Exploring the structure of digital literacy competence assessed using authentic software applications

Reichert, Frank; Zhang, James; Law, Nancy; Wong, Gary; Torre, Jimmy de la

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Reichert, F., Zhang, J., Law, N., Wong, G., & Torre, J. d. l. (2020). Exploring the structure of digital literacy competence assessed using authentic software applications. *Educational Technology Research and Development*, *68*(6), 2991-3013. https://doi.org/10.1007/s11423-020-09825-x

Education Tech Research Dev (2020) 68:2991–3013
https://doi.org/10.1007/s11423-020-09825-x

RESEARCH ARTICLE

Exploring the structure of digital literacy competence assessed using authentic software applications

Frank Reichert[1] · James Zhang[2] · Nancy W. Y. Law[3] · Gary K. W. Wong[4] · Jimmy de la Torre[5]

Accepted: 10 September 2020 / Published online: 25 September 2020

Abstract

Digital literacy competence (DL) is an important capacity for students' learning in a rapidly changing world. However, little is known about the empirical structure of DL. In this paper, we review major DL assessment frameworks and explore the dimensionality of DL from an empirical perspective using assessment data collected using authentic software applications, rather than simulated assessment environments. Secondary analysis on representative data collected from primary and secondary school students in Hong Kong using unidimensional and multidimensional item response theory reveals a general dimension of digital literacy performance and four specific, tool-dependent dimensions. These specific DL dimensions are defined by the software applications that students use and capture commonality among students' performance that is due to their familiarity with the assessment tools and contexts. The design of DL assessment is discussed in light of these findings, with particular emphasis on the influence of the nature of digital applications and environments used in assessment on the DL achievement scores measured.

Keywords Assessment · Authentic software applications · Digital literacy · Purpose-built software · Twenty-first century skills

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11423-020-09825-x) contains supplementary material, which is available to authorized users.

✉ Frank Reichert
reichert@hku.hk

[1] Faculty of Education, The University of Hong Kong, Room 217, Runme Shaw Building, Pok Fu Lam Road, Hong Kong, Hong Kong SAR

[2] Faculty of Economics and Business, University of Groningen, Nettelbosje 2, 9747 Groningen, The Netherlands

[3] Faculty of Education, The University of Hong Kong, Room 111A, Runme Shaw Building, Pok Fu Lam Road, Hong Kong, Hong Kong SAR

[4] Faculty of Education, The University of Hong Kong, Room 113, Runme Shaw Building, Pok Fu Lam Road, Hong Kong, Hong Kong SAR

[5] Faculty of Education, The University of Hong Kong, Room 520, Meng Wah Complex, Pok Fu Lam Road, Hong Kong, Hong Kong SAR

Introduction

Digital representation of information and its communication through digital technologies has transformed the way we work, learn, express ourselves, and even the way we think, as humans are challenged by the cognitive, informational, technological, and socio-emotional demands of the digital world (List et al. 2020; Littlejohn et al. 2012). Digital technologies are ubiquitous and require a certain level of literacy in their usage. Unsurprisingly, digital literacy competence (DL) has become an essential capacity in daily life and important for lifelong learning in our rapidly changing world (Rohatgi et al. 2016). Thus researchers, politicians and practitioners alike consider the acquisition of DL to be as important as learning to read and write using non-digital means (Bawden 2008). Against this background, frameworks for the teaching and learning of DL as well as its assessment have emerged in the educational literature.

There is much overlap in the knowledge, skills, and attitudes deemed to be necessary in order to use digital technology for work, leisure and well-being in the twenty-first century (Siddiq et al. 2017; van Laar et al. 2017; Voogt and Roblin 2012), and the labels used to refer to such competence are various. It appears that these labels change over time with the pervasiveness of digital technology adoption in the society. Computer literacy was among the earliest and refers to more general knowledge and skills required to understand and operate computers and computer applications (Bawden 2008). Information literacy has a broader scope, referring to an individual's ability to locate and use information from a variety of sources, including but not limited to computers (Bawden 2008). The concept of computer and information literacy (CIL) combines both literacies and is conceptually synonymous to DL (Siddiq et al. 2016). Despite the diverse labels used, there is a distinct set of competences in the literature that converge around the retrieval and processing of information via new technologies, communication through these technologies, as well as the production of content using information and communication technologies (ICT; Siddiq et al. 2016). Further, irrespective of the labels used, the various frameworks refer to these necessary competences as different dimensions. DL is the term preferred in this paper, because it refers to all kinds of digital devices of different form factors as well as digital environments that operate across hardware and software platforms.

In addition to frameworks of DL, there have also been studies assessing DL over the last two decades, including several high-profile national and international assessment studies, such as the Australian National Assessment Program for ICT literacy (NAP–ICT) and the International Computer and Information Literacy Study (ICILS). All of these studies have adopted a multidimensional assessment framework, although the exact number of dimensions and the labels used for these dimensions may differ. On the other hand, the empirical findings from most of these studies have indicated DL to be a unidimensional construct (e.g. Aesaert et al. 2014; Gebhardt and Schulz 2015). This tension prompted us to explore the question *whether DL is a unidimensional or a multidimensional construct* [research question (RQ1)].

In addition, DL is often implicitly conceptualized as a generic competence that is independent of the devices, tasks and technologies used. However, software applications differ in interface design, and the same functions can be achieved using different devices or applications (Fraillon 2018). In real-life situations, one may use the same application to achieve multiple kinds of functions (e.g. one can use a spreadsheet to manage and organize information as well as to present and communicate information). On the other hand, the focus and skills in performing the same function may differ depending on the technology

used (e.g. communication using email, a messaging application, social media or a presentation software requires different operational skills and understanding). Hence, this analysis further examined the question *whether a person's DL as performed in authentic software applications is generic*, i.e. independent of the digital technology used in an assessment (RQ2).

With the pervasive adoption of digital technology in homes and schools, even young children are exposed to and make use of digital technology for entertainment, communication with family members, and learning. Thus, it is reasonable to expect that primary age children would have developed DL. However, studies that have assessed primary students' DL are relatively rare. The present study also explored *whether the dimensionality and tool (in)dependence of the DL measured differ between primary and secondary students* (RQ3).

To address the identified research questions, we first reviewed the major existing DL assessment frameworks to understand how these conceptualized DL and to further elaborate the research challenges addressed in this study. This was followed by a secondary analysis of data collected from a territory-wide assessment study that administered the same assessment instrument to primary and secondary school students in Hong Kong. The results prompted us to discuss issues of tool and task dependence in the assessment of DL, and we conclude with suggestions for further research on the development of valid DL assessments.

Conceptual frameworks of DL

There have been numerous studies of DL in recent years. Voogt and Roblin (2012) identified about 178 publications on DL and selected eight competence *frameworks* for review and analysis because of their wider recognition. These researchers found that the frameworks converged to a common set of twenty-first century competences, despite the differences in terminologies and procedures for synthesis. In an effort to provide a unified and comprehensive DL framework, the European Commission synthesized existing conceptualizations of DL to establish the Digital Competence Framework (DIGCOMP) onto which other DL frameworks can be mapped. The current DIGCOMP 2.1 comprises 5 competence areas (information and data literacy, communication and collaboration, digital content creation, digital safety, and problem solving), which are further differentiated into 21 sub-areas (Carretero et al. 2017). It is important to note that DIGCOMP, as with the frameworks that it references, depicts DL as a generic competence that is independent of the devices or specific software technologies used.

However, software applications differ in interface design and how the same functions can be achieved using different devices or applications (Fraillon 2018; Haßler et al. 2016). In view of the fact that all of the popularly adopted DL frameworks have been developed in countries with relatively high national incomes, the United Nations Educational, Scientific and Cultural Organization (UNESCO) commissioned a study to develop a Global Framework of Reference on Digital Literacy (GFDL) that is relevant for different country contexts, and remains so as technology and its use in the society changes over time (Law et al. 2018). The GFDL study collected examples of digital technology usage in major areas of socioeconomic activity from 47 countries to identify the DL competences needed to satisfactorily accomplish the tasks in the usage examples. The identified competences were also mapped onto the DIGCOMP 2.1 framework (Woo and Law 2021). The results showed that by and large, the DIGCOMP 2.1 framework is comprehensive enough to encompass the

functional competences identified in the diverse contexts represented by the usage examples, with an important caveat: The same DL may be accomplished in very different ways depending on the nature and sophistication of the devices and software applications used. For example, financial transaction services via SMS messages on a mobile phone are only available in low-income countries. Thus, knowledge of how to select and operate the hardware and software available in a certain (geographical) context is in itself an important DL. The study further found that there are often specialized digital technologies for different employment sectors, and that operating specialized digital technologies for a particular field is an important DL. Thus, findings from this study challenge the assumption that DL is technology and task context independent (Law et al. 2018).

Assessment of DL

Dimensionality of DL in assessment frameworks

This study aimed at an empirical analysis of the dimensionality of DL. Therefore, we reviewed the assessment frameworks of DL performance-based assessment studies involving relatively large samples. One of these studies is the ICILS, which is the only international comparative achievement study series that focuses on students' DL-related achievement. The first ICILS was conducted in 2013 and assessed grade 8 students' "ability to use computers to investigate, create and communicate in order to participate effectively at home, at school, in the workplace, and in the community" (Fraillon et al. 2013, p. 18). The ICILS 2013 assessment framework comprised two strands of generic, tool-independent DL constructs, one on collecting and managing information and the other on producing and exchanging information, each further differentiated into three and four aspects, respectively (Fraillon et al. 2013). The assessment constructs at the aspects level were largely retained in ICILS 2018, but restructured into four strands: understanding computer use; gathering information; producing information; and digital communication (Fraillon et al. 2019).

Another international DL-related assessment is the Program for the International Assessment of Adult Competencies (PIAAC) Problem-solving in Technology-rich Environments. This assessment targets adults' ability to use "digital technology, communication tools, and networks to acquire and evaluate information, communicate with others, and perform practical tasks." (OECD 2012, p. 47) PIAAC collected data in a range of countries between 2011 and 2017, and a second cycle will be conducted starting in 2021.

We also identified three DL assessments that were conducted at the national level with a relatively large scale. Australia is a pioneer in this respect as it introduced its triennial NAP-ICT in 2005 (ACARA 2018; MCEETYA 2005). Another DL assessment is "iSkills", developed in the United States to measure college-age students' "ability to appropriately use digital technology, communication tools, and/or networks to solve information problems in order to function in an information society" (Katz 2007, p. 4). Finally, the Information Literacy Performance Assessment (ILPA) study that was conducted in Hong Kong conceptualized DL as the competences required to solve problems effectively using digital means, comprising both cognitive (e.g. literacy, numeracy, problem-solving skills) and technical components (e.g. basic knowledge of hardware, software, and networks; Law et al. 2009).

Dimensionality of assessed DL

Table 1 summarizes the key information from the assessment studies reviewed above. The table shows that all of the assessment frameworks conceptualized two or more theoretical dimensions to guide the design of the DL instruments, but with the exception of ILPA, reported only one score for overall DL achievement. On the other hand, the underpinning analyses that provided the rationale for reporting the overall DL achievement using a single score was only reported in ICILS 2013 (Gebhardt and Schulz 2015) and NAP-ICT (ACER 2008).

While a variety of DL-related assessments other than those listed in Table 1 have been conducted, these have been less influential (Siddiq et al. 2016). They have often focused on fewer aspects of DL, such as information retrieval and/or digital communication (e.g. Aesaert et al. 2014; Spisak 2018), and involved smaller samples (e.g. Markauskaite 2007; Porat et al. 2018; Siddiq et al. 2017). The empirical evidence from smaller scale studies appears to be more varied, but overall still suggests that DL may be unidimensional. For example, Aesaert et al. (2014) conducted a nonlinear factor analysis and found that a unidimensional model fitted their DL data better than a model with multiple factors; many cross-loadings occurred when a two-dimensional model was estimated. Claro et al. (2012) aimed to measure three dimensions but they found that information and communication did not split into two dimensions, whereas the ethical dimension measuring awareness (rather than applied performance) appeared as a separate factor. In analyzing DL data from Korean students, Kim et al. (2019) further found that DL may form a single higher-order factor with sub-dimensions.

There are also studies that provide empirical support for DL as a multidimensional construct. For example, the "Learning in Digital Networks" DL test proposed four DL dimensions (Wilson et al. 2017): functioning as a consumer in networks (e.g. managing information), functioning as a producer in networks (e.g. creating digital products), social capital through networks (e.g. moderating communication), and intellectual capital through networks (e.g. understanding how social networks operate). That test utilized authentic assessment tasks using existing web-based tools (e.g. Google docs), and a multidimensional model was found to fit the data better than a unidimensional model (Siddiq et al. 2017; Wilson et al. 2017).

In a nuanced analysis, Ihme et al. (2017) provided evidence that challenges both the conceptual frameworks *and* the unidimensional measurements of DL. Their analysis examined the relations among the assessment items used in ICILS 2013 to investigate whether these items form one unidimensional DL construct or need to be scaled as a multidimensional construct reflecting several DL dimensions. In analyzing the data from 12 European countries, these researchers noticed a factor structure that showed a possible confounding of item types and item content. Their results gave support for a three-dimensional structure of DL in ICILS 2013, each comprising specific item content and task type: theoretical computer knowledge measured with non-interactive tasks; basic skills requiring procedural knowledge that are measured via interactive simulation tasks; and productive tasks requiring strategic knowledge measured through purpose-built software applications.

It has to be noted that in studies suggesting a multidimensional DL structure, the dimensions were highly correlated (Ihme et al. 2017; Siddiq et al. 2017; Wilson et al. 2017) such that reporting a unidimensional DL scale would be justified. However, these findings also alerted us of the need to be more cautious in adopting unidimensional

Table 1 Conceptualized assessment constructs, task format, and dimensionality of DL achievement reported in selected large-scale DL studies

Study	Assessees targeted	Dimensions in assessment framework	Task formats and technology platform used	Dimensions
ICILS 2013	International, 8th grade students	*2 strands:* collecting and managing information; producing and exchanging information	Multiple choice items, short text responses, and interactive tasks in real-world scenarios using a purpose-built virtual environment	1 DL score
ICILS 2018	International, 8th grade students	*4 strands:* understanding computer use; gathering information; producing information; digital communication	Multiple choice items, short text responses, and interactive tasks in real-world scenarios using a purpose-built virtual environment	1 DL score
ILPA	Hong Kong, 5th and 8th grade students	*7 aspects:* define; access; manage; integrate; create; communicate; evaluate	Authentic tasks using authentic software applications such as the Office suite and browsers	1 DL score (plus numeric scores for each dimension)
iSkills	USA, high school seniors and college students	*7 aspects:* define; access; evaluate; manage; integrate; create; communicate	Authentic tasks using simulated software applications	1 DL score
NAP-ICT	Australia, 6th and 10th grade students	*3 strands:* working with information; creating and sharing information; using ICT responsibly	Multiple choice items, short text responses, and interactive tasks in real-world scenarios using a purpose-built virtual environment	1 DL score
PIAAC PS-TRE	International, 16–65-year olds	*3 dimensions:* tasks (context, complexity, explicitness); technology (devices, applications, functions); cognition (agency and autonomy)	Authentic tasks using simulated software applications	1 DL score

scales stemming from comprehensive DL assessments. Moreover, performance estimates for multiple DL dimensions provide finer-grained feedback to educators and learners enabling more focused training and interventions (Wilson et al. 2018). Further explorations of other DL assessments are also required to examine the dimensionality of DL along aspects other than the conceptual frameworks (Bundsgaard 2019; Ihme et al. 2017).

Task and technology (in)dependence of DL assessments

The above review indicates that the dimensionality of the measured DL may be connected with the design of the assessment instruments and the technologies used in their delivery. Siddiq et al. (2016) reviewed the designs adopted in DL performance assessments. They identified 38 tests from 17 countries. A majority of the reviewed assessments (24 tests) utilized constrained response formats such as multiple choice or fill-in-the-blanks tasks. Only three of the reviewed tests involved solely performance tasks set in an authentic digital environment. These three studies employed either qualitative or mixed-methods designs with small samples. Nine other tests contained dynamic tasks with interactive stimulus and response formats, but most of these tests were administered to relatively small and unrepresentative samples.

In terms of the depictions of DL in popularly known frameworks reviewed earlier, and in the design and description of assessment platforms and instruments, there has been an implicit assumption that DL is a generic competence independent of the devices, tasks and technologies used. Most large-scale studies that have utilized performance-based assessments in a virtual environment relied on a purpose-built digital environment mimicking existing software applications at a lower level of complexity (e.g. ACARA 2018; Fraillon et al. 2014; Hohlfeld et al. 2013; Katz 2007). Each of the five large-scale assessments listed in Table 1 emphasized authenticity, and four of them used simulated software and/or purpose-built digital platforms to mimic scenarios and software applications that individuals may encounter in real-world situations (sometimes supplemented with additional constrained response items). This is aligned with the conceptualization of DL as a generic competence and the assumption that task completion would be independent of the specific software applications used for the task. Large-scale performance-based assessments of DL using authentic software applications commonplace in everyday life have been rare, and ILPA is one such exception that adopted real-world software applications commonly used at the time of the study to deliver the assessment (Law et al. 2009).

As competences develop through real-life experiences, it can be argued that assessments need to reflect real-life situations as closely as possible to examine dimension-specific competences (see also Fraillon 2018). How far DL as assessed through virtual environments specifically developed for the purpose of assessment reflects a person's competence when interacting with authentic software available in the commercial market remains unclear. Yet recent evidence, such as moderate correlations among DL performance shown with different commercial software tools in a longitudinal study (Lazonder et al. 2020) and the task dependence identified among ICILS items (Ihme et al. 2017), challenge the assumptions of task and tool independence of DL.

Research challenges to be addressed

Our review of assessment systems showed the existence of different assumptions about the competences comprising DL, the task dependence (or otherwise) of DL, as well as whether the technology platform used (i.e. authentic tools or simulated environments) matters. This led us to identify two research challenges that need to be addressed: First, *is DL a unidimensional or a multidimensional construct?* (RQ1) Second, *is DL a generic competence that is stable across task and tool contexts?* (RQ2)

As mentioned in the introduction, there is also a lack of studies on the development of DL from primary to secondary school age. It is possible that the answers to RQ1 and RQ2 could differ depending on the maturity of the assessee. Thus, this study also explored the following research question: *are there discrepancies between primary and secondary students regarding the dimensionality of DL and/or whether DL is a generic competence?* (RQ3) Note that a valid assessment enables researchers to compare performance differences across age cohorts.

The present study investigated these research questions through a secondary analysis of the underexplored data collected from ILPA using authentic software applications popularly used in school, personal and business contexts.

Research design and methodology

Context and design of the study

Our research was based on the ILPA study data from Hong Kong (Law et al. 2007, 2009). ILPA was commissioned in 2006 by the Hong Kong SAR government and conducted in 2007 as part of a territory-wide evaluation of the implementation of the "IT in Education Strategic Plans" since 1998 (Law et al. 2009). The assessment developers adopted the OECD's definition of information literacy (referred to as DL in this paper) as:

"The interest, attitude and ability of individuals to appropriately use digital technology and communication tools to access, manage, integrate and evaluate information, construct new knowledge and communicate with others in order to participate effectively in society." (Lennon et al. 2003, p. 8).

ILPA assessed two age cohorts of students. For grade 5, three assessments were administered on DL in addressing problems in *general* everyday life contexts (referred to as technical DL), in learning Chinese language (Chinese language DL), and mathematics (mathematics DL). Three separate assessments were also administered at grade 8: technical DL, Chinese language DL, and scientific DL. Here we only discuss the design of the technical DL assessment instrument, which was identical for both age cohorts of students as the researchers considered it possible that the differences in technical DL across students may be dependent on their experience in digital technology use both within and outside of school, and less influenced by their academic knowledge and skills. Microsoft® Windows Terminal Server (WTS) was used to provide the same technology environment to all students, who took the assessment in the computer laboratories in their own schools. In completing the assessment tasks, students were directed to use common Microsoft® Office applications and a web browser available through the WTS. This way, students were able to demonstrate their ability

Table 2 Definition of the seven aspects in the ILPA framework (Law et al. 2007, p. 7)

Aspect	Definition
Define	Using ICT tools to identify and appropriately represent information needs
Access	Collecting and/or retrieving information in digital environments
Manage	Using ICT tools to apply an existing organizational or classification scheme for information
Integrate	Interpreting and representing information, such as by using ICT tools to synthesize, summarize, compare and contrast information from multiple sources
Create	Adapting, applying, designing or inventing information in ICT environments
Communicate	Communicating information properly in its context (audience and media) in ICT environments
Evaluate	Judging the degree to which information satisfies the needs of the task in ICT environments, including determining authority, bias and timeliness of materials

to address the DL tasks using a technology interface and tools that were available and commonly adopted in schools and at home during the time of the assessment.

Although the ILPA study was conducted in 2007, we utilized these data for several reasons. First, the seven aspects covered in ILPA, shown in Table 2, are still important today, as suggested by Law et al.'s (2018) review of DL frameworks. These aspects overlap substantially with the DL dimensions included in the assessment frameworks adopted in ICILS and PIAAC, as shown in Table 1.

Second, the assessment instruments used in large-scale international studies often remain confidential, making secondary analyses to investigate the dimensionality of DL unfeasible. Assessments that provide details of the assessment tasks were often conducted with small and unrepresentative samples (e.g. Siddiq et al. 2017; Porat et al., 2018). In contrast, the ILPA data enabled us to conduct secondary analysis of a large random sample with full details of each item.

A third, valuable feature of the ILPA data is that rather than mimicking existing software applications at a lower level of complexity, it utilized authentic software applications that students would use in their daily lives. This contrasts with most large-scale DL assessments, which utilized simulated instead of real-world software applications (e.g. Fraillon et al. 2019; OECD 2016). Thus, the ILPA data may have more value in providing information on students' DL in handling everyday tasks than assessments utilizing purpose-built software tools. The tools used in ILPA also appear to be more contemporary than assessments using purpose-built simulated tools. However, as the primary goal of ILPA was to evaluate an education policy, no thorough dimensional analysis was conducted on the collected data.

An additional benefit of analyzing the ILPA data is that the same assessment instrument was administered to a cohort of primary school students *and* a cohort of secondary students. Dimensional analysis of DL across two age cohorts allows for a more robust examination. If the same dimensional structure were to be found across cohorts, we would have more confidence in the accuracy and generalizability of the results.

Given these benefits and the context of the ILPA study, secondary analysis was conducted on the assessment data to answer the three RQs that were summarized in the previous section.

Participants

The ILPA study sampled P5 (equivalent to grade 5) and S2 (equivalent to grade 8) students in the 2006/2007 academic year as the target populations. Stratified random samples of schools were drawn based on school size and the mean academic ability of students in those schools, yielding a sample of 40 primary and 33 secondary schools (i.e. school response rates of 27% and 24%).[1] At each school, one full class was randomly sampled, which resulted in an assessment of 1320 primary and 1302 secondary school students (i.e. student response rates of 99% and 100%). Each student had to complete only two of the three ILPA assessment modules administered at the relevant grade level. Hence, only two thirds of the sampled students took the technical DL assessment, and the analyses reported below were carried out on data from the 830 P5 and 823 S2 students who took the technical DL assessment module. Most P5 students were 10 or 11 years old (90%; 1% of P5 students were younger than 10 years and 6% were 12 years or above), 46% were female and 51% were male. Most S2 students were 13 or 14 years old (85%) and some were 15 years or above (11%; 2% of the S2 students were younger than 13 years). Among the S2 students, 42% were female and 56% were male.[2]

Instrumentation

Assessment instruments were developed around the seven aspects in Table 2 to assess DL outcomes at P5 and S2 levels. For each assessment module, students had 45 min to complete all tasks. In the technical DL module, students were asked to plan a trip for their grandparents to visit Hong Kong. The module required students to complete four assessment tasks: search for tourist information on the Internet; reorganize search results on scenic information in a Microsoft® Word document; create a Microsoft® PowerPoint presentation on the proposed trip; and discuss scenic spots in an online forum. Table 3 provides a brief summary of all subtasks; a comprehensive summary of all tasks can be found in online Appendix 1. All instruments were piloted rigorously before the main data collection, which was conducted from December 2006 to April 2007.

In the main study, assessment data were collected online in schools and scored by human markers using scoring rubrics, except for a few items that could be scored automatically. The scoring rubric (see online Appendix 2) defined four performance levels: "novice," "basic," "proficient," and "advanced." Some items differentiated only between novice and basic; a few between novice, basic, and proficient; and others between all four performance levels. An inter-coder reliability of $r = .98$ across all schools shows that the scoring was very consistent and reliable.

[1] School response rates were affected by: a lack of time; confusion as there was another evaluation on the aforementioned education strategy; as well as a lack of school infrastructure and technical support in schools.

[2] Age and/or gender information was unavailable for 3% of the P5 students and for 2% of the S2 students.

Table 3 Task description and aspects of technical DL performance (Law et al. 2007, p. 32)

Question	Task description	Aspect	Score
Q1 Students were asked to search 2 scenic spots from the Internet			
Q1.1	Identify appropriate search engine	Access	2
Q1.2	Define appropriate search terms	Define	3
Q1.3	Identify appropriate websites	Access	1
Q1.4.1a, Q1.4.2a	Access information on appropriate scenic spots from websites	Access	3 (each)
Q1.4.1b, Q1.4.2b	Evaluate reasons to support the suggested scenic spots	Evaluate	3 (each)
Q2 Students were asked to edit a Word® document for their groupmates			
Q2.1	Save a document to an appropriate folder	Manage	1
Q2.2	Reorganize the information as required	Manage	6[a]
Q2.3	Design and enhance the document formatting using proper tool functions	Create	3
Q3 Students were asked to create a PowerPoint® for presentation			
Q3.1	Save a document to an appropriate folder	Manage	1
Q3.2	Summarize information found on the Internet	Integrate	6[a]
Q3.3	Evaluate and retrieve appropriate information found on the Internet	Evaluate	6[a]
Q3.4	Design and enhance the presentation using proper tool functions	Create	6[b]
Q.4 Students were asked to post ideas and discuss with their classmates in the forum			
Q4	Post ideas and discuss with students in the forum	Communicate	3

[a]Partial (non-integer) scores were possible for this task

[b]Although scores up to six were possible, no student obtained a score higher than three

Data analysis

Recoding

Due to the scoring strategy some assessment items had up to 13 categories (i.e. any full or half point between zero and a maximum of six points). This can cause estimation errors owing to empty cells in bivariate frequency tables. Moreover, the performance levels assessed by the different items varied, measuring different ranges of the four performance levels, which made it more difficult to interpret the results and to compare performance across items. To enable robust dimensional analyses and stable estimations, we therefore recoded all items into dichotomous variables distinguishing between novice (coded "0") and at least basic (coded "1") performance. In addition, items Q3.2 and Q3.3 were collapsed into a new variable due to an empty cell in their bivariate frequency table (if performance on at least one of both variables was at the novice level, the new variable was coded zero, otherwise it was coded one).

Item response modeling

Item response theory (IRT) has been recognized as a particularly useful technique to model competences. In IRT, an individual's competence and item characteristics determine the probability of a correct response to a particular assessment item (Embretson and Reise 2000). An advantage of IRT models is that the item parameters are independent of the sampled respondents, and the person parameters are independent of the selected items (Embretson and Reise 2000). Furthermore, IRT has been designed specifically to model binary measurement items (de Ayala 2013), and researchers increasingly and successfully implement IRT analyses to answer questions related to ICT in education (e.g. Aesaert et al. 2014; Claro et al. 2012; Siddiq et al. 2017; Wilson et al. 2017).

Researchers often choose a one-parameter logistic (1PL or Rasch) model or a two-parameter logistic (2PL or 2PL Birnbaum) model when analyzing binary items. 1PL models only include a difficulty parameter for each item which locates the items on the competence scale, where the probability of a correct response is 50% (more difficult items require higher levels of competence). 2PL models assume that items additionally vary in terms of the extent to which they can discriminate among respondents with different ability levels. Therefore, 2PL models also estimate a discrimination parameter for each item, which indicates how well an item separates individuals into different competence levels (de Ayala 2013).

While common IRT models require that the analyzed items are sufficiently unidimensional, multidimensional IRT (MIRT) enables researchers to analyze more complex data structures. The assumption that all items measure the same and only one competence can be unrealistic in reality and MIRT can be a solution in these situations (Aesaert et al. 2014; Siddiq et al. 2017); it enables researchers to examine the (hypothesized or unknown) structure of a set of binary items (Chalmers 2012). Consequently, MIRT may provide a more realistic approach for investigating the dimensionality of DL.

Analysis strategy

We conducted a series of IRT and MIRT models for the whole student sample (indicated as "total sample"), and separately for P5 and S2 students, using "mirt" in R (Chalmers et al. 2017). We first fitted a unidimensional IRT model to examine whether technical DL can be understood as one construct (RQ1). This was in line with other studies (e.g. Aesaert et al. 2014; Gebhardt and Schulz 2015), and with the pragmatic approach chosen by the investigators of the original ILPA study, who summed the scores across all items to measure overall information literacy (Law et al. 2007). In addition, a series of exploratory MIRT models was fitted to examine whether technical DL needs to be conceptualized as a multidimensional construct, and if so, how many dimensions it comprises. Exploratory models can be employed to compare multidimensional DL models with a hypothesized unidimensional model of DL and/or when a hypothesized dimensional model does not yield acceptable fit. The loading matrices of the multidimensional DL models were rotated to find the most parsimonious structure in which each item ideally reflects exactly one dimension as indicated by the size of their loadings (Chalmers 2012), and to make the dimensions more clearly interpretable and distinguishable from each other. An orthogonal rotation procedure (Varimax) with uncorrelated dimensions was chosen due to the intention to extract clear and distinct patterns.

When an MIRT model fits the data better than a unidimensional IRT model, it is necessary to interpret the dimensions based on the content of each item, the task to which the item belongs, and the software application used. A bifactor model that combines a unidimensional model with a multi-dimensional model can be a useful follow-up analysis as it allows us to maintain a unidimensional structure, which has been utilized in other studies (Claro et al. 2012; Fraillon et al. 2014), while accounting for significant coherence among groups of items. It is a hierarchical MIRT model which estimates a *general dimension* measured by all items, thus accounting for *overlap among all items* and representing general DL performance as the latent variable of central interest (von Davier and Khorramdel 2013). In addition, a bifactor model accounts for local dependencies among identified subsets of items. These subsets are also called "specific," "residual," or "unique" dimensions, because they capture commonality ("specific variance") among items not accounted for by the general dimension; these *independent common latent factors* consequently reflect *additional associations between subsets of items* (Gibbons et al. 2007). This approach is particularly useful when subsets of items measure separate dimensions that might depend on the tool or task (also known as "common method bias;" Podsakoff et al. 2003). Furthermore, a bifactor model is well-suited to increase the parsimony of the dimensional structure (Chalmers 2012; Gibbons et al. 2007). If the initial unidimensional IRT model only inadequately reflects the data, a bifactor model linking the unidimensional IRT model and the MIRT model would enable us to answer both RQ1 and RQ2.

Multiple-group analysis was employed to test whether measurement invariance of DL can be established across P5 and S2 students (RQ3). The existence of measurement invariance provides evidence for the validity of a model by implying that the dimensional structure of DL is identical in both age cohorts. Furthermore, measurement invariance is a requirement for mean comparisons across groups based on the same measurement instrument. To do this, we examined the bifactor model for each cohort of students and then placed constraints on both models to determine whether the dimensions measured the same construct in both groups and if the performance levels were comparable between P5 and S2 students. A fully constrained model was estimated as the baseline model (full measurement invariance: all parameters, including the mean levels of performance, were constrained and identical across P5 and S2 students), and compared against a model with only different performance levels in both groups (unconstrained means) as well as against a model with complete measurement non-invariance (i.e., all parameters could differ between P5 and S2 students, which would imply non-comparability of the performance scores).

All analyses were conducted using two-parameter logistic models (i.e. item difficulty and item discrimination could vary between items).[3] The models were estimated by means of the quasi-Monte Carlo Expectation–Maximization algorithm. This method maximizes local approximations of the likelihood function using an iterative procedure to return a solution that most likely reflects the data (Chalmers 2012); it can handle higher-dimensional models effectively and more accurately than standard Expectation–Maximization estimations (Chalmers et al. 2017).

[3] One-parametric models were estimated but were inferior to two-parametric models.

Table 4 Fit of unidimensional and multidimensional IRT models

Dimensions	LL	$\Delta\chi^2$	SABIC	RMSEA	SRMSR	CFI
P5 students						
1	−4,993		10,086	.151	.137	.700
2	−4,621	745.37***	9,387	.113	.073	.859
3	−4,437	367.61***	9,061	.074	.048	.951
4	**−4,397**	**78.74*** **	**9,022**	**.061**	**.065**	**.974**
5	−4,427	−59.91	9,117	.060	.091	.981
6	−4,382	91.46***	9,058	.056	.110	.988
S2 students						
1	−4,138		8,376	.142	.156	.680
2	−3,866	545.71***	7,876	.083	.072	.910
3	−3,747	237.62***	7,681	.050	.043	.973
4	**−3,719**	**56.00*** **	**7,664**	**.034**	**.046**	**.990**
5	−3,728	−17.91	7,717	.039	.060	.990
6	−3,738	−21.68	7,771	.085	.117	.967

The overall analysis yielded the same pattern. LL: Log-Likelihood; Δx^2: X^2 difference between two models

SABIC sample-size adjusted Bayesian Information Criterion, *RMSEA* Root Mean-Square Error of Approximation, *SRMSR* Standardized Root Mean Square Residual, *CFI* Comparative Fit Index

*$p < .05$, **$p < .01$, ***$p < .001$

Table 5 Dimensional structure of the exploratory four-dimensional MIRT model (loadings)

	P5 students					S2 students				
	F1	F2	F3	F4	h^2	F1	F2	F3	F4	h^2
Q1.1	**0.34**	0.21	0.22	0.13	0.23	**0.45**	0.27	0.13	0.15	0.32
Q1.2	**0.85**	0.39	0.22	0.01	0.92	**0.63**	0.42	0.22	0.12	0.63
Q1.3	**0.65**	0.40	0.19	0.03	0.62	**0.74**	0.20	0.29	0.13	0.69
Q1.4.1a	0.09	**0.98**	0.02	0.14	1.00	0.12	**0.84**	0.23	0.08	0.77
Q1.4.1b	0.25	**0.86**	0.15	0.17	0.84	0.06	**0.70**	0.11	0.08	0.52
Q1.4.2a	0.30	**0.78**	0.33	0.23	0.87	0.13	**0.98**	0.03	0.13	0.99
Q1.4.2b	0.36	**0.74**	0.32	0.24	0.83	0.12	**0.98**	0.12	0.12	1.00
Q2.1	0.12	0.26	**0.84**	0.37	0.92	0.19	0.24	**0.86**	0.38	0.98
Q2.2	0.26	0.19	**0.85**	0.42	1.00	0.13	0.05	**0.91**	0.39	1.00
Q2.3	0.07	0.01	**0.59**	0.11	0.36	0.00	0.34	**0.47**	0.09	0.34
Q3.1	0.01	0.05	0.32	**0.94**	0.99	0.08	0.10	0.32	**0.94**	1.00
Q3.23	0.08	0.32	0.22	**0.92**	1.00	0.07	0.15	0.32	**0.92**	0.97
Q3.4	-0.03	0.20	0.15	**0.81**	0.72	0.06	0.15	0.14	**0.63**	0.44
Q4	0.02	0.05	0.08	0.28	0.09	0.00	-0.09	0.12	0.20	0.07

The overall analysis yielded the same pattern. Q3.23 is recoded from Q3.2 and Q3.3. F1, F2, F3, and F4 refer to dimensions one to four (*F1*: online information seeking; *F2*: knowledge-based information seeking; *F3*: word processing; *F4*: digital presentation); h^2: commonality

Results

Unidimensional and multidimensional models

First, we inspected unidimensional IRT models to examine whether technical DL could be construed as a single construct. However, these measurement models did not fit the data well (see Table 4): technical DL as measured in ILPA was apparently not a unidimensional construct but may be comprised of several dimensions. Therefore, a series of exploratory MIRT models was estimated to examine whether technical DL needs to be conceptualized as a multidimensional construct, and if so, how many dimensions technical DL comprises. The four-dimensional exploratory MIRT model provided the best and acceptable model-data fit (rows printed in bold in Table 4) and we concluded that DL as measured in the ILPA study likely tapped four dimensions.

The structure of the four-dimensional models was further examined using Varimax rotation. The analysis showed that each of the four dimensions comprised at least three items with acceptable loadings on at least one dimension (shown in bold in Table 5). Two of these items (Q1.1 and Q2.3) had quite low commonalities (i.e. below .50, which means that the four dimensions together accounted for less than half of the variance in these two items; MacCallum et al. 1999), despite their acceptable loadings. As the average commonality was above .70, we decided to drop only the last item from subsequent analyses, as it did not match any of the four dimensions.[4]

Table 5 also suggests that the dimensions primarily related to the tasks and tools utilized in ILPA. Specifically, the first three items represented one dimension that reflects students' *online information seeking* competence. The next four items, although related to information seeking, were quite *knowledge-based* in that these items could be easily answered by students without any search for information if they are well-aware of scenic spots in their city. Hence, it makes sense that these items formed a separate dimension that is independent of the search for information on the Internet. The next three items, which required students to process information in Microsoft® Word, formed another dimension labelled *word-processing DL*. The following three items stemmed from the task in which students were asked to process information using Microsoft® PowerPoint and were labelled *digital presentation*. Finally, and as mentioned before, the last item on communicating information online did not load on any dimension.

Bifactor models

We then examined the structure of thirteen items in a confirmatory bifactor model combining a unidimensional model with the four-dimensional model. This procedure yielded a model with a global dimension and four specific dimensions (reported for the overall sample; multiple group comparisons are shown below). The bifactor model had an acceptable to good model fit (RMSEA = .052, CFI = .978, SABIC = 15,963) and performed better than a simple four-dimensional model without a general dimension (RMSEA = .096, CFI = .909, SABIC = 16,334). The general dimension and the four specific dimensions together accounted for a moderate 71.4% of the total variance. This bifactor model was therefore

[4] The average commonality would have been below that threshold for secondary school students had we not dropped the last item.

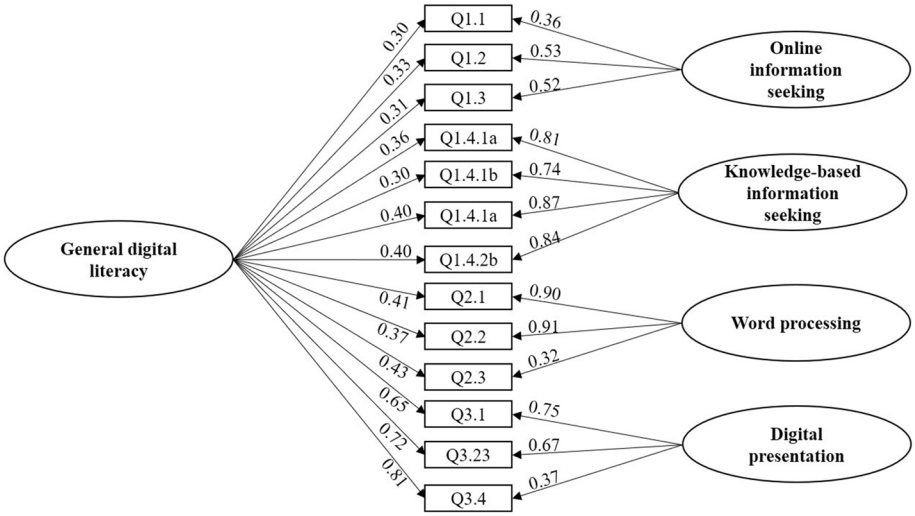

Fig. 1 Bifactor model with unconstrained means. Note. Numbers at the arrows are loadings, which are constrained to be the same in both age cohorts

Table 6 Reliabilities of the general and specific dimensions (bifactor model with unconstrained means)

Dimension	ω	ω_H / ω_{HS}	*H*
General DL	.93	.58	.84
Online information seeking	.59	.41	.48
Knowledge-based information seeking	.94	.78	.90
Word processing	.89	.67	.90
Digital presentation	.97	.39	.69

Omega (ω) is the proportion of observed variance attributable to all modeled sources of common variance; omega hierarchical (ω_H) estimates the proportion of variance attributable to general DL and omega hierarchical subscale (ω_{HS}) reflects the unique variance attributable to each specific dimension; and construct reliability (*H*) indicates how well each dimension is represented by its indicators (Rodriguez et al. 2016)

further examined using multiple-group analysis to examine its invariance between P5 versus S2 students.

Multiple-group models

Multiple-group analysis was employed to test whether measurement invariance of DL could be established across P5 and S2 students. The assumption of full measurement invariance was not quite supported (RMSEA = .056, CFI = .914, SABIC = 15,963), but a model where all parameters except the mean levels of student performance were held equal showed a much better fit to the data (RMSEA = .048, CFI = .931, SABIC = 15,710). Although a model with complete measurement non-invariance (which would imply that students' performance cannot be directly compared between both age cohorts) performed

even better than a partially invariant model with non-invariant means (RMSEA = .045, CFI = .960, SABIC = 15,678), the partially constrained model performed just fine and yielded acceptable model fit. Hence, the same dimensional structure of DL was adopted for both age cohorts (Fig. 1).

Table 6 shows that the general dimension and three specific dimensions (knowledge-based information seeking, word processing, and digital presentation) were highly reliable. Online information seeking was somewhat less reliably measured. The ratio of ω_H to ω also shows that only 62.2% of the reliable variance in the total scores was due to general DL, aiding the interpretation of DL as a multidimensional (bifactor) construct. An examination of ω_{HS} further indicated that the common variance among the items declined when separating it from the variance for general DL. Specifically, much of the common variance among the items measuring digital presentation was due to general DL. Therefore, its reliability mostly is attributable to individual differences on general DL.

The adequate fit of the partially constrained model justified comparing the performance levels between P5 and S2 students based on this model. Although students in the older cohort performed much better on average compared to younger students, the mean differences of the standardized scores suggest huge overlaps in the performance levels of P5 and S2 students (positive mean differences imply that S2 students performed better than P5 students). This difference was particularly evident for the general DL performance ($\Delta M \geq .99$), as well as for general information seeking on the Internet ($\Delta M \geq 1.26$). The difference was less pronounced for knowledge-based information seeking ($\Delta M = .23$), as well as for word processing-related DL ($\Delta M = .30$). Surprisingly, P5 students seemed to perform slightly better than S2 students with respect to presentation-related DL ($\Delta M = -.41$). This "peculiarity" was due to the general dimension of DL though: As shown in Fig. 1 and in Table 6 (ω_{HS}), the three variables reflecting the presentation-related DL were hugely influenced by the general dimension *when assuming partial measurement invariance* ($f_{ig} \geq .65$). Therefore, the differences in these items were already captured in the general dimension, and the presentation-specific dimension of DL as a residual factor influenced the performance in the respective items beyond what was already explained by the general dimension. Also, it is noteworthy that item Q3.4, of which the entire range of possible scores was not even nearly exhausted (Table 2), had a substantially smaller loading on the specific dimension in comparison to its loading on the general dimension ($f_{i4} = .37$ vs. $f_{ig} = .81$), as well as compared to the other two items that reflect presentation-related DL ($f_{i4} \geq .67$ vs $f_{ig} \leq .72$). Therefore, these three items mainly measured the general dimension, and the specific dimension was primarily influenced by only two of the three presentation-related items. However, this peculiarity can also be interpreted as the partial measurement invariance model might imposing too strong an assumption.

Discussion and conclusions

Dimensionality and tool dependence of DL

Our exploration of the ILPA data yielded results that shed new light on the complexities in assessing DL which to-date have not been uncovered. It also raises questions for further research on the dimensional structure of DL in relation to the apparent non-neutrality of tool use on achievement as defined in generic constructs.

While we confirmed the likely existence of a general or "global" dimension reflecting the "essence" of DL, the specific "residual" dimensions identified do not map well onto existing DL frameworks (RQ1). Instead, those dimensions seem to be defined by the software applications that students used in the assessment; they capture commonality among students' performance that seems to be due to their familiarity with the assessment tools and/or the context (RQ2). These results extend recent research that indicate item types may constitute measurable DL dimensions (Ihme et al. 2017), suggesting that the measurement of DL performance may need to take account of the specific software applications used in the assessment. The existing theoretical frameworks on the dimensionality of DL may need revision, as most have assumed that DL is a generic competence, similar to reading literacy and numeracy. This is particularly problematic for DL assessment frameworks, as these have not sufficiently considered the possible effect of software applications and contexts. Although performances in different item types or software applications are likely to be positively associated with each other (Ihme et al. 2017; Lazonder et al. 2020) and, as in ILPA, may form a general or higher-order factor (Ihme et al. 2017; Kim et al. 2019), knowledge about tool-specific performance can enable teachers and learners to better channel their educational efforts (Wilson et al. 2018).

The finding that a general latent DL is supplemented by several tool-specific DLs thus also has implications for educators. While it is important to understand DL as a comprehensive competence covering multiple dimensions across tools and contexts, educators should make use of multiple tools for learning in educational contexts and when assessing learning outcomes via digital means, as tool-specific performances may not always correlate highly (see Lazonder et al. 2020). This particularly applies to DL, but assessments of other literacies, such as mathematical competences, writing ability etc., might also be biased if teachers use digital tools with which some students are less familiar than other students. Importantly, reliance on only one specific tool is particularly likely to provide biased results of students' DL and should be avoided unless the assessment of a tool-specific competence is the explicit goal. Teachers as well as researchers who develop digital assessments, including but not limited to DL assessments, need to consider potential tool dependencies in the analysis and when interpreting assessment results.

Finally, schools prepare young people for participation in society, and DL should thus constitute an important educational outcome. Our study indicates that familiarity with commonly used tools contributes importantly to the preparation of students for their future lives and jobs. Hence, teachers, school leaders, and education policy decision-makers need to provide access to the prevalent tools to develop the fluency and confidence needed for accomplishing tasks using digital tools in the context of everyday and context specific problem-solving scenarios as an integral part of DL educational programs. The development of technological fluency as an integral part of DL is also essential, because individuals need to develop the capacity to use digital tools they have not encountered to accomplish tasks in new contexts, as digital devices and applications change rapidly.

Cohort differences in DL performance and tool dependence of DL

Our analysis shows that the dimensions of DL are reflected by the same items in two different cohorts of students, which lends support to the validity of the bifactor structure of DL (RQ3). Overall, S2 students performed better than the younger cohort of P5 students. However, these differences varied across the DL dimensions, and there was a substantial overlap in the performance profile of the primary and secondary student samples. Both findings

compare well with other studies which found overlap in the performance levels of students in different grades (ACARA 2018; Jin et al. 2020; Lazonder et al. 2020).

The greatest performance gap between the two age cohorts in our analysis was in the area of online information seeking. When ILPA was conducted in 2007, Internet-enabled mobile devices such as smartphones were much less common, and most P5 and S2 students at the time probably accessed the Internet via computers at school or at home. Further, the search for information on the Internet as a learning task set by teachers at school was probably more likely for S2 students. It is not clear to us that we would be able to observe similar differences in performance for this same cluster of items between the two age cohorts if the assessment was conducted today, when many P5 students in Hong Kong have access to an Internet-enabled smartphone (see also Law et al. 2018).

Another question is whether these same performance patterns would have been observed if the assessment was not conducted with Microsoft® Office but on unfamiliar software applications that have similar functions, as in the case of the ICILS (Fraillon et al. 2019, 2014). There are likely some generic and transferable skills in completing the assessment tasks that are tool independent, but how can such generic competences be disentangled from students' competence and familiarity with a specific tool? Further, what are the exact constructs that are being measured if the assessment environment does not adopt real-life tools? A technology environment that is built solely for the purpose of assessment is likely to be relatively primitive, modeled on dated versions of tools that were popular in the market, and hence cannot assess students' ability to handle tasks using common software applications that have more complex functions and sophisticated interfaces at the time of assessment. Questions of construct validity are inherent whichever type of tool is adopted in an assessment of DL (Fraillon 2018). Software applications and devices that are available on the market and used by individuals in their daily lives can enable them to achieve the same DL in different ways (Law et al. 2018). This also begs the question of how one handles the tension between the generic formulation of DL in curriculum and assessment frameworks, and the fact that DL can only be demonstrated through specific tool use contexts.

We argue that decisions about assessment design should take account of students' prior experiences with the devices and software applications involved. Statistical solutions exist that can take test takers' familiarity with software tools and applications into account. Future research could investigate the impact of tool-specific assessment tasks on DL performance using contemporary software applications beyond Microsoft® Office, and a longitudinal study design could examine how generic and specific DLs develop over time. One challenge for such an endeavor would be the establishment of measurement invariance across all cohorts, as the findings of the current analysis tentatively indicate that—despite the structural similarities of DL among primary and secondary school students—some measurement properties of identical tasks and items might be non-invariant across different age cohorts.

Limitations

A few limitations of our study need to be addressed. First, the ILPA was conducted a decade ago. While only few studies have collected data from primary *and* secondary students using relatively large samples (e.g. ACARA 2018), which clearly is a strength of the ILPA, many technological changes have taken place during this period, which limits the generalizability of its findings to today's society. However, the utilization of authentic, real-world software applications in the assessment lets the ILPA study appear more contemporary

than other assessments, and despite significant changes Microsoft® Office is still one of the most commonly used applications. Moreover, most of the aspects assessed in the ILPA are constantly included in governmental and non-governmental notions of DL (Law et al. 2018), providing evidence of the continuing importance of these aspects.

Second, the assessment tasks had a fairly small item pool of primarily information-related DL, which somewhat constrained our analyses. The recommendation for future studies is to clearly define the aspects of DL that shall be assessed and to use a larger set of items in order to have adequate data points to establish the reliability of each of the dimensional scales. Third, and related to the previous point, the ILPA dataset included only one item measuring digital communication. Our model did not properly capture communication using digital means as this singular item performed poorly in the analyses. Therefore, researchers are advised to include multiple items measuring a common dimension in their instruments.

Conclusion

Secondary analysis of representative data collected from primary and secondary school students in Hong Kong using unidimensional and multidimensional IRT shows that DL is not a generic competence that is independent of the tasks and tools used (see also Ihme et al. 2017). It most likely comprises a general dimension as well as specific dimensions that appear to be tool-dependent, suggesting that conceptualizations of DL as an essentially unidimensional competence must take account of tool-specific factors on DL performance. These findings shed light on the importance of developing context-relevant instruments for assessing DL. Good practices in designing DL assessment instruments should consider respondents' experiences and fluency with digital tools (Law et al. 2018). Future work should also consider the relationship between respondents' subject matter knowledge and familiarity with discipline-specific tools and their demonstrated DL in discipline-specific problem-solving contexts. These are essential aspects that all future studies need to take into consideration, for instance, when deciding about the use of large versus small assessment tasks, or whether existing tools and platforms versus tools and platforms that are just developed for the purpose of the assessment should be used.

Acknowledgments This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region [Project No. T44-707/16-N]. The authors also would like to acknowledge the support received from the Centre for Information Technology in Education at the Faculty of Education of The University of Hong Kong.

Author contributions FR drafted the manuscript, supervised the statistical analysis and contributed with the interpretation of the results; JZ conducted the statistical analysis and contributed with the writing and the interpretation of the results; NL was the idea originator of the paper, contributed with the literature review, the results interpretation and the writing of the paper; GW assisted with the literature review and the writing of the manuscript; JdlT coordinated the statistical analysis and contributed with the interpretation of the results and the writing.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Aesaert, K., van Nijlen, D., Vanderlinde, R., & van Braak, J. (2014). Direct measures of digital information processing and communication skills in primary education: Using item response theory for the development and validation of an ICT competence scale. *Computers & Education, 76*, 168–181.

Australian Council for Educational Research (ACER). (2008). *National Assessment Program Information and Communication Technology literacy 2005 year 6 and 10: Technical report*. Sydney: ACER.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2018). *NAP sample assessment ICT literacy: Years 6 & 10*. Sydney: ACARA.

Bawden, D. (2008). Origins and concepts of digital literacy. In C. Lankshear & M. Knobel (Eds.), *Digital literacies: Concepts, policies and practices* (pp. 17–32). New York: Peter Lang.

Bundsgaard, J. (2019). DIF as a pedagogical tool: Analysis of item characteristics in ICILS to understand what students are struggling with. *Large-scale Assessments in Education, 7*, 69.

Carretero, S., Vuorikari, R., & Punie, Y. (2017). *DigComp 2.1: The Digital Competence Framework for Citizens with eight proficiency levels and examples of use* (No. JRC106281). Joint research centre (Seville site). Retrieved from https://publications.jrc.ec.europa.eu/repository/bitstream/JRC-106281/web-digcomp2.1pdf_(online).pdf.

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K.-H., Falk, C. F., & Meade, A. (2017). Multidimensional item response theory: Package "mirt". Version 1.23. Retrieved from https://cran.r-project.org/web/packages/mirt/mirt.pdf.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Claro, M., Preiss, D. D., San Martín, E., Jara, I., Hinostroza, J. E., Valenzuela, S., et al. (2012). Assessment of 21st century ICT skills in Chile: Test design and results from high school level students. *Computers & Education, 59*, 1042–1053.

de Ayala, R. J. (2013). The IRT tradition and its applications. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 1, pp. 144–169). New York: Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum.

Fraillon, J. (2018). International large-scale computer-based studies on information technology literacy in education. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *Second handbook of information technology in primary and secondary education* (pp. 1161–1179). Cham: Springer.

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *International Computer and Information Literacy Study 2018: Assessment framework*. Amsterdam: IEA.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age. The IEA International Computer and Information Literacy Study*. Cham: Springer.

Fraillon, J., Schulz, W., & Ainley, J. (2013). *International Computer and Information Literacy Study: Assessment framework*. Amsterdam: IEA.

Gebhardt, E., & Schulz, W. (2015). Scaling procedures for ICILS test items. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt (Eds.), *ICILS 2013 technical report* (pp. 155–175). Amsterdam: IEA.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., et al. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4–19.

Haßler, B., Major, L., & Hennessy, S. (2016). Tablet use in schools: A critical review of the evidence for learning outcomes. *Journal of Computer Assisted Learning, 32*, 139–156.

Hohlfeld, T. N., Ritzhaupt, A. D., & Barron, A. E. (2013). Are gender differences in perceived and demonstrated technology literacy significant? it depends on the model. *Educational Technology Research and Development, 61*, 639–663.

Ihme, J. M., Senkbeil, M., Goldhammer, F., & Gerick, J. (2017). Assessment of computer and information literacy in ICILS 2013: Do different item types measure the same construct? *European Educational Research Journal, 16*, 716–732.

Jin, K.-Y., Reichert, F., Cagasan, L. P., de La Torre, J., & Law, N. (2020). Measuring digital literacy across three age cohorts: Exploring test dimensionality and performance differences. *Computers & Education, 157*, 103968. https://doi.org/10.1016/j.compedu.2020.103968.

Katz, I. R. (2007). Testing information literacy in digital environments: ETS' iSkills assessment. *Information Technology and Libraries, 26*, 3–12.

Kim, H. S., Ahn, S. H., & Kim, C. M. (2019). A new ICT literacy test for elementary and middle school students in Republic of Korea. *Asia-Pacific Education Researcher, 28*, 203–212.

Law, N., Lee, Y., & Yuen, A. H. K. (2009). The impact of ICT in education policies on teacher practices and student outcomes in Hong Kong. In F. Scheuermann & F. Pedró (Eds.), *Assessing the effects of ICT in education: Indicators, criteria and benchmarks for international comparisons* (pp. 143–164). Luxembourg: Joint Research Centre European Commission.

Law, N., Woo, D., de la Torre, J., & Wong, G. (2018). A global framework of reference on digital literacy skills for indicator 4.4.2. Montréal: UNESCO Institute of Statistics. Retrieved from https://uis.unesco.org/sites/default/files/documents/ip51-global-framework-reference-digital-literacy-skills-2018-en.pdf

Law, N., Yuen, A. H. K., Shum, M. S. K., & Lee, Y. (2007). Final report on phase (II) study on evaluating the effectiveness of the "Empowering Learning and Teaching with Information Technology" strategy (2004/2007). Hong Kong: Education Bureau HKSAR.

Lazonder, A. W., Walraven, A., Gijlers, H., & Janssen, N. (2020). Longitudinal assessment of digital literacy in children: Findings from a large Dutch single-school study. *Computers & Education, 143*, 103681. https://doi.org/10.1016/j.compedu.2019.103681.

Lennon, M., Kirsch, I., Davier, M. von, Wagner, M., & Yamamoto, K. (2003). Feasibility study for the PISA ICT literacy assessment: Report to network A. OECD. Retrieved from https://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/33699866.pdf

List, A., Brante, E. W., & Klee, H. L. (2020). A framework of pre-service teachers' conceptions about digital literacy: Comparing the United States and Sweden. *Computers & Education, 148*, 1–20.

Littlejohn, A., Beetham, H., & McGill, L. (2012). Learning at the digital frontier: A review of digital literacies in theory and practice. *Journal of Computer Assisted Learning, 28*, 547–556.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84–99.

Markauskaite, L. (2007). Exploring the structure of trainee teachers' ICT literacy: The main components of, and relationships between, general cognitive and technical capabilities. *Educational Technology Research and Development, 55*, 547–572.

Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA). (2005). National Assessment Program Information and Communication Technology literacy 2005 years 6 and 10: An assessment domain for ICT literacy. MCEETYA. Retrieved from https://www.iste.org/docs/pdfs/australia_ict_assessment.pdf?sfvrsn=2.

Organisation for Economic Co-operation and Development (OECD) (Ed.). (2016). Technical report of the survey of adult skills *(PIAAC)* (2nd ed.). OECD. Retrieved from https://www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf

Organisation for Economic Co-operation and Development (OECD). (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills*. Paris: OECD.

Podsakoff, P. M., MacKenzie, S., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research. A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879–903.

Porat, E., Blau, I., & Barak, A. (2018). Measuring digital literacies: Junior high-school students' perceived competencies versus actual performance. *Computers & Education, 126*, 23–36.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*, 137–150.

Rohatgi, A., Scherer, R., & Hatlevik, O. E. (2016). The role of ICT self-efficacy for students' ICT use and their achievement in a computer and information literacy test. *Computers & Education, 102*, 103–116.

Siddiq, F., Gochyyev, P., & Wilson, M. (2017). Learning in digital networks—ICT literacy: A novel assessment of students' 21st century skills. *Computers & Education, 109*, 11–37.

Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., & Scherer, R. (2016). Taking a future perspective by learning from the past–a systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. *Educational Research Review, 19*, 58–84.

Spisak, J. R. (2018). In J. R. Spisak (Ed.), *Secondary student information literacy self-efficacy vs. performance*. Richmond: Virginia Commonwealth University.

van Laar, E., van Deursen, A. J., van Dijk, J. A., & de Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in Human Behavior, 72*, 577–588.

von Davier, M., & Khorramdel, L. (2013). Differentiating response styles and construct-related responses: A new IRT approach using bifactor and second-order models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Presentations from the 77th annual psychometric society meeting* (pp. 463–487). New York: Springer.

Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies, 44*, 299–332.

Wilson, M., Gochyyev, P., & Scalise, K. (2017). Modeling data from collaborative assessments: Learning in digital interactive social networks. *Journal of Educational Measurement, 54*, 85–102.

Wilson, M., Scalise, K., & Gochyyev, P. (2018). Domain modelling for advanced learning environments: The BEAR assessment system software. *Educational Psychology*. https://doi.org/10.1080/01443410.2018.1481934.

Woo, D., & Law, N. (2021). A methodology for deploying a digital literacy framework for diverse socioeconomic and sector contexts. To appear. In I. A. Lubin (Ed.), *ICT-enabled learning ecologies: Representation and sustainability across contexts*. New York: Routledge.

Frank Reichert is a Research Assistant Professor at the Faculty of Education at the University of Hong Kong. He previously held academic positions in Germany and Australia and was a 2016 Spencer Postdoctoral Fellow of the National Academy of Education, Washington, DC, in the United States.

James Zhang is a PhD Candidate at the Faculty of Economics and Business, University of Groningen. His research interests mainly lie in auditing and corporate governance, and general statistical applications in business research.

Nancy W. Y. Law is a Professor in the Unit of Teacher Education and Learning Leadership and founding honorary director of the Centre for Information Technology in Education (CITE) at the Faculty of Education, University of Hong Kong. She was the lead author of the International Report on the Second International Information Technology in Education Study (SITES) 2006, and Hong Kong National Research Coordinator for the ICILS 2013 study.

Gary K. W. Wong is an Assistant Professor in the Faculty of Education at the University of Hong Kong. His research interests include cognitive development of children and computational thinking education.

Jimmy de la Torre is a Professor in the Unit of Human Communication, Development, and Information Sciences at the Faculty of Education, University of Hong Kong. His research interests include methodological developments and applications of item response theory and cognitive diagnosis models, and the use of assessments to inform instruction and learning.