thinking skills" (p.449)—Nickerson (1988), and Perkins & Salomon's 'Teaching for transfer' (1988). These scholars contend that the issue of transfer must be addressed whether CT is taught using a general or discipline-embedded approach.

Nickerson (1988) noted that teaching CT separate from content runs the risk that learners gain some comprehension of CT principles but may "fail to connect that knowledge to the many situations in life in which it could be useful," and that a risk of teaching the same aspect of thinking only within the context of a standard course is that the student will "fail to abstract from the situation what is really context independent and again will not transfer what has been learned to other contexts" (p. 34).

This study shares the assumption that CT development revolves around the issue of transfer and aligns with the leading research by focusing on the extent to which skills facilitated in CT-oriented curricula effectively transfer toward evaluating information and media literacy tasks based on situations faced in daily interactions with social media and online news.

2.2.3 Critical thinking efficacy

A central driving question to this research asks: ‘How do we know the explicit facilitation of CT skills is actually developing the desired CT skills in learners?’ The exploration applies the criteria and approach of CT to the teaching and learning of CT, asking fundamental questions about the ways in which the skill sets theoretically developed through CT instruction translate into real world CT outcomes. From a working understanding of ‘efficacy’ to mean “the ability to produce a desired or intended result,” (Simpson & Weiner, 2018) the term ‘critical thinking efficacy’ seems to best encapsulate the overriding concept under investigation. While the meta-analyses of various intervention studies in CT and recent scholarship on media literacy reveal that an effort into CT efficacy has begun, searches in both EBSCO the International Education Research Database for the phrase revealed no relevant returns with the words in that order with that specified or intended meaning. This may be indicative of the recentness with which CT efficacy has become an active concept in academia. For a working definition, ‘CT efficacy’ denotes consideration of the extent to which the facilitation of CT skills produces the desired or intended results of its efforts. The definition is broader than the scope of this research, as it does not limit the term to the explicit facilitation of CT by external actors and can include ‘self-taught’ CT skills development. The former remains the more specified subject of CT efficacy under present investigation.

2.3 Information and media literacy

The materials utilised as measurement tools for this study consider the development of CT skills which fall under the canopy of ‘information literacy’ generally, ‘media literary’ more specifically, and even more specifically within digital contexts representative of young peoples’ daily interaction with social media and online news.

The SHEG prefers the term ‘civic online reasoning’ to avoid ambiguity and to retain specificity, as explained below. Indicative of the timeliness of the topic of this study to the advent of ‘fake news’ permeating the news cycle, during the writing of this paper a publication titled ‘The promises, challenges, and futures of media literacy’ by the Data & Society Research Institute offer an evaluation of media literacy efforts to date with contextualisation to the current media landscape. The authors state that:

> In general, there is a lack of comprehensive evaluation data of media literacy efforts. Some research shows that media literacy efforts can have little-to-no impact for certain materials, or even produce harmful conditions of overconfidence. The longitudinal nature of both assessing and updating media literacy programs makes this a perennial struggle. (Bulger & Davidson, 2018, p. 3-4)

The authors begin their exploration by stating that media literacy is conceived "as a process or set of skills based on critical thinking" (p.3) and directly tie their working definition of 'media literacy' to a "skill set that promotes critical engagement with messages produced by the media" (p.4). For a general working definition, media literacy is the "active inquiry and critical thinking about the messages we receive and create" (Hobbs & Jensen, 2009) and most proponents of media literacy emphasise this connection to CT.

The authors quote SHEG researchers Wineburg and McGrew, who comment on the dangers inherent to the Internet opening "the floodgates to misinformation, fake news, and rank propaganda masquerading as dispassionate analysis" (as cited in Bulger & Davison, 2018, p. 5). The report concludes with a series of open questions, some of which overlap with the research questions and implications of this study. Questions such as 'Can media literacy even be successful in preparing citizens to deal with 'fake news' and information?' and 'Are traditional media literacy practices (e.g., verification and fact-checking) impractical in everyday media consumption?' (p. 21) show congruent enquiry into CT efficacy.

Civic online reasoning

The SHEG use the term 'civic online reasoning' to differentiate the set of practices developed through their assessment tasks from broader understandings of media literacy which might include learning competencies from learning how to type to advanced programming and coding. To the SHEG, 'civic online reasoning' is "a more narrowly focused term to describe how to evaluate and use online information to make decisions about social and political matters than the larger field of media literacy" (McGrew, Breakstone, Ortega, Smith & Wineburg, 2018, p. 5) such as those found in descriptions from the National Association for Media Literacy Education (2007) and the National Council for the Social Studies (2016).

For the purposes of this investigation, given the established understandings of media literacy to include development in CT skill sets, the terms ‘civic online reasoning’ and information, media or digital literacy are transposable.

2.4 Curricula under consideration

Given that three-fourths of the student participants in this study predominantly attended the Finnish education system prior to entering the IBDP programme (see 5.2), both the Finnish National Core Curricula (NCC) and the IBDP are examined in this section. Due to the comparative value of the original cohorts utilised in the SHEG study, the U.S. curriculum for the state of California guiding those student participants is also explored and analysed.

2.4.1 International Baccalaureate

The IB is a non-profit foundation which provides educational curricula for students from age 3 to 19. It is comprised of four programmes which share a “focus on teaching students to think critically and independently, and how to inquire with care and logic” (‘About the IB’, 2018). Both the area of enquiry for this study and broader issues of CT transfer and lifelong learning are addressed on the IB’s homepage, where it describes itself as designed to prepare students for “a world where facts and fiction merge in the news, and where asking the right questions is a crucial skill that will allow them to flourish long after they’ve left our programmes” (2018).

Founded fifty years ago in Geneva, Switzerland, the IB was born out of UNESCO (Maurette, 1948) in 1968 with the pre-university diploma programme for students aged 16 to 19, expanded in the 1990s to include the middle and primary years programmes, and most recently added vocational training for high school students who are not planning to pursue academic higher education. The coverage if the IB is vast: at least one of these programmes is administered in almost 5,000 schools in over 150 countries (International Baccalaureate Organization [IBO], 2018), which to date has resulted in 1.5 million graduates from around the world (IBO, 2018).

The IB undergoes consistent revision and often makes substantial changes to its curricula. Course guides are published with a ‘first assessment’ date two years after the date of

publication and are typically revised every half-decade to decade depending on the course. Since the revision years for courses are staggered across the curricula, this results in some guides adhering to recent changes while others remain several years from implementing new approaches and structures. The general trend over the last decade has been the development and strengthening of 'Approaches to teaching and learning' (ATL) and a stronger integration of Theory of Knowledge (TOK)—the IBDP's course explicitly covering CT skills—into subject course designs.

These top-down changes are evidenced in the IBDP course guides. A major change during the previous revision cycle was to explicitly discuss how TOK relates to the course area subject and to provide a list of TOK guiding questions specific to the academic discipline in each course guide. Now a decade later, the most recent revisions in 2016 (for assessments beginning in 2018) reveal course guides containing an updated curriculum model, a new section on ATL as it relates to the specific course, and TOK being more explicitly integrated into course instructional and assessment structures for each subject in the programme.

Approaches to teaching and learning

The IB's ATL are "deliberate strategies, skills and attitudes that permeate the IB teaching and learning environment" (IBO, 2013, p. 1), including skills development in engaging students as enquirers and thinkers (p. 2). Given the IB's extensive network, determining how these strategies are utilised in practice becomes challenging. Implementation of ATL requires some level of oversight by the IB regarding teacher performance which strives to balance quality assurance with teacher autonomy in course design and delivery.

As the IB continues to grow exponentially around the world, the last half-decade has seen the introduction of ATL and more explicit support for skills in teaching students not only what to learn, but also how to learn (p. 1). ATL establishes the IB's approaches to learning skills as: 1) thinking 2) communications 3) social 4) self-management and 5) research. The approaches to teaching skills are: 1) based on enquiry; 2) focused on conceptual understanding; 3) developed in local and global contexts; 4) focused on effective teamwork and collaboration; 5) differentiated to meet the needs of all learners; and 6) informed by formative and summative assessment (p. 1).

Diploma Programme

The DP is an intensive, two-year programme designed for high school grades 11 and 12 that includes "a strong emphasis on encouraging students to develop intercultural understanding, open-mindedness, and the attitudes necessary for them to respect and evaluate a range of points of view" (p. 2). There are six groups from which IBDP students select their coursework, taken at either a standard or higher level: 1) Studies in language and literature; 2) Language acquisition; 3) Individuals and societies; 4) Sciences; 5) Mathematics; and 6) Arts. Groups have several courses available to choose from, but many schools are only resourced enough to offer a few, whereas better resourced schools may offer more. Taking a course from the arts is optional and some schools may not offer any subjects from this group; rather, candidates may 'double up' with a course selection from Groups 1 – 5.

Courses are graded on a 1 – 7 scale and candidates must earn at least 24 points to receive the diploma. Up to three additional points are made available through successful completion of the core subjects (TOK, Extended Essay, and Creativity, Activity, Service), meaning that the highest score an IBDP graduate can achieve is 45 points.[3] Assessments are evaluated both externally by experts in the academic field of study and internally by the classroom teacher. They are criterion-based with the criteria by which they are assessed published in the course guides and made available to students throughout the duration of their studies.

The current graphical depiction of the IBDP presents its curricular approach through a series of concentric circles and corresponding text. The outer circle is the IBDP and the overriding concept of 'international-mindedness'. The next concentric circle holds the six subject area groups. The next circle represents core requirements: TOK, the extended essay and creativity, activity, service. TOK is described in more detail below; the extended essay is a 4000-word research paper on a topic of the student's choosing and submitted within a specified course area to demonstrate competency in the subject's methodological practices; creativity, activity, service requires involvement in extracurricular activities related to artful expression, physical activity, and meaningful volunteer work. The next concentric circle represents the newly added ATL, with the innermost circles representing the individual learner.

[3] This is achieved by only a few hundred, or .25% of eligible IBDP candidates, on average per year.

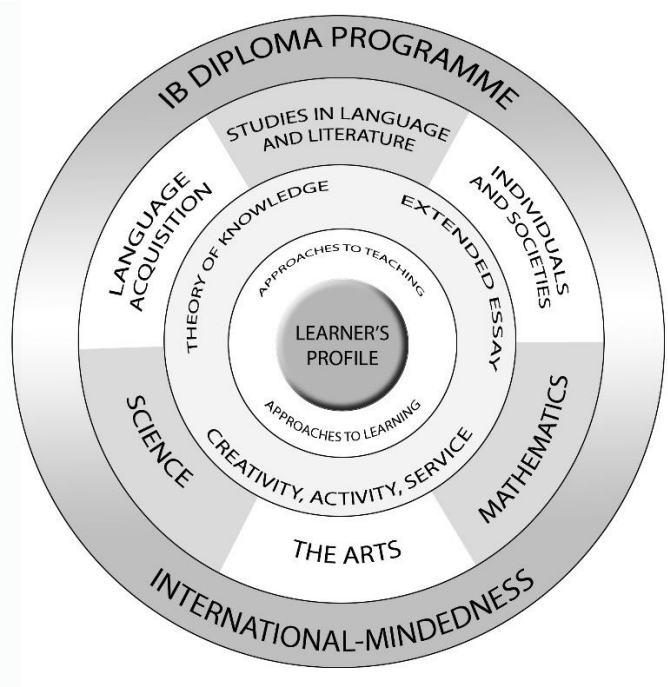


Figure 1. IB Diploma Programme design.

Theory of Knowledge

The IB describes TOK as an explicit course in CT skills development which invites learners to enquire into the process of how knowledge is generated and received rather than to acquire a particular body of knowledge. It is a uniquely adapted, skills-based introductory course in epistemology that prioritises the individual learner's voice. The fundamental question explored in TOK is 'How do I/we/they know that?' which is explored through a framework of knowledge questions and claims generated by 'areas of knowledge' (AOK) through 'ways of knowing' (WOK) and the resulting process of interplay between them (see Figure 2). In TOK, students learn how to critically examine external knowledge claims while developing skills in how to generate, justify, and defend their own knowledge questions and claims.

Figure 2. IB Theory of Knowledge design.

The two-year TOK course has two assessments: a presentation wherein students select a real-world situation from which they recognise and extract a knowledge claim for consideration, develop the knowledge claim by broadening its implications, then apply the knowledge claim to other real-world situations. This internal assessment is evaluated by the students' TOK teacher and accounts for one-third of their TOK score. The remaining score results from students responding to a choice of one from six 'prescribed titles' in their graduating year, during which time they have one month to write a 1600-word essay response. This external assessment is evaluated by examiners hired by the IBDP with no affiliation to the school.

The are several elements of TOK course design which address developing skill sets of CT transfer. For example, TOK encourages students to make connections across all the subject areas in their DP coursework while actively reflecting on the transfer of CT skills developed in TOK to these subject areas. A recent study in Australia (Cole, Gannon, Ullman & Rooney, 2015) had current DP students at differing levels of study rate their CT skills. Students indicated gains in the use of CT skills between the two successive years of the DP, with second-year DP students reporting a significantly greater likelihood of using a range of CT skills in their daily interactions.

Perceived TOK outcome	Mean score (SD)	Comparison of means (Year 11-12)
Critical-thinking skills	Year 11: 4.23 (*1.25*)	$t = 3.79$ $p < .001$
	Year 12: 4.50 (*1.27*)	

Table 2. Mean differences in students' perceived outcomes of TOK study (Cole, Gannon, Ullman & Rooney, 2015).

Examples of instructional material for TOK which offer preparation in the specific skill sets covered in the SHEG assessment tasks can be found in a commercially published course companion, which poses a series of critical questions for students to ask regarding the credibility of various sources of information including those found online and the relevance of any statistics, maps, graphs, or photographs utilised as evidence (Dombrowski, Rotenberg, & Bick, 2013, pp. 219-220). Additional questions for digital media, developed from Dombrowski's 'TOK and "fake news": 3 tips, 2 downloads, and 3 resources,' was published as a supplement in 2017. This revision includes the consideration of sponsored content amongst other skills (sec. 2, para. 5) specific to digital media literacy covered in the SHEG assessment tasks replicated in Finland.

Mixed infusion approach

In consideration of Ennis' (1989) classification for CT instructional approaches, the IBDP's overall design is mixed by combining the general with infusion approaches. The TOK course is specifically designed to facilitate CT skills development which are taught explicitly and as a separate subject from the other academic disciplines (general approach). The subject guides for each course within the IBDP curriculum include connections to TOK which are made to be taught and assessed explicitly by subject area teachers (infusion approach).

While this determination applies to the curricular approach, it is plausible that a randomly selected IBDP subject area classroom anywhere in the world could be seen following the immersion approach in practice; for example, if an IBDP subject teacher does not explicitly cover TOK in the course instruction. This was possible and likely more common under previous course guides, but the recent revisions to IBDP structure have integrated TOK into the content and assessment structures to the extent that the current IBDP curricula treats all

subject teachers as also being teachers of TOK, thereby aligning it with the mixed infusion approach in terms of its curricular design, structure, and intended implementation.

2.4.2 Finnish education system

Although the intended focus of this examination was to consider the IBDP's approach to facilitating CT skills development, the role of the Finnish education system as a variable of interest became apparent upon collection and analysation of the demographic data (5.2). While the role of educational background in the study's participants will be examined in more detail in the methods (4) and results (5) sections, the exploration necessitates providing some foundational information on the Finnish education system and approach to better understand its potentially causal relationship to the measured outcomes of the task materials.

It is common in popular literature to find the Finnish education system lauded as one of the world's best ('Finland overview', 2018; Jackson, 2016; Williams-Grut, 2016), with the PISA, or Programme for International Student Assessment (OECD, 2018), showing Finland consistently performing amongst the top countries. Transitioning from an inequitable system with unexceptional outcomes, Finland underwent major reform around three decades ago with a shift to an overriding focus on equality and has been branded in terms of its perceived quality since. Finland has been the focus of increased international interest by educationalists during the past two decades to explore how it has achieved measurable success while simultaneously rejecting global education reform trends such as core subject study, competition, standardisation, test-based accountability and control (Anderson, 2011).

Some features of the Finnish education system which are often highlighted by way of distinction (and possible contribution toward its evidential success) are the relatively late age at which students begin formal study; the emphasis on play, creativity, and joy of learning; comparatively longer times spent in outdoor activities; the rarity of formal measurements such as exams or standardised testing; a de-emphasis on homework; the absence of tracking or segregating students by ability or other factours; the small gap between stronger and weaker performing students; strength of vocational schooling; strong teacher training and preparation; highly competitive teacher education programmes; high level of teacher autonomy with little concept of 'accountability'; a relatively low amount of teacher 'contact time'; time set aside for teachers' professional development; near total lack of private or for-

profit schooling; and a largely decentralised approach with the national curriculum providing broad guidelines for implementation (Taylor, 2011).

Compulsory schooling begins at the age of seven and lasts for nine years. It is provided in a single structure system called basic education which includes grades 1 – 9. Education is free for students as are learning materials, meals, health and welfare services and transport from home to school. At age 16 students can choose to stop attending school or pursue one of two upper secondary institutes: vocational education which prepares students for work, or upper secondary school which prepares students for university. The curricula guiding both basic education and upper secondary education is determined by the Finnish National Agency for Education. Brief descriptions of the curricula are explored below with an analysis of which category for approaches to CT instruction they follow.

National Core Curricula

According to the European Commission (2018), the NCC for compulsory basic education (NCCBE) and general upper secondary education (NCCGUSE) includes the objectives and core contents of different subjects. The NCCBE was revised at the end of 2014, with schools implementing the new curricula in August 2016 for grades 1 – 6 and by 2019 for grades 7 – 9. Key aims of the reform include 'transversal competencies' such as 'thinking and learning to learn' with multiliteracy and ICT skills.

The syllabus for the NCCBE includes the following subjects which all students study: Mother tongue and literature (Finnish or Swedish); other national language, foreign languages, environmental studies, health education, religion or ethics, history, social studies, mathematics, physics, chemistry, biology, geography, physical education, music, visual arts, craft, and home economics. The syllabus for the NCCGUSE includes similar coursework to the NCCBE, with the addition of psychology, philosophy, thematic studies, and 'worldview' added to the religion/culture and ethics course. The NCCGUSE requires a minimum of 75 study courses divided into compulsory, specialisation or applied subjects. Students take around 50 compulsory courses and at least 10 specialisation courses for advanced study in courses of their choosing. The applied courses follow an interdisciplinary approach by exploring a problem through various subject methodologies.

These subject areas are surrounded by cross-curricular themes, such as technology and society "to enhance the student's understanding of the interaction between technology and societal development" (European Commission, 2018, Sec. 3, para. 5). Teachers are given autonomy over methods and materials and "are recommended to guide students to use new technologies such as digital learning materials and environments" (Sec. 4, para. 1). There are 17 upper secondary schools in Finland which offer instruction leading to the IB diploma.

Mixed immersion approach

Regarding Ennis' (1989) classification for CT instructional approaches, the Finnish education system becomes more difficult to categorise due to the high level of autonomy provided to both teachers and local school implementation. The NCC makes clear that CT skills are at least implicit across certain subject areas—the specific term 'critical thinking' is found twenty times embedded in over ten different courses in the NCCBE (2014) and fourteen times embedded in over ten different courses in the NCCGUSE (2015)—but the level of teacher autonomy and lack of standardised assessments make it challenging to ascertain the extent to which individual teachers explicitly develop CT skills in learners.

In outlining its educational mission and goals, the NCCBE lists seven transversal competences—defined as entities "consisting of knowledge, skills, values, attitudes and will" (p. 36)—which guide the curriculum. Included in these are: 'Thinking and learning to learn' (T1); 'Managing daily life' (T3); and 'Multiteracy' (T4) which is defined as one's abilities to "obtain, combine, modify, produce, present and evaluate information in different modes, in different contexts and situations, and by using various tools. Multiliteracy supports the development of critical thinking and learning skills" (p. 39).

In a segment of the NCCBE titled 'The conception of learning' (sec. 2.3) which parallels much of the philosophies guiding the IB curriculum is the following reference to CT:

> While acquiring new knowledge and skills, the pupils learn to reflect on their learning, experiences and emotions… [Learning] involves doing things alone and together, thinking, planning and exploring, and assessing these processes in a versatile manner… [This] promotes the pupils' skills in creative and critical thinking and problem-solving and their ability to understand different viewpoints… (p. 27)

This passage provides further parallels to the metacognitive practices of the IB:

> Learning is diverse and connected to the content to be learnt, time and place. Developing the learning-to-learn skills lays the foundation for goal-oriented and lifelong learning. The pupils are thus guided in becoming aware of their personal ways of learning and using this knowledge to promote their own learning. Pupils who are aware of and responsible for their learning processes will increasingly learn self-regulation. During the learning process, they learn working and thinking skills and practise anticipating and planning the various stages of learning. (p. 27)

The compulsory courses in the NCC in worldview and ethics and the philosophy course at the upper secondary level support "development of [student] capacity for thinking… critically as well as their learning-to-learn skills" (p. 249). The first course in ethics is called 'Worldview and critical thinking' and is designed to analyse and evaluate how worldviews are formed, to critically examine beliefs both internally and externally, and to critically examine information provided by different media. The course aims for a student to "utilise what he or she has learned in different subjects in shaping his or her worldview as well as to distinguish between scientific, non-scientific, and unscientific perceptions of the world" (p. 261). In this course, CT is understood as "a self-correcting activity that seeks reason and perceives connections and is sensitive to different situations" (p. 249). It encourages an open-minded and reflective attitude, as per denotative qualities of CT skills development.

The compulsory philosophy course in the NCCGUSE is designed for students to rationally consider arguments given to justify views. Amongst its objectives are to identify problems and provide alternative solutions in relation to current issues; distinguish conceptually, analyse and evaluate information, particularly differing statements, their meanings, and justifications; master the basic skills of consistent argumentation; rely on one's own thinking; critically evaluate one's own thinking; and reflect on limitations in different fields of science and daily life, including "when dealing with unreliable and conflicting information" (p. 216). The course is assessed according to one's ability "to critically analyse and problematise information, perceive and specify its conceptual structure, and present a justified judgement on the topic" (p. 217). It supports and develops students' abilities to assess their thinking and

encourages them to plan and develop their learning, as per denotative qualities of CT skills development.

In addition to these being courses designed to explicitly facilitate CT skills development, direct attention to the skill sets measured on the SHEG tasks are also evident in the worldview course where students examine daily information processing and the impact of media, politics, science, and the entertainment industry. In ethics, students are instructed in "the operating logic of the media and other sources of information, critical thinking related to worldviews aimed at these as well as analysis of argumentation related to worldviews from different viewpoints" (p. 261). Lastly, "due to its nature that involves asking questions and looking for justifications," the philosophy course "helps in perceiving and structuring the constantly increasing information overflow of today" (p. 216).

In consideration of this overview, Finland's curricular structure follows the mixed immersion approach. CT skills are explicitly taught and assessed in the compulsory courses of ethics, worldview, and philosophy (general approach) with evidence of CT implicitly embedded throughout the curricula (immersion approach). There remains the possibility of meeting criteria for the mixed infusion approach at the classroom level, dependent upon the individual teacher who in Finland retains autonomy over instructional approaches to learning.

2.4.3 U.S. education system

Comparative studies looking at the 'U.S. education system' are confronted with the problem that curricula are decentralised to the state level with no national curriculum. While there is some legislative oversight from the national government which has increased in the past decades to include the development of the Common Core State Standards' (CCSS), it is left to each state to decide whether to implement national standards. The state of California where the final SHEG materials were administered selected to follow the CCSS and utilises the standards in conjunction with curricular frameworks to guide their state's education system.

Since an implication of this study is to encourage more caution in education science toward overgeneralising implications of study results (see 6), the researcher would normally caution against generalising implications from the SHEG study beyond coverage of the curriculum to the state level. However, given the scope of the prototyping phase for the materials which

covered thousands of participants over eight different states with outcomes which did not significantly differ from the finalised tasks (McGrew, et al., 2018), this offers more generalisability toward the U.S. public education system overall, though this remains limited by the tasks in the wider administrations still being in prototype phase. The demographics and sample size by which the original SHEG study can be generalised from administration of the finalised tasks are included in the methods (4) and results (5) sections.

Common Core State Standards and curricular frameworks

In 2010, states across the U.S. were offered the option to implement national standards in English and math for K-12 public education systems. The CCSS described for the first time what students should know and do at each grade level across the country so that there would be state-by-state standardisation should a student change schools or move to a different state. The State Board of Education for California adapted the CCSS to cover language arts and literacy in history/social studies, science, and technical subjects and developed additional state standards for English language development, career technical education, computer science, health education, history/social science, model school library, physical education, science, the visual and performing arts, and world languages. In 2016, California developed curricular frameworks to guide implementation of the standards. These frameworks cover content and pedagogical approaches, assessment, access and equity, and skills development for learning in the 21st century for language arts, English language development, science, mathematics, history/social science, health, physical education, foreign language, and the visual and performing arts.

Immersion approach

The term 'critical thinking' does not appear in the CCSS, nor is there a separate course within the curricula which aims to facilitate CT skill sets in isolation of the subject areas, but there is evidence of CT skills development being implicitly embedded into the course areas. In the college and career readiness anchor standards for reading, one finds a CCSS standard wherein students are asked to "delineate and evaluate the argument and specific claims in a text, including the validity of the reasoning as well as the relevance and sufficiency of the evidence" and in the reading standards for informational texts, to assess "whether reasoning is sound and the evidence is relevant and sufficient; recognize when irrelevant evidence is

introduced" (California Department of Education [CADOE], 2013, p. 10). These extracts align with the denotative qualities of CT skills development and exhibit relevance to the skill sets measured by the SHEG assessment tasks.

The most explicit mention of CT skills development is in the frameworks, where the term 'critical thinking' is found 30 times in the history/social science framework and 20 times in the English language arts / English language development framework. The latter has a full chapter dedicated to 'learning in the 21st century' that includes a section titled 'critical thinking skills' and further sections dedicated to fostering global awareness, digital citizenship, and understanding multimedia text (CADOE, 2015, p. 937). It is unclear how these skills are intended to be integrated into the curriculum while absent from the CCSS.

With CT skills coverage embedded into the curricular framework and no evidence of explicit CT instruction provided, California is most appropriately considered as following the immersion approach with the possibility that if subject teachers cover CT skills explicitly in their coursework then it would follow an infusion approach for those individual classrooms. In contrast to both the Finnish education system and the IBDP, there are no separate courses wherein CT skills are clearly and explicitly facilitated as per the general approach.

2.5 Applying Ennis' classification for CT approaches to curricula

When the curricula under consideration for this study are analysed to determine which of Ennis' (1989) approaches to CT instruction apply, the CCSS emerges with CT skills being immersed into the subject areas implicitly, although an individual subject teacher could choose to cover CT skills explicitly as per the infusion approach. The NCC follows the mixed immersion model by facilitating CT skills separately through ethics, worldview and philosophy and by implicitly embedding CT skills into courses; although, again, a subject teacher could individually infuse CT skills explicitly. The IBDP, meanwhile, is mixed infusion due to its approach of emerging its separate CT course into all the subject areas explicitly. This overview is synthesised in Table 3.

CT Approach	Condition	Curriculum	Teacher Autonomy
General	Explicit	-	-
Infusion	Explicit	-	CCSS

Immersion	Implicit	CCSS	NCC
Mixed	Explicit + Explicit	IBDP	NCC
	Explicit + Implicit	NCC	-

Table 3. Ennis' (1989) classification for CT applied to curricula.

Since the categories established by Ennis (1989) are exclusive—with 'mixed' by definition requiring the general approach with either the infusion or immersion approaches—it is of further conceptual assistance to divide the 'general' approach, where a separate course in CT is employed, from an 'embedded' approach where CT skills are integrated into the course subjects per the infusion and immersion (and by extension, mixed) approaches.

Table 4 provides a clearer division between the IBDP and NCC, both of which follow the general approach by explicitly teaching CT as an independent course, from the CCSS, which follows an immersion only approach by implicitly embedding CT into the subject areas. As evidenced in the current literature, the teaching of CT explicitly has been observed as producing stronger outcomes to implicitly embedding CT, which is at best inconclusive in its relationship to positive CT learning outcomes.

CT Separate	**CT Embedded**	**CT Approach**	**Curriculum**
Yes	None	General only	-
Yes	Explicit	Mixed Infusion	IBDP
Yes	Implicit	Mixed Immersion	NCC
No	Explicit	Infusion only	-
No	Implicit	Immersion only	CCSS
No	None	None	-

Table 4. Ennis' (1989) classification with CT taught separately or embedded.

Research Questions

The central question guiding this research asks:

- To what extent are CT skills developed in the classroom transferring to external contexts such as everyday interactions with social media and online news?

That is a more specific question to broader considerations regarding CT efficacy, such as:

- How do we know the facilitation of CT skills is producing its intended result?

This broader question influenced the study design toward comparing two cohorts with differing variables of age, grade level, and curricula exposure to measure potential variation in the performance outcomes as well as influencing the analysis of data derived from the differing curricula guiding the original participants from those utilised in the replication.

Directly replicating task materials from an existing study to determine the extent to which the results may be generalised to other socioeducational contexts led to the generation of additional research questions:

- Would students from differing socioeducational backgrounds reveal a variation in performance outcomes on a replication of the same assessment tasks?
- What relationship can be determined between curricula which explicitly facilitates CT skills development with their relative performance outcomes on the assessments?
- What can be generalised from the overall results and what are their implications?

Methodology

Task assessments used to measure the extent to which students exhibit media literacy skills in online contexts were requested from the SHEG, which provided access to them for replication. While the tasks and results were made available to the public at the time of the SHEG study's publication (McGrew et al., 2018), when the measurement in Finland was conducted the SHEG had publicly released Task 4 from the high school level with limited information on their results in an executive summary (Wineburg, S., McGrew, S., Breakstone, J., & Ortega, 2016). This was expressly published in some media, but the remaining four tasks at the high school level were kept confidential. In addition to providing access to their tasks and evaluation rubrics, the SHEG provided additional information on their results under the condition that the data not be publicly shared until publication of their January 2018 journal article (J. Breakstone, personal communication, November 2, 2017).

4.1 Replication

In their systematic review of intervention studies on CT instruction, Tiruneh et al. (2014) conclude that evaluation of the effectiveness of CT instruction could be influenced by the type of CT measures employed in a study. For example, two infusion approach studies by Anderson, Howe, Soden, Halliday, & Low (2001) and Bensely & Haynes (1995) reported differing outcomes when they utilised the same teaching strategies and research design but differed in CT measurements. The authors noted that some variations on CT outcomes could be explained by the multiple-choice format of the standardised CT measure being utilised. In a study by Plath, English, Connors, & Beveridge (1999) in which two CT measures were utilised together, significant CT improvement was revealed on the measure that required students to respond to open-ended items rather than in a multiple-choice format (Tiruneh et al., p. 8).

In addition to employing a standardised assessment with open-ended items, the current study reduces CT measurement type as a confounding variable by replicating the assessment tasks from a prior study. An inspiration for the overall design and approach for this study came

from the meta-analysis, 'Facts are more important than novelty: Replication in the education sciences' (Makel & Plucker, 2014) which analysed the top 100 peer-reviewed education journals and found that only .13% of articles were replicated studies. While most replications revealed results supportive of the original studies, they were less likely to replicate successfully when authorship differed between the original and replicating articles (2014). The authors argue for the importance of third-party, direct replications for education research to improve its ability to shape policies and practices.

This study is not a direct third-party replication as per Makel & Plucker's suggestion as the materials were administered on a different study group but does directly replicate the original materials in similar fashion to the original study. The benefit of replicating the same materials in differing socioeducational contexts and curricula is that it helps researchers determine the strength of generalisability of the implications from the original study. If a replicated study reveals similar results, the generalisability of the original study is strengthened, whereas if a replicated study differs in its results then confounding variables present a threat to overall generalisability. Since the confounding variables introduced in this replication are the age and grade level of one group entering the IBDP in Finland and another exiting the DP and the socioeducational curricula amongst the U.S. cohort and the IBDP school in Finland, these are the variables of interest for further consideration of the results.

4.1.1 Stanford History Education Group task development

The assessment tasks were developed by the SHEG over three phases of an 18-month period which covered 12 states and led to the collection of 7,804 responses at the middle school, high school, and college levels. Sites for field-testing included under-resourced, inner-city schools in Los Angeles to well-resourced schools in the suburbs of Minneapolis. Five assessments were finalised for each level, with computer-based testing at the college level and paper-and-pencil tasks covering digital content at the middle and high school levels. While the paper-and-pencil nature of the assessments at the high school level make for a more feasible replication, the SHEG addresses the issue of using paper-and-pencil tasks to measure digital literacies with an OECD study (2015) establishing that important abilities for judging online sources can be effectively measured offline (Wineburg et al., p. 6).

The final high school level assessment tasks were administered to 348 students in participant groups of between 170 – 176 across three districts in California. Each district had diverse populations with a free and reduced student lunch rate of 36%, 55%, and 68%. Students were given 30 minutes to complete packets of three tasks which were randomly divided so that half of the students in each class completed one packet of three tasks while the other half completed a packet of three different tasks (McGrew et al., 2018, p. 8).

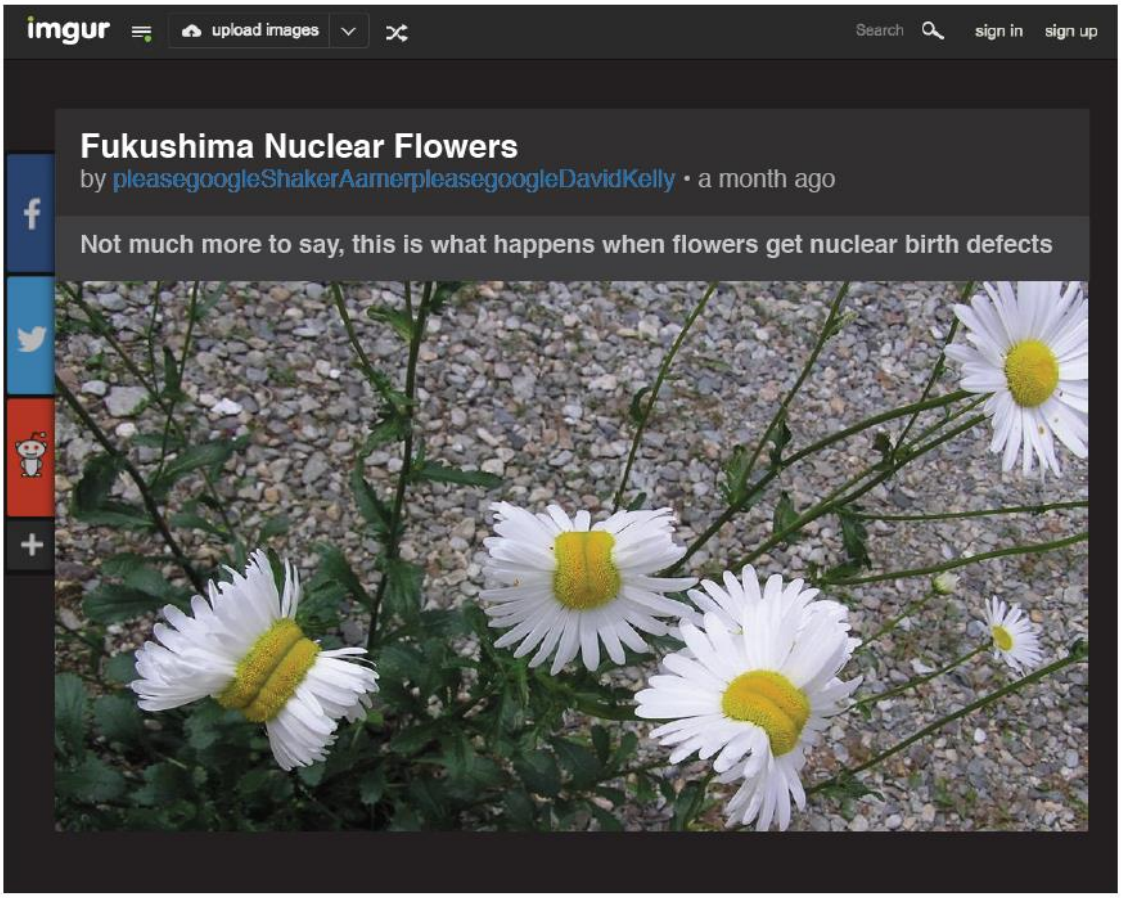
The following are the final task topics with brief descriptions (p. 6) and task sample sizes as provided by the SHEG (J. Breakstone, personal communication, April 19, 2018):

Task	Topic	Description	SHEG Sample Size
1	Argument analysis	Students read two comments in response to a news article and explain which commenter makes a stronger argument.	172
2	News on Facebook	Students explain which of two news posts (one from a verified account, one not) is a better source.	173
3	Facebook argument	Students explain which poster in a Facebook conversation provides stronger evidence about gun laws.	175
4	Evaluating evidence	Students evaluate the strength of evidence in a photograph posted on Imgur.	170
5	Comparing articles	Students explain which of two sources (one sponsored content, one traditional news) is a more reliable source about climate change.	176

Table 5. SHEG high school assessment task descriptions with sample size.

The five finalised tasks with corresponding evaluation rubrics for the high school level are found in Appendix A. As an example, the fourth task, which was released in the SHEG's executive summary (Wineburg et al., 2016) prior to their study's publication (McGrew et al., 2018) and explored in more detail in their study, is:

On March 11, 2011, there was a large nuclear disaster at the Fukushima Daiichi Nuclear Power Plant in Japan. This image was posted on Imgur, a photo sharing website, in July 2015.



Does this post provide strong evidence about the conditions near the Fukushima Daiichi Power Plant? Explain your reasoning.

Stanford History Education Group

Task 4. Evaluating evidence (Wineburg et al., 2016, p. 16).

Assessment rubrics were created by the SHEG with categories determining performance levels of beginning, emerging, or mastery. Descriptors are included at each level for an assessor to identify, evaluate, and categorise student responses.

EVALUATING EVIDENCE RUBRIC

MASTERY	Student argues the post does not provide strong evidence and questions the source of the post (e.g., we don't know anything about the author of the post) and/or the source of the photograph (e.g., we don't know where the photo was taken).
EMERGING	Student argues that the post does not provide strong evidence, but the explanation does not consider the source of the post or the source of the photograph, *or* the explanation is incomplete.
BEGINNING	Student argues that the post provides strong evidence or uses incorrect or incoherent reasoning.

Task 4. Evaluating evidence rubric with descriptors for mastery, emerging, and beginning performance levels (Wineburg et al., 2016, p. 18).

Sample responses with a discussion on the elements which placed the response at the corresponding level are also provided for each performance outcome level:

SAMPLE RESPONSES

MASTERY

This student questions the source of the photo, arguing that there is no way to know whether the photo was actually taken near the plant or if the mutations were a result of nuclear radiation.

No, it does not provide strong evidence about the conditions near the Fukushima Daiichi power plant. It does not provide strong evidence because it could just be a mutation in the plant. There also isn't evidence that this is near the Fukushima Daiichi power plant.

This student questions the source of the post, arguing that we know nothing about the poster's credentials or whether the evidence was doctored.

No, it does not really provide strong evidence. A photo posted by a stranger online has little credibility. This photo could very easily be Photoshopped or stolen from another completely different source; we have no idea given this information, which makes it an unreliable source.

EMERGING

This student begins to question both the photo and the source of the post but does not fully explain his thinking.

This post does not provide strong evidence about conditions near the power plant. They just put a picture of a flower. Plus the poster has a strange username.

This student critiques the evidence by arguing that it could have been digitally altered but does not offer any further explanation or critique of the evidence.

No, because this picture could be Photoshopped.

BEGINNING

This student accepts the evidence at face value, arguing that it provides visual proof of the effects of the nuclear disaster.

This post does provide strong evidence because it shows how the small and beautiful things were affected greatly, that they look and grow completely different than they are supposed to. Additionally, it suggests what such a disaster could do to humans.

Although this student argues that the post does not provide strong evidence, she still accepts the photo as evidence and simply wants more evidence about other damage caused by the radiation.

No, this photo does not provide strong evidence because it only shows a small portion of the damage and effects caused by the nuclear disaster.

Task 4. Evaluating evidence sample responses with discussion (Wineburg et al., 2016, p. 18-19).

That this task was made publicly available and was reported on in media prior to replication posed a threat since participants could have been exposed to it prior to task administration. This was considered in the pre-survey by asking students to list any previous awareness of the tasks or materials and through direct in-person enquiry to the participant groups asking whether they had seen, read, or heard of these tasks beforehand.

4.1.2. Validity

In the prototyping phase, the SHEG researchers began by administering 56 tasks in a product design method that sought user testing for revision and improvement. In the validation phase, extensive piloting and qualitative data through 'think aloud' interviews were collected by hundreds of participants to establish cognitive validity, defined as "the relationship between what an assessment seeks to measure and what it actually does" (Wineburg et al., 2016, p. 5). During the final phase of field testing, thousands of responses were collected along with teacher consultations until 15 assessments were finalised, with the SHEG concluding: "Together with the findings from the cognitive validity interviews, we are confident that our assessments reflect key competencies that students should possess" (p. 5).

4.1.3 Reliability

An inter-rater familiar with the testing materials was trained on the evaluation criteria and evaluated a 20% sample distributed evenly across the five tasks from both cohorts. Inter-rater agreement was 86% for the pre-IB cohort (Cohen's $\kappa = 0.77$) and 84% for the IB2 cohort (Cohen's $\kappa = 0.76$). This is lower than the 97% inter-rater agreement (Cohen's $\kappa = 0.92$) reported by the SHEG but still within a level indicating a strength of agreement and a high probability of returning similar results from other external raters.

A principal component analysis was conducted to explore the extent to which the five tasks may be measuring distinct or similar concepts. According to the SHEG's study, tasks were developed around three conceptual questions intended to measure: 1) What is the evidence? 2) Who is behind the information? and 3) What do other sources say? (p. 10 – 17). The question 'What do other sources say?' only applies to the college level and is not utilised within the middle or high school tasks. At the high school level, three of the tasks—Tasks 1, 3, and 4—are designed to answer, 'What is the evidence?'; Task 2 is designed to answer,

‘Who is behind the information?’; and Task 5 is designed to answer both ‘What is the evidence?’ and ‘Who is behind the information?’. This is represented by Table 6.

What is the evidence?	Who is behind the information?	Both
Task 1: Argument analysis Task 3: Facebook argument Task 4: Evaluating evidence	Task 5: Comparing articles	Task 2: News on Facebook

Table 6. Conceptual questions for measurement framework aligned with tasks.

The principal component analysis on these five tasks resulted in the following component matrix and plot in rotated space:

Component Matrix[a]

	Component	
	1	2
Task 3 FB argument	.789	
Task 4 Evaluating evidence	.626	.460
Task 1 Argument analysis	.475	
Task 5 Comparing articles		.728
Task 2 News on FB	.575	-.598

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Table 7: Principle component analysis of SHEG tasks matrix.

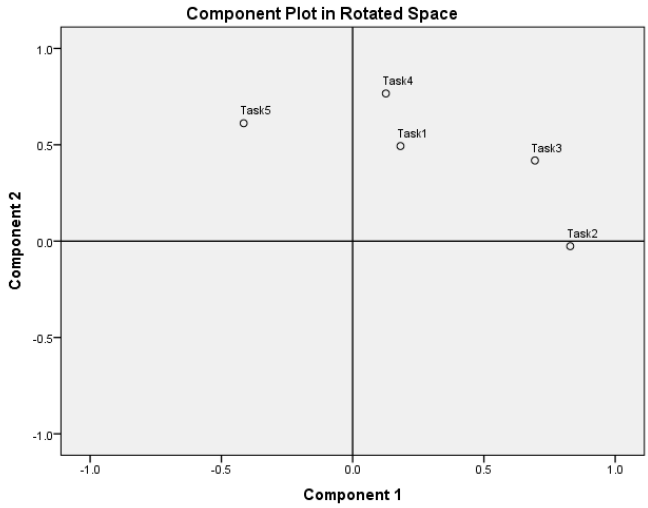


Table 8: Component plot in rotated space for the SHEG tasks.

For the component matrix, the division between two components resulted in Task 5 with .728 on component 2 and no result for component 1. As Task 5 was designed to ask, 'Who is behind the information?' then component 2 should theoretically align with this conceptual question. Task 3 resulting in .789 on component 1 and no result for component 2 and Task 1 resulting in .475 in component 1 and no result for component 2 support the notion that these tasks may be asking a different question, so that component 1 is likely 'What is the evidence?'. Task 4, however, which was intended to measure 'What is the evidence?' only, resulted in scores of .626 in component 1 and .460 in component 2. Task 2, designed to ask both conceptual questions, resulted in .575 on component 1 and -.598 on component 2.

The component plot in rotated space indicates that Tasks 1, 3, and 4 somewhat mix with the other questions on two dimensions; meanwhile, Task 2 and Task 5 appear rather distinct, suggesting they are on different dimensions as per the SHEG design. Task 2 was designed to ask both conceptual questions represented in the other tasks, which should indicate overlap with Task 5, yet Task 2 graphs close to zero on one plane and the strongest on another, indicating that it may be primarily asking one question over the other.

4.2 Task threats

Tasks were administered in a trial session amongst graduate students in education science to test for clarity of understanding, cultural bias of materials, and the timing necessary for task administration. While the participants in the trial were older and at a higher level of education than the study participants, they shared characteristics of being from diverse backgrounds by way of nationality, culture, and language. Of the five researchers who participated in the trial session, three came from languages other than English and all came from differing sociocultural backgrounds. During these sessions, surveys were developed to collect demographic data from participants and to provide control for potential threats to the research (Appendix C).

Replicating the tasks in a differing sociocultural environment introduces the possibility of a language barrier for students who are not native English speakers. In the case of the school in Finland utilised for testing, the students were studying in an English-medium curriculum and are assessed in English in all but their native language literature course. Further

considering the overriding U.S.-centricity of the materials is the possibility that participants may not have enough cultural-specific background knowledge to accurately reveal CT skills of analysis and evaluation. The difficulty inherent to measuring thinking skills rather than domain-specific knowledge requires a delineation of when tasks require knowledge[4] rather than thinking. Each task was analysed for potential threats with these considerations in mind.

Task 1: Argument analysis

This task presents contrasting arguments posted in the comments section of a major news website responding to an article about a social media application that allows users within a small radius to read and post comments anonymously. As described in the task, some schools have had problems with students using the application to harass peers. In the first argument the context of the school being American is mentioned, and in the opposing argument Winston Churchill is quoted as an authority figure to support the position. That the schools under discussion are context-specific to the U.S. does not appear pertinent to understanding the overall argument, as the scenario could otherwise apply elsewhere. Given that the quote is provided by the weaker argument in a manner that is irrelevant and off-topic, it becomes less relevant that a respondent knows and recognises Winston Churchill to make the correct determination; his title of 'prime minister' is provided to ensure understanding that, at least, the subject of the quote is one of authority. The presentation of arguments is otherwise universal and does not appear dependent upon cultural knowledge for effective evaluation.

Task 2: News on Facebook

This task contains two posts from accounts with identical logos claiming to be Fox News, both of which are announcing Donald Trump's candidacy for president of the U.S. Post A by 'Fox News' was posted on June 16, has a blue checkmark next to its name, was edited, and includes an image of Donald Trump. Post B by 'Fox News The FB Page' was posted on August 28, does not have a blue checkmark, was not edited, and includes an image of a tweet by 'Donald J. Trump @realDonaldTrump' which reads: "Today I officially declared my

[4] While the term 'knowledge' can be as multi-faceted and difficult to define as CT, here it is used in the meaning developed by Bloom's classification (1956) as indicating the recall of factual information, which is distinct from the thinking skills utilised in cognitive operations for analysis and evaluation.

candidacy for President of the United States. I will also be producing a new reality show for NBC. The 'Presidential Apprentice' to premier this fall! youtu.be/q_q61B-DyPk."

The task runs the risk of decay by being administered nearly a year after the election which Donald Trump won, where participants could draw from their *post hoc* knowledge that no such reality show has since taken form to guide them toward the correct determination. While the topic being centred around the U.S. presidential election as covered through a post on social media by a U.S. news outlet runs the risk of cultural bias, U.S. politics and news media remain influential throughout Europe and indeed much of the world. None of the participants in the trial reported lack of clarity due to U.S.-centricity, nor is an in-depth understanding of U.S. politics necessary to be successful on the task since the main skill sets being assessed are disassociated with those elements.

The task according to the SHEG rubric measures knowledge over thinking skills with regard to the blue verification checkmark to indicate account verification from social media sites such as Facebook, Twitter, Instagram, and Youtube which distinguish genuine accounts—particularly those of celebrities and companies—from unofficial accounts. It also offers other elements which remain specific to the context of universal media literacy which are less to do with the U.S.-centric context of the task. For example, considering the names of the pages 'Fox News' versus 'Fox News the FB Page,' and that many students in the SHEG study were swayed by the verified account being published at an earlier date or being edited as reasons for finding it untrustworthy, which are irrelevant indicators toward determining authenticity in the scenario materials.

Task 3: Facebook argument

This task contains competing arguments in the form of comments posted on Facebook about gun control in the U.S. The first comment says, "Another mass shooting in America - we need to pass strict gun laws like Australia has. According to this, Australia hasn't had a single mass shooting since they tightened their gun laws in 1996" and includes a link to a news article from the *The New York Times* titled 'How a Conservative-Led Australia Ended Mass Killings'. The article itself is not accessible in the task. Another user responds, "Actually, tighter gun laws won't stop mass shootings" with an accompanying graph, 'Guide to Gun Control', by the Minnesota Gun Owners political action committee. The graph shows various

proposed gun laws with markings to indicate that none of them would have prevented a series of mass shootings in the U.S.

The SHEG rubric for this task requires students at the mastery level to explain why evidence presented by a gun owners' political action committee is weaker than an article by a prestigious news organisation. As such, the task does reward knowledge of *The New York Times* as a credible and indeed prestigious source of news. Knowledge of what a political action committee does is less important than understanding that any organisation exclusive for gun owners would be more likely to contain bias against stricter gun control laws.

There remain other skills by which students could evaluate this task. For example, by observing that the stronger argument conflates the correlation of Australia's gun laws to its reduction in mass killings with definitive causation, and (interestingly, in relation to the implications of the present study) that what occurred in one sociocultural environment would naturally replicate to another. The second argument also suffers from being hypothetical in an *ex post facto* reasoning by stating that something would not have occurred even if certain laws were in place, whereas the first argument considers the strength of correlation on events which indeed occurred.

While such a response would be more based on reasoning skills and an arguably more sophisticated consideration than merely knowing and accepting the credibility of *The New York Times* over a meme designed by a special interest group, the SHEG rubric is specific about assessing a correct response which considers these elements of support in lieu of considering the source of information at the emerging rather than mastery level. A specific example with a student response is provided:

EMERGING

This student focuses on the qualities of the evidence that Anya and Grace provide—Anya's evidence is based on past events while Grace's evidence is hypothetical—and does not critique the sources of the information.

I feel like Anya had stronger evidence because she stated a fact about Australia that she can researcn while Grace practically just stated that gun laws won't stop shootings and made a graph without facts.

Task 4. Evaluating evidence extract from the SHEG rubric showing an example response at the emerging level.

It is perhaps indicative of the SHEG being dedicated to history education that the assessment criteria prioritise approaches more common to an historian, such as focusing on sources of the information. While this researcher may disagree with marking a response which indicates thinking skills over knowledge on tasks considered to measure 'civic online reasoning', for assessing the tasks the level descriptors as developed by the SHEG are utilised in the replication to ensure comparability with the U.S. results. Overall, there remain context clues beyond knowledge of *The New York Times* as a reputable news organisation to make the correct determination, and a reasonable expectation that the average Finnish student would be as aware or unaware of *The New York Times* as a high school student in the U.S.

Task 4: Evaluating evidence

This task presents a picture posted on an image sharing website which depicts mutated flowers. A text written by SHEG above the post states: "On March 11, 2011, there was a large nuclear disaster at the Fukushima Daiichi Nuclear Power Plant in Japan. This image was posted on Imgur, a photo sharing website, in July 2015." The post is titled 'Fukushima Nuclear Flowers' by username 'pleasegoogleShakerAamerpleasegoogleDavidKelly' with the comment, "Not much more to say, this is what happens when flowers get nuclear birth defects." Students are asked to assess whether the post provides strong evidence about the conditions near the Fukushima Daiichi Power Plant with an explanation of their reasoning.

Students performing at the mastery level would consider that one does not know anything about the author of the post and/or that one does not know where the photo was taken. If a student argues that the post does not provide strong evidence but with an incomplete explanation that does not consider the source of the post or photograph or the explanation is incomplete, then it would be assessed at the emerging level. Students arguing that the post provides strong evidence would be performing at the beginning level.

Although the background information provided on the task is brief, it provides enough explanation so that reliance on historical knowledge of the event is not necessary for successful evaluation of the task. It is possible that students could come to the correct determination by relying on existing background knowledge of mutation effects, as well as by demonstrating other thinking skills by considering logical problems with the post's source of evidence; for example, if a student considers that the sample seen in the photograph may

not represent the whole (there are two flowers included in the picture which do not show signs of mutation effects) or that causation is not clearly demonstrated since such mutations can be a naturally occurring phenomenon. Following the SHEG rubric, these responses should be assessed at the emerging level if they do not consider the author of the post or the location where the photo was taken.

Task 5: Comparing articles

This task contains the headlines and accompanying visuals from articles posted on the news site *The Atlantic*. Both articles are about policies to solve global climate change. Article A is titled, 'Why Solving Climate Change Will Be Like Mobilizing for War' with a subtitle, 'And even then, victory is far from guaranteed'. It has a painting of the historical 'Uncle Sam' character[5] looking determined with a wrench in one hand and rolling up his sleeve with the other. In the background are trees and windmills. The article was written by Venkatesh Rao and was published in the science section of *The Atlantic*. Article B is titled 'The Great Transition' with a subtitle 'Saving the world from climate change is all about altering the energy mix, bringing renewable sources online fast enough to keep up with the demand of a growing global population.' It contains a graph of renewable sources of energy—coal, liquids, natural gas, renewables, and nuclear—and a statement that "Today's ideas will be critical for the larger, energy-hungry world of tomorrow." In the background is a drawing of windmills amongst hills and trees near a river, and on the other side of the river industrial buildings, housing and cars. In the top left corner is the logo for the Shell Oil Company with the text: "Sponsor Content: What's This?" and unlike in Article A the URL for the site, https://www.theatlantic.com/sponsored/2015-shell/the-great-transition/595/, is included.

The task is intended to measure a student's ability to identify that an article posted in the science section of a reputable news site is more reliable as a source for learning about policies to solve global climate change than an article of sponsored content by a company vested in fossil fuels. The mastery level of the rubric requires students to provide a clear rationale for why a company with a vested interest in the topic is less reliable. Students who select Article A but do not fully explain why Article B is problematic are assessed at the emerging level. Students who argue that Article B is more reliable or notice that Article B is sponsored

[5] The 'Uncle Sam' character comes from patriotic folklore and is often used to personify the U.S.

content but argue that this makes it more trustworthy are assessed at the beginning level. An example response for the beginning level reads:

> I think Article B is more reliable, because it provides the reader with evidence (quotes) from reliable sources like the *New York Magazine* and provides current facts. In Article A there aren't any quotes and although it's much longer it may not be as helpful...

What is curious about this example is that the student could only have generated this response if he or she was viewing the article digitally; the paper-and-pencil task which was used for replication in this study only included the title, subtitle, and brief explanation of the 'Future,' whereas the full article online includes the entire content of the article with attributed quotes and statistics. Given that the participants from the SHEG study must have had access to the full articles online, this creates a threat wherein they would have had access to more information and context clues—including being able to click on the Shell Oil Company icon to learn about *The Atlantic*'s policy on sponsored content—to make the correct determination than the students in Finland. This was enquired about to the SHEG director, who stated that the final test group used for their measurement were administered the same pencil-and-paper task without online access, and that the results from these administrations were no different from participants who had computer access (J. Breakstone, personal communication, 2017).

As the Shell Oil Company is represented only by its logo on the task, this requires students to know and recognise the company and its vested interests to understand why there would be a conflict of interest. Likewise, the source of the articles, *The Atlantic*, is a magazine with a long publishing history in the U.S. which produces articles of high quality; knowing that the source is reputable might guide a student toward the science article being trustworthy. Since Shell is prominent in Finland, including in the city where the tasks were replicated, there is reasonable expectation that students in the replication test group would be as aware of the oil company as students in the U.S. Similarly, and much like the other SHEG tasks for the high school level, that U.S. sources of information remain prominent in Finland provides a reasonable expectation of similar awareness of *The Atlantic* amongst the test groups.

4.3 Qualitative data

Qualitative data was collected through follow up interviews and email with the TOK teachers, the IBDP coordinator, and the principal of the school under study. Questions included how TOK is facilitated at the school, data collection to consider the comparability between groups, and the ways in which CT skills are facilitated in the Finnish curriculum.

Results

5.1 Administration of tasks

Tasks were administered during two different testing sessions over the winter of 2017 at a school in Finland connected to the researcher's university that is utilised for teacher training and academic research. The first cohort measured, the IB2 (n=25), were students in their second year and nearing completion of their IBDP studies for high school graduation. The second cohort measured, the pre-IB (n=42), were in an IB preparation programme equivalent to the 10th grade. Both administrations occurred over the course of an hour's time, with five minutes for reading the introduction, 45 minutes for task completion, and 10 to 15 minutes for reading and explaining the consent form, discussing the tasks, and answering any questions about the nature of the study. All 67 student participants provided consent for inclusion of their data. During these post-administration discussions, it was established that none of the participants had previously seen the specific tasks or read about the SHEG study.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Argument analysis is the same across categories of Condition.	Independent-Samples Mann-Whitney U Test	.356	Retain the null hypothesis.
2	The distribution of News on FB is the same across categories of Condition.	Independent-Samples Mann-Whitney U Test	.756	Retain the null hypothesis.
3	The distribution of FB argument is the same across categories of Condition.	Independent-Samples Mann-Whitney U Test	.089	Retain the null hypothesis.
4	The distribution of Evaluating evidence is the same across categories of Condition.	Independent-Samples Mann-Whitney U Test	.179	Retain the null hypothesis.
5	The distribution of Comparing articles is the same across categories of Condition.	Independent-Samples Mann-Whitney U Test	.122	Retain the null hypothesis.
6	The distribution of t1345 is the same across categories of Condition.	Independent-Samples Mann-Whitney U Test	.022	Reject the null hypothesis.
7	The distribution of t12345 is the same across categories of Condition.	Independent-Samples Mann-Whitney U Test	.046	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Table 9: Mann-Whitney U Test per task, all tasks, and all tasks except Task 2.

Table 9 shows the results for non-parametric independent samples comparison between pre-IB and IB2 students. While none of the comparisons for individual tasks revealed significant differences, comparisons between combined tasks (with p = .045, or without the somewhat oddly behaving Task 2, p = .022) did show a significant difference between the two groups. The lack of significant outcomes for the individual tasks could be explained by the sample sizes in combination with the fact that there are only three different values.

5.2 Demographic Data

The following demographic data were collected from the IBDP participants in Finland.

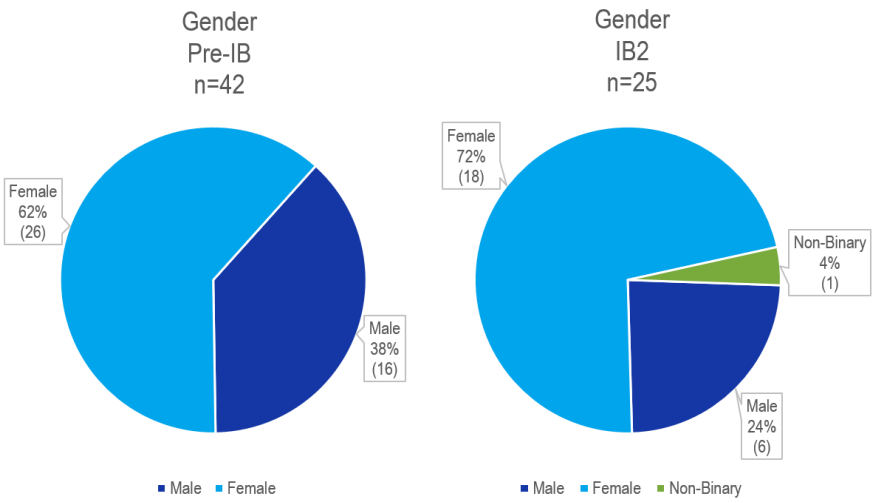


Chart 1. Participant gender distribution.

The cohorts were nearly 2/3rds female in the pre-IB and nearly 3/4ths female in the IB2.

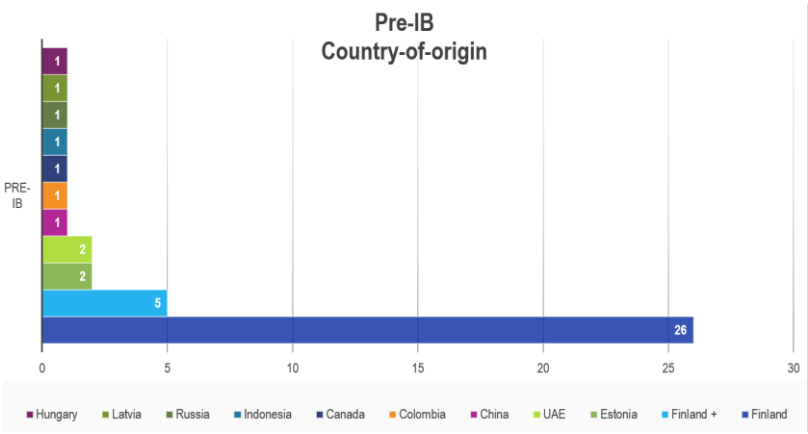


Chart 2. Country-of-origin for the pre-IB.

62% of the pre-IB indicated Finland as their country-of-origin, or 74% when combining those who responded with Finland and an additional country; the remaining 26% are spread fairly evenly amongst individual or pairs of students from nine different countries.

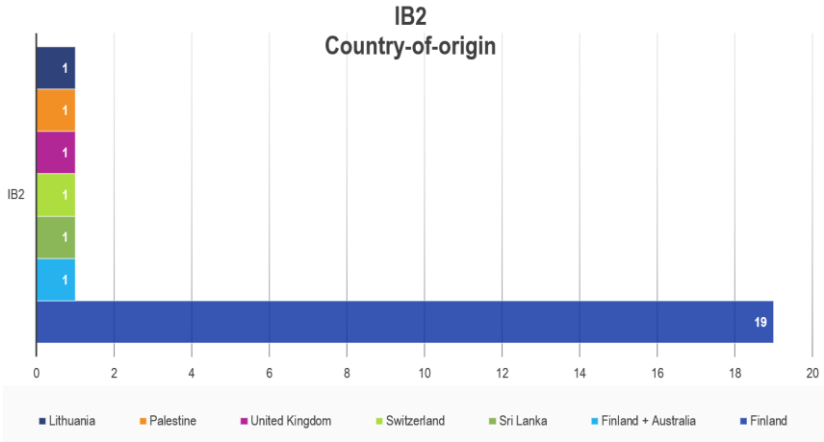


Chart 3. Country-of-origin for the IB2.

76% of the IB2 responded with Finland as their country-of-origin, with the remaining 24% spread evenly amongst individuals from six different countries.

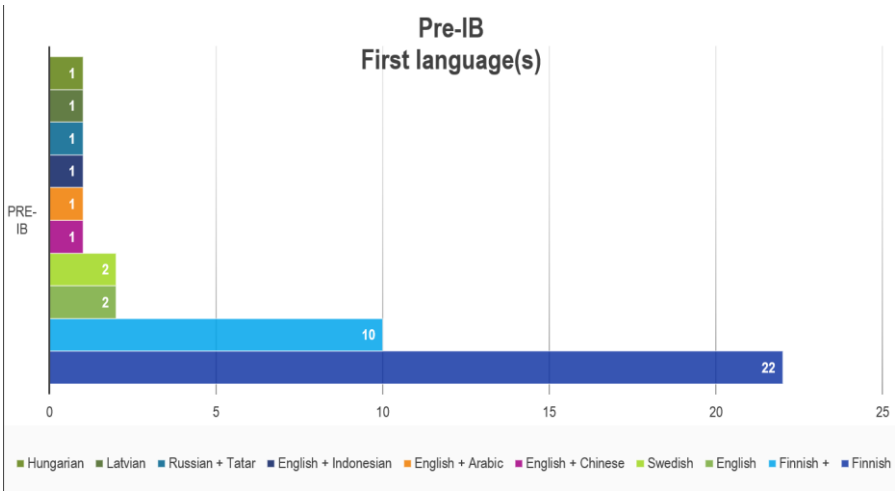


Chart 4. First language for the pre-IB.

52% of the pre-IB listed Finnish as their first language, or 76% when combining those who responded with Finnish and another language; the remaining 24% were spread amongst individual or pairs of students from ten different first language backgrounds.

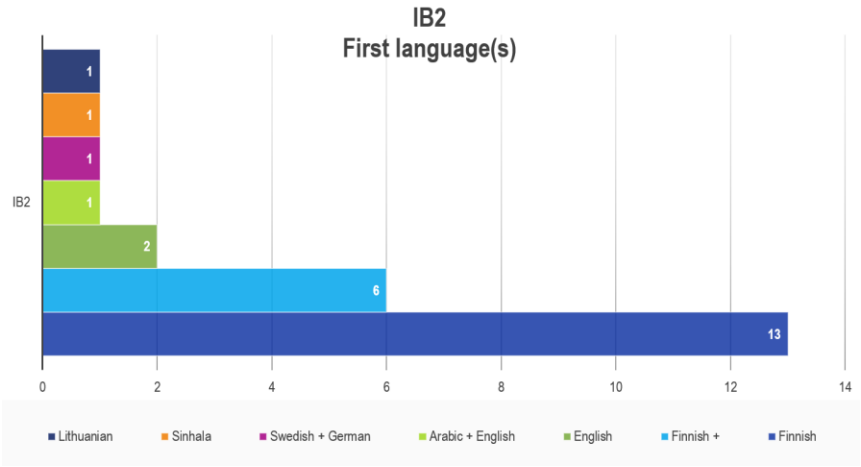


Chart 5. First language for the IB2.

52% of the IB2 listed Finnish as their first language, or 76% when combining those who responded with Finnish and another language; the remaining 24% were spread amongst individual or pairs of students from six different first language backgrounds.

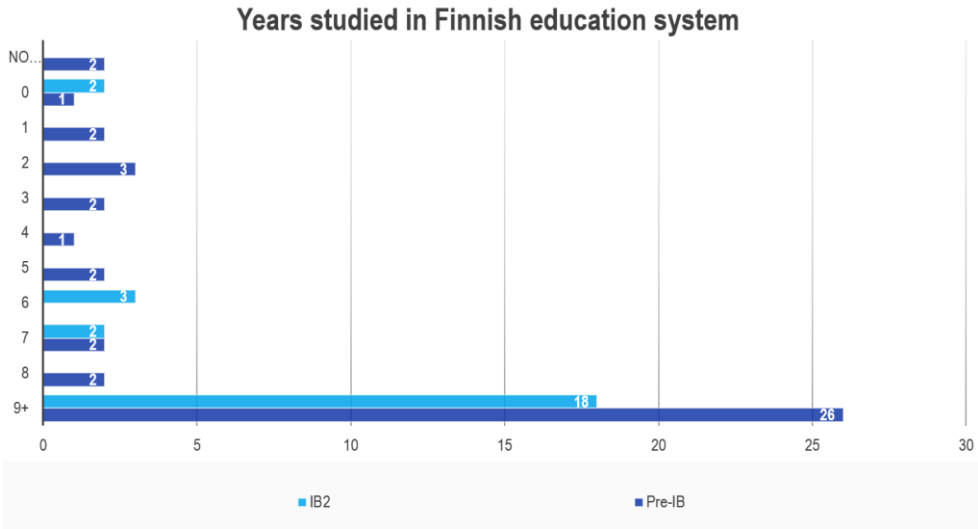


Chart 6. Years studied in the Finnish education system for the IB2 and the pre-IB.

62% of the pre-IB indicated they had their entire schooling within the Finnish education system, with 76% having experienced five or more years within the NCC. Less than 20% had one-to-four years' experience, with only one participant having no previous exposure. 72% of the IB2 indicated that their entire schooling experience was within the NCC, with 92% having experienced five or more years; the remaining 8% were two non-responses.

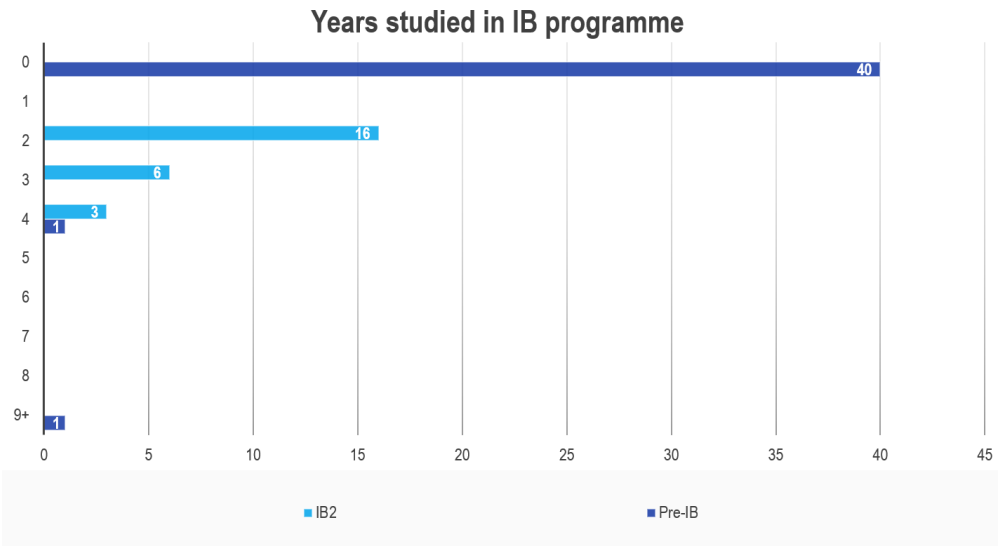


Chart 7. Years studied in the IB programme for the IB2 and pre-IB.

Over 95% of the pre-IB had no prior exposure to studying in an IB programme, with one student indicating four years' previous experience and another indicating that their entire schooling to date had been within the IB curricula. 64% of the IB2 indicated two years' experience in the IB, with an additional 24% indicating three years, within which they may have been including their pre-IB year, indicating 86% who had no previous exposure. Three participants, or 12% of the cohort, indicated four years of previous IB experience.

5.3 Post-Survey data

In the post-survey, 78% of the pre-IB cohort respondents indicated that they were aware of the 'blue checkmark' verification used on many social media networks, while 90% of the IB2 indicated such awareness. 88% of the pre-IB and 71% of the IB2 indicated that they use social media several times a day, with those numbers totaling 93% and 88% respectively of respondents who indicated using social media at least once a day. Only one respondent amongst both cohorts indicated using social media once a week or less, and none of the respondents indicated that they do not use social media at all. Respondents typically listed several sites such as Instagram, Twitter, Facebook, Reddit, and Youtube amongst the social media networks they frequent.

Fewer students from both cohorts indicated that they read or watch news media online, with almost half of the pre-IB (47%) and just over three-fifths of the IB2 (67%) reporting their frequency as at least once a day, and just over 40% of the pre-IB and 29% of the IB2 who indicate consuming news media online a few times a week. Only 6% of both cohorts indicated their frequency of online news consumption as less than once a week, with two students from the pre-IB and none from the IB2 indicating that they never consume news media online. Most students listed multiple sources for their news such as CNN, BBC, *The New York Times*, and prominent Finnish sources such as the *Ilta-sanomat* and *Iltalehti* at the national level and the *Helsinginsanomat* and *Turunsanomat* more locally, as well as Yle the public broadcasting service. Many participants indicated familiarity with the topics such as knowing Donald Trump or *The New York Times*, while none indicated awareness of the tasks.

5.4 Results on tasks

Results from the task administration allow for descriptive data of the pre-IB and IB2 viewed in isolation of other results, comparing performance outcomes of the pre-IB to the IB2, and comparing both the pre-IB and IB2 cohorts to those measured in the SHEG study.

5.4.1 Descriptive data

The overall results on each task of the pre-IB and IB2 when taken separately provide an overall 'snapshot' of descriptive data to analyse as per what the SHEG study provided.

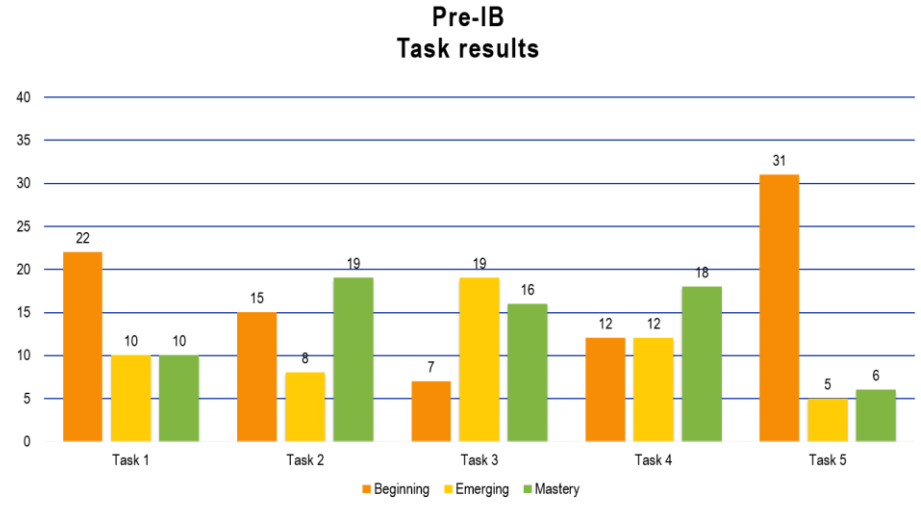


Chart 8. Results of the Pre-IB on the five SHEG high school tasks.

One indicator of student success on each task could be whether the cohort performed more at the mastery or more at the beginning/emerging levels overall. In this view, the pre-IB taken individually reveals mixed performance outcomes. The pre-IB had more participants perform at the mastery level than any other level on Task 2 and Task 4, performed highest at the emerging level on Task 3 with low outcomes at the beginning level, and appeared challenged by Task 1 and Task 5 where performance outcomes where highest at the beginning level with similarly low outcomes at the emerging and mastery levels.

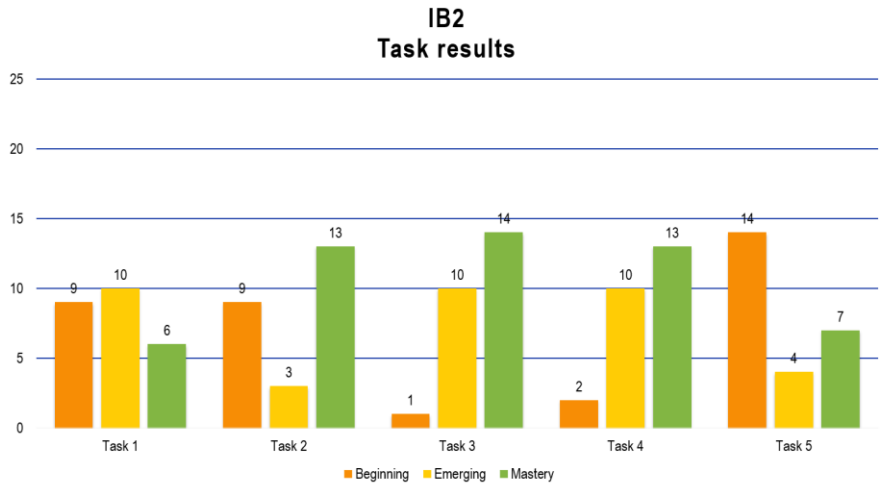


Chart 9. Results of the IB2 on the five SHEG high school tasks.

Descriptive data from the IB2 results reveal highest performance outcomes at the mastery level on Tasks 2, 3, and 4, the latter two of which have notably low performance outcomes at the beginning level. The IB2 performed highest at the emerging level on Task 1 and, like the pre-IB, were challenged by Task 5 with the highest outcomes at the beginning level.

5.4.2 Comparing outcomes of the pre-IB to the IB2

Both cohorts performed highest at mastery on Task 2 and Task 4, with the IB2 additionally performing highest at mastery on Task 3. Task 1 and Task 5 remained the most challenging tasks for both groups. To increase comparative value, assessment outcomes by the pre-IB and IB2 can be viewed as percentages to allow analysis for any differentials in performance outcome between the two cohorts on each task.

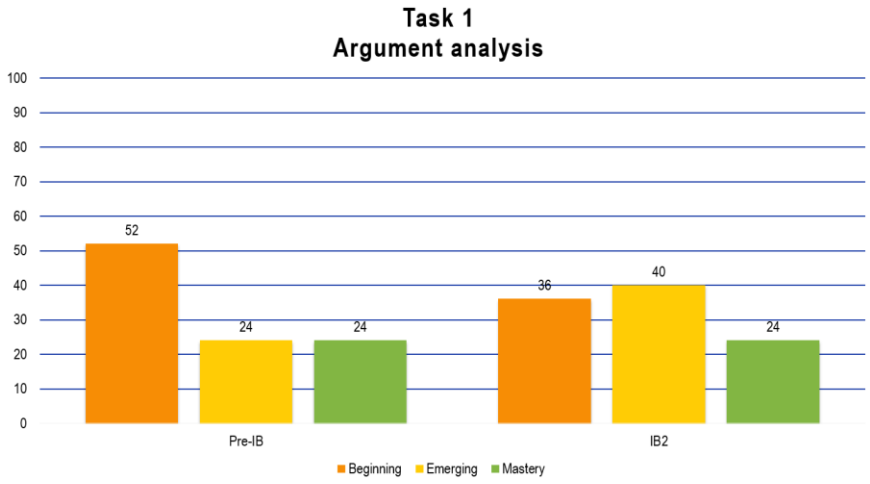


Chart 10. Results of the pre-IB and IB2 on Task 1.

The pre-IB and IB2 cohorts performed at the same percentage at the mastery level, with the IB2 performing slightly higher at emerging and lower at beginning.

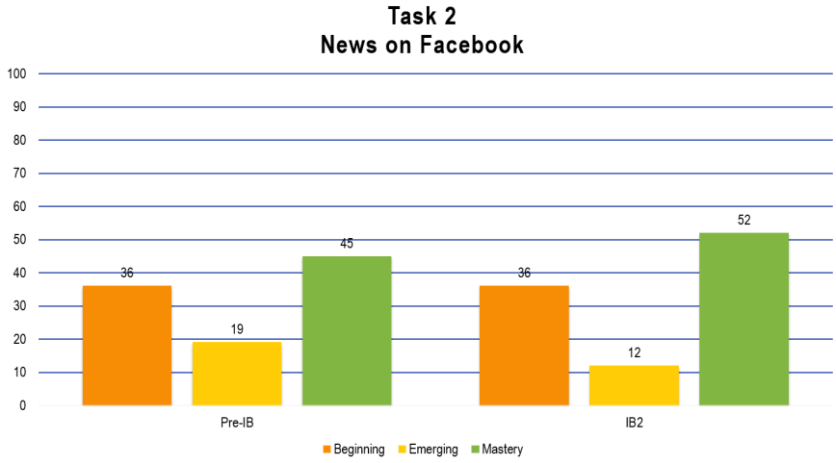


Chart 11. Results of the pre-IB and IB2 on Task 2.

The IB2 performed higher at the mastery level, lower at emerging, and the same at the beginning level compared with the pre-IB.

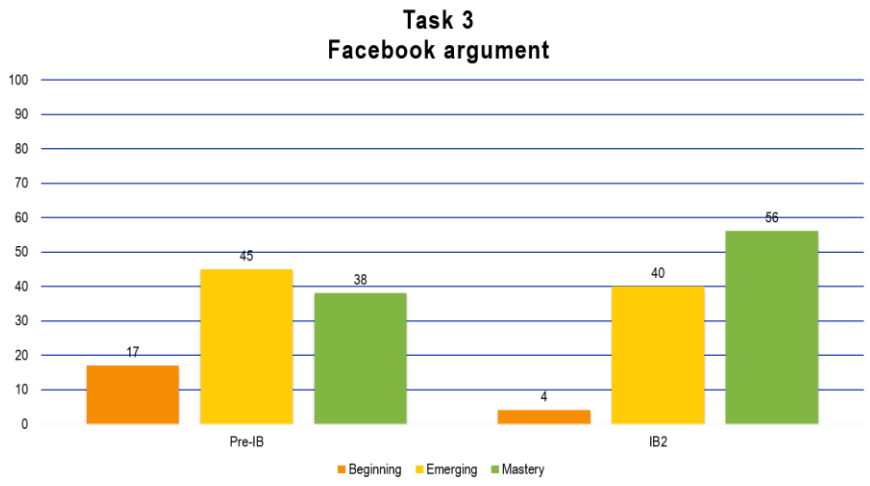


Chart 12. Results of the pre-IB and IB2 on Task 3.

The IB2 performed higher at the mastery level, lower at emerging, and lower at the beginning level than the pre-IB on Task 3.

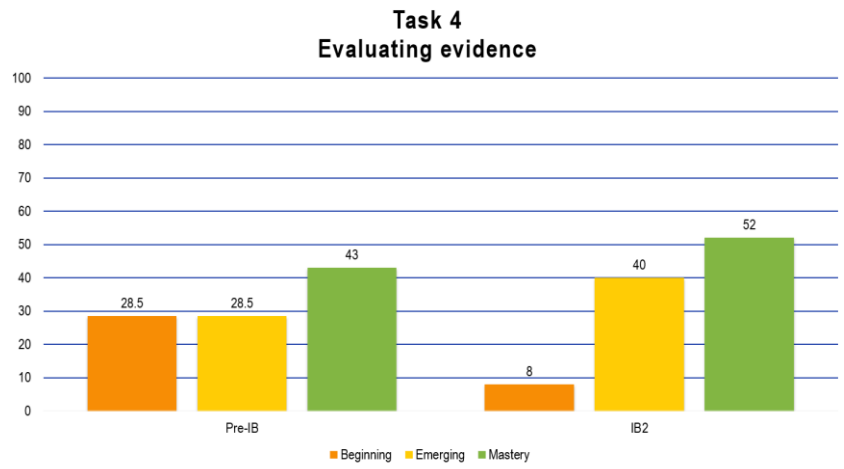


Chart 13. Results of the pre-IB and IB2 on Task 4.

The IB2 performed higher at the mastery level, higher at emerging, and lower at the beginning level than the pre-IB on Task 4.

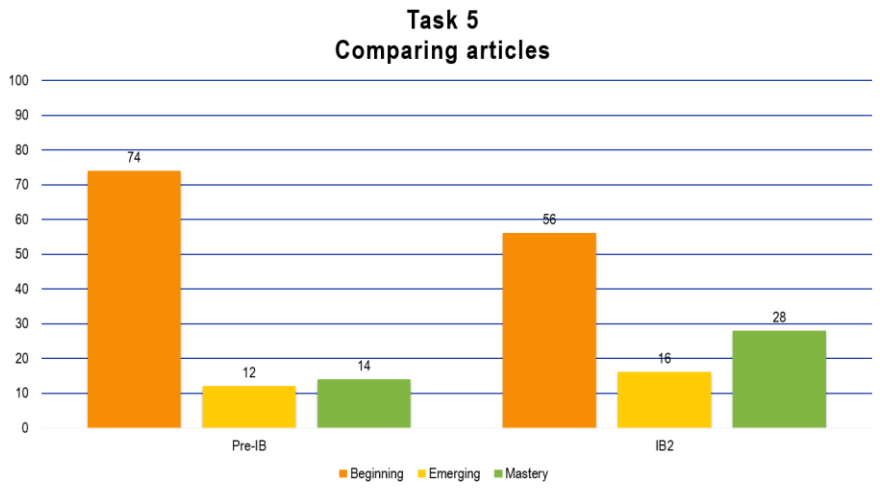


Chart 14. Results of the pre-IB and IB2 on Task 5.

The IB2 performed higher at the mastery level, higher at emerging, and lower at the beginning level than the pre-IB on Task 5.

5.4.3 Regression analysis

Apart from comparing the pre-IB and IB2 students on their task performance, it is also of interest to see whether background variables from the survey data can predict some of the students' results on the tasks. For this purpose, regression models were run to predict outcome variation from survey data collected. The results indicate that outcome variation per task is not explained by gender, language background, years attended in the IB, or frequency of both social media use and online news consumption. Prior knowledge of the blue verification checkmark found in social media appears to predict the level of success on Task 2, $F(1, 58) = 6.38$, p-value = .01, $R^2 = .09$, as would be expected. There is an indication that performance on Task 4 might be explained by country-of-origin, $F(1, 65) = 4.80$, p-value = .03, $R^2 = .07$ and years studied in the Finnish educational system may also explain performance outcomes on Task 4, $F(1, 63) = 8.14$, p-value = .01, $R^2 = .11$, although it should be considered that this variable also partially takes the two-year school difference between the pre-IB and IB2 into account.

The variables which resulted in $p < .05$ when tested against all five tasks combined were years attended in the Finnish education system, p-value = .00, and frequency of social media use, p-value = .03. Testing all eight variables against Task 2 results in $F(8, 50) = 2.22$, p-

value = .04, R^2 = .26 with statistically significant coefficients on knowledge of the blue checkmark, *p*-value = .01, and frequency of news media consumption online, *p*-value = .03.

When the pre-IB cohort is isolated as a testing condition, all eight survey variables against all five tasks result in $F(8, 30) = 3.05$, *p*-value = .01, R^2 = .45 with a statistically significant coefficient on frequency of social media use, *p*-value = .00. When the IB2 is isolated as a testing condition, none of the variables are determined as being statistically significant.

The pre-IB isolated against just Task 2 results in $F(8, 30) = 2.56$, *p*-value = .03, R^2 = .41 with statistically significant coefficients on gender, *p*-value = .03 and frequency of news media consumption online, *p*-value = .02. When the IB2 is isolated as a testing condition, none of the variables are determined as being statistically significant. Various other combinations of variables and tasks did not result in significant outcomes.

5.4.4 Comparing outcomes of the U.S. to the pre-IB and IB2

The comparative value continues to increase when the results from the SHEG study, which are expressed only in percentages, are viewed alongside the pre-IB and IB2 per task. Although for these studies the raw data is not available, some indication on the differences between groups can be obtained from χ^2 comparisons that reveal pattern differences which, in turn, can be qualitatively interpreted based on the pattern of the distributions in the different groups. Since no data were available on the individual student responses, composite score comparisons could not be conducted. Distributions of the U.S. cohorts were created based on percentages and reported sample sizes for the different tasks, where the number of task participants is multiplied by the percentage results per task and then rounded to achieve a real number to account for the differential in sample sizes between groups.

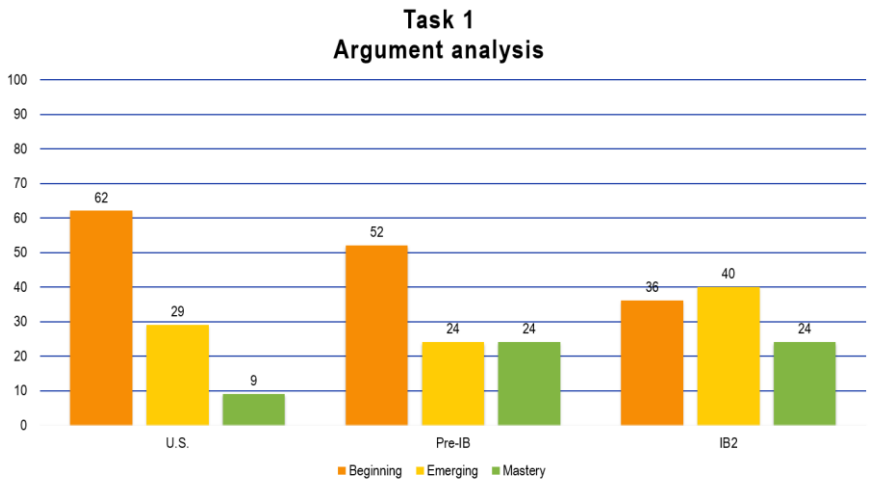


Chart 15. Results of the U.S., the pre-IB, and the IB2 on Task 1.

The pre-IB and IB2 performed at lower beginning levels than the U.S., with the pre-IB performing lower and the IB2 performing higher at emerging, and both performing equally higher at mastery. The differences between the pre-IB and the U.S., $\chi^2 = 7.45$, $p = .02$, and the IB2 and the U.S, $\chi^2 = 8.18$, $p = .02$, are statistically significant.

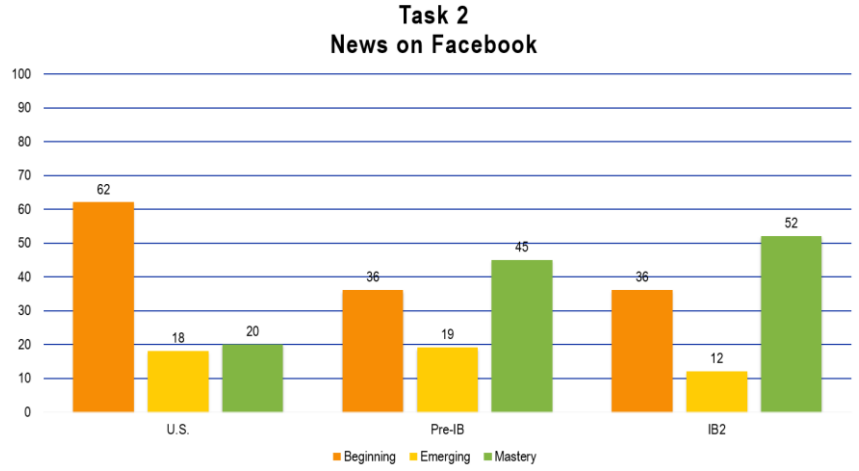


Chart 16. Results of the U.S., the pre-IB, and the IB2 on Task 2.

The pre-IB and IB2 performed at equally lower beginning levels, with the pre-IB performing slightly higher and the IB2 performing lower at emerging; both performed higher at the mastery level. The differences between the pre-IB and the U.S., $\chi^2 = 12.50$, $p = .00$, and the IB2 and the U.S, $\chi^2 = 12.17$, $p = .00$, are statistically significant.

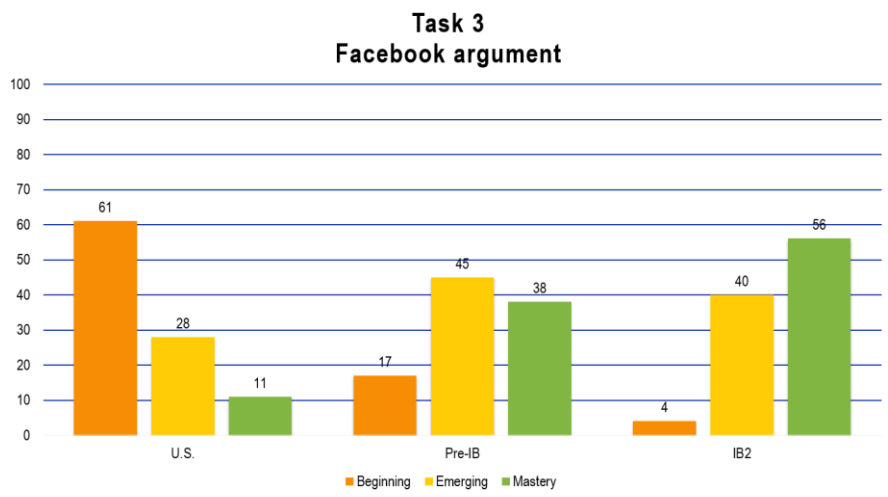


Chart 17. Results of the U.S., the pre-IB, and the IB2 on Task 3.

The pre-IB and IB2 performed at lower beginning levels, higher emerging, and higher mastery levels than the U.S. The differences between the pre-IB and the U.S., $\chi^2 = 31.55$, $p = .00$, and the IB2 and the U.S., $\chi^2 = 41.31$, $p = .00$, are statistically significant.

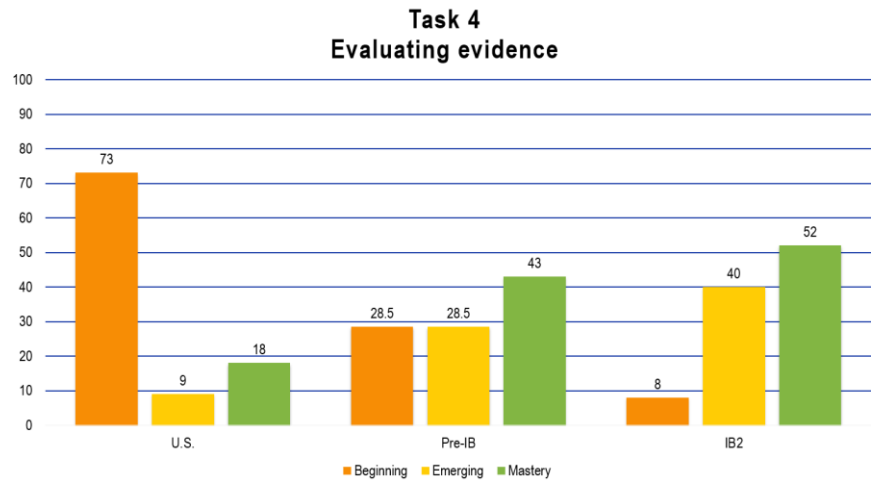


Chart 18. Results of the U.S., the pre-IB, and the IB2 on Task 4.

The pre-IB and IB2 performed at lower beginning, higher emerging, and higher mastery levels than the U.S. The differences between the pre-IB and the U.S., $\chi^2 = 29.48$, $p = .00$, and the IB2 and the U.S., $\chi^2 = 41.76$, $p = .00$, are statistically significant.

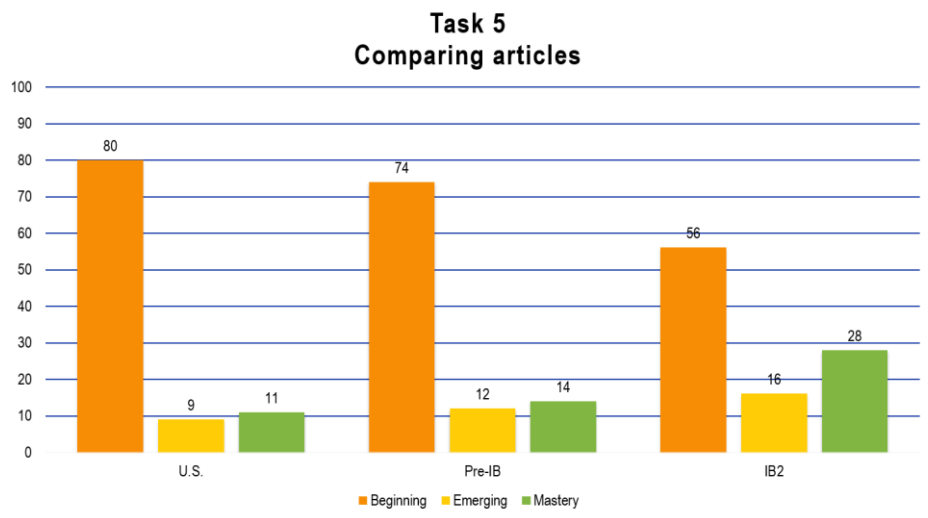


Chart 19. Results of the U.S., the pre-IB, and the IB2 on Task 5.

The pre-IB performed slightly lower and the IB2 performed lower at the beginning level than the U.S. and both performed higher at the emerging and mastery levels. The difference between the pre-IB and the U.S. cohorts is not significant, $\chi^2 = .81$, $p = .67$; however, the difference between the IB2 and the U.S. is statistically significant, $\chi^2 = 7.71$, $p = .02$.

Overview

The cohorts of students entering and exiting the IBDP at a school in Finland outperformed the cohorts of students in the U.S. at the mastery level and revealed lower outcomes at the beginning level across all five tasks. The differences were statistically significant when comparing the cohorts measured at the high school level by the SHEG in California with the students measured in Finland, whether combining the pre-IB and IB2 cohorts together or comparing them separately. This was true on all tasks except Task 5, where the difference did not reach a level of statistical significance between the U.S. cohorts and the pre-IB or combined cohorts with *p*-values of .67 and .09, respectively, but was statistically significant when comparing the IB2 with the U.S. cohorts with a *p*-value of .02. This suggests that while the difference in outcomes between the pre-IB and IB2 did not reach a level of statistical significance on all tasks, there remains a measurable difference which occurred between them, in some cases nearing differential values which indicate significance.

Discussion

As this study is interested in exploring the extent to which students at an IBDP school in Finland demonstrate skills in CT efficacy and transfer, a first consideration is how and to what level of success this can be determined. CT transfer can be evidenced by measuring students' skills in contexts which are external from the classroom; in the case of this study, with established tasks designed to measure skill sets in digital media literacy taken from actual online situations. Measuring CT efficacy, meanwhile, requires considering how and in what ways the intended outcomes of CT instruction are being met, wherein the curricula under which the participants study reveal the approach and intention of the CT instruction received. From the literature review emerged Ennis' (1989) classification for approaches to CT instruction which provided a framework for considering a school's curricular approach to CT against existing scholarship on the various approaches' effectiveness.

With over three-fourths of intervention studies examining the CT-embedded approaches of either immersion or infusion (Abrami et al., 2008; Behar-Horenstein & Niu, 2011), there remains less evidence toward their effectiveness and indeed some evidence toward their facilitating no measurable effect compared to the general and mixed approaches that facilitate CT skills development in a separate course (Tiruneh et al., 2014). While there remains work to determine the extent to which any one approach may be more conclusively effective over another, the results of this study concur with previous scholarship in the field to suggest that approaches explicitly facilitating CT as a course separate from subject area integration reveal stronger outcomes than those which implicitly embed CT into subject area coursework.

The conclusions of the SHEG in their study offer a strong warning that CT efficacy is largely not occurring amongst the students measured in the U.S. "Overall," the authors state, "young people's ability to reason about information on the internet can be summed up in one word: bleak" (Wineburg et al., 2016, p. 4). While the SHEG could make this determination against reasonable expectations of what constitutes adequate or desired performance, this study benefits from the comparative value of replicating the tasks, wherein performance can be further evaluated against the existing SHEG results. The results from the students measured in Finland could be considered mixed if taken in isolation—with Task 1 and Task 5 revealing higher outcomes at the beginning level—but when compared to the U.S., the IBDP students

in Finland revealed consistently superiour outcomes, at times reaching levels of mastery equivalent to the level at which the U.S. students had performed as beginners.

While the drastic differences in outcomes between the U.S. cohorts and those measured for this study are revealing, there were also differences—albeit to a much smaller degree—between the two cohorts measured separately in Finland. The differences between the pre-IB cohort preparing to enter the DP and the IB2 cohort preparing to conclude their studies in the DP are statistically significant when all the tasks are grouped together, with the IB2 revealing higher outcomes than the pre-IB on each task. Given three possible results—that the IB2 would have performed better, the same, or worse—one would expect that the cohort which had spent nearly two years in an intensive mixed infusion CT environment to, at least, not perform worse than the younger cohort which has yet to enter the programme. The differential in outcomes by the IB2 comparative to the pre-IB are made more relevant under consideration that the SHEG, in response to direct enquiry, claims that there were no measurable differences between grade levels in their study (J. Breakstone, personal communication, April 19, 2018).

The main line of enquiry for these results becomes an exploration of factours which could plausibly and feasibly explain these differences while remaining cognisant of both the limitations of the study (6.1) and the ways in which the results may inform future research (6.3). While Ennis' (1989) classification for approaches to CT provides a viable area of causal exploration, there remain other factours worth considering. Although it has not received as much media attention as the PISA results mentioned in 2.4.2, the OECD's Programme for the International Assessment of Adult Competencies (PIAAC) includes measurements of 'problem solving in technology-rich environments' in addition to testing for literacy and numeracy skills. The defining parameter for this measurement is "the capacity to access, interpret and analyse information found, transformed and communicated in digital environments" (OECD, 2012), with interpretation and analytical skills touching on the same CT elements found in the SHEG tasks. One advantage of considering PIAAC results, which are focused on the working-age population between ages 16 and 65, is that it provides insight into the issue of CT transfer for lifelong learning.

The first, and as of now only, PIAAC measurement took place over 2011 – 2012 with the results published in 2013. Like the PISA results, Finland performed well comparative to other nations, with only Japan performing higher in numeracy and literacy and only neighbouring Sweden scoring higher in problem solving within technology-rich environments. While only 8.4% of adults revealed proficiency at the highest achievement level, this is comparative to an average of 5.8% of adults in all participating countries. Further, 33.2% attained the second highest proficiency level in digital problem solving compared with the overall country average of 28.2%, and 61.9% of Finland's younger adults aged 16 – 24 achieved the top two levels compared with 50.7% of young adults across all participating countries. This is only 1.5% below Korea, where young adults attained the highest scores in problem solving, and 24.3% higher than the U.S. where young adults attained the lowest scores (2013).

Further revealing of Finland's success specific to CT skills development comes from the International Civic and Citizenship Study (ICCS) conducted by the International Association for the Evaluation of Educational Achievement. When asked to prioritise three of the most important aims of civic and citizenship education, 82% of teachers polled in Finland indicated 'promoting student independent and critical thinking' as a major aim (Schulz, et al., 2017, p.59). This was not only the top choice by Finnish teachers by over 25% from the second highest aim selected, but it was rather notably also the highest percentage of any participating country (of which teachers from only two other nations selected 'independent and critical thinking' by more than 10% of the ICCS average). These data may also indicate a likelihood of subject area teachers in Finland explicitly facilitating CT skills development as per the mixed infusion model, though this would require further research to establish.

Indeed, as will be discussed in 6.1, the role of individual teachers and any mitigating impact they may have on student learning beyond the design or intent of the surrounding curriculum should be considered extraneous variables when attempting to generalise from the results of this study. In lieu of such, the role of ATL in the IB and 'conception of learning' in the NCC provide some indication of the role of teacher training and the types of instructional approaches inherent to each curriculum. With California, while teacher licensing and requirements are mandated at the state level, any professional development training which guide instructional practices are decentralised to levels as local as the individual school.

The integrity of the tasks must also be considered if the results are to impart meaning and value. While the principle component analysis should be interpreted with caution and the dimensions which were identified in the analysis do not directly align with the conceptual design proposed by the SHEG, removing Task 2 which primarily rewards knowledge over thinking skills does provide a total score distribution closer to normality for both the pre-IB and IB2 cohorts. Taking this into consideration, the data seem to suggest that the IBDP further develops CT skills. For example, taking Tasks 1, 3, and 4 (designed to ask 'What is the evidence?'), the IB2 had two students—or, 8% of the cohort—who performed at the beginning level on two of these tasks and none who performed at the beginning level on all three, whereas the pre-IB had nine students—or, over 20% of the cohort—at the beginning level on at least two of these three tasks and three students, or 7%, at beginning on all three. Conversely, whereas 60%, of students in the IB2 cohort did not perform at the beginning level on any of these three tasks, this percentage was half that at 30% for the pre-IB. While over 20% of the pre-IB cohort did not achieve mastery level on any of the five tasks, all of students in the IB2 cohort achieved mastery on at least one task. Only two of the 67 participants achieved beginning levels on all tasks, both of whom were in the pre-IB cohort.

A natural challenge to the implication that these results may indicate the IBDP fostering such development is the possibility that the students of the IB2 cohort are naturally stronger performers, with other extraneous variables aside from the curriculum affecting their higher performance outcomes. This was accounted for by two metrics which help solidify the similarities of the cohorts: entrance exam results for entry to the pre-IB programme, and the IBDP scores from previous graduating classes. The entrance exams are created locally by the school with a focus on language and literary analysis. The lowest accepted score for the IB2 cohort was 14.81, with the lowest score for the pre-IB at a very similar 14.53 (J. Valtanen, personal communication, February 26, 2018). The average IB scores for the school in the past five years showed little variation, between 32 – 35, with 95 – 100% of the candidates earning the diploma. While this indicates that the school consistently performs above the IB world average of 30 points and ~80% diploma pass rate (IBO, 2018), it confirms that there is little by way of academic variation amongst the cohorts to explain the internal differentials.

6.1 Limitations

A major limitation to the research is the academic excellence of the participant groups measured comparative to both Finland and other IBDP schools, so that generalisation beyond the school measured becomes appropriately challenged. In addition to Finland's already outstanding outcomes in various educational measurements, the students admitted into the IBDP at the school under study represent those who perform above average within a country that performs above international average. As such, the results between both cohorts at the IBDP school in Finland and the sampled schools in the U.S.—while at times drastic in their differences—are not quite as surprising. As will be discussed in 6.3, this invites further testing amongst more normative performing schools under various curricula to continue narrowing down the variables which appear most likely to influence the outcomes.

Due to budget constraints, the 11th grade students who had concluded the pre-IB programme and were partially through their first year of the IBDP (or, the 'IB1 cohort') could not be measured, which could have added to the study of progression within the programme. During the administration of the IB2 cohort, a snowstorm reduced the participant group by half, so that the estimated 50 students were only 25 on the day of task administration. The potential mitigating role of the individual teacher is also not well accounted for in and should be taken into consideration for further studies related to the topic, tasks, and area of enquiry. It is of further importance to not conflate the measurement of two separate cohorts in different stages of study from the advantages inherent to a pretest-posttest design conducted on a singular cohort over time. While such a design offers advantages in terms of causal determination over the design of this study, the establishment of the similarity of the cohorts under study helps alleviate some of the limitations connected to this study's methodological design.

Considering threats to the tasks discussed in the methodology section (4) such as language barrier, cultural bias, and differing testing conditions, the results appear to neutralise many of these concerns given the superiour performance outcomes by the students in Finland over those in the U.S. Post-survey data reveal that students' digital habits and ways in which they consume online information align with the assumptions guiding the SHEG tasks: all participants in the Finland cohorts indicated that they are users of social media, with the vast majority indicating daily usage amongst several different networks.

6.2 Recommendations

One concrete recommendation which emerges from this study is for educational curricula of any socioeducational context to consider implementing explicit coursework in CT as a separate and compulsory component of established curricula along with core subjects such as language, literature, social studies, science, and mathematics. While further studies are required to determine the extent to which it may prove even more advantageous to explicitly embed CT into the course areas as per the mixed infusion approach, the research is clear in suggesting that curricula which implement a specific course in CT as per the general and mixed approaches reveal higher outcomes in CT skills development. Ideally, and given CT's increased importance in the development of what are often referred to as '21st century skills' CT should: 1) be heavily structured into teacher training programmes; 2) include separate licensing and certification for CT teachers; and 3) become a permanent fixture within curricula such as is found in the design and structure of the IBDP and the NCC.

In addition to increasing development of CT skills generally through explicit coursework, there appears to be a growing demand for explicitly developing CT skills specific to digital media literacy. While developments such as the adaptation in the TOK course companion to specifically address skills to determine 'fake news' from genuine media (Dombrowski, 2017) and a recent initiative in Finland to send professional journalists to schools to share their expertise on journalistic practices and social responsibility to help further facilitate skills in media literacy (Koponen, 2018b) reveal explicit action toward further developing the skill sets in CT transfer specific to online environments, there remain "a lack of comprehensive evaluation data of media literacy efforts" (Bulger & Davidson, 2018, p. 3-4) which require further scrutiny and scholarship. The results of this study and the existing scholarship seem to indicate that such efforts toward explicitly facilitating these skills will continue to result in stronger development for determining the credibility and reliability of online information.

Other considerations are methodologically related to approaches to education science generally, which include advocating for more intervention studies into the general and mixed approaches to CT instruction, and for increased replication of studies across differing socioeducational environments. Regarding the former, it would greatly benefit research efforts in the field to separate the category of mixed infusion from mixed immersion for

better comparative evaluation, particularly given that approaches which explicitly facilitate CT skills appear to lead to stronger outcomes. The preset study provides an example of the benefits for the latter by introducing the socioeducational environment of an IBDP school in Finland as a confounding variable from which the results of the U.S. cohort from the SHEG study should not be generalised beyond its own socioeducational context.

6.3 Further research

With an existing deficit in CT intervention studies, particularly at the upper high school level, there exists an apparent need for more research on what approaches for teaching and learning CT remain most effective. Follow up studies which could both challenge or reinforce the initial implications of this study include replicating the materials across more average performing Finnish upper secondary students to include at least grades 10 and 12. This would separate the IBDP and the NCC as variables of interest for influencing the higher outcomes to better consider the differential between the pre-IB and IB2. Should Finnish students in the NCCGUSE overall perform along the lines of the U.S. cohort, this would weaken the possibility of the Finnish curriculum and its mixed immersion approach positively effecting the outcomes. Should the Finnish students perform equal to or greater than the IBDP students in Finland, the IBDP and its mixed infusion approach are weakened as being considered additional influences on the stronger outcomes beyond the effectiveness of the NCC. Should NCCGUSE students perform slightly lower than the IBDP cohort—which would be this researcher's prediction—this could indicate the relative strength of the NCC's mixed approach against the implicit approach such as that found in the SHEG study, yet not as effective as the outcomes from the IBDP's mixed infusion approach.

The next logical replication would be at an average-performing IBDP school, preferably in the U.S., to isolate Finland as a confounding variable of interest and focus on the potential effectiveness of the IBDP and possibly by extension the mixed infusion approach. To maximise comparability with the SHEG study, a sample of IBDP students in California—or, at minimum, a state which has similarly adopted the CCSS with an immersion approach—should be studied. It was also determined that CT is an explicit component of the educational objectives in other countries and some U.S. states (Ennis, 2018, p.165; Silva, 2009, p. 630),

in which case the value of replicating the study to further isolate the effectiveness of explicit instruction in these states and other national curricula and socioeducational environments would provide further value to the overall implications.

Other natural extensions to this study include replicating the materials under curricula other than the IBDP, NCC and CCSS which follow the mixed infusion or mixed immersion approaches to test for outcome correlations amongst the approaches. This would help determine the extent to which the categorical approach can be generalised beyond its curriculum of implementation.

Conclusion

The driving research question for this thesis considered the extent to which CT skills developed in classroom environments may have effectively transferred to the external contexts of everyday interactions with social media and online news. Subsequent enquiries considered how learners from different socioeducational backgrounds would perform on the same assessment tasks, the relationship between curricula which explicitly facilitates CT skills development and their relative performance outcomes on the assessments, and a consideration of the implications and generalisibility of the overall results.

In many ways the strongest contribution of this study toward existing knowledge on CT efficacy and transfer is the metaphorical opening of doors for future research efforts to consider the existing confounding variables so that the emerging implications may be effectively strengthened or challenged. This study began with a focus on the ways in which the IBDP may facilitate skills development in CT transfer to daily online contexts, and yet following the evidence soon uncovered the major contributing role of the Finnish education system. While the initial results show evidential progress from the pre-IB cohort to the IB2 cohort, comparison of both cohorts to the U.S results reported by the SHEG reveal differentials which are far more statistically significant and worthy of consideration. This invited investigation into the variables of interest which could account for such differentials.

While the limitations of the study—most pertinently the superiour academic performance of the participants studied comparative to other students in both Finnish and IBDP education systems which already produce internationally renowned results—should be taken into consideration, the results overall provide evidence that CT skills developed through explicit instruction, especially by treating CT as a separate subject, reveal stronger outcomes on tasks designed to measure digital and media literacy in upper level high school students at an IBDP school in Finland. Replicating tasks in a differing socioeducational environment and examining the stronger outcomes offers caution against generalising from studies which have not been tested under additional socioeducational environments. In congruence with existing scholarship, the curricula which explicitly facilitate CT skills revealed stronger outcomes, with further indication that the more explicit the approach—such as mixed infusion with explicit CT separate and explicit CT embedded—the stronger the resultant outcomes.

References

Ab Kadir, M. (2007). Critical thinking: A family resemblance in conceptions. *Journal of Education and Human Development, 1*(2), 1-11.

Abrami, P., Bernard, R., Borokhovski, E., Wade, A., Surkes, M., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78*(4), 1102-1134. http://dx.doi.org/10.3102/0034654308326084

About the IB. (2005 – 2018). *The International Baccalaureate*. Retrieved from http://www.ibo.org/en/about-the-ib/.

Anderson, J. (2011, December 12). From Finland, an intriguing school-reform model. *The New York Times*. Retrieved from https://www.nytimes.com/2011/12/13/education/from-finland-an-intriguing-school-reform-model.html?src=me&ref=general

Armstrong, P. (2018). Bloom's taxonomy. *Center for Teaching.* Vanderbilt University. Retrieved from https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/

Aro, J. (2016) The cyberspace war: propaganda and trolling as warfare tools. *European View, 15*(21), 121-132. https://doi.org/10.1007/s12290-016-0395-5

Australian curriculum, assessment and reporting authority. (n.d.). Critical and creative thinking learning continuum. *Australian Curriculum*. Retrieved from https://www.australiancurriculum.du.au/media/1072/general-capabilities-creative-and-critical-thinking-learning-continuum.pdf

Bee, S. Black, A., Cassels, P., Drysdale, E., Erhardt, E., Free, T., . . . Taub, M. (Writers), Allen, & Pennolino, P. (Directors). (2017, October 11). Finnish fake news [Television series segment]. In S. Bee, J. Jones, T. Hernandez, M. Kahn, A. Camillo, P. King (Executive Producers) & T. Hall (Producer), *Full frontal with Samantha Bee*. New York City, NY: TBS.

Behar-Horenstein, L. S., & Niu, L. (2011). Teaching critical thinking skills in higher education: A review of the literature. *Journal of College Teaching & Learning, 8*(2), 25-42.

Bensley, D. A., & Haynes, C. (1995). The acquisition of general purpose strategic knowledge for argumentation. *Teaching of Psychology, 22*(1), 41-45. http://dx.doi.org/10.1207/s153280 23top2201_13

Bloom, B., Engelhart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: David McKay Company.

Bulger, M. & Davison, P. (2018). The promises, challenges, and futures of media literacy.

Data & Society Research Institute. Retrieved from https://datasociety. net/output/the-promises-challenges-and-futures-of-media-literacy/

California Department of Education. (2017). History–social science framework for California Public schools. *Curriculum Framework and Evaluation Criteria Committee*. Retrieved from https://www.cde.ca.gov/ci/hs/cf/documents/hssframeworkwhole.pdf

California Department of Education. (2015). English language arts / English language development framework for California public schools. *Curriculum Framework and Evaluation Criteria Committee*. Retrieved from https://www.cde.ca.gov/be/st/ss/ documents/finale laccssstandards.pdf

California Department of Education. (2013, March). English language arts & literacy in history/social studies, science, and technical subjects. *California Common Core State Standards*. Retrieved from https://www.cde.ca.gov/be/st/ss/documents/ finalelaccss standards.pdf

Cole, D., Gannon, S., Ullman, J., & Rooney, P. (2015). Critical thinking skills in the International Baccalaureate's "Theory of Knowledge" subject: Findings from an Australian study. *Australian Journal of Education, 59*(3).

Dewey, J. (1910). *How we think*. New York City, NY: D. C. Heath & Co.

Dombrowski, E. (2017, March 27). TOK and 'fake news': 3 tips, 2 downloads, and 3 resources. Theory of Knowledge blog, *Oxford Education Blog*, March 27, 2017. Retrieved from https://educationblog.oup. com/category/theory-of-knowledge

Dombrowski, E., Rotenberg, L. & Bick, M. (2013). Should I believe it? A guide to evaluating knowledge claims. *International Baccalaureate Diploma Programme Theory of Knowledge course companion*. Oxford University Press, 219-220.

Domonoske, C. (2016, November 23). Students have 'dismaying' inability to tell fake news from real, study finds. *National Public Radio*. Retrieved from https://www.npr.org/ sections/thetwoway/2016/11/23/503129818/study-finds-students-have-dismaying-inability-to-tell-fake-news-from-real

Ennis, R. (2018). Critical thinking across the curriculum: A vision. *Topoi, 37*(1), 165-184. https://doi.org/10.1007/s11245-016-9401-4

Ennis, R. (1991). Critical thinking: A streamlined conception. *Teaching Philosophy 14*(1), 4-18. DOI:10.5840/inquiryctnews201126215

Ennis, R. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18*(3), 4-10. http://dx.doi.org/10.3102/0013189X018003004

Ennis, R. (1987). A taxonomy of critical thinking dispositions and abilities. In J. Baron, & R. Sternberg (Eds.), *Teaching thinking skills*, 9-26. New York: Freeman and Company.

European Commission. (2018). Key features of the education system. Finnish National Agency for Education, *Ministry of Education and Culture*. Retrieved from https: //eacea.ec.europa .eu/national-policies/eurydice/content/finland_en

European Commission. (2018). Single structure education (Integrated primary and lower secondary education). Finnish National Agency for Education, *Ministry of Education and Culture*. Retrieved from https://eacea.ec.europa.eu/national-policies/eurydice/content/ single-structure-education-integrated-primary-and-lower-secondary-education-11_en

European Commission. (2018). Teaching and learning in general upper secondary education. Finnish National Agency for Education, *Ministry of Education and Culture*. Retrieved from https://eacea.ec.europa.eu/nationalpolicies/eurydice/ content/teaching-and-learning-general-upper-secondary-education-15_en

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

Fillion, N. & Martelli, D. (2017, November 10). Overview of critical thinking in the BC K-12 curriculum revisions. *Simon Fraser University*, British Columbia, Canada. Retrieved from https://www.sfu.ca/content/sfu/conferences/criticalthinking/position paper/_jcr_content/main_content/download/file.res/Critical%20Thinking%20position%20paper.pdf

Finland overview. (2018). Center on international education benchmarking. *The National Center on Education and the Economy*. Retrieved from http://ncee.org/what-we-do/center-on-international-education-benchmarking/top-performingcountries/ finland-overview/

Finnish National Agency for Education. (n.d.). Basic education. *Ministry of Education and Culture*. Retrieved from http://minedu.fi/en/basic-education

Finnish National Board of Education. (2016). *National core curriculum for basic education 2014*. Helsinki, Finland: Author.

Finnish National Board of Education. (2016). *National core curriculum for general upper secondary schools 2015*. Helsinki, Finland: Author.

Gewertz, C. (2008, October 13). States press ahead on 21st century skills. *Education Week, 28*(8), 21-23.

Glaser, E. (1941). *An experiment in the development of critical thinking*. New York City,

NY: Bureau of Publications, Teachers College, Columbia University.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39*(2), 93-104. http://dx.doi.org/10.1037/0003-066X.39.2.93

Glevey, K. (2008). Thinking skills in England's National Curriculum. *Improving Schools 11*(2), 115–125. DOI: 10.1177/1365480208091104

Halpern, D. (1998). Teaching critical thinking for transfer across domains. Dispositions, skills, structure training, and metacognitive monitoring. *The American Psychologist, 53*(4), 449-455. http://dx.doi.org/10.1037/0003-066X.53.4.449

Hobbs, R. & Jensen, A. (2009). The past, present and future of media literacy education. *Journal of Media Literacy Education 1*(1), 1 -11.

Hunt, E. (2016, December 17). What is fake news? How to spot it and what you can do to stop it. *The Guardian*. Retrieved from https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate

Ingram, M. (2016, November 23). Most students can't tell the difference between fake and real news, study finds. *Fortune*. Retrieved from http://fortune.com/2016/11/23/stanford-fake-news/

International Baccalaureate. (2018). *Programmes*. Retrieved from http://www.ibo.org/en/programmes/

International Baccalaureate. (2018). *Diploma Programme statistical bulletin*. Retrieved from https://www.ibo.org/about-the-ib/facts-and-figures/statistical-bulletins/diploma-programme-statistical-bulletin/

International Baccalaureate Organization. (2017). The IB Diploma Programme statistical bulletin, May 2017 examination session. *IB Global Centre*. Cardiff, Wales: U.K. Retrieved from http://www.ibo.org/contentassets/bc850970f4e54b87828f83c7976a4db6/ dp-statistical-bulletin-may-2017-en.pdf

International Baccalaureate Organization. (2016, July 5). *Almost 150,000 IB graduates celebrate 2016 exam results*. Retrieved from https://www.ibo.org/news/news-list/almost-150000-ib-graduates-celebrate-2016-exam-results/

International Baccalaureate Organization. (2014). Approaches to teaching and learning in the International Baccalaureate (IB) Diploma Programme. *Approaches to teaching and learning*. Retrieved from http://www.ibo.org/globalassets/digital-tookit/flyers-and-artworks/approaches-to-teaching-learning-dp-en.pdf

Jackson, A. (2016, November 21). Finland has one of the world's best education systems.

Here's how it compares to the US. *World Economic Forum*. Retrieved from https://www.weforum.org/agenda/2016/11/finland-has-one-of-the-worlds-best-education-systems-four-ways-it-beats-the-us

Koponen, H. (2018a, January 24). Finland remains resistant to 'fake news', disinformation. *International Press Institute*. Retrieved from https://ipi.media/finland-remains-resistant-to-fake-news-disinformation/

Koponen, H. (2018b, January 24). New Finnish project brings journalists to schools to teach media literacy. *International Press Institute*. Retrieved from https://ipi.media/new-finnish-project-brings-journalists-to-schools-to-teach-media-literacy/

Larsson, K. (2017). Understanding and teaching critical thinking—A new approach. *International Journal of Educational Research, 84*, 32-42. DOI:10.1016/j.ijer.2017.05.004

Leen, C., Hong, H., Kwan, F., & Ying, T.(2014). Creative and critical thinking in Singapore schools. *Office of Education Research, National Institute of Education, Nanyang Technological University*. Retrieved from https://www.nie.edu.sg/docs/default-source/nie-working-papers/niewp2_final-for-web_v2.pdf?sfvrsn=2

Lipman, M. (1988). Critical thinking - What can it be? *Educational Leadership, 45*(1), 38-43.

Logos and programme models. (2018). *International Baccalaureate Organization*. Retrieved from http://www.ibo.org/digital-toolkit/logos-and-programme-models/

Makel, M. & Plucker, J. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*(6), 304-316. https://doi-org.ezproxy.utu.fi/10.3102/0013189X14545513

Maurette, M.T. (1948, October 4). Educational Techniques for Peace: Do they exist? *The United Nations Educational, Scientific and Cultural Organization*: Paris. Retrieved from http://unesdoc.unesco.org/images/0015/001582/158270eb.pdf

Mayer, R. (1992). Teaching of thinking skills in the sciences and mathematics. In D. Halpern (Ed.), *Enhancing Thinking Skills in the Sciences and Mathematics* (pp. 95-116). New Jersey: Lawrence Erlbaum Associates.

McGrew, S., Breakstone, J., Ortega, T., Smith, M. & Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory & Research in Social Education*. https://doi.org/10.1080/00933104.2017.1416320

McGrew, S., Ortega, T., Breakstone, J., & Wineburg, S. (2017). The challenge that's bigger than fake news: Civic reasoning in a social media environment. *American Educator*. Retrieved from https://www.aft.org/sites/files/periodicals/fall2017mcgrew.pdf

McPeck, J. (1990). Critical thinking and subject specificity: A reply to Ennis. *Educational Researcher, 19*(4), 10-12. http://dx.doi.org/10.3102/0013189X019004010

McPeck, J. (1981). *Critical thinking and education*. Oxford: Martin Robinson.

Mulnix, J. (2012). Thinking critically about critical thinking. *Educational Philosophy and Theory, 44*(5), 464-479. https://doi.org/10.1111/j.1469-5812.2010.00673.x

National Association for Media Literacy Education. (2007). *Core principles of media literacy education in the United States*. Retrieved from https://namle.net/publication s/core principles

National Council for the Social Studies. (2016). Media literacy. *Social Education, 80*, 183–185.

Nickerson, R. (1988). On improving thinking through instruction. *Review of Research in Education, 15*(3), 3-57. http://www.jstor.org/stable/1167360

Ontario Ministry of Education. (2015). The Ontario curriculum grades 11 and 12. *Canadian and World Studies*. Toronto, ON: Queen's Printer for Ontario. Retrieved from www.edu.gov. on.ca/eng/curriculum/secondary/2015cws11and12.pdf

Organisation for Economic Cooperation and Development. (2018). *PISA 2015 Results in Focus.* Retrieved from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Organisation for Economic Cooperation and Development. (2015). *Students, computers, and learning: Making the connection*. Paris: OECD Publishing.

Organisation for Exonomic Cooperation and Development. (2013). Country note: Finland. *Survey of adult skills first results*. Retrieved from http://www.oecd.org/skills/ piaac/Country%20note%20-%20Finland.pdf

Orszag, A. (2015). Exploring Finnish university students' perceived level of critical thinking. (Master's thesis). *University of Jyväskylä*, Jyväskylä, Finland.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students. A third decade of research, Vol. 2*. San Francisco: Jossey-Bass.

Paul, R. (1995). *Critical thinking: What every person needs to survive in a rapidly changing world*. Santa Rosa, CA: Foundation for Critical Thinking.

Perkins, D., & Salomon, G. (1988). Teaching for transfer. *Educational Leadership, 46*(1), 22-32. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=afh& AN=85248 29&site=ehost-live&scope=site

Peters, M. (2017). The information wars, fake news and the end of globalization.

Educational Philosophy and Theory. DOI: 10.1080/00131857.2017.1417200

Plath, D., English, B., Connors, L., & Beveridge, A. (1999). Evaluating the outcomes of intensive critical thinking instruction for social work students. *Social Work Education: The International Journal, 18*(2), 207-217. http://dx.doi.org/10.1080/02615479911220201

Resnick, L. B. (1987). *Education and Learning to Think*. Washington DC: National Academy Press.

Rosendahl, J. & Forsell, T. (2016, October 19). Finland sees propaganda attack from former master Russia. *Reuters*. Retrieved from https://www.reuters.com/article/us-finland-russia-informationattacks/finland-sees-propaganda-attack-from-former-master-russia-idUSKCN 12J197

Saiz, C. & Nieto, A. (2010). Critical thinking: A question of aptitude and attitude? *Inquiry: Critical Thinking Across the Disciplines, 25*(2), 19-26.

Schulten, K. & Brown, A. (2017, January 19). Evaluating sources in a 'post-truth' world: Ideas for teaching and learning about fake news. *The New York Times*. Retrieved from https://www.nytimes.com/2017/01/19/learning/lesson-plans/evaluating-sources-in-a-post-truth-world-ideas-for-teaching-and-learning-about-fake-news.html

Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2017). Becoming citizens in a changing world: International civic and citizenship education study 2016 international report. *International Association for the Evaluation of Educational Achievement.* Retrieved from https://iccs.iea.nl/fileadmin/user_upload/Editor_Group/Downloads/ICCS_2016_International_report.pdf

Shellenbarger, S. (2016, November 21). Most students don't know when news is fake, Stanford study finds. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/most-students-dont-know-when-news-is-fake-stanford-study-finds-1479752576

Silva, E. (2009, May 1). Measuring skills for 21st-century learning. *Phi Delta Kappan, 90*(9), 630-634. https://doi.org/10.1177/003172170909000905.

Simpson, J., Weiner, E., & Oxford University Press. (2018). *The Oxford English Dictionary*. Oxford: Clarendon Press.

Standish, R. (2017, March 1). Why is Finland able to fend off Putin's information war? *Foreign Policy*. Retrieved from http://foreignpolicy.com/2017/03/01/why-is-finland-able-to-fend-off-putins-information-war/

Survey: Majority of Finns feel effects of fake news. (2017, November 14). *Yle*. Retrieved from https://yle.fi/uutiset/osasto/news/survey_majority_of_finns_feel_effects_of_fake_news/9931091

Taylor, A. (2011, December 14). 26 Amazing facts about Finland's unorthodox education System. *Business Insider*. Retrieved from http://www.businessinsider.com/finland-education-school-2011-12?r=US&IR=T

Tiruneh, D., Verburgh, A. & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies 4*(1), 1-17. http://dx.doi.org/10.5539/hes.v4n1p1

Trem, K. (2014). Curriculum theory, conflicting visions and enduring concerns. *Education and Training, 56*(4), pp.360-363. https://doi-org.ezproxy.utu.fi/10.1108/ET-10-2013 - 0120

US experts gird Finnish officials for information war. (2016, January 22). *Yle*. Retrieved from https://yle.fi/uutiset/osasto/news/us_experts_gird_finnish_officials_for_information_war/8616336

Utriainen, J., Marttunen, M., Kallio, E., & Tynjälä, P. (2017). University applicants' critical thinking skills: The case of the Finnish educational sciences. *Scandinavian Journal of Educational Research, 61*(6), 629-649. doi:10.1080/00313831.2016.1173092

Williams-Grut, O. (2016, 18 November). The 11 best school systems in the world. *Business Insider*. Retrieved from http://www.businessinsider.com/wef-ranking-of-best-school-systems-in-the-world-2016-2016-11?r=UK&IR=T

Wineburg, S., McGrew, S., Breakstone, J. & Ortega, T. (2016). Evaluating Information: The Cornerstone of Civic Online Reasoning. *Stanford Digital Repository*. Retrieved from http://purl.stanf ord.edu/fv751yt5934

Wineburg, S. & McGrew, S. (2016, December 13). Most teens can't tell fake from real news. *PBS News Hour*. Retrieved from https://www.pbs.org/newshour/education/column-students-cant-google-way-truth

APPENDIX A

Stanford History Education Group assessment tasks for high school

A major news website published an article about Yik Yak, a social media application that allows users within a 1.5-mile radius to read and post comments anonymously. Some schools have had problems with students using Yik Yak to harass and threaten their peers. The comments below were posted in response to the article.

Comments

rabbits882 Kalamazoo, MI • 4 days ago

I think Yik Yak is worse than old school forms of gossip. As reported in the article, Yik Yak employees have had to disable Yik Yak at nearly all American middle schools and high schools. Plus the article states that Yik Yak "has been used to threaten mass violence on more than a dozen college campuses." It says that these threats have even led to police investigations! Even though old school gossip can hurt, I think the fact that people can post on Yik Yak publicly and anonymously leads it to affect entire schools much more easily and much more dangerously.

fortuneseeker3 New York City, NY • 2 days ago

There's nothing wrong with Yik Yak. For ages humans have gossiped and said cruel things about one another. As Prime Minister Winston Churchill said, "Criticism may not be agreeable, but it is necessary." This app is simply the newest technology young people are using to do the same thing. If you prefer not to read what others think of you, you're not obligated to download the app. The government collects citizens' private phone and Internet records all the time, but we're worried about some little app?

Question: Did rabbits882 or fortuneseeker3 make a stronger argument? Explain.

Stanford History Education Group

Task 1: Argument analysis. Students read two comments in response to a news article and explain which commenter makes a stronger argument.

ARGUMENT ANALYSIS RUBRIC

MASTERY	Student argues that rabbits882 made a stronger argument and provides a thorough and accurate explanation based on the fact that rabbits882 used evidence from the article or that fortuneseeker3 did not provide relevant evidence.
EMERGING	Student argues that rabbits882 made a stronger argument but provides a vague or incomplete explanation.
BEGINNING	Student argues that fortuneseeker3 made a stronger argument *or* argues that rabbits882 made a stronger argument but provides an incoherent or incorrect explanation.

SAMPLE RESPONSES

MASTERY

This student compares the evidence that rabbits882 and fortuneseeker3 used in order to show that rabbits882 made the stronger argument.

Rabbits882 makes a stronger argument because she uses evidence that is both compelling and directly relevant (from the article vs. an unrelated quote taken out of context). In contrast, fortuneseeker3 doesn't really do anything but generalize without support, and doesn't show us the harms of disabling Yik Yak in comparison to the clear harms shown by rabbits882 (mass violence).

EMERGING

This student argues that rabbits882 made a stronger argument but does not provide a specific explanation about what makes this commenter's evidence stronger than fortuneseeeker3's evidence.

> Rabbits882 made a stronger argument because he brought evidence to support his claim. He also explained the evidence to some degree. The other commenter just said his opinion without using evidence.

This student argues that fortuneseeker3's evidence does not support his claim but does not explain why or how that compares to the evidence that rabbits882 provides.

> I think rabbits882 made a stronger argument because his organization of thoughts is more clear and connected. He makes a clear claim and supports it, while Fortuneseeker3 is very scattered and uses evidence that isn't supportive of his claim.

BEGINNING

This student argues that fortuneseeker3 provided stronger evidence and thus had a stronger argument.

> Fortuneseeker3 made a stronger argument because there is a quote from someone in history that supports the claim. Rabbits882, however, gives more of his opinion by using the words "I think."

Instead of evaluating the quality of the evidence and overall argument, the student selects rabbits882's comment because it matches his own opinion.

> Rabbits882 made a stronger argument because he emphasizes the damage that yikyak has created. The fact that humans have been gossiping for as long as we can remember doesn't make it okay to do. Besides, there's a huge difference between constructive criticism and verbal harassment.

Stanford History Education Group

sheg.stanford.edu

Task 1: Argument analysis rubric.

Facebook Post A

Facebook Post B

Facebook Post A / **Facebook Post B** (circle one) is a more trustworthy source about Donald Trump's decision to run for president because _____

_____

_____

_____

_____

_____

Stanford History Education Group

Task 2: News on Facebook. Students explain which of two news posts (one from a verified account, one not) is a better source.

NEWS ON FACEBOOK RUBRIC

MASTERY	Student identifies the blue check mark and explains that Post A is a more trustworthy source because it is from the verified *Fox News* Facebook page.
EMERGING	Student identifies the blue check mark but does not explain its significance; *or* identifies Post A as the actual *Fox News* account but does not provide an explanation.
BEGINNING	Student selects Post B as more trustworthy *or* provides an irrelevant explanation for selecting Post A.

SAMPLE RESPONSES

MASTERY

This student identifies the authentic *Fox News* account and explains how he knows it is a verified account.

As you can see, there is a checkmark beside the Facebook name "Fox News." That lets you know that it is the real Fox News, unlike the other. Anyone can make a fake account about Fox News, but the check mark next to the username lets you know they are the real deal.

EMERGING

This student identifies the blue check mark but does not explain its significance.

Post A is more trustworthy because it has a check mark by its name.

This student correctly identifies the authentic *Fox News* account but does not explain how she made this determination.

Post A is more trustworthy because it is from a reliable and official news website.

BEGINNING

This student argues that Post B is more trustworthy because it presents evidence from Trump himself.

Post B is more reliable because it is showing his tweet, which comes from Trump, so that makes it more trustworthy. This is evidence that comes straight from him.

This student identifies Post A as the more trustworthy source but defends his choice by referencing irrelevant factors.

Post A because Trump explains that he is going to make our country great again, and it shows a picture of him smiling. The other post does not say anything about what he's going to do once he becomes president except that he's going to be one of the presidential candidates.

Task 2: News on Facebook rubric.

Examine the Facebook conversation below and answer the question that follows.

Did Anya or Grace provide stronger evidence to support her argument? Explain.

Stanford History Education Group

Task 3: Facebook argument. Students explain which poster in a Facebook conversation makes a stronger argument about gun laws.

FACEBOOK ARGUMENT RUBRIC

MASTERY	Student chooses Anya's evidence or rejects Grace's evidence and explains why evidence presented by a gun owners' PAC is weaker than an article by a prestigious news organization.
EMERGING	Student chooses Anya's evidence or rejects Grace's evidence but provides limited reasoning about the source of information.
BEGINNING	Student's explanation ignores the sources of evidence and instead compares the evidence presented by Grace and Anya on other grounds.

SAMPLE RESPONSES

MASTERY

This student compares Anya's and Grace's sources of evidence and provides a sound argument for why Anya's evidence is stronger.

Anya provided stronger evidence with her article. All Grace did was show a vague chart. Her evidence is also biased because the chart is from a website run by Minnesota gun owners. Anya's evidence is from a well-known professional newspaper company.

EMERGING

This student focuses on the qualities of the evidence that Anya and Grace provide—Anya's evidence is based on past events while Grace's evidence is hypothetical—and does not critique the sources of the information.

I feel like Anya had stronger evidence because she stated a fact about Australia that she can research while Grace practically just stated that gun laws won't stop shootings and made a graph without facts.

BEGINNING

This student does not discuss the sources of Anya's and Grace's evidence and instead is swayed by Grace's evidence because it provides a chart.

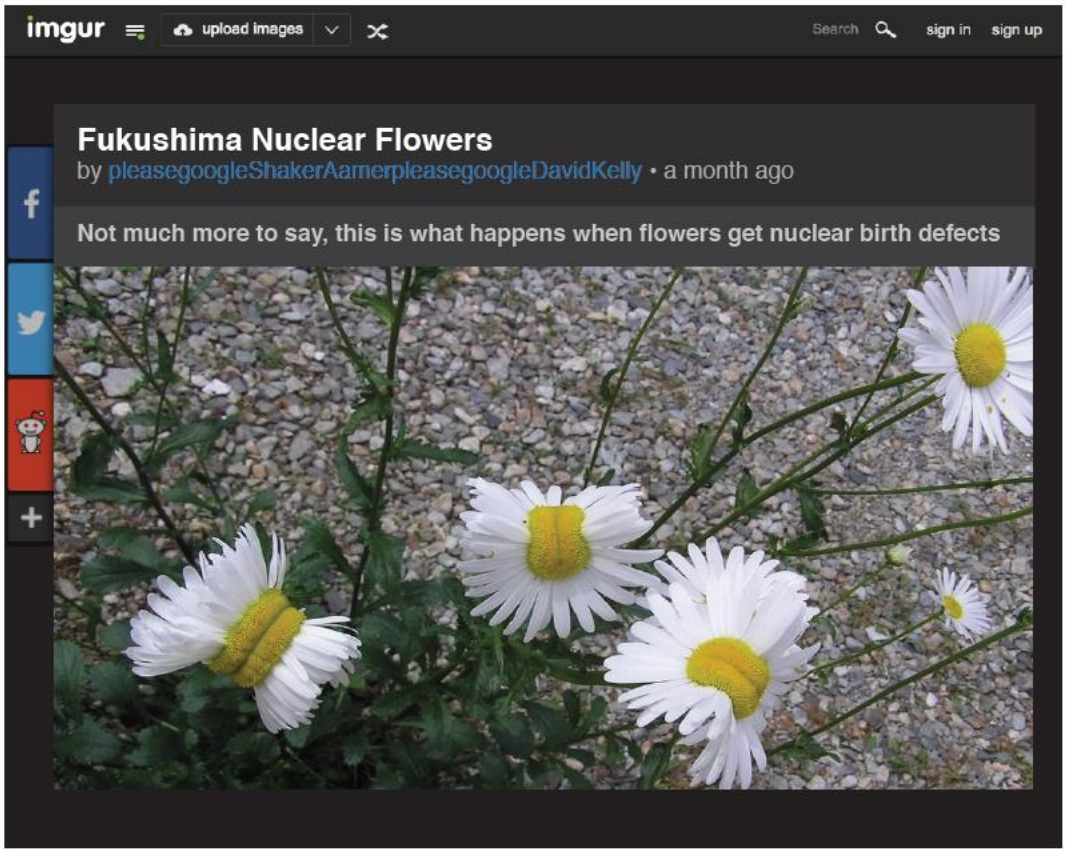
Grace provided stronger evidence to support her argument by not only stating her opinion but giving factual evidence in the form of a chart to win her case.

Stanford History Education Group

sheg.stanford.edu

Task 3: Facebook argument rubric.

On March 11, 2011, there was a large nuclear disaster at the Fukushima Daiichi Nuclear Power Plant in Japan. This image was posted on Imgur, a photo sharing website, in July 2015.



Does this post provide strong evidence about the conditions near the Fukushima Daiichi Power Plant? Explain your reasoning.

Stanford History Education Group

Task 4: Evaluating evidence. Students evaluate the strength of evidence in a photograph posted on Imgur.

EVALUATING EVIDENCE RUBRIC

MASTERY	Student argues the post does not provide strong evidence and questions the source of the post (e.g., we don't know anything about the author of the post) and/or the source of the photograph (e.g., we don't know where the photo was taken).
EMERGING	Student argues that the post does not provide strong evidence, but the explanation does not consider the source of the post or the source of the photograph, *or* the explanation is incomplete.
BEGINNING	Student argues that the post provides strong evidence or uses incorrect or incoherent reasoning.

SAMPLE RESPONSES

MASTERY

This student questions the source of the photo, arguing that there is no way to know whether the photo was actually taken near the plant or if the mutations were a result of nuclear radiation.

> No, it does not provide strong evidence about the conditions near the Fukushima Daiichi power plant. It does not provide strong evidence because it could just be a mutation in the plant. There also isn't evidence that this is near the Fukushima Daiichi power plant.

This student questions the source of the post, arguing that we know nothing about the poster's credentials or whether the evidence was doctored.

> No, it does not really provide strong evidence. A photo posted by a stranger online has little credibility. This photo could very easily be Photoshopped or stolen from another completely different source; we have no idea given this information, which makes it an unreliable source.

EMERGING

This student begins to question both the photo and the source of the post but does not fully explain his thinking.

> This post does not provide strong evidence about conditions near the power plant. They just put a picture of a flower. Plus the poster has a strange username.

This student critiques the evidence by arguing that it could have been digitally altered but does not offer any further explanation or critique of the evidence.

> No, because this picture could be Photoshopped.

BEGINNING

This student accepts the evidence at face value, arguing that it provides visual proof of the effects of the nuclear disaster.

> This post does provide strong evidence because it shows how the small and beautiful things were affected greatly, that they look and grow completely different than they are supposed to. Additionally, it suggests what such a disaster could do to humans.

Although this student argues that the post does not provide strong evidence, she still accepts the photo as evidence and simply wants more evidence about other damage caused by the radiation.

> No, this photo does not provide strong evidence because it only shows a small portion of the damage and effects caused by the nuclear disaster.

Stanford History Education Group

sheg.stanford.edu

Task 4: Evaluating evidence rubric.

The following headlines appeared on *The Atlantic*, a news website. Both of their accompanying articles are about policies to solve global climate change.

A

Why Solving Climate Change Will Be Like Mobilizing for War

And even then, victory is far from guaranteed

VENKATESH RAO | SCIENCE

B

Is Article A or Article B a more reliable source for learning about policies to solve global climate change? Explain. _____

_____

_____

_____

_____

Stanford History Education Group

Task 5: Comparing articles. Students explain which of two sources (one sponsored content, one traditional news) is a more reliable source about global climate change.

COMPARING ARTICLES RUBRIC

This task presents students with links to two articles about climate change and asks which is a more reliable source about that topic. One article is a traditional news story appearing in the "Science" section, while the other is a native advertisement sponsored by Shell Oil Company. Successful students will recognize the conflict of interest inherent in Shell's sponsorship of an article about climate change and, based on that, choose the article that appeared in the "Science" section as the more reliable source.

MASTERY	Student identifies that Article B is sponsored by a company with a vested interest in the article's topic. Student provides a clear rationale for why this makes the article less reliable.
EMERGING	Student identifies Article B as sponsored content and explains that this makes it less reliable as a source, but the explanation is limited.
BEGINNING	Student does not identify the sponsored content as a relevant consideration *or* identifies the sponsored content but argues that it is the more reliable source.

SAMPLE RESPONSES

MASTERY

This student argues that Article B has a clear conflict of interest because it is sponsored by an oil company.

> Article A gives several outside sources to expand upon certain points it makes, bringing other ideas and viewpoints into the mix to contribute to one overall message. Article B is content that was sponsored by a company. Shell is an energy corporation, and they therefore have a reason to disseminate misinformation regarding the use of alternative energy and the value of fossil fuels. Why would Shell sponsor anything that makes a case against their business? There is a straightforward, money-motivated agenda behind Article B.

EMERGING

This student selects Article A, but does not fully explain why Article B is problematic.

> Article A is a more reliable source because it is science content, Article B is sponsored content.

BEGINNING

This student selects Article B because it includes more quotes and facts.

> I think Article B is more reliable, because it provides the reader with evidence (quotes) from reliable sources like the *New York Magazine* and provides current facts. In Article A there aren't any quotes and although it's much longer it may not be as helpful as Article B.

This student notices that Article B is sponsored content but argues that this makes it more trustworthy.

> I chose Article B because it's sponsored so not only does the writer agree with what they're saying - an entire company does enough to sponsor them. The article also gives a better idea of what needs to be done to fix this problem than relating it to "mobilizing for war."

Task 5: Comparing articles rubric.

APPENDIX B
Study script

My name is Shane Horn, a student in a master's degree programme for international students at the University of Turku. Before coming to Finland to study education I have been an international teacher, including for Theory of Knowledge, which is the inspiration for my thesis research.

The materials for this study include five tasks which relate to digital media. We can discuss more fully the nature and purpose of the tasks and my research after completion. Your teachers may use these and additional materials for future lessons, but for these tasks you will not receive any grades or individual feedback. The results are associated with participant numbers and not your names.

Before we begin, there is a short survey to collect some demographic data. While the information from this survey is valuable to the research, it is optional for you to fill out. Your teachers will not see the survey results so will not be able to associate the information with your task results.

*** Hand out surveys.**

Regarding your first language or languages, this is what you consider to be your native language or 'mother tongue'. If you were raised in more than one language equally, please include them.

If you attended the IB middle years programme or the IB primary years programme before the diploma programme, please include your total years of IB experience. If your only experience with the IB is the diploma programme, please put "two" for how many years you have studied in the IB.

Lastly, please write in your course selections for each IB diploma programme group and indicate whether you are taking it at the higher or standard level by circling 'HL' or 'SL'.

When you are finished with the survey please set it aside but do not turn the paper over.

As the tasks are handed out, please leave them face down until you are told to begin.

*** Hand out tasks**

There are five tasks for you to complete, which should take about five to ten minutes each. Your responses should be able to fit into the space provided, but if you require more space, feel free to use additional space or the back of the paper.

As you complete one task, please move on to the next. When you have completed all five tasks, please raise your hand to indicate you are done.

Are there any questions? You may begin.

*** When students raise their hand upon completion, they are quietly asked to turn over the survey and complete the post-survey. When all tasks and survey are complete, they are collected together.**

APPENDIX C

Survey Questions

Gender: _____

Country-of-Origin: _____

First Language(s): _____

For how many years have you studied in the Finnish education system?

For how many years have you studied in the International Baccalaureate programme?

Enrollment (please circle which section you are attending):

Turku International School or Turun normaalikoulu

What courses are you taking in the IBDP (please write in) and at what level (please circle)?

Group 1 – Language and Literature: _____ (HL / SL)

Group 2 – Language Acquisition: _____ (HL / SL)

Group 3 – Individuals and Societies: _____ (HL / SL)

Group 4 – Sciences: _____ (HL / SL)

Group 5 – Mathematics: _____ (HL / SL)

Group 6 – The Arts (or other): _____ (HL / SL)

Post-Survey Questions

Are you aware of the 'blue checkmark' verification used on many social media networks? (please circle) Yes / No

How often would you say you use social media (Facebook, Twitter, Reddit, VK, etc.)?

1. Several times a day
2. Once daily
3. A few times a week
4. Once a week or less
5. Never

What are some social media networks that you use? (If none, leave blank.)

How often would you say you consume (i.e., read or watch) news media online?

1. Several times a day
2. Once daily
3. A few times a week
4. Once a week or less
5. Never

What are the main sources from which you receive your online news?

(If none, leave blank.)

Were you previously aware of these topics, materials, or tasks before today?

If yes, how so?

APPENDIX D
Consent form

TURUN YLIOPISTO
UNIVERSITY OF TURKU

December 2017
University of Turku, Department of Teacher Education and The Centre for Learning Research

Topic: Consent Form
Permission for Research Data

Dear Participant,

The Department of Teacher Education and The Centre for Learning Research at the University of Turku has for many years been researching the learning of young people in collaboration with teachers and schools. A central aim of this effort is to contribute to improved learning gains of students here and around the world.

Research being conducted for a thesis project in the Master's Degree Programme for Learning, Learning Environments, and Educational Systems at the University of Turku is focused on the extent to which critical thinking skills developed in the classroom may transfer to real world contexts, such as the digital media with which young people interact daily. The IBDP course Theory of Knowledge was chosen for testing due to its being a critical thinking course designed to facilitate such skills.

Task materials used in the research were developed, validated, and field tested by researchers at the Stanford History Education Group (SHEG), who for the past two years have administered the tasks to thousands of middle school, high school, and university students across the United States. You can access an executive summary of the original study by the SHEG at: https://purl.stanford.edu/fv751yt5934.

Agreements of conducting the study have been made with the teachers and the principal. The tasks are administered under the supervision of the researcher and teachers on the school premises. The data collected is associated with participant numbers and any identifying information will not be published at any time. All information will be handled confidentially according to the Personal Data Act.

With permission, collecting your results on the tasks will lead to the first comparative data outside of the U.S., as well as the first exploration into the ways explicitly studying critical thinking such as through the Theory of Knowledge course may affect the learning outcomes. Your permission is sought to include your anonymous data in the results, including the possibility of publication in a scientific journal.

If you feel positively toward this research and agree to have the data included in the study's results, kindly fill out the Consent Form found on the reverse side.

Best Regards,

Shane Horn

Master's Degree Programme in Learning, Learning Environments, and Educational Systems

University of Turku, Department of Teacher Education and Center for Learning Research

I understand the nature of the study and

- ☐ **AGREE** to my data being included in the research results.
- ☐ **DO NOT AGREE** to my data being included in the research results.

_____ _____

Date Signature