Education and Information Technologies
https://doi.org/10.1007/s10639-025-13592-8

Measuring digital literacy across ages and over time: Development and validation of a performance-based assessment

Qianqian Pan[1] · Frank Reichert[2] · Qianru Liang[3] · Jimmy de la Torre[4] · Nancy Law[5]

Received: 26 June 2024 / Accepted: 24 April 2025

Abstract

Measuring digital literacy (DL) across ages and tracking its growth over time have remained challenging in the area of digital literacy assessment. The current analysis reports on the psychometric properties of a performance-based Digital Literacy Assessment (DLA) instrument grounded in the DigComp 2.1 framework. Utilising a longitudinal cohort study design, the DLA was administered to Hong Kong students across three age cohorts, from lower primary to upper secondary, over a two-year period. Data were collected in 2019, before the COVID-19 pandemic, and in 2021, during the pandemic. The analysis provides validity and reliability evidence for using the DLA in longitudinal studies to assess DL from late childhood through late adolescence. The results further suggest that students' DL improved with grade level, with secondary students outperforming primary students but also displaying greater variability in scores. Over the two years, students generally demonstrated improvement in DL while inter-individual differences in DL growth rates widened. These findings indicate the widening of digital divides and highlight the need to investigate factors that contribute to diversity in DL development. In conclusion, our study provides evidence for the robustness of the DLA as an instrument to assess DL growth across ages and over time. Further, the DLA allowed us to uncover the substantial overlap in DL ability across different age groups and the widening second-level digital divide as children move into higher grades, and that the digital divide aggravated during the COVID-19 pandemic. Implications and challenges to the learning and assessment of DL are discussed.

Keywords Twenty-first century skills · Digital divide · Digital literacy · Information literacy · Elementary education · Secondary education

Extended author information available on the last page of the article

Published online: 12 June 2025

1 Introduction

The rapid advancement of educational technology has significantly accelerated the expansion of online learning, underscoring the imperative for education systems to equip students with robust digital literacy (DL) skills. The COVID-19 pandemic further illuminated existing digital divides among children and adolescents, revealing disparities not only in access to digital devices but also in DL (Law et al., 2023). These disparities, often termed the "second-level digital divide," influence how individuals utilise digital technologies (Scheerder et al., 2017).

Comprehensive DL, encompassing various competences required to use information and communication technology (ICT) effectively, is essential for mitigating educational inequalities (Engzell et al., 2021). Yet despite the recognized importance of DL, there is a paucity of research examining its development across different age groups over time. Existing assessment instruments predominantly target narrow age ranges (e.g., Fraillon et al., 2015, 2018a; Law et al., 2007; Ockwell et al., 2019) and often emphasize technical skills over other critical areas, such as safe and responsible technology use and problem-solving (Godaert et al., 2022). This limited scope restricts the utility of these instruments in longitudinal studies aimed at tracking DL progression. The longitudinal measurement of DL is crucial for educators and policymakers to monitor students' developmental trajectories, identify trends and patterns, and design effective curricula and programs (Tolboom & van Rooyen, 2021). Comparing students' DL trajectories can offer insights into digital divides, elucidating when and how gaps in DL emerge and evolve and enabling educators and institutions to allocate resources efficiently to address digital divides.

However, capturing DL development over time presents significant challenges, such as developing assessment instruments that are appropriate at different ages. Challenges for longitudinal studies also arise due to rapid changes in digital technologies (software and hardware), which can outdate assessment tasks and tools for measuring DL more quickly than in other domains (Reichert et al., 2023; Wong et al., 2023). Hence, assessment tools that encompass a broad age range and diverse content areas and whose assessment tasks are applicable in the current technology environment are required.

The present research aimed to address this need by developing a performance-based Digital Literacy Assessment (DLA) that overcomes the limitations of the more commonly used self-report DL assessments and is tailored for a longitudinal cohort study. In the initial phase, referred to as Time 1 (T1), the DLA was developed and validated cross-sectionally, with findings published by Jin et al. (2020). The current study represents the subsequent phase, Time 2 (T2), focusing on validating the updated DLA and modelling students' DL growth over time. Using a longitudinal cohort design overcomes the limitations of cross-sectional studies and enables us to examine both age-related changes and cohort-specific differences in DL (Raudenbush & Chan, 1992).

The design of the DLA was guided by the Digital Competence Framework for Citizens 2.1 (DigComp 2.1; Carretero et al., 2017), which defines DL across five competence areas: information and data literacy, communication and collaboration, digital content creation, safety, and problem-solving. This comprehensive framework addresses previously noted limitations in content scope, making it well-suited for developing a DL assessment for longitudinal research with several age groups. Moreover, DigComp has been identified as the "most comprehensive and widely used framework for general digital skills" (Bashir & Miyamoto, 2020, p. 9). It has also been instrumental in conceptualizing DL across various educational systems, as evidenced by its application in several empirical studies (Buchan et al., 2024; Mattar et al., 2022; Reichert et al., 2023).

In summary, this study addresses shortcomings in the existing literature by providing a comprehensive tool for assessing DL development across a wide age range. The DLA proposed and examined in the current study is particularly timely and well-suited to assess DL performance and track DL development over time, as it (1) adopts the comprehensive DigComp framework, which is particularly suitable for evaluating DL; (2) implements a performance-based measurement of DL (instead of a self-report); and (3) measures DL across an extensive age range, spanning from late childhood through to late adolescence. By leveraging the breadth of DigComp 2.1 and adopting a longitudinal cohort study design to examine the evolution of students' DL over time, this study facilitates a deeper understanding of how DL evolves, which can inform targeted educational interventions to address digital divides.

2 Background

In this section, we review the research literature that underpinned the current study, including challenges in developing a longitudinal DL assessment, literature on existing DL assessments, and the development of students' DL.

2.1 Measuring digital literacy

2.1.1 Digital literacy

A variety of labels have been used to describe the knowledge, skills and attitudes that individuals need to function effectively in a digital society. These labels include terms such as digital literacy, ICT literacy, media literacy and digital competence. While these terms may differ in emphasis, they have significant overlaps in their focus on developing competence for navigating through different digital environments. For the purposes of this study, we have focused on assessing students' knowledge and skills in digital functioning and have therefore chosen to use the term digital literacy (DL). We follow UNESCO's definition of DL as "the ability to access, manage, understand, integrate, communicate, evaluate and create information safely and appropriately using digital technologies for employment, decent work and entrepreneurship" (Law et al., 2018, p. 6).

2.1.2 Digital literacy assessments

Digital literacy (DL) assessments can be categorised into two types: self-report and performance-based assessments. A recent review has shown that self-reported DL assessments are more prevalent (Mattar et al., 2022). However, self-report assessments can be limited by the influence of social stereotypes and socioeconomic background, leading to an inaccurate reflection of actual skills and competences (Aesaert & van Braak, 2015; Siddiq & Scherer, 2019). Additionally, research showing that self-report measures of DL correlate only moderately with actual DL skills (e.g., Pan et al., 2022; Schwarz et al., 2024) indicates that self-report assessments may not adequately capture DL competence. To overcome these limitations, some international and nationwide studies have developed performance-based DL assessments that require students to complete specific tasks to demonstrate their DL (e.g., the International Computer and Information Literacy Study [ICILS], Fraillon et al., 2015; the Australian National Assessment Program: Information and Communication Technology Literacy [NAP-ICTL], Fraillon et al., 2018b; or the German National Educational Panel Study – Computer Literacy [NEPS-CL], Senkbeil et al., 2014). The present study also adopts a performance assessment for the measurement of DL.

Performance assessment of DL also has its challenges. The effectiveness of performance-based DL assessments may be influenced by various aspects of the testing environment, such as internet connectivity, as well as students' backgrounds, including their access to online information beyond the assessment setting (Aesaert & van Braak, 2015; Heitink, 2018). Moreover, in small- or medium-scale DL assessments, researchers often fail to provide adequate development procedures; these studies often also lack rigorous psychometric analyses (see reviews by Mattar et al., 2022; Siddiq et al., 2016). These limitations compromise the validity of the test results and impede the ability of other researchers or practitioners to utilise the assessment instruments.

Furthermore, current DL assessments in K-12 education primarily focus on specific age groups, with a significant emphasis on secondary education. Research on the digital competences of primary school students is very limited (e.g., Godaert et al., 2022; Kong et al., 2019; Siddiq et al., 2016). For example, ICILS and the American National Assessment of Educational Progress in Technology and Engineering Literacy (NAEP-TEL) are designed for eighth-grade students, who are typically around 13.5 years old (Fraillon et al., 2015; WestEd, 2018). The Australian NAP-ICTL is one of the very few studies assessing primary (sixth-grade) and secondary (tenth-grade) students (Ainley et al., 2008). In particular, there is a gap in validated assessments for students below sixth grade, with very few exceptions, such as the NEPS-CL, which covers a wider age range, including 3rd, 6th, 9th, and 12th graders (Weinert et al., 2011). It is worth noting that the NEPS-CL assessments were conducted using a paper-and-pencil format until 2017, and there are currently no available results for its computer-based version (Senkbeil & Ihme, 2021, p. 12). Additionally, Pedaste et al. (2023) developed the "Digitest," an assessment focused on digital competence for learning with respect to relevant attitudes, skills, and legal behaviour in online settings. Although the Digitest takes an innovative approach, its reliance on a modest sample of 863 Estonian students and the exclusive focus on DL within learning scenarios may limit the generalizability of its results. Therefore, existing DL assessment studies may be restricted in their ability to help researchers fully grasp the DL divide across age groups.

Another issue related to DL assessment is its dimensionality. DL is widely conceptualised as a multidimensional construct, and most DL assessments are designed to measure digital skills and competences as a set of multiple (sub-)competences or literacies. For instance, the NAP-ICTL has organised the measurement content into three strands: Strand A focuses on working with information, Strand B on creating and sharing information, and Strand C on using ICT responsibly (Ainley et al., 2008). In the ICILS, the assessment framework includes four strands, namely: understanding computer use, gathering information, producing information, and digital communication (Fraillon & Duckworth, 2023). However, the results of most empirical studies suggest that the empirically measured DL is a unidimensional construct. Consequently, DL has often been summarised on one continuous scale rather than as separate scores for multiple competences or literacies (Fraillon et al., 2015, 2018a; Ockwell et al., 2019; Senkbeil et al., 2014).

2.2 Measuring digital literacy over time

It is desirable to examine the development of students' digital skills and competences over time, as this offers more benefits than relying solely on one-off assessments. Monitoring students' developmental trajectories allows educators and policymakers to identify trends and patterns in the development of DL skills, enabling the design of more effective DL curricula and programmes (Tolboom & van Rooyen, 2021). In addition, comparing students' DL trajectories can offer insights into the digital divide in DL, such as whether, how and at what age gaps in DL begin to widen or otherwise. However, measuring DL over time presents a significant challenge. DL assessments should reflect the prevalent technologies used by the population while maintaining comparability over time. We examine two specific challenges and existing solutions in detail.

2.2.1 Challenge 1: Technology evolves rapidly

Measuring DL over time can be challenging due to the ever-evolving nature of technology (e.g., social media being more popularly used by children and youth) and emerging phenomena (e.g., recent shifts towards online classes). There is thus a continuous need for DL assessments to consider the digital tools with which individuals may be familiar when the test is conducted and to include new scenarios and technologies (Wong et al., 2023). For example, as students increasingly take online courses through different platforms, the assessment of digital etiquette may need to include the awareness of muting themselves during lectures. Due to the rapid changes in digital technologies (e.g., the addition of new features or the replacement of outdated tools) and the ubiquity of these technologies, DL assessments must be updated frequently to incorporate these changes to remain relevant (Wong et al., 2023).

Some DL assessment programmes have extended their assessment frameworks to deal with the fast-evolving digital technologies. For example, ICILS changed from two strands in 2013 (Strand 1: collecting and managing information and Strand 2: producing and exchanging information) to four strands in 2018 (Fraillon & Duckworth, 2023; Fraillon et al., 2015; Ockwell et al., 2019), as noted above. Other DL assessment programmes have developed new items in each round of administration, where new items that replace outdated items reflect newer aspects in digital technologies that have emerged between two assessment rounds (e.g., NAP-ICT; Ainley et al., 2016).

2.2.2 Challenge 2: Comparing DL scores over time despite measurement instrument changes

One challenge when updating DL assessment frameworks or adding new items to existing assessments is ensuring the comparability of DL scores over time. It is important to ensure that DL assessments measure the same DL construct over time so that changes in DL scores can be accurately interpreted as changes in individuals' DL (Ainley et al., 2016).

To overcome the challenge of including new items, most DL assessment programmes adopt a "common-item design" and apply mathematical linking and equating techniques in the item response theory (IRT) framework to achieve the comparability of DL scores. These psychometric approaches place the DL scores estimated from different measurement occasions on the same scale, thus facilitating comparisons of DL scores over time even though the scores may stem from modified versions of a DL assessment (e.g., Ainley et al., 2008; Ockwell et al., 2019).

Figure 1 illustrates a hypothetical common item design for measuring DL across two measurement occasions (Time 1 and Time 2). This assessment design integrates unique item sets for each occasion with a common set administered at both times to ensure consistency and comparability of scores. The Time 2 unique items can measure new aspects of DL based on the updated assessment framework. The common items serve as a link between the two measures, allowing Time 1 and Time 2 test scores to be placed on the same scale as long as item parameters of these common items, such

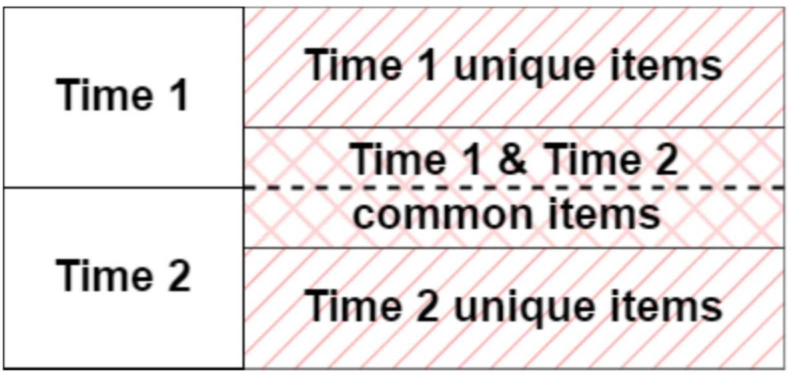


Fig. 1 Hypothetical example of the common item design

as item difficulty and discrimination, remain consistent over time. This design is crucial for maintaining score comparability across different assessment cycles, a criterion known as the invariance property of common items (e.g., Kolen & Brennan, 2004).

To maintain score reliability, the set of common items should represent a mini-version of the operational tests, with a minimum length of 20% of the operational test (Dorans et al., 2010, p. 20; Kolen & Brennan, 2004; von Davier et al., 2004). However, the stability of these item parameters needs to be verified through psychometric analysis. If an item's difficulty or discrimination varies over time—known as item parameter drift (IPD)—the comparability of scores from different testing cycles may be compromised (Goldstein, 1983; Kim et al., 2008).

2.2.3 Longitudinal digital literacy assessment studies

Longitudinal studies examining the development of students' DL over time are notably scarce, primarily due to the challenges in creating adaptable and comprehensive assessment tools. Our review identified only two methodologically sound longitudinal studies that used performance-based assessment instruments to examine individual DL development in school-aged students. The German National Educational Panel Study (NEPS) is a large-scale initiative that aims to track individuals' development in several areas, including DL, across their lifespan (Weinert et al., 2011). Gnambs (2021) analysed two ICT assessments from NEPS involving a representative sample of German ninth-grade students over three years (2010 and 2013); his research identified an average increase in students' ICT literacy over that period. However, these tests were only administered to secondary school students and used paper-and-pencil formats instead of a technology-based delivery. Moreover, given the rapid evolution of ICT, the applicability of these findings to current generations may be limited.

Research examining primary school children's DL development is particularly limited. A longitudinal study involving Dutch fifth and sixth graders revealed that students' digital skills improved over time (Lazonder et al., 2020). However, that study covered a less comprehensive set of digital competences, focusing exclusively on students' ability to collect and safely use data (assessed via paper-and-pencil test), as well as their operational proficiency in Microsoft Word and PowerPoint. This approach is somewhat restricted in terms of software applications, and it may not comprehensively capture the multifaceted nature of DL, which also includes, for example, communicating via digital devices.

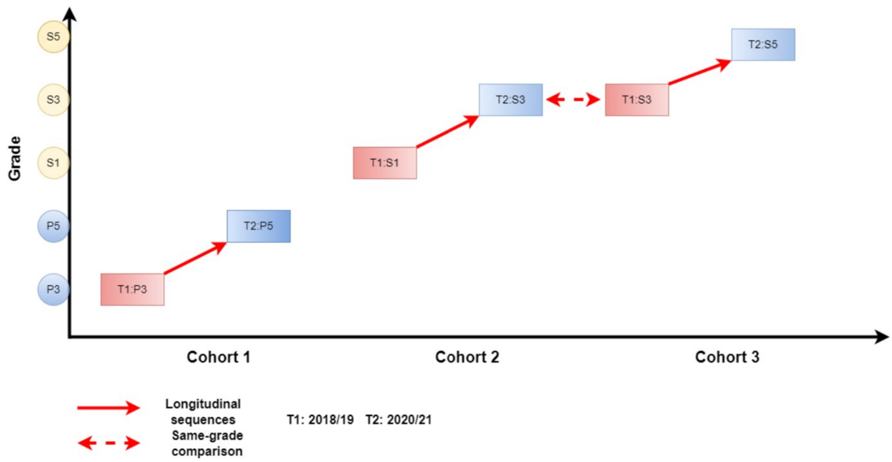
Finally, some researchers have tracked students' digital skills over time using self-report digital skills assessments, such as the ySkills study that was conducted in six European countries between 2021 and 2023 and showed that students' perceived digital skills improved over time (Machackova et al., 2024). However, self-report data may be influenced by social desirability biases and stereotypes and do not adequately reflect actual competence, as previously discussed.

In summary, few studies have simultaneously investigated DL growth in both primary and secondary school students, and no longitudinal study has previously used a psychometrically sound performance-based DL assessment tool across primary and secondary school ages. This gap highlights the need for research that encompasses a broader age range to provide a more comprehensive understanding of how DL evolves throughout different educational stages.

2.3 The present study

The current study addressed the outlined challenges and research gaps in its development of a performance-based instrument suitable for measuring DL performance at several ages and longitudinally. This study was part of a larger project entitled "Learning and Assessment for Digital Citizenship" that aimed to foster the understanding of DL development in children and youth. This goal required developing a DL assessment that could help reveal important insights about the digital divide at various levels and, thereby, contribute significant implications for learning in the digital age. Therefore, this study endeavoured to address the outlined challenges by developing a DL assessment that (1) is based on a comprehensive DL framework (DigComp 2.1), (2) covers a wider age range than existing assessments, and (3) measures DL in the same individuals at different times. It adopted a longitudinal cohort study design, which included two rounds of data collection across three cohorts, as shown in Fig. 2.

Assessing the same students on multiple occasions enables researchers to identify changes in DL performance and to understand when, how, and at what age disparities in DL may begin to manifest. An accelerated longitudinal design that samples different cohorts and measures their DL over a two-year period was implemented in the present study, as it offers a more nuanced perspective than a cross-sectional snapshot. Specifically, the described longitudinal cohort design allows researchers to examine developmental changes over an extended age range within a shorter timeframe, it helps to reduce participant dropout rates due to



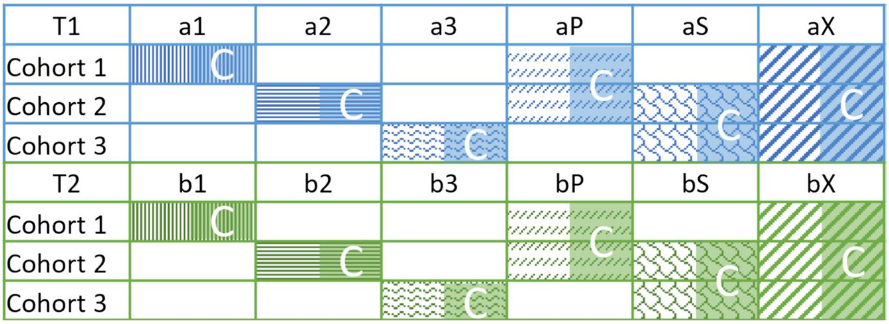
P3: Primary 3, P5: Primary 5, S1: Secondary 1, S3: Secondary 3, S5: Secondary 5. T1: T1 Study (2018/19), T2: T2 Study (2020/21).

Fig. 2 Diagrammatic representation of the research design a shorter period over which the same individuals must commit to participating in a study, and it allows distinguishing between age-related changes and cohort-specific influences (Raudenbush & Chan, 1992).

In the first round of data collection (the T1 Study, conducted during the 2018/19 school year), a total of 80 items were assembled into three test forms (referred to as T1 DLA) and administered to three cohorts of students (Cohort 1: Primary 3 [grade 3, usually 8–9 years old], Cohort 2: Secondary 1 [grade 7, usually 12–13 years old], Cohort 3: Secondary 3 [grade 9, usually 14–15 years old]), respectively, with cross-sectional common items for linking the performance of different cohorts (see Fig. 3). The design and robustness of the DLA instrument for comparison across the three age groups assessed in T1 have been previously reported (Jin et al., 2020). In short, the data of the T1 Study supported a unidimensional construct of DL, and the psychometric properties of the test items were sound. A detailed description of the instrument development and its validation for the T1 Study can be found in Jin et al. (2020).

The current study (T2 Study) was designed as an extension of the T1 Study's DLA to accommodate students' growth in DL over two years. The T2 DLA incorporated both original items from the T1 DLA and new items to reflect the expanded content range and increased difficulty level appropriate for older students, now in Primary 5 (Grade 5), Secondary 3 (Grade 9), and Secondary 5 (Grade 11). This approach ensures comparability between T1 and T2 scores while addressing the expected advancement in DL skills.

The T2 data collection occurred during the 2020/21 academic year amidst the COVID-19 pandemic. By this time, students had undergone significant changes in their daily lives, including extensive online schooling and increased digital engagement. The new items were designed to capture recent technological shifts



T1: T1 Study, T2: T2 Study. a1: T1 unique items for Cohort 1, a2: T1 unique items for Cohort 2, a3: T1 unique items for Cohort 3, aP: T1 common items for Cohorts 1 & 2, aS: T1 common items for Cohorts 2 & 3, aX: T1 common items for all cohorts. b1: T2 unique items for Cohort 1, b2: T2 unique items for Cohort 2, b3: T2 unique items for Cohort 3, bP: T2 common items for Cohorts 1 & 2, bS: T2 common items for Cohorts 2 & 3, bX: T2 common items for all cohorts. Shaded box with C: longitudinally common items (i.e., items in the shaded boxes with C shown in the same column were identical in T1 and T2).

Fig. 3 Design of T1 and T2 digital literacy assessment test forms and emerging phenomena, such as the transition from platforms like Skype to Zoom for online education.

The present study examines the extension of the original DLA instrument with respect to its validity for comparing DL across age cohorts two years later and measuring DL growth over time. Consequently, the current analysis focuses on the T2 DLA and seeks to answer the following research questions (RQs) (see Fig. 4 for details):

RQ 1: Is the dimensionality of the DL measure time-invariant (i.e., does each DLA measure one [unidimensional] construct, or do the T1 and T2 DLAs measure multiple digital literacies)?
RQ 2: Can the estimated DL scores of the T2 DLA be compared reliably and validly across the three age cohorts (i.e., cross-sectional measurement invariance across age cohorts)?
RQ 3: Can the estimated DL scores be compared reliably and validly longitudinally between the T1 DLA and the T2 DLA (i.e., longitudinal measurement invariance over time)?
RQ 4: Using the DLA instrument developed, what was found about students' DL progress over time? That is, (a) what were the differences in DL across the three *age cohorts* (i.e., how did DL scores differ across different grade levels), and (b) how did students' DL *change over time* (i.e., how do the DL scores differ between T1 and T2 for individual students)?

3 Method

3.1 Instrumentation

Based on the common item design, the T2 DLA included a set of items from the T1 DLA and a set of new items. The items that were identical in both the T1 and T2 DLAs (i.e., longitudinal common items) aimed to place the T1 and T2 DL scores onto the same scale. The newly developed items aimed to (1) extend the content and difficulty coverage for the older participants (e.g., Cohort 3 students would be 16–17 years old at T2, but the T1 DLA only tested students up to 14–15 years of age) and (2) reflect the rapid changes of digital technology in students' learning and their lives over the two years from 2019 to 2021 (e.g., video conferencing tools such as Zoom were widely used in schools).

Similar to the T1 DLA, the T2 DLA comprised three test forms. Initially, a total of 101 items were mapped onto the five competence areas in the DigComp 2.1 framework. The three test forms were administered to the three student cohorts (Cohort 1: Primary [P5], average age = 10.6 [SD = 0.80], Cohort 2: Secondary [S3] average age = 14.7 [SD = 1.02], and Cohort 3: Secondary [S5], average age = 16.8 [SD = 1.42]) from April to July 2021 and included some common items to link the test forms across the three grade levels (i.e., cross-sectional common items).

The T2 DLA includes traditional item types (multiple choice [MC] items, see Fig. 5a), technology-enhanced items (TEIs; such as drag-and-drop items, see Fig. 5b),

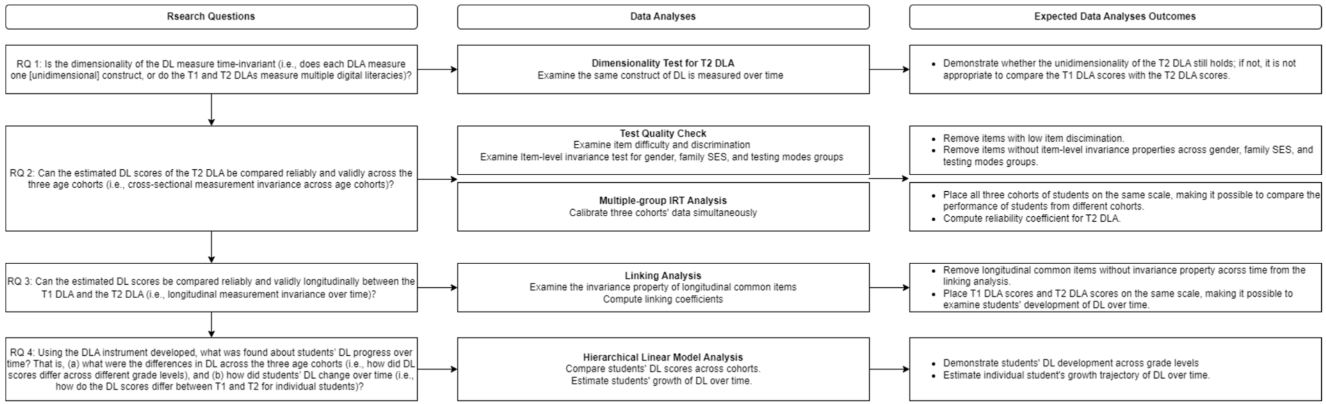


Fig. 4 Mapping of research questions, data analysis methods, and expected outcomes

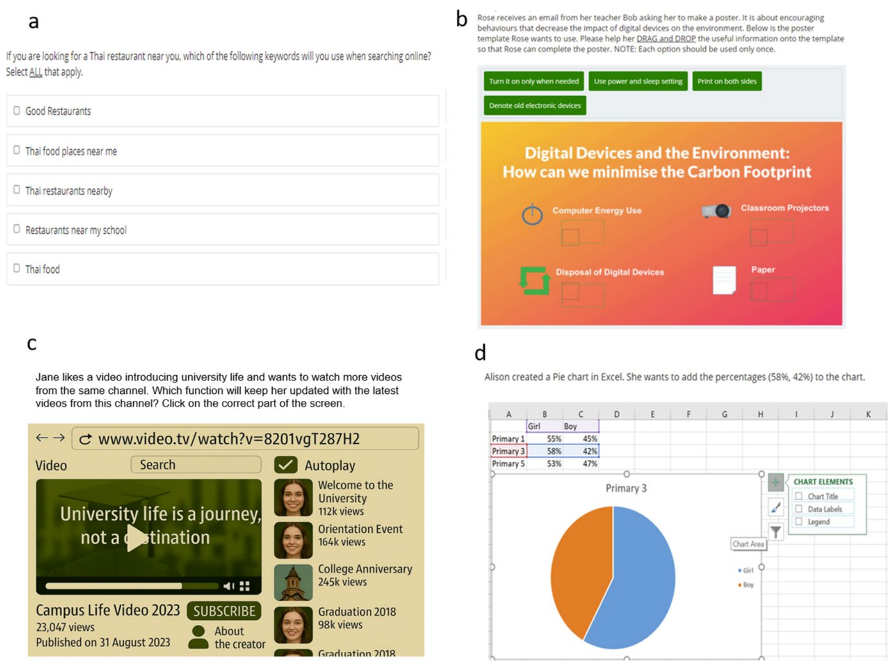


Fig. 5a and 5b were created by the authors and their project team. Fig. 5c was generated using the assistance of Microsoft Copilot to illustrate the interactive item type while ensuring compliance with intellectual property rights. Fig. 5d is a screenshot taken from Microsoft Excel 2019 (Version 1808, Build 10384.20008) [Computer software]. © Microsoft Corporation

Fig. 5 Sample items from the DLA

and interactive items (e.g., Fig. 5c, which requires clicking on a specific position in a screenshot to indicate the correct action, and Fig. 5d, which mimics Microsoft Excel). Most items (94 out of 101 items) were scored dichotomously (0 or 1), while seven items were scored polytomously (0, 1, or 2, with 1 reflecting partial credit).

In addition, we collected students' background information (e.g., gender, parental education levels, home possessions, etc.) to conduct differential item functioning (DIF) analysis based on students' gender, parental educational levels, and family socioeconomic status (SES). We followed the approach used in the Programme for International Student Assessment (PISA, Kastberg et al., 2021) to calculate family SES scores. This procedure involved using a two-parameter logistic item response theory (2PL IRT) model for dichotomous items and the Graded Response Model (GRM) for polytomous items. These models helped identify low, moderate, and high SES groups for comparison, categorising scores $\leq 33\%$ as low, scores $\geq 67\%$ as high, and the remainder as moderate. Further details are available in the supplementary materials (Appendix 1).

3.2 Sample

A stratified random sampling method was employed to ensure the sample was representative of the broader student population. At T1, four of the 18 districts of the Hong Kong Special Administrative Region (SAR) were randomly selected to achieve geographical and socioeconomic diversity. Within each district, schools were randomly sampled, and a total of 18 primary and 14 secondary schools participated in the study. Two classes at the respective grade levels were randomly selected from each of these schools, and all students in the sampled classes were invited to participate in our study. As a result, the T1 DLA was completed by 2,046 students.

In 2021, 23 schools (12 primary and 11 secondary schools) that had participated in T1 also participated in the T2 data collection. In both the T1 and T2 studies, Cohorts 2 and 3 were sampled from the same secondary schools. A total of 1,971 students completed the T2 DLA (see Table 1 for a detailed breakdown). Intact classes were sampled in 2019, and for ease of test administration, some schools chose intact classes in 2021 for the T2 data collection. However, some students from the T1 sample had changed classes by 2021. Thus, besides the attrition of some students, the T2 sample also included new students who had not participated in the T1 Study. About 45% of all students who completed the DLA in T2 had also participated in T1. The institutional review board of The University of Hong Kong approved all study procedures, and informed consent was obtained before data collection from school principals, students, and parents (passive or active, depending on students' age).

3.3 Administration procedures

The T2 DLA was delivered to students on the customised Open edX platform [https://openedx.org/], an open-source online learning management system (LMS) that enables the administration of online assessments. Due to the long periods of school suspension during the 2020/21 school year, it was challenging to organise onsite data collections in the T2 Study. Therefore, the research team piloted and refined two additional modes of data collection to maximise participation in T2. In addition to the onsite mode, schools could also choose for their students to participate in the study via an online supported or an online self-directed mode (see Table 2). A careful statistical analysis showed that the three modes of assessment were comparable (Pan et al., 2022).

Table 1 Number of participating schools, classes, and students in T1 (2019) and T2 (2021)

	Schools		Classes		Students		
Cohort	2019	2021	2019	2021	2019	2021	Common
1	18	12	39	48	750	507	234
2	14	11	27	39	715	839	389
3			29	38	581	625	264
Total	32	23	95	125	2046	1971	887

Common: students who completed both the T1 and T2 DLAs

Table 2 Sample sizes for the three testing modes by cohort

	Primary 5		Secondary 3		Secondary 5	
Testing Mode	*N*	%	*N*	%	*N*	%
Online support	111	21.89	288	34.45	250	40.00
Onsite support	388	76.53	441	52.75	300	48.00
Self-directed	8	1.58	107	12.80	75	12.00
Total	507		836		625	

3.4 Data analysis

3.4.1 Psychometric analysis

A series of psychometric analyses was conducted to address the research questions, as shown in Fig. 4. Item Response Theory (IRT) was employed in the current study due to its robust capability in vertical scaling (Briggs & Weeks, 2009). Vertical scaling is essential for placing assessments from various grade levels onto one scale, thereby enabling the accurate measurement of students' DL development over time. The IRT model facilitates this by providing sample-independent item parameters, ensuring that item characteristics remain consistent across different populations and time points. Measuring invariance via DIF tests within the IRT framework is also possible, and this invariance allows for the meaningful longitudinal comparison of student abilities, as it accounts for variations in test forms and sample groups (McArdle et al., 2009).

Dimensionality test for T2 DLA To determine the dimensionality of the DLA (RQ1), two IRT models were fitted to the T2 Study data for each cohort: (1) a unidimensional IRT and (2) a multidimensional IRT (MIRT) with five correlated factors representing subdomains of DL as per the DigComp 2.1 framework. In this study, the Graded Response Model (GRM) was used for polytomous items, while 2PL IRT models were used for binary items.[1] The DETECT statistics (Stout et al., 1996) can detect violations of unidimensionality and were also computed to confirm the dimensional structure of the T2 DLA.

Test quality check To address RQ2, firstly, item discrimination analyses were performed for each cohort. Items with low discrimination (< 0.30) were discarded to ensure the ability to differentiate between students with varying levels of DL.

Secondly, DIF tests were conducted to verify the fairness and unbiasedness of the T2 DLA scores (Holland & Thayer, 1986). DIF tests identify whether performance

[1] In the preliminary analysis, we compared the 2PL and a three-parameter logistic IRT model (3PL) and the likelihood ratio test (LRT), Akaike's information criterion (AIC; Akaike, 1998) and the Bayesian information criterion (BIC; Schwarz, 1978). All showed that the 2PL was better suited for the data than the 3PL.

differences on items are influenced by factors unrelated to DL competence, such as gender or parental education. This is crucial for valid group comparisons on multi-item measures. DIF was assessed across gender, testing modes, parental education levels, and family SES, ensuring item invariance for all cross-sectional common items across three cohorts. Items showing DIF were adjusted or removed as necessary to maintain score comparability.

Multiple-group IRT analysis A multiple-group IRT model was applied to calibrate data from all cohorts (RQ2) simultaneously, with Cohort 3 as the reference. The common items between cohorts were anchor items that ensured the item parameters were calibrated on the same scale. Item difficulties of common items that functioned consistently across cohorts were constrained to be equal, while the difficulties of items identified as exhibiting DIF were not constrained. Plausible values for DL scores were estimated via the expected a posteriori method, enhancing measurement precision and reliability (Chalmers, 2012). The mean of these values represented the final DL score for subsequent analyses.

Linking analysis To link T1 and T2 DLA scores (RQ3), we first verified the invariance of the construct over time. A multiple-group bifactor model was used to detect any shifts in the DL construct (Li & Lissitz, 2012). Longitudinal common items that failed to maintain invariance were excluded from linking analyses. The Stocking-Lord method was employed to align T2 scores with T1.

All psychometric analyses were conducted in R using the libraries *"mirt"* for IRT modelling (Chalmers, 2012), *"sirt"* for dimensionality analysis (Robitzsch, 2022), and *"equatIRT"* for estimating linking constants (Battauz, 2015). Further technical details can be found in the supplementary materials (Appendix 2).

3.4.2 Examining DL development over time

The linked T1 and T2 DLA scores were used to examine how students' DL developed over time and to identify possible changes in the digital divide in DL. First, we analysed the two cross-sectional datasets (T1 and T2) separately to examine DL progression across grade levels for three cohorts (P3, S1, S3 in T1 and P5, S3, S5 in T2, respectively; RQ4a). Due to the multilevel structure of the data (students were nested within cohorts and schools), the hierarchical linear model (HLM) was used to account for between-student dependence in the same cohort and same school (Hoffman, 2015, p. 12). We adopted a random intercept three-level HLM, which was a composite of the individual level, cohort level, and school level. Hence, the school means of students' respective DL scores were allowed to differ among schools. Since Cohorts 2 and 3 were sampled from the same secondary schools, different DL cohort means were also possible between cohorts. Technical details can be found in the supplementary materials (Appendix 3).

Second, we estimated the growth trajectories of DL over time for students who completed both the T1 and the T2 DLA using the HLM (RQ4b). We modelled each of the three cohorts separately, and the results included the intercept (i.e., the initial level of students' DL) and the growth rates of students' DL for each cohort. The HLM analysis was performed using the R package *"lme4"* (Bates et al., 2015). The technical details can be found in the supplementary materials (Appendix 4).

4 Results

4.1 Dimensionality of the T2 DLA

DL was found to be unidimensional using the T1 DLA (Jin et al., 2020). The T2 DLA analysis also showed that the best model for each cohort was the unidimensional model of DL, as indicated by the AIC, BIC and DETECT statistics shown in Table 3. This model was, therefore, adopted for all cohorts. As both T1 and T2 DLA results were unidimensional, it was possible to conduct a link analysis between the T1 and T2 DL scores.

4.2 T2 DLA test quality

Based on the unidimensional IRT model, we examined item discriminations for each cohort. The results showed that six items had low item discriminations (< 0.30). These six items were excluded from the subsequent analyses. Table 4 displays the number of remaining items for each DigComp 2.1 competence area. For these remaining 95 items of the T2 DLA, further analyses showed no gender DIF, testing mode DIF, or SES-related DIF for all three cohorts. However, six cross-sectional common items showed a lack of invariance between adjacent cohorts. Hence, these items were not modelled as link items but instead freely estimated in each cohort in

Table 3 Results of the multidimensionality analysis of T2 DLA

	AIC	BIC	DETECT
Cohort 1			
Unidimensionality	26499.91	26888.93	0.23
Multidimensionality	26710.13	27141.44	
Cohort 2			
Unidimensionality	45399.63	45896.14	0.18
Multidimensionality	46227.53	46747.68	
Cohort 3			
Unidimensionality	35397.83	35872.67	0.24
Multidimensionality	36471.44	36964.03	

DETECT < 0.20: essential unidimensionality; 0.20 < DETECT < 0.40: weak multidimensionality (Jang & Roussos, 2007; Zhang, 2007)

Table 4 Item distributions of the 2019 and 2021 DLAs mapped to the DigComp 2.1 framework

Competence Areas	Sub-competences	T1 DLA	T2 DLA*
1. Information and data literacy	1.1 Browsing, searching, filtering data, information and digital content	5	4
	1.2 Evaluating data, information and digital content	4	4
	1.3 Managing data, information and digital content	6	4
2. Communication and collaboration	2.1 Interacting through digital technologies	5	3
	2.2 Sharing through digital technologies	8	6
	2.3 Engaging in citizenship through digital technologies	3	4 (5)
	2.4 Collaborating through digital technologies	0	5
	2.5 Netiquette	4	3 (4)
	2.6 Managing digital identity	2	4 (5)
3. Digital content creation	3.1 Developing digital content	4	1
	3.2 Integrating and re-elaborating digital content	0	4
	3.3 Copyright and licenses	3	3
	3.4 Programming	0	11 (12)
4. Safety	4.1 Protecting devices	7	6
	4.2 Protecting personal data and privacy	11	6 (7)
	4.3 Protecting health and wellbeing	5	2
	4.4 Protecting the environment	1	4
5. Problem solving	5.1 Solving technical problems	11	7 (8)
	5.2 Identifying needs and technological responses	0	6
	5.3 Creatively using digital technologies	0	4
	5.4 Identifying digital competence gaps	1	4
Total		80	95 (101)

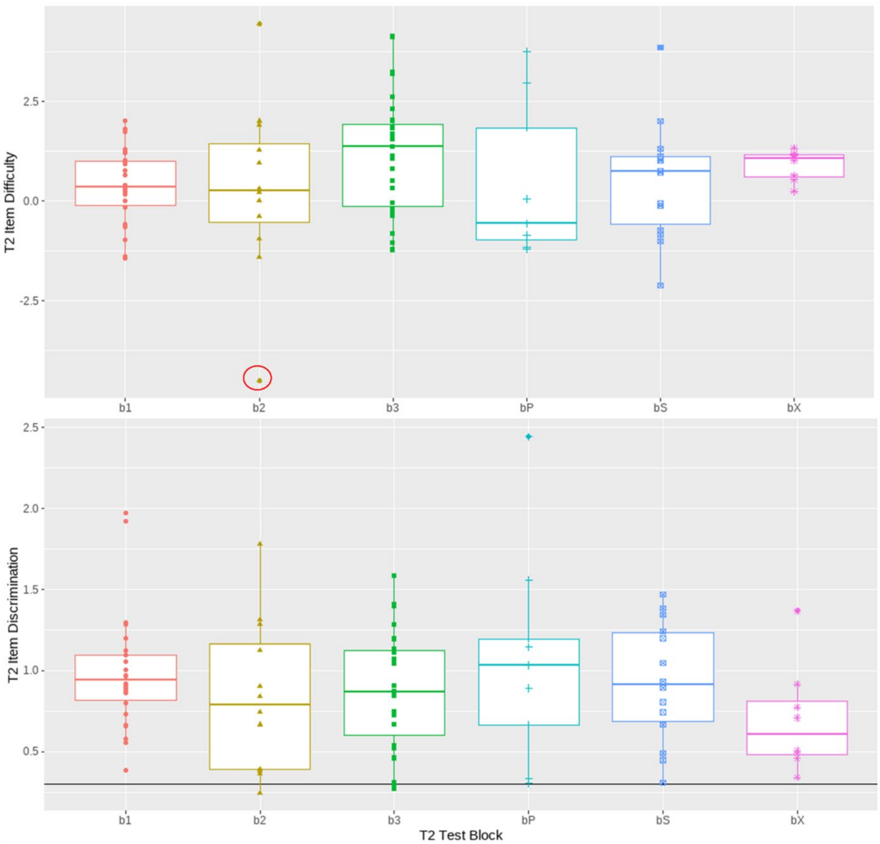
*The numbers in brackets represent the original number of items prior to the T2 psychometric analysis

the subsequent multiple-group analyses. As a result, a total of nine common items were shared by Cohort 1 and Cohort 2, 16 common items by Cohort 2 and Cohort 3, and an additional eight common items across all three cohorts. Table 5 shows the number of items for each cohort mapped to the DigComp 2.1 framework.

The parameter estimates of the 95 items from the multiple-group IRT model are presented in Fig. 6. In general, the T2 DLA had a wide range of item difficulties. The average item difficulties of the test forms were as follows: Cohort 1: 0.51, Cohort 2: 0.63 (after the removal of one outlier, as demonstrated in Fig. 6; this was 0.51 before removing the outlier), and Cohort 3: 0.97, which increased with increasing grade levels of the target cohorts. Hence, the item selections were appropriate

Table 5 Item distributions of T2 DLA mapped to the DigComp 2.1 framework per cohort

Competence Areas	Cohort 1	Cohort 2	Cohort 3
1. Information and data literacy	8	7	8
2. Communication and collaboration	10	11	12
3. Digital content creation	8	11	10
4. Safety	10	10	10
5. Problem solving	9	9	11
Total	45	48	51



An extremely easy item of b2 (item difficulty = -4.5) was circled in red. Three items (below the black line) had a low discrimination (< 0.30). b1: T2 unique items for cohort 1, b2: T2 unique items for cohort 2, b3: T2 unique items for cohort 3, bP: T2 common items for cohort 1 & 2, bS: T2 common items for cohort 2 & 3, bX: T2 common items for all cohorts.

Fig. 6 Equated item difficulties and item discriminations of T2 DLA dichotomous items by test blocks

Table 6 Distribution of longitudinally linked items

Competence Areas	Number of Items	Number of Linked Items
1. Information and data literacy	12	4
2. Communication and collaboration	25	8
3. Digital content creation	19	3
4. Safety	18	11
5. Problem solving	21	4

for the broad age range of students targeted. In addition, the majority of items had moderate to high item discriminations (> 0.65; Fig. 6[2]) (Baker, 2001).

The reliability of the estimated expected a posteriori (EAP) score of the T2 DLA was 0.91, indicating a relatively small measurement error. Taken together, the psychometric evidence supported the conclusion that the T2 DLA provided reliable and unbiased DL scores for all students to be used in further analysis.

4.3 Linking T1 and T2 DLAs

Before linking the T1 and T2 DLAs, construct shift was examined. The time-specific bifactor T2 explained only a minimal percentage of the total variance of each item for all cohorts (ranging from 0.004% to 0.018%). This result supports the assumption that the construct of DL did not shift over time, thus allowing the linking of the T1 and T2 DLAs.

Regarding the invariance property of longitudinal common items, a total of 33 longitudinal common items were tested to examine whether they were invariant over time. The results showed that three items did not hold the invariance property. Thus, 32% of the T2 DLA items (30 out of 95) were suitable for estimating the linking constants between the T1 and T2 DLAs, which is above the rule of thumb of 20% of the total test length required (Angoff, 1984; Shea & Norcini, 1995). Table 6 shows that these 30 items cover all sub-domains, which is in line with the recommendation (Kolen & Brennan, 2004). Finally, Table 7 shows the results of the linkage. The positive intercept (B) indicates that T2 students had higher DL scores, whereas the slope (A) was above 1, indicating a greater variability of the DL scores at T2 compared to T1.

4.4 Students' DL development

Having confirmed the unidimensionality of the DLA across T1 and T2, and established the instrument's reliability and validity for comparing students' DL across age cohorts and over time, it is now possible to report in this section the findings regarding students' DL development between 2019 and 2021.

[2] As the majority of items were dichotomous, only dichotomous item difficulties were plotted.

Table 7 Summary of linking results

Linking Constants			
		A (Slope)	*B* (Intercept)
Stocking-Lord		1.40	1.08
Equated Linking Item Parameter			
		Item Discrimination	Item Difficulty
Mean (SD)	From (minimum)	1.13 (0.54)	−0.52 (0.83)
	To (maximum)	1.32 (0.55)	0.24 (1.11)
Equated Estimated DL Scores			
		T1 DLA	T2 DLA
Mean (SD)		0.00 (0.95)	1.08 (1.40)
Range		[−2.91,2.64]	[−2.87,5.12]

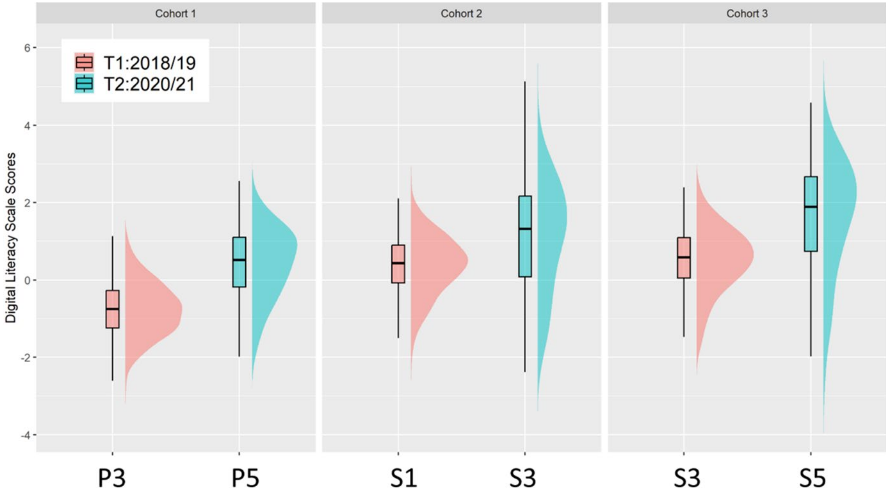
4.4.1 Differences in DL across grade levels and changes in DL over time

Figure 7 displays students' DL scores for all three cohorts at the two time points. The figure shows that students' overall DL scores and inter-individual variation increased over time. The exact means and standard deviations of the DL scores estimated from the IRT model are displayed in Table 8. Hierarchical Linear Modelling (HLM) was subsequently applied to statistically examine cohort differences and students' growth over time.

The HLM allowed us to account for within-cohort and within-school variations in order to examine cohort differences in students' DL scores at Time 1 (T1) and Time 2 (T2). The first half of Table 9 presents the estimated average DL scores for each cohort. At T1, P3 students' DL scores were significantly lower than zero,[3] while both S1 and S3 students' scores were significantly higher than zero. At T2, P5 students' DL scores did not differ significantly from zero, while both S3 and S5 students' scores were significantly above zero. The second half of Table 9 illustrates the differences between cohorts at the same time point. At T1, the difference between S1 (Cohort 2) and S3 (Cohort 3) students was not statistically significant. Conversely, at T2, S5 students (Cohort 3) scored significantly higher than S3 students (Cohort 2). Overall, secondary students consistently demonstrated significantly higher DL scores in comparison to primary students at both time points.

In addition, Fig. 7 shows that, on average, P5 students had slightly higher DL at T2 than S1 students at T1, and S3 students at T2 had higher DL than S3 students at T1. We, therefore, used HLM to investigate whether these differences in the means of cohorts of similar ages between T1 and T2 were statistically significant. The results showed that P5 students (Cohort 1) at T2 had achieved similar DL performance as S1 students (Cohort 2) at T1 ($DL_{T2P5} - DL_{T1S1} = 0.03, SE = 0.15, p = .84$). In addition, the results revealed that S3 students (Cohort 3) at T1 had significantly lower DL scores than S3 students (Cohort 2) at T2 ($DL_{T1S3} - DL_{T2S3} = -0.58, SE = 0.21, p = .02$). That is, students

[3] Zero represents the mean across the three cohorts at T1.



P3: Primary 3, P5: Primary 5, S1: Secondary 1, S3: Secondary 3, S5: Secondary 5. T1: T1 Study, T2: T2 Study. The density plots may extend beyond data boundaries to provide a smooth density curve.

Fig. 7 Raincloud plots of students' digital literacy scale scores by cohort over two years

at T2 outperformed students of the same ages or even older at T1, as shown in the case of Cohorts 1 and 2 students. In addition, we can see from the raincloud plots that the DL divides for each cohort increased between T1 and T2.

4.4.2 Individual DL growth

A total of 887 students completed both the T1 and T2 DLAs (see Table 1), which allowed us to examine their individual growth in DL over the two years. As the students who completed both waves represented only a fraction of all participants, we also compared the T1 DL scores of students who only participated in the T1 Study with those who took part in both the T1 and T2 DLAs to see if there were any statistical differences between these two samples. We found that the secondary school students who participated in both the T1 and T2 DLAs achieved higher scores in T1 than those who only participated in T1. However, there was no significant difference in scores among these two groups of students in the primary school sample. These

Table 8 Mean and SD of students' DL across time

	Cohort 1		Cohort 2		Cohort 3		Total	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
T1	−0.77	0.68	0.36	0.76	0.52	0.81	0.00	0.95
T2	0.42	0.90	1.09	1.42	1.61	1.47	1.08	1.40

Means and *SDs* were calculated from the IRT model. To ensure model identification, the overall mean of DL across all three cohorts was fixed to 0 at T1

Table 9 Estimated mean DL scores at T1 and T2 from HLM

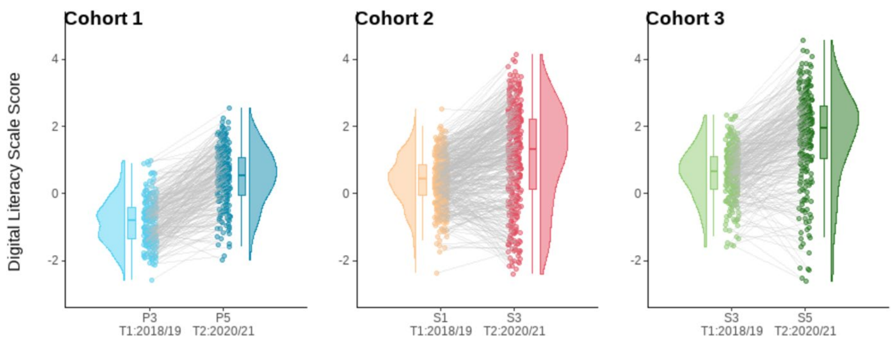
	T1				T2			
Fixed Effects	Grade	*Estimated Mean*	*SE*	*p*	Grade	*Estimated Mean*	*SE*	*p*
Cohort 1	P3	−0.77	0.06	**	P5	0.40	0.21	
Cohort 2	S1	0.37	0.10	**	S3	1.12	0.22	**
Cohort 3	S3	0.51	0.10	**	S5	1.61	0.22	**
	T1 Estimated Mean Differences				*T2 Estimated Mean Differences*			
Grade Difference	P3 vs S1	−1.14	0.13	**	P5 vs S3	−0.72	0.31	*
	P3 vs S3	−1.28	0.13	**	P5 vs S5	−1.21	0.31	**
	S1 vs S3	−0.14	0.10		S3 vs S5	−0.49	0.17	*

Estimated *Mean* and *SE* are from HLM. P3: Primary 3, P5: Primary 5, S1: Secondary 1, S3: Secondary 3, S5: Secondary 5. T1: T1 Study, T2: T2 Study. The p-value for the estimated mean is used to test whether the estimated mean is different from 0; the p-value for the grade difference is used to test whether the grade difference is different from 0. $^{**}p < 0.01$, $^{*}p < 0.05$

results indicate that a larger proportion of lower DL-performing secondary students in the T1 sample did not participate in T2.

Figure 8 displays the growth trajectories of all common students. Overall, all three cohorts showed improvements in their DL scores over the two years, as indicated by the uplifted boxplots. However, as shown by the spaghetti plots, not all students improved over time. Rather, some students regressed over time, and this proportion was higher for secondary students. This finding provides a more nuanced understanding of the increased DL divide across the three cohorts: the interpersonal variation increased in all cohorts, especially among secondary school students, and it was not simply due to differences in progression but also due to regression found in some cases.

Table 10 presents the estimated intercept (i.e., the initial DL level) and growth rates of common students in the three cohorts as estimated by the HLMs. On average,



P3: Primary 3, P5: Primary 5, S1: Secondary 1, S3: Secondary 3, S5: Secondary 5. T1: T1 Study, T2: T2 Study. The density plots may extend beyond data boundaries to provide a smooth density curve.

Fig. 8 Common students' digital literacy scale scores over time across cohorts

Table 10 Estimated initial levels and growth rates of common students across time

	Cohort 1 P3 → P5	Cohort 2 S1 → S3	Cohort 3 S3 → S5
Estimated Fixed Effects	Coef. (*SE*)	Coef. (*SE*)	Coef. (*SE*)
Intercept	−0.81 (0.11)**	0.42 (0.21)	0.61 (0.20)*
Growth Rate	1.33 (0.05)**	0.72 (0.06)**	1.10 (0.07)**
Correlation Between Intercept and Growth Rate	−0.23**	−0.15**	−0.17**

P3: Primary 3, P5: Primary 5, S1: Secondary 1, S3: Secondary 3, S5: Secondary 5. T1: T1 Study, T2: T2 Study. ***p* < 0.01, **p* < 0.05

Cohort 1 students improved by 1.33 points, Cohort 2 students improved by 0.72 points, and Cohort 3 students improved by 1.10 points. Furthermore, Cohort 1 students had the largest and Cohort 2 students had the smallest average growth rates, respectively.

We also examined the correlations between the intercept and the growth rates in the three cohorts. For Cohort 1, the correlation was −0.23, and it was −0.15 and −0.17 for Cohort 2 and Cohort 3, respectively. These negative correlations suggest that students' growth rates tended to be smaller the higher their initial (T1) DL scores were, especially for primary school students.

5 Discussion

Education systems worldwide must equip young people with digital literacy (DL) to prepare them for life in the digital era and to prevent widening DL divides that could hinder effective participation in society. Valid DL assessments are essential to support this effort. However, most existing assessments target a narrow age range, rely on one-time measurements that cannot track individual growth, and/or focus only on a narrow set of DL competences.

This study developed a comprehensive Digital Literacy Assessment (DLA) to address these gaps. The DLA was designed to measure DL longitudinally across three student age cohorts over two years, in 2019 and 2021. The first assessment (T1) occurred before the COVID-19 pandemic (Reichert et al., 2020a), while the second (T2) took place during the pandemic after a year of online learning (Law et al., 2022). Statistical analyses revealed three key findings: (1) students' DL achievement increased with age and grade level, (2) inter-individual differences in DL scores (the second-level digital divide) widened over time, and (3) the fastest average growth in DL was observed among primary school students.

In the following, we address our findings in reverse order, starting with the results pertaining to DL development, particularly the identified cohort differences and individual growth in DL over time and their implications. Subsequently, we discuss the psychometric quality of the DLA with respect to the measurement challenges that researchers must address when developing DL assessments.

5.1 DL development

5.1.1 DL differences across grade levels

Secondary school students showed higher DL levels than primary students, consistent with prior research (Fraillon et al., 2018a; Reichert et al., 2020b). Older students usually have more experience using digital devices, and they self-report better technical and operational digital skills as well as digital knowledge than younger students (Machackova et al., 2024). Furthermore, Hong Kong's Technology Education Curriculum Guides (The Curriculum Development Council, 2017) emphasise different competences across grade levels. Primary education focuses on exploring technological concepts and becoming aware of technological developments and their societal impacts. At the junior secondary level (Secondary 1 to 3), the emphasis shifts to offering a broad and balanced technology education curriculum. At the senior secondary level (Secondary 4 to 6), schools are expected to provide electives on technology that enable students to specialise for lifelong learning or workforce preparation. This structured progression reflects increasing complexity and depth in technology-related learning objectives and the expectation that secondary school students reach higher levels of DL.

In our study, Secondary 5 students (Cohort 3) significantly outperformed Secondary 3 students (Cohort 2) in the T2 assessment, whereas two years prior, when students were in Secondary 3 and Secondary 1, respectively, there was no significant difference between the two cohorts. This trend suggests that middle adolescence, corresponding to the transition from Secondary 3 to Secondary 5, may be a pivotal period for DL development. The transition from focusing on processing and presenting information to engaging in problem-solving using IT tools at the upper secondary level reflects heightened DL requirements and may explain this shift. However, further research is necessary to explore factors facilitating DL development in greater detail. For example, our finding contrasts with self-reported DL progression in a six-country study in Europe, where older adolescents (15–17 years of age in 2021) on average self-reported slightly higher levels of digital skills in 2021, 2022, and 2023, while younger adolescents (aged 12–14 years in 2021) on average self-reported slightly greater improvements over that period (Machackova et al., 2024). This discrepancy suggests caution when relying on self-reports of DL; it also highlights the need to study how educational contexts influence DL development, as curricula in different regions may vary in their emphasis on DL at different grade levels.

Additionally, inter-person variations in DL were largest in Cohort 3, indicating greater diversity in DL among older students, which reflects a digital divide in DL. This finding raises concerns about the possible consequences resulting from this divide, such as lower productivity and wellbeing, limited learning gains, or lack of societal participation (Aydin, 2021; Livingstone et al., 2023). The COVID-19 pandemic may have contributed to these disparities, as disadvantaged students experienced the greatest setbacks during the pandemic (Tan et al., 2025). Early identification and prevention of digital inequalities are crucial to prevent them from compounding over time. The DLA developed in this research can facilitate early identification and help track the efficacy of efforts in raising students' DL. However, further research is needed to identify the factors contributing to DL divides.

Finally, students in T2 outperformed their T1 counterparts at the same grade level (e.g., the primary students—Cohort 1—at T2, when they were only in P5, reached a DL level comparable to S1 students at T1). This result is possibly due to the prolonged online schooling during the COVID-19 period, supporting claims that increased and/or more sophisticated digital technology use might enhance DL development (Chiu, 2023; Yu, 2022). Consequently, offering opportunities to engage in online activities may bolster DL (Aydin, 2021; Jara et al., 2015; Livingstone et al., 2019, 2023). However, the present study was not designed to measure the impacts of the pandemic on students' DL, and it cannot establish a causal relationship between the pandemic and the observed increase in DL scores.

5.1.2 DL growth over time

This study estimated the growth trajectories of students' DL over a period of two years. The positive growth rates show that students' DL improved overall, which is consistent with other longitudinal studies (Gnambs, 2021; Lazonder et al., 2020), including those based on self-reported digital skills (Machackova et al., 2024). The negative correlations between students' T1 DL scores and the DL growth rates further suggest that students with lower initial DL may have a greater potential to improve. Although previous studies have shown that prior DL is positively associated with later DL (Hübner et al., 2023), the current analysis highlights that students with lower initial DL can catch up, especially in primary school. Consequently, intervention programmes may be an effective means to quickly address DL deficits among students with comparatively low DL and to narrow the digital divide.

The DLA developed in this study demonstrated measurement invariance across grade levels and time points, allowing us to examine students' growth in DL accurately. A significant contribution of this study is the novel insight that growth trajectories differ among students in primary, lower secondary, and upper secondary schools. While such differences are known for other subjects (e.g., Lee, 2010), the current study adds that DL growth is more accelerated among primary students. Hence, educational efforts to raise DL and address (anticipated or existing) inequities in DL should start early. Further studies are needed to investigate the factors influencing DL growth trajectories throughout childhood and adolescence and to identify factors associated with both positive and negative DL growth. For example, previous studies have also shown that students' DL is influenced by family SES and access to technological devices (Livingstone & Helsper, 2007; Livingstone et al., 2015; Ren et al., 2022; van Dijk, 2006). However, the mere possession of digital devices or extensive online engagement does not guarantee higher DL (Park & Burford, 2013). Hence, it is recommended that educators permit students to engage in more online learning activities, especially those related to creativity and problem-solving, as active engagement through creating original content or solving personal or social problems can facilitate DL (Chang & Kuo, 2025). Future studies can then explore how students' background variables and online learning at school influence DL growth over time using the DLA.

5.2 Challenges in measuring DL over time

While the DLA offers the unique opportunity to study DL longitudinally, the psychometric results also offer insights for researchers interested in developing their own instruments. Developing DL assessments that enable researchers to examine DL development over time requires addressing several measurement challenges. These challenges include the need to keep DL instruments up to date in a rapidly changing technology environment and establishing DL scores that can be compared over time to compare different age cohorts, track individual growth, and examine gaps in DL across ages and over time. In the following subsections, we discuss these challenges and how they were addressed in developing the DLA that afforded the findings reported earlier. Future research could further expand the usability of the DLA by developing additional items to measure DL on the same scale in lower primary school and among college students, allowing even more comprehensive tracking of DL development across childhood, adolescence, and early adulthood.

5.2.1 Maintaining construct invariance when measuring DL over time

As new contexts for DL emerge (e.g., taking online classes), assessment instruments must be updated to measure digital competences that individuals actually need at the time (Wong et al., 2023), while following the same assessment framework. Outdated items must be retired, and new ones developed to ensure the validity of the measured construct. For example, as Skype became less popular and Zoom became a mainstream videoconferencing app, items related to the proper use of Zoom were added to the T2 DLA.

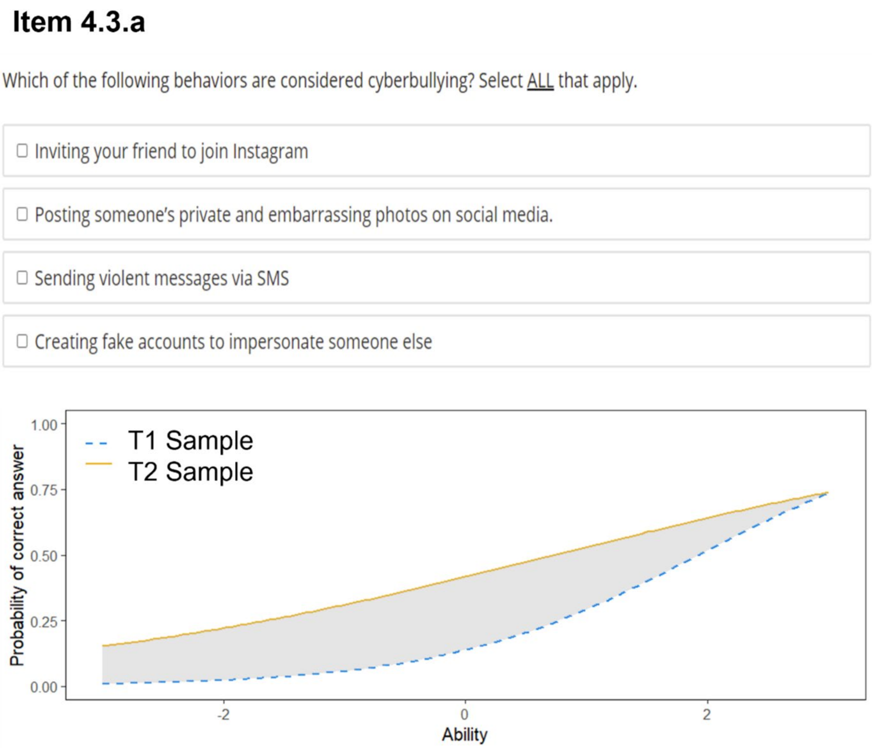
However, adding new items may lead to construct shifts and pose challenges to comparing scores over time (e.g., Martineau, 2006). Digital assessments can be influenced by prevailing technology or devices and familiarity with these technologies (Wong et al., 2023). Thus, routine tests for both construct- and item-level invariance are crucial. This study confirmed that the DLA maintained its unidimensionality at T2, forming the basis for linking the T1 and T2 DLA scores and enabling comparisons over time. The finding of unidimensionality is consistent with other research (Fraillon et al., 2015, 2018a, 2019; Senkbeil et al., 2014) and allows for creating shorter DLA versions for use in applied research to reduce the testing time (Purpura et al., 2015; Reichert et al., 2020b).

5.2.2 Examining item parameter drift when measuring DL over time

To accurately evaluate students' progress and compare gaps in DL over time, this study utilised a "common item design" to link scores across different assessment periods. This approach requires a proportion of items to remain constant over time and assumes their psychometric properties do not vary (Ainley et al., 2008; Ockwell et al., 2019). The predetermined number of consistent items between T1 and T2 fulfilled the guideline of at least 20% common items for a unidimensional construct (Angoff, 1984; Shea & Norcini, 1995).

Our analysis indicated that most items maintained their stability over time, affirming the effectiveness of the DLA in tracking individual growth. However, a few instances of significant item parameter drift (IPD) were identified, likely influenced by changes in curricula, societal norms, or technological advancements. For example, an item evaluating students' ability to identify cyberbullying demonstrated increased ease in T2, as illustrated by the item characteristic curve in Fig. 9. This finding suggests improved awareness of cyberbullying, possibly due to anti-cyberbullying efforts, which is recognised locally and worldwide as a critical issue (Castellanos et al., 2021; Chen, 2018; HKFYG, 2019; Strohmeier & Gradinger, 2022; Zhu et al., 2021; Zonta, 2023).

This case underscores the challenges associated with using a common-item design scale to track individuals' DL growth, as items could become easier (e.g., when the measured behaviours are more commonly exercised across society) or harder (e.g., when the measured skill is not required because respective applications have become obsolete) due to societal or technological shifts (Wong et al., 2023). Psychometric analyses on the longitudinal measurement invariance of DL instruments are required to maintain accuracy and reliability in assessing DL development over time.



The blue dashed line represents the item characteristic curve (ICC) of item 4.3.a from the T1 study. The yellow solid line represents the ICC of item 4.3.a from the T2 study.

Fig. 9 Example of an item with IPD

5.2.3 Ensuring test quality

Test quality was evaluated through item discrimination and difficulty, assessment fairness, and test reliability. Six items with low discrimination were excluded, ensuring that the DLA can differentiate between students' DL competences. The T2 DLA displayed a wide range of item difficulties, accommodating students from primary to secondary school levels. DIF analyses revealed fairness across gender, family SES, and testing modes, attesting to the DLA's suitability for investigating the digital divide in DL. Additionally, the DLA demonstrated reliability across a wide age range, making it a robust tool for assessing discrepancies in DL and tracking individual growth over time.

The analysis demonstrates the high quality of the T2 DLA instrument, showing that its scores have sufficient reliability for subsequent analyses. The DLA provides unbiased estimates of primary and secondary students' DL, allowing performance comparisons across age cohorts. The findings indicate that developing longitudinal DL assessment instruments that reliably and validly assess DL across different testing modes and social groups is feasible, enabling the examination of individual growth over time.

Consequently, the DLA offers unique opportunities for schools and researchers to examine factors influencing students' DL performance and development over time and factors contributing to digital divides in DL. Such research could inform targeted educational interventions, such as curriculum enhancements or DL training programs, tailored to students' specific needs (Aydin, 2021). Moreover, the DLA provides a means for schools to monitor and evaluate their DL initiatives, enabling program refinements as needed. Expanding the DLA beyond the current age groups also appears feasible.

6 Conclusions and limitations

This paper addresses challenges in developing a valid longitudinal DL assessment through rigorous procedures and psychometric analyses. It is the first study that (1) developed a performance-based DL assessment (instead of merely relying on the more common self-report measures of DL), which was (2) applied to explore DL differences and growth across three age cohorts using a longitudinal design (instead of using a less informative cross-sectional study) that (3) included students at both primary and secondary school levels (instead of focusing on a narrower age range). The comprehensive validation procedures that have been implemented address the deficiencies of some existing DL assessments, which frequently lack sufficient psychometric evidence to support their usage, as indicated in a recent review (Mattar et al., 2022). The comprehensive DLA, which has been demonstrated to be unidimensional across age groups and time, effectively measures DL development from late childhood to late adolescence.

By developing and validating the DLA designed for longitudinal application across a broad age range, from late childhood to late adolescence, this study significantly contributes to the field of DL assessment. Few existing DL assessments possess strong psychometric properties while also enabling the simultaneous examination of

DL growth among both primary and secondary school students (Chourio-Acevedo et al., 2024; Godaert et al., 2022). Unlike previous assessments that often focus on specific age groups or rely solely on cross-sectional data or self-reported digital skills, the DLA facilitates the tracking of individual DL development over time, providing valuable insights into growth patterns and the widening digital divide as students progress through their education. The longitudinal design of this study enables the observation of DL trajectories, offering empirical evidence on how DL evolves, which is crucial for informing targeted educational interventions.

However, some limitations should be acknowledged. Firstly, sample attrition occurred as not all T1 students participated in the T2 Study due to class allocations, potentially affecting the representativeness of DL growth for secondary students. Consequently, further research with more relatively constant samples over time is necessary to explore the development of DL in secondary school students more comprehensively. Researchers may also consider applying matching methods to impute missing data when attrition is problematic. Secondly, the current DLA is relatively long; thus, developing a shorter or a computerised adaptive version of the DLA with sufficient psychometric properties could enhance administration efficiency and student engagement. Thirdly, a single DL score has limitations in providing detailed diagnostic information. Future research could explore the use of longitudinal cognitive diagnostic models for the assessment to have more formative value (e.g., Liang et al., 2023; Madison & Bradshaw, 2018). Fourthly, while the current study demonstrates that the interindividual variation in DL increases as students age, it does not identify the variables responsible for this widening digital divide. Further investigation is required to explore the factors that explain and, if possible, identify ways to minimize this variability. It is important to note that COVID-19 occurred between the T1 and T2 studies. Consequently, we lack evidence to determine whether and to what extent the identified patterns of differences and changes were influenced by the COVID-19 period. Future studies conducted during an uninterrupted period could establish how DL normally grows over time. Finally, Hong Kong is a highly developed region with very high levels of digital device ownership and Internet penetration (International Telecommunication Union, 2025). DL performance levels and growth rates may differ in regions with lower penetration rates, where students likely also face greater challenges adapting to online learning. Future research should gather comparative data in regions with differential digital access to examine the validity of the DLA and to explore DL development in contexts where ICT connectivity rates are lower.

Despite these limitations, this study contributes to effective DL measurement and offers insights that can inform the creation of DL curricula and interventions. This research successfully developed a fair DL assessment across a wide age range. The high psychometric quality of the DLA yields reliable and comparable DL scores for 8- to 17-year-olds, an age range rarely covered in cross-sectional DL assessment studies and even less commonly in longitudinal research, making it an invaluable assessment tool for stakeholders, especially in regions with high levels of ICT connectivity. Longitudinal tracking of primary and secondary students' DL development revealed continued growth and potential for faster improvement among students with lower DL. Thus, adequate educational interventions can likely narrow gaps in DL performance. The widening DL gaps and varying developmental trajectories identified in this study require particular attention. Specifically, growth in DL skills appears to be faster at a younger age, and inter-individual differences in DL performance seem to increase with age. To avoid further inequities down the road, educational efforts to address gaps in DL should be implemented as early as possible and, ideally, start already in primary school.

In conclusion, this study emphasises the need for longitudinal investigations into individual differences and DL growth in early adolescence over an extended period. The DLA provides a valuable tool to aid longitudinal studies into DL development across ages and over time. It is well-suited for further inquiries into the digital divide to examine, for example, how differences in access to and uses of ICT, individuals' familiarity with digital technologies, and social, economic, or cultural factors influence DL development or the potential outcomes of discrepancies in DL.

This study also highlights the widening digital divide in DL, especially as students age. Consequently, early educational interventions are needed to address DL gaps, ideally in primary school, before these gaps start to widen. Implementing educational efforts early on can help mitigate the disparities in students' DL growth as they progress through their education. In addition, regular monitoring of students' DL is needed to identify potential setbacks in DL and to provide timely intervention. These issues underscore the need for reliable and valid assessment instruments, which are paramount for tracking DL across various age groups and over time.

However, changes to prevalent technologies can also influence the specific digital skills required to prosper in the digital world and affect the psychometric properties (e.g., item difficulty) of previously validated assessment items and tasks. To ensure the assessment remains valid, such developments need to be constantly monitored and inform the necessity of developing new and retiring outdated items. Future research and development are also essential to keep the DLA, in particular, updated in response to rapid technological advancements. However, it should be noted that validating assessment instruments takes time, and sudden changes at a societal level that may rapidly accelerate the development of specific skills or behaviours (e.g., video conferencing skills), such as a pandemic sparking online learning, can be challenging to address quickly. Nonetheless, understanding and acknowledging the technological and societal contexts in which DL assessments are administered is essential to rigorous research and reporting.

From a policy perspective, implementing a regular, large-scale DL assessment at the national or regional level could provide a more systematic approach to monitoring students' DL. Such assessment programs would help policymakers track progress over time, identify gaps in digital skills development, and inform targeted interventions to enhance students' preparedness for an increasingly digitalized society. Establishing a structured DL monitoring system could also support the timely implementation and modification of evidence-based curriculum adjustments and teacher training programs addressing urgent needs and tailored to emerging technological demands.

Finally, there is still a need to understand DL development at even younger ages and beyond the formal school system into adulthood. In addition, further research is needed to examine the factors influencing DL development and cross-cohort differences in DL growth. A key policy priority could be addressing the persistent digital divide at an early stage. Targeted initiatives such as providing digital learning subsidies, improving broadband infrastructure in primary schools, and enhancing teacher training in DL could support the development of DL in students. Exploring these areas will provide a more comprehensive understanding of DL development and facilitate the creation of effective strategies to bridge the digital divide.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10639-025-13592-8.

Acknowledgements We thank all participating schools and students for providing valuable data for this study.

Author contribution Conceptualization: Qianqian PAN, Frank REICHERT, Qianru LIANG, Jimmy DE LA TORRE, Nancy LAW.

Methodology: Qianqian PAN, Frank REICHERT, Jimmy DE LA TORRE, Nancy LAW.

Formal analysis and investigation: Qianqian PAN, Qianru LIANG.

Writing—original draft preparation: Qianqian PAN, Frank REICHERT.

Writing—review and editing: Qianqian PAN, Frank REICHERT, Qianru LIANG, Jimmy DE LA TORRE, Nancy LAW.

Funding acquisition: Nancy LAW.

Resources: Nancy LAW.

Supervision: Jimmy DE LA TORRE, Nancy LAW.

Funding This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China, under its Theme-based Research Scheme [Project No. T44-707/16-N].

Data availability The full Digital Literacy Assessment instrument cannot be made available publicly to maintain test security. The empirical data supporting the findings of this study will be made available upon reasonable request.

Declarations

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

References

Aesaert, K., & van Braak, J. (2015). Gender and socioeconomic related differences in performance based ICT competences. *Computers & Education, 84*, 8–25. https://doi.org/10.1016/j.compedu.2014.12.017

Ainley, J., Fraillon, J., Freeman, J., & Murphy, M. (2008). *National Assessment Program: Information Communication and Technology Literacy: Year 6 and year 10: Technical report 2005.* Australian Council for Educational Research. https://research.acer.edu.au/ict_literacy/5. Accessed 3 May 2025.

Ainley, J., Schulz, W., & Fraillon, J. (2016). *A global measure of digital and ICT literacy skills.* UNESCO. https://research.acer.edu.au/ict_literacy/12. Accessed 3 May 2025.
Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15
Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.
Aydin, M. (2021). Does the digital divide matter? Factors and conditions that promote ICT literacy. *Telematics and Informatics, 58*, 101536. https://doi.org/10.1016/j.tele.2020.101536
Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.
Bashir, S., & Miyamoto, K. (2020). *Digital skills: Frameworks and programs.* World Bank. https://openknowledge.worldbank.org/handle/10986/35080
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01
Battauz, M. (2015). equateIRT: An *R* package for IRT test equating. *Journal of Statistical Software, 68*(7), 1–22. https://doi.org/10.18637/jss.v068.i07
Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3–14. https://doi.org/10.1111/j.1745-3992.2009.00158.x
Buchan, M. C., Bhawra, J., & Katapally, T. R. (2024). Navigating the digital world: Development of an evidence-based digital literacy program and assessment tool for youth. *Smart Learning Environments, 11*(1), 8. https://doi.org/10.1186/s40561-024-00293-x
Carretero, S., Vuorikari, R., & Punie, Y. (2017). *DigComp 2.1: The digital competence framework for citizens with eight proficiency levels and examples of use*. Publications Office of the European Union. https://doi.org/10.2760/38842
Castellanos, A., Ortega-Ruipérez, B., & Aparisi, D. (2021). Teachers' perspectives on cyberbullying: A cross-cultural study. *International Journal of Environmental Research and Public Health, 19*(1), 257. https://doi.org/10.3390/ijerph19010257
Chalmers, R. P. (2012). **mirt**: A multidimensional item response theory package for the *R* environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06
Chang, C.-Y., & Kuo, H.-C. (2025). The development and validation of the digital literacy questionnaire and the evaluation of students' digital literacy. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-13216-7
Chen, J.-K. (2018). Cyberbullying among secondary school students in Hong Kong. *The Hong Kong Journal of Social Work, 52*(1/2), 49–62. https://doi.org/10.1142/S0219246218000050
Chiu, T. K. F. (2023). Student engagement in K-12 online learning amid COVID-19: A qualitative approach from a self-determination theory perspective. *Interactive Learning Environments, 31*(6), 3326–3339. https://doi.org/10.1080/10494820.2021.1926289
Chourio-Acevedo, L., Köhler, J., Coscarelli, C., Gacitúa, D., Proaño-Ríos, V., & González-Ibáñez, R. (2024). *Information literacy development and assessment at school level: A systematic review of the literature* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2404.19020
Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series, 2010*(2), i–41. https://doi.org/10.1002/j.2333-8504.2010.tb02236.x
Engzell, P., Frey, A., & Verhagen, M. D. (2021). Learning loss due to school closures during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences, 118*(17), e2022376118. https://doi.org/10.1073/pnas.2022376118
Fraillon, J., & Duckworth, D. (2023). Computer and information literacy framework. In J. Fraillon & M. Rožman (Eds.), *IEA International Computer and Information Literacy Study 2023: Assessment framework* (pp. 25–41). Springer.
Fraillon, J., Schulz, W., Friedman, T., Ainley, J., & Gebhardt, E. (2015). *International Computer and Information Literacy Study: ICILS 2013 technical report.* IEA. https://research.acer.edu.au/ict_literacy/10. Accessed 3 May 2025.
Fraillon, J., Gebhardt, E., Friedman, T., Duckworth, D., O'Malley, K., & Vemon, K. (2018a). *NAP sample assessment ICT literacy: Years 6 and 10.* Australian Curriculum, Assessment and Reporting Authority. https://research.acer.edu.au/ict_literacy/13. Accessed 3 May 2025.
Fraillon, J., Gebhardt, E., O'Malley, K., Murphy, M., Kwong, R., Duckworth, D., Friedman, T., Fallas, J., & Ockwell, L. (2018b). *NAP Sample Assessment ICT Literacy (technical report).* Australian

Curriculum, Assessment and Reporting Authority. https://www.nap.edu.au/docs/default-source/default-document-library/nap-ictl17_tr_final.pdf. Accessed 3 May 2025.
Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *IEA International Computer and Information Literacy Study 2018 assessment framework*. Springer.
Gnambs, T. (2021). The development of gender differences in information and communication technology (ICT) literacy in middle adolescence. *Computers in Human Behavior, 114*, 106533. https://doi.org/10.1016/j.chb.2020.106533
Godaert, E., Aesaert, K., Voogt, J., & van Braak, J. (2022). Assessment of students' digital competences in primary school: A systematic review. *Education and Information Technologies, 27*(7), 9953–10011. https://doi.org/10.1007/s10639-022-11020-9
Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20*(4), 369–377.
Heitink, M. (2018). *Eliciting teachers' and students' technological competences* [PhD, University of Twente]. https://doi.org/10.3990/1.9789036546850
HKFYG (The Hong Kong Federation of Youth Groups). (2019, October 4). Youth Hong Kong: From physical to digital bullying. HKFYG. https://hkfyg.org.hk/en/2019/10/04/youth-hong-kong-from-physical-to-digital-bullying-2/. Accessed 3 May 2025.
Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge.
Holland, P. W., & Thayer, D. T. (1986, April). *Differential item performance and the Mantel-Haenszel Procedure*. [Paper presentation]. 67th American Educational Research Association (AERA) Annual Meeting, San Francisco, CA. https://eric.ed.gov/?id=ED272577. Accessed 3 May 2025.
Hübner, N., Fahrbach, T., Lachner, A., & Scherer, R. (2023). What predicts students' future ICT literacy? Evidence from a large-scale study conducted in different stages of secondary school. *Computers & Education, 203*, 104847. https://doi.org/10.1016/j.compedu.2023.104847
International Telecommunication Union (ITU). (2025). ITU DataHub ICT Statistics. Indicators: Connectivity (Year: 2023). ITU. https://datahub.itu.int/. Accessed 3 May 2025.
Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement, 44*(1), 1–21. https://doi.org/10.1111/j.1745-3984.2007.00024.x
Jara, I., Claro, M., Hinostroza, J. E., San Martín, E., Rodríguez, P., Cabello, T., Ibieta, A., & Labbé, C. (2015). Understanding factors related to Chilean students' digital skills: A mixed methods analysis. *Computers & Education, 88*, 387–398. https://doi.org/10.1016/j.compedu.2015.07.016
Jin, K.-J., Reichert, F., Cagasan, L. P. [Jr.], de la Torre, J., & Law, N. (2020). Measuring digital literacy across three age cohorts: Exploring the dimensionality and performance differences. *Computers & Education*, *157*, 103968. https://doi.org/10.1016/j.compedu.2020.103968
Kastberg, D., Cummings, L., Ferraro, D., & Perkins, R. C. (2021). *Technical report and user guide for the 2018 Program for International Student Assessment (PISA).* (NCES 2021–011). National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2021011. Accessed 3 May 2025.
Kim, D.-I., Choi, S. W., Lee, G., & Um, K. R. (2008). A comparison of the common-item and random-groups equating designs using empirical data. *International Journal of Selection and Assessment, 16*(2), 83–92. https://doi.org/10.1111/j.1468-2389.2008.00413.x
Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). Springer
Kong, S.-C., Wang, Y.-Q., & Lai, M. (2019). Development and validation of an instrument for measuring digital empowerment of primary school students. *Proceedings of the ACM Conference on Global Computing Education*, (172–177). Association for Computing Machinery. https://doi.org/10.1145/3300115.3309523
Law, N. W. Y., 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. https://www.edb.gov.hk/attachment/en/edu-system/primary-secondary/applicable-to-primary-secondary/it-in-edu/phase%20ii%20study_final%20report.pdf
Law, N., Woo, D., de la, J., & Wong, G. (2018). *A global framework of reference on digital literacy skills for indicator 4.4.2*. UNESCO Institute of Statistics. https://uis.unesco.org/sites/default/files/documents/ip51-global-framework-reference-digital-literacy-skills-2018-en.pdf. Accessed 3 May 2025.
Law, N., Pan Q., Tao, S., Liang, Q., Chen, L. L., Rao, N., Reichert, F., & de la Torre, J. (2022). *Hong Kong students' digital citizenship development 2019-2021: Findings from a longitudinal study.* The University of Hong Kong. https://doi.org/10.25442/hku.22085738.v1

Law, E. L.-C., Vostanis, P., & O'Reilly, M. J. (2023). Insights from impacts of the digital divide on children in five majority world countries during the COVID-19 pandemic. *Behaviour & Information Technology, 42*(15), 2696–2715. https://doi.org/10.1080/0144929X.2022.2141136

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

Lee, J. (2010). Tripartite growth trajectories of reading and math achievement: Tracking national academic progress at primary, middle, and high school levels. *American Educational Research Journal, 47*(4), 800–832. https://doi.org/10.3102/0002831210365009

Li, Y., & Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement, 36*(1), 3–20. https://doi.org/10.1177/0146621611432864

Liang, Q., de la Torre, J., & Law, N. (2023). Latent transition cognitive diagnosis model with covariates: A three-step approach. *Journal of Educational and Behavioral Statistics, 48*(3), 690–718. https://doi.org/10.3102/10769986231163320

Livingstone, S., & Helsper, E. (2007). Gradations in digital inclusion: Children, young people and the digital divide. *New Media & Society, 9*(4), 671–696. https://doi.org/10.1177/1461444807080335

Livingstone, S., Mascheroni, G., & Stoilova, M. (2023). The outcomes of gaining digital skills for young people's lives and wellbeing: A systematic evidence review. *New Media & Society, 25*(5), 1176–1202. https://doi.org/10.1177/14614448211043189

Livingstone, S., Mascheroni, G., & Staksrud, E. (2015). *Developing a framework for researching children's online risks and opportunities in Europe*. EU Kids Online, The London School of Economics and Political Science. https://eprints.lse.ac.uk/64470/. Accessed 3 May 2025

Livingstone, S., Kardefelt-Winther, D., & Kanchev, P. (2019). *Is there a ladder of children's online participation? Findings from three Global Kids Online countries* (Innocenti Research Brief No. 2019–02). UNICEF Office of Research – Innocenti. https://www.unicef-irc.org/publications/1019-ladder-of-childrens-online-participation-findings-from-three-gko-countries.html. Accessed 3 May 2025

Machackova, H., Jaron Bedrosova, M., Tolochko, P., Muzik, M., Waechter, N., & Boomgaarden, H. (2024). Digital skills among children and youth: A report from a three-wave longitudinal study in six European countries. Zenodo. https://doi.org/10.5281/ZENODO.11173075

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika, 83*(4), 963–990. https://doi.org/10.1007/s11336-018-9638-5

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31*(1), 35–62. https://doi.org/10.3102/10769986031001035

Mattar, J., Ramos, D. K., & Lucas, M. R. (2022). DigComp-based digital competence assessment tools: Literature review and instrument analysis. *Education and Information Technologies, 27*(8), 10843–10867. https://doi.org/10.1007/s10639-022-11034-3

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14*(2), 126–149. https://doi.org/10.1037/a0015857

Ockwell, L., Daraganov, A., & Schulz, W. (2019). Scaling procedures for ICILS 2018 test items. In J. Fraillon, J. Ainley, W. Schulz, D. Duckworth, & T. Friedman (Eds.), *IEA International Computer and Information Literacy Study 2018 assessment framework* (pp. 133–158). Springer.

Pan, Q., Reichert, F., de la Torre, J., & Law, N. (2022). Measuring digital literacy during the Covid-19 pandemic: Experiences with remote assessment in Hong Kong. *Educational Measurement: Issues and Practice*, *41*(1), 46–50. https://doi.org/10.1111/emip.12498

Park, S., & Burford, S. (2013). A longitudinal study on the uses of mobile tablet devices and changes in digital media literacy of young adults. *Educational Media International, 50*(4), 266–280. https://doi.org/10.1080/09523987.2013.862365

Pedaste, M., Kallas, K., & Baucal, A. (2023). Digital competence test for learning in schools: Development of items and scales. *Computers & Education, 203*, 104830. https://doi.org/10.1016/j.compedu.2023.104830

Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. *School Psychology Review, 44*(1), 41–59. https://doi.org/10.17105/SPR44-1.41-59

Raudenbush, S. W., & Chan, W.-S. (1992). Growth curve analysis in accelerated longitudinal designs. *Journal of Research in Crime and Delinquency, 29*(4), 387–411. https://doi.org/10.1177/0022427892029004001

Reichert, F., Lam, P., Loh, E. K. Y., & Law, N. (2020a). *Hong Kong students' digital citizenship development: Initial findings.* The University of Hong Kong. https://doi.org/10.25442/hku.22085726.v1

Reichert, F., Zhang, D. J., Law, N. W. Y., Wong, G. K. W., & de la Torre, J. (2020b). 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

Reichert, F., Pan, Q., & Chen, L. (2023). *Digital literacy assessment* (ED/GEMR/MRT 2023/P1/25). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386202.locale=en. Accessed 3 May 2025.

Ren, W., Zhu, X., & Yang, J. (2022). The SES-based difference of adolescents' digital skills and usages: An explanation from family cultural capital. *Computers & Education, 177*, 104382. https://doi.org/10.1016/j.compedu.2021.104382

Robitzsch, A. (2022). *sirt: Supplementary Item Response Theory Models* (Version 3.12–66) [R]. https://CRAN.R-project.org/package=sirt. Accessed 3 May 2025.

Scheerder, A., van Deursen, A., & van Dijk, J. (2017). Determinants of Internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. *Telematics and Informatics, 34*(8), 1607–1624. https://doi.org/10.1016/j.tele.2017.07.007

Schwarz, S., Bieg, T., Svecnik, E., Schmölz, A., Geppert, C., & Gerdenitsch, C. (2024). Digital Competence Scale (DCS): A short self-assessment instrument for measuring digital competences. *Nordic Journal of Digital Literacy, 19*(3), 126–143. https://doi.org/10.18261/njdl.19.3.2

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://www.jstor.org/stable/2958889. Accessed 3 May 2025.

Senkbeil, M., & Ihme, J. M. (2021). *NEPS technical report for computer literacy: Scaling results of Starting Cohort 3 for Grade 12* (NEPS Survey Paper No. 90). Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/NEPS:SP90:1.0

Senkbeil, M., Ihme, J. M., & Adrian, E. D. (2014). *NEPS technical report for computer literacy – Scaling results of Starting Cohort 3 in Grade 6 (wave 2)* (NEPS Working Paper No. 39). National Educational Panel Study. https://www.lifbi.de/Portals/2/Working%20Papers/WP_XXXIX.pdf. Accessed 3 May 2025.

Shea, J., & Norcini, J. (1995/2012). Equating. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices.* (pp. 253–287). Buros Center for Testing. https://digitalcommons.unl.edu/buroslicensure/16. Accessed 3 May 2025.

Siddiq, F., & Scherer, R. (2019). Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. *Educational Research Review, 27*, 205–217. https://doi.org/10.1016/j.edurev.2019.03.007

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. https://doi.org/10.1016/j.edurev.2016.05.002

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331–354. https://doi.org/10.1177/014662169602000403

Strohmeier, D., & Gradinger, P. (2022). Cyberbullying and cyber victimization as online risks for children and adolescents. *European Psychologist, 27*(2), 141–150. https://doi.org/10.1027/1016-9040/a000479

Tan, C. Y., Jang, S. T., Lam, S. M., An, A. Q., & Venus Lo, U. K. (2025). Teaching and learning challenges due to the COVID-19 pandemic: A systematic review. *Educational Research Review*, 100667. https://doi.org/10.1016/j.edurev.2025.100667

The Curriculum Development Council. (2017). *Technology Education: Key learning area curriculum guide (Primary 1—Secondary 6)*. The Education Bureau HKSAR Government. https://www.edb.gov.hk/en/curriculum-development/kla/technology-edu/curriculum-doc/index.html. Accessed 3 May 2025.

Tolboom, J., & van Rooyen, L. (Eds.). (2021). *Digital literacy: Development and implementation in European countries*. CIDREE yearbook 2021. SLO, Netherlands Institute for Curriculum Development. https://www.cidree.org/wp-content/uploads/2021/12/cidree_yearbook_2021_digital_literacy_2.pdf. Accessed 3 May 2025.

van Dijk, J. A. G. M. (2006). Digital divide research, achievements and shortcomings. *Poetics, 34*(4–5), 221–235. https://doi.org/10.1016/j.poetic.2006.05.004

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. Springer. https://doi.org/10.1007/b97446

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, 14*(S2), 67–86. https://doi.org/10.1007/s11618-011-0182-7

WestEd. (2018). *Technology and Engineering Literacy framework for the 2018 National Assessment of Educational Progress*. National Assessment Governing Board, U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED594359.pdf. Accessed 3 May 2025.

Wong, G. K.-W., Reichert, F., & Law, N. (2023). Reorienting the assessment of digital literacy in the twenty-first century: A product-lifecycle and experience dependence perspective. *Educational Technology Research and Development, 71*(6), 2389–2412. https://doi.org/10.1007/s11423-023-10278-1

Yu, Z. (2022). Sustaining student roles, digital literacy, learning achievements, and motivation in online learning environments during the COVID-19 pandemic. *Sustainability, 14*(8), 4388. https://doi.org/10.3390/su14084388

Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika, 72*(1), 69–91. https://doi.org/10.1007/s11336-004-1257-7

Zhu, C., Huang, S., Evans, R., & Zhang, W. (2021). Cyberbullying among adolescents and children: A comprehensive review of the global situation, risk factors, and preventive measures. *Frontiers in Public Health, 9*, 634909. https://doi.org/10.3389/fpubh.2021.634909

Zonta Club of Hong Kong East (Zonta). (2023). *Say NO to Cyber Bullying Video Competition*. Zonta. https://m21.hk/zontaHKE2023/?language=eng. Accessed 3 May 2025.

Authors and Affiliations

Qianqian Pan[1] · Frank Reichert[2] · Qianru Liang[3] · Jimmy de la Torre[4] · Nancy Law[5]

✉ Frank Reichert
reichert@hku.hk

Qianqian Pan
qianqian.pan@nie.edu.sg

Qianru Liang
qr.liang@outlook.com

Jimmy de la Torre
j.delatorre@hku.hk

Nancy Law
nlaw@hku.hk

1 Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University, Singapore, Singapore

2 Academic Unit of Social Contexts and Policies of Education, Faculty of Education, The University of Hong Kong, Pok Fu Lam Road, Hong Kong, Hong Kong SAR China

3 Guangdong Institute of Smart Education, Jinan University, Guangzhou, China

4 Academic Unit of Human Communication, Learning, and Development, Faculty of Education, The University of Hong Kong, Pok Fu Lam Road, Hong Kong, Hong Kong SAR China

5 Academic Unit of Mathematics, Science, and Technology, Faculty of Education, The University of Hong Kong, Pok Fu Lam Road, Hong Kong, Hong Kong SAR China