Developing evaluation tools for assessing the educational potential of apps for preschool children in the UK

Joanna Kolak[a], Sarah. H. Norgate[a], Padraic Monaghan[b,c] & Gemma Taylor[a]

[a] *Department of Psychology, School of Health and Society, University of Salford, Salford, M5 4WT, UK*

[b] *Department of Psychology, Lancaster University, Lancaster, UK*

[c] *Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, Netherlands*

Corresponding details:

Joanna Kolak, Department of Psychology, School of Health and Society, University of Salford, Salford, M5 4WT, UK. Email: j.kolak@salford.ac.uk. Tel: +44 (0) 161 295 5083

Twitter: @joanna_kolak

Biographical Notes

Joanna Kolak is a Researcher in Psychology at the University of Salford and a PhD student at the Faculty of Psychology at the University of Warsaw. Her research focusses on monolingual and bilingual language acquisition and children's learning from digital media.

Gemma Taylor, Ph.D. is a lecturer in Developmental Psychology at the University of Salford. Her research focusses on the impact of screen media on children's early cognitive development. Twitter: @Gemma_Taylor1

Sarah H. Norgate, Ph.D. was formerly a Reader in Applied Developmental Psychology at the University of Salford during the research. Sarah is an author and has edited her second book 'Flexible Work: Designing Our Healthier Future Lives' (Routledge Taylor and Francis, 2020) with Professor Cary Cooper which concerns the future of flexible work including remote work in an era where our work life balance is navigated in a digital connected world. Twitter: @sarahnorgate

Padraic Monaghan is Professor of Cognition at Lancaster University. His research focusses on language acquisition, language evolution, and reading development.

Developing evaluation tools for assessing the educational potential of apps for preschool children in the UK

Selecting high quality apps can be challenging for caregivers and educators. We here develop tools evaluating educational potential of apps for preschool children. In Study 1, we developed two complementary evaluation tools tailored to different audiences. We grounded them in developmental theory and linked them to research on children's experience with digital media. In Study 2 we applied these tools to a wide sample of apps in order to illustrate their use and to address the role of cost in quality of educational apps. There are concerns that a social disadvantage may lead to a digital disadvantage, an "app gap". We thus applied our tools to the most popular free (N=19) and paid (N=24) apps targeting preschoolers. We found that the "app gap" associated with cost is only related to some aesthetic features of apps rather than any observable educational advantage proffered by paid apps. Our study adds a novel contribution to the research on children's apps by developing tools to be used across a wide range of audiences, providing the first description of the quantity of app design features during app use and evaluating the educational potential of free and paid apps.

Word count: 193

Keywords: educational technology; app evaluation; touchscreen apps; app features; app design; app gap; early years; children

Introduction

Touchscreen devices are increasingly popular among children under the age of 5 (e.g., Chen & Adler, 2019). An estimated 80,000 apps claim to be 'educational' (Healthy Children, 2018) within the context of an unregulated market. Yet, there is a consensus among researchers that the majority of children's apps advertised as "educational" lack educational value and any foundation in research (Ólaffson et al., 2013). This means that informed decisions about which apps are high quality can be challenging for parents and educators (Livingstone et al., 2018) who could potentially benefit from an app evaluation tool based on early years learning theory. An app evaluation tool could also benefit app developers who want to ensure that the products they create include high quality features.

To date, a number of authors have proposed evaluation tools[1] to assess educational potential and design of apps for children (Callaghan & Reich, 2018; Chau, 2014; Department for Education, 2019; Highfield & Goodwin, 2013; Hirsh-Pasek et al., 2015; Lee & Cherner, 2015; Lee & Kim, 2015; McManis & Parks, 2011; Papadakis & Kalogiannakis, 2017; Papadakis et al., 2018; Shoukry, Sturm, & Galal-Edeem, 2015; Walker, 2011). We summarise the most recent (2015 – 2020) evaluation tools in Table 1.

[1] We use the term "evaluation tool" to refer to rubrics, frameworks and schemes for consistency throughout the paper.

Table 1. Summary of evaluation tools available to assess educational value of apps for young children

Authors	Aim of the tool	Main group that the tool targets	Key areas of evaluation	Methods of evaluation	Tool validated?	Theoretical underpinning	No. and type of apps tested using the tool, and findings	Limitations
Shoukry et al. (2015)	Pre-MEGa framework - for designing and evaluating mobile educational games for preschoolers	App developers Users who compare products	A set of heuristics for combining play, learning, usability and mobility requirements. 23 categories.	The categories were devised in order to be incorporated into an available rating system.	No, it was intended as a theoretical framework.	Literature on children's use of digital media	No apps tested	Some guidelines are unclear (e.g. 'Offers some uncertain outcomes') The tool is not fully developed yet for target users. Includes some subjective criteria, e.g. 'attractive, fun, humourful', which cannot be measured objectively by users.
Lee & Kim (2015)	Evaluating educational potential of apps	Educators	Four main areas: Teaching & learning, Screen design, Technology, Economy & ethics 23 close-ended questions.	Answering yes or no to the questions.	Yes, using exploratory factor analysis.	Pedagogical and learning literature	No apps tested	Adjustments are needed for the too to be effective in the target population. Does not allow to quantify the app features. Includes some subjective criteria, e.g. 'Is an app funny and interesting, exciting and imaginative?', which cannot be measured objectively by users.
Lee & Cherner (2015)	Evaluating the educational potential of instructional apps	Educators, researchers, app developers	Three main domains: Instruction, Design, Engagement. 24 dimensions.	Dimensions rated on 1-5 scale; indicator descriptors to each score provided.	Yes, by conducting face and content validity study.	Literature on evaluating technologies for educational purposes; some literature on learning	No apps tested	Apps have to be classified based or their purpose before the rating, which limits comparisons of apps more broadly to apps within the same category. Uses technical language and requires knowledge about specific frameworks, e.g. Webb's Depth of Knowledge. Measures child's enjoyment

								through a question (e.g., 'Will the app's content likely appeal to its targeted audience?').
Hirsh-Pasek et al. (2015)	Evaluating the educational potential of apps	Researchers, educators, designers, caregivers	Four pillars of learning: Active, engaged, meaningful, and socially interactive learning.	Rating an app as 'low', 'medium', or 'high' on each of the four pillars, and on the learning goal.	No	Guided by the Science of Learning framework (Bransford et al., 1999).	The use demonstrated on three apps.	Practical use requires an in-depth understanding of the science of learning. The framework does not have items and descriptors which makes it difficult to use in a systematic way.
Papadakis et al. (2017)	REVEAC tool = Rubric for evaluating educational potential of apps	Educators	Four main areas: Educational content, Design, Functionality, Technical characteristics. 18 criteria.	Criteria rated on a 1-4 scale; indicator descriptors to each score provided.	Yes, by conducting content validity, internal consistency and convergent validity study.	Literature on children's interactions with digital media	No apps tested	Less than half of the criteria directly linked to the educational content of apps. Some criteria measure user's engagement/emotional excitement (e.g., 'emotionally excites the child') rather than being objective.
Papadakis et al (2018)	Using an evaluation tool developed by Papadakis et al. (2017) (REVEAC) to conduct app review on Android Google Play Greek educational	Educators	See Papadakis et al (2017) .	See Papadakis et al. (2017)	See Papadakis et al. (2017)	See Papadakis et al. (2017)	40 math and literacy apps evaluated. Only 2 apps scored higher than the average rubric score. Discrepancy between the rubric scores and the website rating system. Most apps invited drill-and-practice style, none provided customisation	The evaluation was based only on the presence or absence of a feature (does not allow for quantifying app features). Review focused exclusively on Android free apps.

	ebook and game apps							
Callaghan & Reich (2018)	Coding and analysing content of preschool apps; comparing the free vs paid apps; and Apple vs Android apps	Educators, caregivers, app developers	4 main areas: Simplicity and clarity of goals, Feedback and rewards, Structure of challenge, Mobile app-based interactions. Over 70 codes.	Using the codes to code app features in order to conduct conventional and directed content analysis.	Yes, inter-rater reliability.	Developmental and learning literature, literature on digital heuristics	171 math and literacy apps coded and analysed. Few apps provided developmentally appropriate guidance (e.g. apps rarely provided scaffolded feedback); few differences between free vs paid apps (paid apps repeated instructions, used prizes as rewards and increased/decreased in challenge more often than free apps)	The evaluation based only on the presence or absence of a feature. Review focused only on math and literacy apps.
Departme nt for Education (England) (2019)	Evaluating communicati on, language and literacy apps for early years	App developers, local authorities, caregivers, educators	Five main categories: Educational content, in-app support for learning and development, interaction, engagement, design/functionali ty. 28 criteria.	Criteria rated on a 1-4 scale; indicator descriptors to each score provided.	No	British Early Years Foundation Stage framework	No apps tested	Uses technical language that is not appropriate for all target users, e.g., caregivers. Some items require specific knowledge about child development (e.g. developmentally appropriate and effective practice in the development of young children). Some of the descriptors are not specific enough (e.g., 'App likely to engage the child'). Includes some subjective criteria, e.g. 'promotes fun, enjoyment, exciting interactions.'

As can be seen in Table 1, there are a number of limitations with the existing tools, some of which were identified by the authors themselves. Specifically, almost all the tools have a long list of criteria (18 - 70+ items) which makes app evaluation time consuming and not practical. The majority of the tools lack examples from children's apps that could allow an in-depth understanding of the descriptors. The descriptors of the items are often not specific enough; they include ambiguous or unclear terminology. Some of the tools also lack theoretical underpinning; they do not draw clear links to developmental theory. Only two of the tools had the content validity assessed, and none of the content validity assessments involved caregivers as participants.

Importantly, only three out of eight tools were aimed at caregivers. Given that preschool aged children use touchscreen devices frequently (e.g., according to Ofcom (2019), children aged 3-4 years living in the UK spend 48 minutes per weekday playing games on a touchscreen device), it is crucial to help parents select good quality apps for their children. The majority of the tools have not been applied to a wide range of apps in order to demonstrate their use. However, the tools that were applied to a sample of apps did not allow for quantifying the app features during app use and were applied to math and literacy apps only. Moreover, some of the tools include subjective criteria, which is difficult to objectively measure by an adult. Therefore, there is a need for a new improved tool that could address those limitations.

The aim of this paper was to create two complementary evaluation tools (adapted to the needs of different audiences) assessing the educational potential of apps for pre-schoolers:

1. A thorough and user-friendly tool accessible by a wide audience: app developers, researchers, caregivers and educators;
2. A tool for researchers that could be used for a more in-depth evaluation by allowing to quantify app features during app use.

Based on the previous literature on app evaluation tools, we propose a set of principles that should guide the development of such tools:

(a) Be informed by the developmental theory and research on children's learning in the context of digital media;

(b) Draw clear links to previously developed tools;

(c) Be brief, have a simple set of clearly described criteria and clear directions on the scoring system;

(d) Focus solely on the objectively measurable factors;

(e) Be applied to a wide variety of apps to demonstrate their use;

(f) Be validated by conducting content validity and inter-rater reliability.

In building the content of our tools, we relied in particular on the British (Department for Education, 2017) and American (Early Childhood Learning and Knowledge Centre, 2015), early years frameworks, which state that preschool children's development should be supported in the areas of cognitive, academic, social-emotional and physical skills.

In the following section, we identify key areas that an evaluation tool ought to include based on previous literature on app evaluation tools, developmental research and theory, and evidence of children's learning from digital media. We also outline a further set of quantity of app features indicators.

Key areas contributing to the educational value of apps

Learning

Learning within an app should be guided by a specific learning goal targeting early skills development relevant to each age and stage (Callaghan & Reich, 2018; Hirsh-Pasek et al., 2015). Educational apps should promote meaningful and authentic learning rather than rote learning, and teach skills transferrable to real life (e.g., Hirsh-Pasek et al., 2015; Papadakis et

al., 2017). Learning should also be cognitively active and involve problem solving, i.e., reasoning, thinking and using creative skills (Aladé et al., 2016; Hirsh-Pasek et al., 2015).

Not all of the previous app evaluation tools included the criteria related to meaningful learning and solving problems. We believe that these features are critical to the learning being deeper, authentic and transferable to real life.

Feedback

Feedback plays a critical role in supporting educational performance (e.g., Mulliner & Tucker, 2017; Schwartz et al., 2016). Specific, meaningful, timely and structured feedback drives child's engagement in the activity (e.g., Hirsh-Pasek et al., 2015; Walker, 2011). Moreover, feedback should reinforce the learning goal and scaffold users' understanding of how to improve (see, e.g., Callaghan & Reich, 2018). All the previous app evaluation tools pointed to the significance of feedback. However, not all of them described explicitly how feedback should be presented by providing relevant examples from the apps.

Social interactions

Social interactions support learning from the very early stages of development (see Hirsh-Pasek et al., 2015, for a summary). Social demonstrations enhanced learning in a touchscreen puzzle task in a group of 2.5- and 3-year-olds (Zimmermann et al., 2017). Apps can involve "parasocial" interactions with animated characters present onscreen, which offer symbolic experiences that can be beneficial for children's social and cognitive development (e.g., Calvert, 2015).

Only some of the previous app evaluation tools recommended the presence of high quality parasocial interactions in the apps. In our tool we specify how the parasocial character should be interacting with the child in order to support learning.

Activity structure

Apps which give the opportunity for exploratory use alongside structured activities, might increase children's intrinsic motivation and engagement. Child autonomy and the sense of agency when using interactive media is crucial for the learning process (e.g., Kirkorian, 2018; Papadakis & Kalogiannakis, 2017). Pre-schoolers who could select their learning experience in a tablet game outperformed those who had no control over the order of presentation of the material (Partridge et al., 2015).

Importantly, almost none of the previous evaluation tools allowed assessing whether apps promote exploratory use

Narrative

Media content that is embedded in an entertaining narrative integrated at the heart of the story can benefit children's learning (e.g., Dingwall & Aldridge, 2006). Content directly linked to a narrative of a television program is recalled better than content which is irrelevant to the storyline (Fisch, 2004).

Although the role of narrative for children's learning has been established by previous research, almost none of the evaluation tools included the presence of narrative in assessment criteria.

Language

Appropriately designed digital media can be a valuable source of language input for young children. The presence of good quality language is crucial for educational potential (Rowe, 2012). Studies using lab-designed apps have shown that children aged 2-4 are able to

learn labels for novel objects (Kirkorian, 2018; Russo-Johnson et al., 2017) or for real-world objects (Dore et al., 2019)

While two of the previous evaluation tools mentioned language as part of some other criteria, none of them focussed on assessing the quality of language directly. We fill in this gap in our tool.

Adjustable content

To ensure effective learning, the difficulty level of an app should be automatically adjusted to users' performance (e.g., Callaghan & Reich, 2018). Specifically, each level of an activity should build on the knowledge gained in earlier levels, and increase hints and feedback if a user makes repeated errors (e.g., Revelle, 2013).

The majority of the previous tools included adjustable content in their evaluation criteria, and following the theoretical motivation outlined above, we also include it in our tool.

App design

As highlighted in the previous evaluation tools, (e.g., Lee & Kim, 2015), app's design should be simple and consistent, style of letters and pictures should be clear, and the arrangement of operating buttons should be appropriate. Unnecessary advertisement, additional in-app purchases and slowly loading content may impede learning. App should also be easy to use and always responsive to touch interactions.

All the previous app evaluation tools included app design in their criteria. We also acknowledged its importance for enhancing children's learning experience.

Quantity of app features indicators

The following section presents the indicators for the quantity of app features. For certain features, it is crucial to estimate how often a given feature occurs during app use, in order to determine whether children's learning environment is age appropriate and not overly complex. None of the previous evaluation tools enabled measuring the proportion or frequency of different app features during app use. Thus, the way we measure app features in our quantitative tool is novel.

Touch gestures

The direct manipulation interaction facilitates pre-schoolers' learning from touchscreen media, yet most educational apps only support tap (99% of apps) and drag (56% of apps; Nacher et al., 2015). Nacher et al., (2015) found that infants aged 2-3 perform one-finger rotation and two-finger scale up and down successfully, but find double tap, long press and two-finger rotation challenging. Russo-Johnson et al. (2017) reported that 2-4-year-old children from low SES families learned more novel object labels when dragging objects versus tapping them, perhaps because tapping is a response that does not require active attention.

Active learning

High quality apps should provide opportunities for active cognition, e.g. making cognitively challenging decisions, and solving problems (e.g., Hirsh-Pasek et al., 2015). Cognitive activities in contrast to stimulus-reaction activities during app use encourage active cognition, while variability across learning encounters has a potential to facilitate learning (e.g., Thiessen, 2011). Thus, a variety of activity goals might contribute to the app being more cognitively active.

Complexity of the learning environment

Background visual, background sound and other app interactions available on the screen contribute to the complexity of learning environment. Cognitive Theory of Multimedia Learning (Mayer, 2005, 2014) envisions that the child's learning might be unsuccessful if the software includes too much extraneous material. Sound effects and animation interfered with story comprehension and event sequencing in children aged 3-6, when compared with paper books (see Reich et al., 2016, for a review). Additional interactions present on the screen alongside the main task can decrease child's engagement in the app (Hirsh-Pasek et al., 2015).

Feedback

In addition to looking at feedback qualitatively and evaluating its meaningfulness, we can also look at it quantitatively and assess its occurrence in the app, its delivery method (audio, onscreen) and its content (ostensive feedback vs other feedback). Interactive media may enhance learning if they promote contingent responses or guide visual attention to relevant information on the screen (Kirkorian, 2018).

App design sophistication

Elements on the screen during app use can either be static, move in a static way, be fully animated or be partly static and partly animated. When learning challenging or novel information, pre-schoolers might benefit more from observing noninteractive video demonstrations than from using interactive media (e.g., Aladé et al., 2016). Furthermore, sound effects and animation in ebooks can interfere with story comprehension in children aged 3-6 years, when compared with paper books (see Reich et al., 2016, for a review).

The present studies

The present paper presents two studies. Study 1 focuses on designing and validating evaluation tools for apps aimed at pre-schoolers (children aged 2-5 years). In order to illustrate the use of our tools, in Study 2 we apply them to apps distinguished in terms of their cost.

Study 1: Designing and validating the evaluation tools

Developing the questionnaire for evaluating the educational potential of apps

First stage: creating a list of items and developing a rating scale

Following the literature reviewed in the introduction we defined 12 concepts (items) to be measured in the questionnaire. We included three indicator descriptors to each item (together with a few examples from the apps to each indicator), such that the app could score between 0 and 2 points for each item. The 12 initially constructed items were: Learning goal, Going beyond rote learning, Solving problems, Feedback, Social interactions, Open-ended, Plotline/narration, Appropriateness of language, Customising, Adjustable content, Suitability of design, Usability.

Second stage: Conducting a content validity study with experts

Once the first version of our questionnaire was designed, we conducted a content validity study. The study was approved by ethical review board at the University of Salford. We followed the procedure outlined by McGartland Rubio, Berg-Weger, Tebb, Lee and Rauch (2003). We recruited three professional design experts (app developers) and three user experts (early years professionals) who shared their feedback on the items' representativeness, clarity and importance in an online survey. The raters were given the following instruction:

"You will be presented with each of the 12 items included in our coding scheme. Please rate each item as follows:

- Please rate the representativeness on a scale of 0 – 4, with 4 being the most representative. Representativeness is the extent to which each item measures the educational potential of children's apps. Space is provided for you to comment on the item or to suggest revisions.
- Please indicate the level of clarity for each item (how clearly the item is worded), also on a four-point scale. Again, please make comments in the space provided.
- On a scale of 1 – 10 please rate the importance of each item for measuring educational potential, with 10 being the most important.

Finally, please evaluate the comprehensiveness of the entire coding scheme by indicating items that should be deleted or added."

We calculated the Content Validity Index (CVI) for each item and for the whole scale (based on its representativeness), following the guidelines described in McGartland Rubio et al. (2003). The CVI for each item was computed by counting the number of experts who rated the item as 3 or 4 and dividing it by the total number of experts. The CVI for the whole questionnaire was obtained by calculating the average CVI across the items. A CVI of at least 0.8 is recommended for new measures. All items in our questionnaire scored either 0.8 or 1, and the CVI for the whole questionnaire was 0.88 (see Table 2).

The raters did not suggest removing any items. They also rated all items high with regards to the items' importance. Consequently, based on the experts' suggestions, we made modifications to the questionnaire. We merged two pairs of items, i.e. Customising and Adjustable content became Adjustable content; Suitability of design and Usability became App design (according to the raters, the descriptions of these two pairs of items overlapped in terms of content). We also added additional examples from the apps to improve the clarity of

the grade descriptors and we reduced the use of technical language in the questionnaire (including rewording some of the items' names, see Table 2).

Third stage: Content validity study with caregivers

After introducing the changes to the questionnaire, we determined whether the tool was comprehensible to caregivers. We recruited six caregivers of children aged 2-5 years to rate the representativeness and clarity of each item and provide further comments. The caregivers were given the same instruction as the experts in the first content validity study. The CVI for the whole tool based on caregivers' ratings was high, 0.75 (see Table 2).

Table 2. Content Validity Index (CVI) for each item and for the whole questionnaire based on the ratings of representativeness by (a) app developers and early years professionals' and (b) caregivers.

App developers and early years professionals		**Caregivers**	
Item	**CVI**	**Item**	**CVI**
Learning goal	1	Learning goal	0.66
Beyond rote learning	0.8	Meaningful learning	1
Solving problems	0.8	Solving problems	0.83
Feedback	1	Feedback	1
Social interactions	0.8	Social interactions	0.33
Open-ended	1	Opportunities for exploration	0.5
Plotline/narration	0.8	Storyline	0.5
Appropriateness of language	1	Quality of language	1
Customizing	0.8	Adjustable content	0.83
Adjustable content	1		
Suitability of design	0.8	App design	0.83
Usability	0.8		
Overall CVI:	**0.88**	**Overall CVI**	**0.75**

Based on the caregivers' comments, we made further modifications to the questionnaire. Most importantly, the participants from both content validity studies pointed out that while social interactions are important for learning, the development of skills for independent learning is also important and social interactions are not congruent with the reasons caregivers might choose apps (see Broekman et al., 2016 for a similar argument). To accommodate this, in our tool we focused on the high quality parasocial interactions in the apps rather than interactions with adults during app use. Our evaluation questionnaire is presented in Table A1 in Supplemental materials.

Developing the coding criteria for quantifying the app features

In addition to the questionnaire that can be easily used by caregivers and educators, we also aimed to develop a tool allowing researchers a more in-depth, quantitative assessment of apps' features.

For the coding criteria, following the literature review outlined in the introduction, we grouped the app features into five broader areas. Each of these areas contains between 1 and 3 coding criteria:

1. Touch gestures
2. Active learning
    a. Activity goal
    b. Activity type
3. Complexity of the learning environment
    a. Screen elements
    b. Background visual
    c. Background sound
    d. Other app interactions
4. Feedback

a. Proportion of feedback
b. Feedback delivery method
c. Feedback content

5. App design sophistication.

For a detailed information on coding instructions and scoring, see Appendix B in Supplemental materials.

Study 2: **Applying the evaluation tools to illustrate their use and to measure the app gap**

In Study 2, we applied the evaluation tools to a sample of paid and free apps in order to illustrate their use and to assess the role of cost on app quality. Digital media is now embedded in family life (Livingstone et al., 2018) and as a result there are concerns that social disadvantage could extend to a digital disadvantage (Vaala et al., 2015; Zhang & Livingstone, 2019), the so called "app gap" (Common Sense Media, 2013). The app gap can be observed, for example, in the availability of devices to go online in the household, caregivers' digital skills and cost of devices (Zhang & Livingstone, 2019). Furthermore, lower socio-economic status parents might not be able to spend substantial quality time with their children (Department for Education, 2020).

It is important to understand whether there are differences between apps that might justify differences in cost. In the present study we focus on a broad distinction between apps that are free at the point of initial access versus apps for which payment at initial access is required. Parents might not be aware of the variety of factors contributing to the app cost (e.g., business decisions that influence app developers' app pricing strategies, including the size of the market, funding opportunities, app's unique selling point) and they might link the higher cost to higher quality of app.

According to the Department for Education research report (2020), children aged 0-5 years living in lower-income households in the UK use educational apps more often than their affluent peers. However, parents in higher income households are more likely to pay for an educational app. It is therefore crucial to establish whether children from less affluent families are disadvantaged with respect to the quality of educational apps that they use.

To the best of our knowledge, to date only one study (Callaghan & Reich, 2018) investigated the differences between educational math and literacy free and paid apps. However, Callaghan and Reich (2018) did not investigate the frequency of app features during app use but limited their analysis solely to identifying whether or not a given feature is present in the app.

Data collection

App selection

We coded 44 of the most popular apps in Google, Amazon and Apple app stores. To be included in this study, apps had to target children aged 2-5 years and feature in the top 10 lists for free and paid apps in each app store. Apps were identified on 7th June 2018. Of these 60 apps, 10 were removed as duplicates and 6 were excluded (5 video-based, which only allowed passive use, 1 unresponsive after installation). The remaining 44 apps were included in the study.

App use

Each app was downloaded and a screen recording was taken while the first author used the app for 5 minutes with a systematic approach to exploring all the features. The 5-minute sample was motivated by practical constraints in terms of the intensity of encoding of the detailed app features in ELAN (described in the coding section), as well as being more practical

for caregivers and educators in appraising an app in an efficient amount of time, based on our evaluation questionnaire.

To maintain parity in approach to data capture across apps, the systematic approach by the first author was to follow all the activities in an order suggested by the app design and to use all the available features on each screen only once.

Coding

Questionnaire for evaluating the educational potential

Each app could score between 0 – 20 points on the educational potential index (between 0-2 points for each of the 10 items, see Table A1 in Supplemental material). 5-minute app screen recordings were assessed individually by the first and last author using the scheme. The discrepancies were discussed and resolved between the coders. Inter-rater reliability was high ($\kappa = .889$, $p < .001$). Internal consistency of the tool was Cronbach's alpha = 0.81, which indicates good internal consistency, further validating the tool.

Coding criteria for quantifying app features

To enable coding for quantifying app features, screen recordings of the app use were coded in ELAN 5.2, software that enables adding annotations to audio and/or video streams. The coder (first author) coded each screen during the app use for the 11 coding categories (see Appendix B in Supplemental material for the details on the coding and scoring). Inter-rater reliability was determined by comparing the coding of the primary coder with the coding of a trained double coder who coded data from 5 apps independently. Inter-rater reliability was $\kappa = 0.917$, $p < .0001$. The small number of discrepancies were resolved by the first coder.

Additionally, in order to determine whether the majority of app features could be captured in 5 minutes of app use (regardless of the person using the app and their style of app

use), we calculated inter-user reliability. This was determined by comparing coded app use data for 5 apps that were also used by a second independent user. Crucially, the second user did not receive any instruction on using the apps. Overall, inter-user reliability was $\kappa = 0.872$, $p < .001$, which shows that the same app features can be captured during 5 minutes of app use, regardless of the user.

Results

To illustrate the use of the tools in practice, we report differences between free and paid apps. This also enables us to determine whether there is an app gap in quality that is reflected in cost, which could contribute to a digital disadvantage. The final sample included 19 free and 24 paid apps (one app was excluded because it was duplicated between two app stores and was listed as free in one store but required payment in the other).

We first report the results from the analysis of the questionnaire for evaluating the educational potential, and then the analyses of coding criteria for quantifying the app features.

Evaluating the educational potential

To test whether there is a difference in educational potential between free and paid apps, a Mann Whitney U-test was performed. The results show that free apps (M = 7.16, SD = 3.70) did not differ from paid apps (M = 6.75, SD = 4.60) on the educational potential index (U=211, Z= -0.405, p=0.685, r = -0.06).

Figure 1 presents cumulative scores for each of the items in the evaluation questionnaire for the whole app sample (0-2 points for each item, 43 apps in the sample; maximum score was 86). Suitability of design and quality of language received the highest scores (58 and 54, respectively), while adjustable content and social interactions appear among those with the lowest scores (8 and 13, respectively).

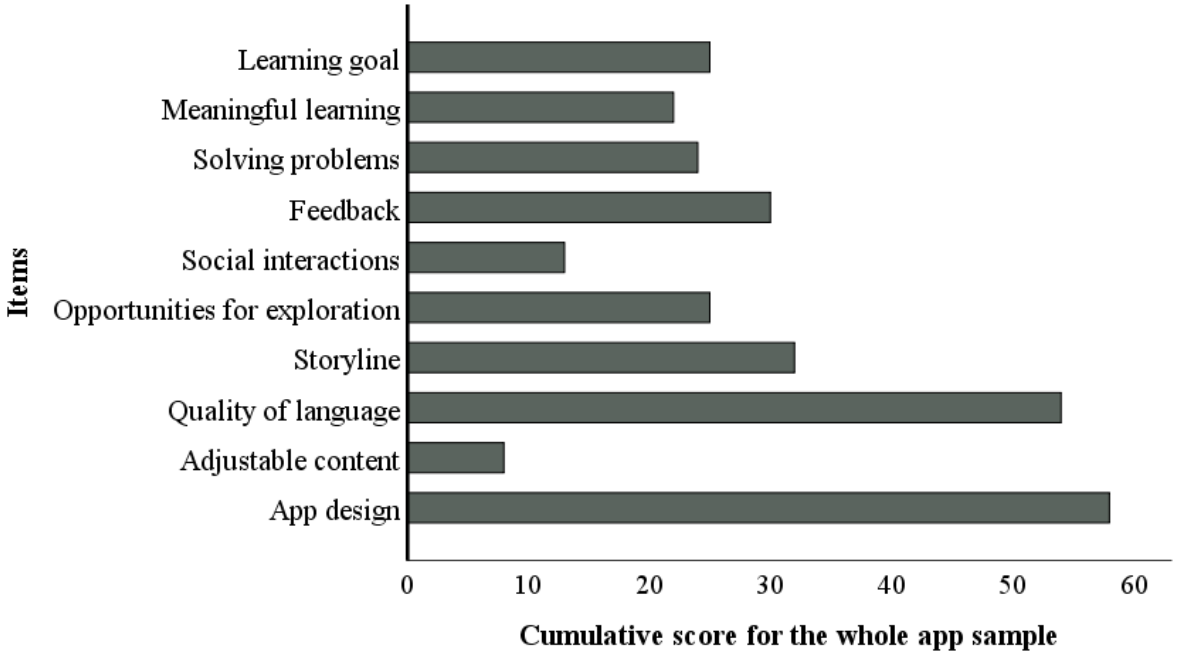

Figure 1. Cumulative scores[2] for all items in the evaluation questionnaire for the whole sample (N = 43).

Quantifying app features: analyses comparing free and paid apps

First, we present the descriptive statistics for app features coded in the study (see Table 3).

Table 3. Descriptive statistics for each app feature for free apps, paid apps and for the whole sample. Freq = frequency.

App feature	**Measure**	**Free apps (N = 19)**	**Paid apps (N= 24) Mean (SD)**	**All apps (N=43)**
Touch gestures				
Touch gestures	Freq of tapping	58.5 (128.8)	34.2 (20.4)	44.95 (86.5)
	Freq of swiping	7.05 (14.1)	3.08 (4.4)	4.84 (10)
	Freq of dragging	7.79 (7.9)	19.04 (18.5)	14.07 (15.7)
	Freq of tracing	1.63 (4.5)	1.46 (4)	1.53 (4.1)

[2] The cumulative scores were not presented separately for the two groups due to the differences in sample size between the groups. We also did not present mean scores for each item for the two groups because each item was measured only on a scale 0-2.

Active learning				
Activity type	Freq of cognitive activities	20.42 (12.0)	20.54 (12.6)	20.49 (12.2)
	Freq of stimulus-reaction activities	9.16 (24.2)	5.25 (5.1)	6.98 (16.4)
Activity goal	Number of different goals	15.1 (7.7)	20.7 (14.2)	18.28 (12)
Complexity of the learning environment				
Screen elements	Mean number on the screen	4.4 (2.2)	6.3 (3.5)	5.5 (3.1)
Background complexity	Proportion of complex background to simple background	0.61 (0.39)	0.77 (0.30)	0.70 (0.35)
Background sound	Freq of no sound	5.81 (9.8)	6.19 (8.8)	6.02 (9.1)
	Freq of simple sound	3.91 (6.3)	5.2 (10.2)	4.63 (8.6)
	Freq of music	12.8 (17.7)	11.2 (16.8)	11.9 (17)
	Freq of complex sound	13.7 (16.4)	11.7 (10.9)	12.6 (13.4
Other app interactions	Mean number on the screen	1.7 (1.2)	2.0 (2.0)	1.9 (1.7)
Feedback				
Presence of feedback	Proportion of feedback to no feedback	0.69 (0.41)	0.84 (0.29)	0.78 (0.35)
Feedback delivery method	Freq of audio	1.79 (2.8)	4.39 (6.0)	3.25 (4.9)
	Freq of onscreen	1.38 (2.1)	1.93 (3.2)	1.68 (2.7)
	Freq of audio & onscreen	4.36 (4.1)	6.67 (12)	5.66 (9.3)
Content of the feedback	Proportion of ostensive feedback compared to other feedback	0.78 (0.33)	0.71 (0.37)	0.74 (0.35)
App design sophistication				
Object property	Freq of static	14.47 (9.6)	16.04 (12)	15.35 (10.9)
	Freq of static movement	2.79 (7.4)	4.83 (9.5)	3.93 (8.6)
	Freq of mixed	3.05 (4.8)	5.58 (10.5)	4.47 (8.5)
	Freq of animation	14.79 (13.2)	6.50 (6.9)	10.16 (10.8)

The analyses comparing free and paid apps are presented in Table 4. Overall, the free and paid apps differed significantly only on two features: (1) the mean number of screen elements, with paid apps having on average more screen elements than free apps; and (2) on object property, with free apps having higher frequency of animation than paid apps, but no differences in other object properties between free and paid apps.

Table 4. Summary of the main analyses.

Area of features	**Compared app feature**	**Aim of analysis**	**Type of analysis**	**Main results/effects**	**Interactions**
Touch gestures	Touch gestures	Comparing the frequency of the four touch gestures:	2 (Cost) x 4 (Touch	Main effect of Cost n.s. ($F < 1$)	Cost x Touch gesture n.s. ($F(3,123)=1.229$,

		• tapping • swiping • dragging • tracing	gesture) ANOVA	Main effect of Touch gesture ($F(3,123)=9.361$, $p<0.0001$, $\eta_p^2=0.186$) tapping > swiping ($p=0.009$) tapping > tracing ($p=0.010$) dragging > swiping ($p=0.043$) dragging > tracing ($p<0.0001$)	$p=0.302$, $\eta_p^2=0.029$)
Active learning	Activity type	Comparing the frequency of cognitive activities and stimulus-reaction activities	2 (Cost) x 2 (Activity type) ANOVA	Main effect of Cost n.s. ($F < 1$)	Cost x Activity n.s. ($F < 1$)
	Activity goal	Comparing the number of unique activity goals	U-Mann Whitney test	No difference ($U=181$, $Z= -1.151$, $p=0.250$, $r= -0.17$)	
Complexity of the learning environment	Screen elements	Comparing the mean number of screen elements	U-Mann Whitney test	Paid apps > free apps ($U=149$, $Z= -1.921$, $p=0.055$, $r = -0.29$)	
	Background complexity	Comparing the proportion of complex background	U-Mann Whitney test	No difference ($U=156$, $Z= -1.793$, $p=0.073$, $r= -0.273$)	
	Background sound	Comparing the frequency of different sound types: • no background sound • simple background sound • music • complex background sound	2 (Cost) x 4 (Background sound) ANOVA	Main effect of Cost n.s. ($F < 1$). Main effect of Background sound ($F(3,123)=3.743$, $p=0.013$, $\eta_p^2=0.084$) complex sound > simple sound ($p=0.024$)	Cost x Background sound n.s. ($F < 1$)
	Other app interactions	Comparing the mean number of other interactions available on a screen	U-Mann Whitney test	No difference ($U=214$, $Z= -0.343$ $p=0.732$, $r= -0.05$)	
Feedback	Presence of feedback	Comparing the proportion of feedback	U-Mann Whitney test	No difference ($U=104$, $Z= -0.922$, $p=0.419$, $r= -0.14$)	

	Feedback delivery method	Comparing the frequency of feedback delivered: • via audio, • onscreen • simultaneously via audio & onscreen	2 (Cost) x 3 (Feedback delivery method) ANOVA	Main effect of Cost n.s. ($F < 1$) Main effect of Feedback delivery n.s. ($F(2,52)=2.373$, $p=0.103$, $\eta_p^2=0.840$)	Feedback x Cost n.s. ($F < 1$)
	Content of the feedback	Comparing the proportion of ostensive feedback	U-Mann Whitney	No difference ($U=94.0$, $Z= -0.634$, $p=0.526$, $r= -0.09$)	
App design sophistication	Object property	Comparing the frequency of different object property: • static • static movement • animation • mixed	2 (Cost) x 4 (Object property) ANOVA	Main effect of Cost n.s. ($F < 1$). Main effect of Object property ($F(3,123)=11.969$, $p<0.0001$, $\eta_p^2=0.226$) static > static movement ($p<0.0001$) static > mixed ($p<0.0001$) animation > static movement ($p<0.021$) animation > mixed ($p=0.027$)	Cost x Object property ($F(3,123)=2.688$, $p=0.049$, $\eta_p^2=0.062$ Free apps higher frequency of animation ($p<0.0001$) than paid apps

Discussion

The primary aim of this paper was to report the design and development of two novel, transparent and comprehensive tools for evaluating the educational potential of apps aimed at 2-5-year-old children. Specifically, a questionnaire aimed at a wide audience, and coding criteria for measuring the quantity of app features aimed at researchers.

The tools were developed specifically for evaluating apps targeting pre-schoolers; they were guided by the early years foundation frameworks and informed by the developmental theory and research on children's learning from digital media. The development of the tools was preceded by a careful analysis of the previously designed evaluation tools. We identified several limitations in the previous tools, such as a long list of

criteria which are not specific enough, no direction to quantify app features, and inclusion of technical language. We designed our tools with the aim to address those limitations. We also demonstrated the use of our tools on a wide range of most popular children's apps. We added a novel contribution to the research on children's apps by evaluating both the educational potential of apps and by providing the first description of the quantity of app design features during app use.

Our tool is the first to have had content validity assessed by caregivers as well as experts. We made further amendments following comments from caregivers to ensure that our tool did not include technical language. The use of examples from existing apps in our tools means that users do not require any existing knowledge of early years education frameworks which was a common limitation of previous tools (Hirsh-Pasek et al., 2015; Lee & Cherner, 2015; Department for Education, 2019). The next step in validating the tools will be to determine how preschool children interact with the apps and evaluate, rather than predict, the educational potential of the children's interactions. A further point for future investigation is also how the various features in apps interact with one another. This is ongoing work in our lab.

Our tool development resulted in a measurement of apps in terms of an educational potential index, which was shown to be high in content validity, internal consistency and in inter-rater reliability. The comparison between free and paid apps on this index did not reveal any difference between apps. It is worth noting that the mean scores on the educational potential index for both groups were rather low (on average less than 10 out of 20). This suggests that the free and paid apps appeared to be equally low in terms of their educational potential, which is consistent with other studies underlining the disparity between the number of self-proclaimed educational apps in the markets and their poor educational value (Chau,

2014; Goodwin & Highfield, 2012; Hirsh-Pasek et al., 2015; Papadakis et al., 2018; Vaala et al., 2015; Schuler, 2012).

The whole app sample showed strength as far as suitability of design and language were concerned (see Figure 1). High scores on suitability of design suggest that the apps were well prepared from the technical perspective. However, the apps showed weakness in terms of the more educational evaluation criteria, such as meaningful learning, offering users problems to solve or having a learning goal, which suggests that they do not offer a meaningful and cognitively active learning experience (in line with Papadakis et al., 2018). The apps in our sample also scored low on social interactions; they rarely encouraged high quality interactions with characters onscreen (in line with Vaala et al., 2015; Papadakis et al., 2018).

Additionally, the apps in our sample scored particularly low on adjustable content. This means that they lacked flexibility in changing the settings and did not tailor content to users' performance. Apps should adjust the content to the user's needs if they intend to increase user's motivation and allow for gradual progress in learning (e.g., Callaghan & Reich, 2018; Papadakis et al., 2017). This finding is again in line with the previous studies, which found that less than 20% (Callaghan & Reich, 2018; Vaala et al.,2015) or none of the reviewed apps (Papadakis et al., 2018) included adjustable content. Overall, our findings highlight the need for developmental psychologists to work with app developers to advance the educational potential of touchscreen apps.

As a secondary aim, we compared the free and paid apps on the coding criteria for quantifying app features in order to assess the "app gap" associated with app cost. The free and paid apps differed only on two features: (1) the number of screen elements, with paid apps having on average more elements on the screen than free apps; and (2) the frequency of animation, with free apps having more animations than paid apps. Considering that only two

differences were observed, it can be concluded that free and paid apps did not differ substantially either in their educational potential or in their features and design. This is partially in line with the content analysis of Callaghan and Reich (2018) who also did not find many differences between free and paid apps with respect to their educational features. Our results suggest that paid apps might not necessarily guarantee a better app quality than free apps, at least based on our app sample.

This study also gives an insight into the educational quality and design features of apps targeting pre-schoolers. Crucially, none of the previous app evaluation reviews quantified the apps features during app use within the evaluated app sample. Thus, our descriptive statistics (see Table 3) are the first ones to present the frequency of various app features during app use, based on a wide sample of apps. In our sample, all apps had higher frequency of cognitive activities than stimulus-reaction activities. Complex sound (two or more sounds playing simultaneously) was more frequent across all the apps than simple sound, which might add to the young children's cognitive processing load while using apps (e.g., Mayer, 2014). The apps had on average 5 screen elements on each screen, and 18 different activity goals during the 5-minute use. On each screen, apart from the target interaction, there were on average 2 additional interactions available. Apps in our sample offered a high proportion of feedback to users' responses (78%), as compared to no feedback during app use (see Callaghan and Reich, 2018, for similar results), and a high proportion of that feedback was ostensive (74%), i.e. referential cues to indicate what is to be learnt. Those characteristics can serve as a reference point for other studies on app features.

Conclusion

In conclusion, we have presented comprehensive evaluation tools based on theories of learning and cognitive development and have shown how they can be implemented in the

analyses of apps available to children. We found that the app gap associated with cost was not an issue in terms of the educational potential for most popular apps currently available. The app gap is instead related to aesthetic features of apps rather than any observable cognitive advantage proffered by paid apps.

Word count: 8,328

Acknowledgements

This work was supported by the Economic and Social Research Council [grant number ES/R004129/1]. The authors would like to thank Eve Bent for help with coding.

References

Aladé, F., Lauricella, A. R., Beaudoin-Ryan, L., & Wartella, E. (2016). Measuring with Murray: Touchscreen technology and preschoolers' STEM learning. *Computers in Human Behavior*, *62*, 433–441. https://doi.org/10.1016/j.chb.2016.03.080

Broekman, F. L., Piotrowski, J. T., Beentjes, H. W. J., & Valkenburg, P. M. (2016). A parental perspective on apps for young children. *Computers in Human Behavior*, *63*, 142–151. https://doi.org/10.1016/j.chb.2016.05.017

Callaghan, M. N., & Reich, S. M. (2018). Are educational preschool apps designed to teach? An analysis of the app market. *Learning, Media and Technology*, *43*(3), 280–293. https://doi.org/10.1080/17439884.2018.1498355

Calvert, S. L. (2015). Children and digital media. M. Bornstein, T. Leventhal, & R. Lerner (Eds.), *Handbook of child psychology and developmental science. (7th, pp. 375-415).*

Chau, C. (2014). Positive technological development for young children in the context of children's mobile apps. PhD Dissertation. USA: Tufts University. Available from: http://ase.tufts.edu/DevTech/resources/Theses/CChau_2014.pdf (Accessed: 5 April 2020)

Chen, W., & Adler, J. L. (2019). Assessment of Screen Exposure in Young Children, 1997 to 2014. *JAMA Pediatrics*, *173*(4), 391–393. https://doi.org/10.1001/jamapediatrics.2018.5546

Common Sense Media (2013). Zero to Eight: Children's Media Use in America. Available from: https://www.commonsensemedia.org/research/zero-to-eight-childrens-media-use-in-america-2013/key-finding-5%3A-reduced-but-persistent-mobile-digital-divide. (Accessed: 20 March 2020).

Department for Education (2017) Statutory Framework for the Early Years Foundation Stage. Available from: https://www.gov.uk/government/publications/early-years-foundation-stage-framework--2 (Accessed: 4 April 2020)

Department for Education (2019). Educational Criteria for early years app. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/822086/Educational_criteria.pdf. (Accessed: 30 March 2020).

Department for Education (2020). Childcare and early years survey of parents 2018, follow-up survey. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/866868/Childcare_and_early_years_survey_of_parents_2018_follow-up_survey.pdf (Accessed: 30 March 2020)

Dingwall, R., & Aldridge, M. (2006). Television wildlife programming as a source of popular scientific information: A case study of evolution. *Public Understanding of Science*, *15*(2), 131–152. https://doi.org/10.1177/0963662506060588

Dore, R. A., Shirilla, M., Hopkins, E., Collins, M., Scott, M., Schatz, J., … Hirsh-Pasek, K. (2019). Education in the app store: Using a mobile game to support U.S. preschoolers' vocabulary learning. *Journal of Children and Media*, 1–20. https://doi.org/10.1080/17482798.2019.1650788

Early Childhood Learning and Knowledge Centre (2015). Early Learning Outcomes Framework. Available from: https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/elof-ohs-framework.pdf (Accessed: 4 April 2020)

Fisch, S. M. (2004). *Children's learning from educational television: Sesame Street and beyond*. Mahwah, NJ: Erlbaum

Goodwin, K., & Highfield, K. (2012). iTouch and iLearn: An examination of "educational" apps. In Early education and technology for children conference (pp. 14-16) (Salt Lake City, Utah, USA).

Healthy Children (2018). Kids & Tech: Tips for Parents in the Digital Age. Available from: https://www.healthychildren.org/English/family-life/Media/Pages/Tips-for-Parents-Digital-Age.aspx (Accessed: 20 July 2020)

Highfield, K., & Goodwin, K. (2013). *Apps for mathematics learning: A review of "educational" apps from the iTunes App Store*. 26.

Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting education in "educational" apps: Lessons from the science of learning. *Psychological Science in the Public Interest*, *16*(1), 3–34. https://doi.org/10.1177/1529100615569721

Kirkorian, H. L. (2018). When and how do interactive digital media help children connect what they see on and off the screen? *Child Development Perspectives*, *12*(3), 210–214. https://doi.org/10.1111/cdep.12290

Lee, C.-Y., & Sloan Cherner, T. (2015). A comprehensive evaluation rubric for assessing instructional apps. *Journal of Information Technology Education: Research*, *14*, 021–053. https://doi.org/10.28945/2097

Lee, J.-S., & Kim, S.-W. (2015). Validation of a tool vvaluating educational apps for smart education. *Journal of Educational Computing Research*, *52*(3), 435–450. https://doi.org/10.1177/0735633115571923

Livingstone, S., Blum-Ross, A., Pavlick, J., & Ólafsson, K. (2018). In the digital home, how do parents support their children and who supports them? Available from: http://www.lse.ac.uk/media-and-communications/assets/documents/research/preparing-for-a-digital-future/P4DF-

Survey-Report-1-In-the-digital-home.pdf. (Accessed: 20 March 2020)

Mayer, R. E. (2005). Cognitive Theory of Multimedia Learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 31–48). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.004

Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learning and Instruction*, *29*, 171–173. https://doi.org/10.1016/j.learninstruc.2013.04.003

McGartland Rubio, D., Berg-Weger, M., Tebb, S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in a social wok research. *Social Work Research, 27*(2), 94-104.

McManis, L.D., & J. Parks. 2011. Evaluating technology for early learners. Ebook and Toolkit. Winston-Salem, NC: Hatch Early Learning.

Mulliner, E., & Tucker, M. (2017). Feedback on feedback practice: Perceptions of students and academics. *Assessment & Evaluation in Higher Education*, *42*(2), 266–288. https://doi.org/10.1080/02602938.2015.1103365

Nacher, V., Jaen, J., Navarro, E., Catala, A., & González, P. (2015). Multi-touch gestures for pre-kindergarten children. *International Journal of Human-Computer Studies*, *73*, 37–51. https://doi.org/10.1016/j.ijhcs.2014.08.004

Ofcom (2019). Children and parents media use and attitudes: annex 1. Children's research annex. Available from https://www.ofcom.org.uk/__data/assets/pdf_file/0027/134892/Children-and-Parents-Media-Use-and-Attitudes-Annex-1.pdf (Accessed: 20 March 2020)

Ólafsson K., Livingstone S., Haddon L. (2013). Children's use of online technologies in Europe: A review of the European evidence base. London, England: EU Kids Online.

Papadakis, S., & Kalogiannakis, M. (2017). Mobile educational applications for children: What educators and parents need to know. *International Journal of Mobile Learning and Organisation*, *11*(3), 256. https://doi.org/10.1504/IJMLO.2017.085338

Papadakis, S., Kalogiannakis, M., & Zaranis, N. (2018). Educational apps from the Android Google Play for Greek preschoolers: A systematic review. *Computers & Education*, *116*, 139–160. https://doi.org/10.1016/j.compedu.2017.09.007

Partridge, E., McGovern, M. G., Yung, A., & Kidd, C. (2015). Young children's self-directed information gathering on touchscreens. In R. Dale et al. (Eds.), *Proc. of CogSci 37*. Austin, TX: Cog. Sci. Society.

Reich, S. M., Yau, J. C., & Warschauer, M. (2016). Tablet-based eBooks for young children: What does the research say? *Journal of Developmental & Behavioral Pediatrics*, *37*(7), 585–591. https://doi.org/10.1097/DBP.0000000000000335

Revelle, G. (2013). Applying developmental theory and research to the creation of educational games. *New Directions for Child and Adolescent Development*, *2013*(139), 31–40. https://doi.org/10.1002/cad.20029

Rosell-Aguilar, F. (2017). State of the app: A taxonomy and framework for evaluating language learning mobile applications. *CALICO journal*, 34(2). https://doi.org/10.1558/cj.27623

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development, 83*, 1762–1774. doi:10.1111/j.1467-8624.2012. 01805.x

Russo-Johnson, C., Troseth, G., Duncan, C., & Mesghina, A. (2017). All tapped out: Touchscreen interactivity and young children's word learning. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00578

Shoukry, L., Sturm, C., & Galal-Edeen, G. H. (2012). Pre-MEGa: a proposed framework for the design and evaluation of preschoolers' mobile educational games. In In the proceedings of The International Conference on Engineering Education, Instructional Technology, Assessment, and E-learning (EIAE 12). Bridgeport: USA

Shuler, C. (2012). iLearn II; an analysis of the education category of the iTunes App Store. New York: The Joan Ganz Cooney Center at Sesame Workshop.

Schwartz, D. L., Tsang, J. M., & Blair, K. P. (2016). *The ABCs of How We Learn: 26 Scientifically Proven Approaches, How They Work, and When to Use Them*. W. W. Norton & Company.

Thiessen, E. D. (2011). When variability matters more than meaning: The effect of lexical forms on use of phonemic contrasts. *Developmental Psychology*, *47*(5), 1448–1458. https://doi.org/10.1037/a0024439

Vaala, S., Ly, A., & Levine, M. H. (2015). Getting a read on the app stores: A market scan and analysis of children's literacy apps. Full report. In Joan Ganz Cooney Center at Sesame Workshop.

Walker, H. (2010). Evaluating the effectiveness of apps for mobile devices. *Journal of Special Education Technology, 26*(4), 59–63.

Zhang, D. & Livingstone, S. (2019). Inequalities in how parents support their children's development with digital technologies. Parenting for a Digital Future: Survey Report. Available from: http://www.lse.ac.uk/media-and-communications/assets/documents/research/preparing-for-a-digital-future/P4DF-Report-4.pdf. (Accessed: 30 March 2020)

Zimmermann, L., Moser, A., Lee, H., Gerhardstein, P., & Barr, R. (2017). The ghost in the touchscreen: Social scaffolds promote learning by toddlers. *Child Development*, *88*(6), 2013–2025. https://doi.org/10.1111/cdev.12683

Supplemental material

Appendix A. Questionnaire for evaluating the educational potential of children's apps

Table A1. Coding items for the questionnaire for evaluating the educational potential of children's apps.

Item	Points		
	2	1	0
Learning goal	There is a clear overall learning goal(s) targeting early skills development, e.g. linking sounds and letters, counting, learning shapes and colours, teaching about people, places and environment (relevant to each age/stage).	There is no clear overall learning goal but some or all activities within the app teach early skills relevant to each age/stage e.g., selecting objects in a particular colour, matching shapes, selecting ingredients to bake a cake.	There is no clear learning goal, e.g. child is avoiding obstacles in a race.
Meaningful learning **(Do not score this item if an app scored 0 for Learning goal)**	In most cases learning is meaningful and has a purpose (relevant to each age/stage); the content is relevant to real life, e.g. child is learning numbers in real-life context, such as selecting and counting the items to be packed in a suitcase before going on holidays, finding the missing word in a sentence, or learning the bedtime routine (brush the character's teeth, take a shower, dry hair)	In most cases, learning occurs outside of a real-life context, e.g. child has to drag the word to the corresponding picture, or is asked questions about real-life knowledge outside of a life context, such as question "What do you use when it's raining" when child has to select the correct image (wellies) on a blank screen, instead of teaching the skill in real-life situation/environment	The app does not promote meaningful learning, e.g. child has to trace the letter or tap on a given letter when it is presented on the screen (as opposed to selecting a correct letter in an array of different letters)
Solving problems	App encourages child to solve problems relevant to each age/stage, which promote reasoning, thinking and creativity, e.g. finding a missing element	App encourages child to solve problems relevant to each age/stage, but the problems are not mentally challenging, e.g. finding two matching elements in a memo game,	The app does not involve problem solving (e.g., avoiding obstacles during a race or collecting gifts during a train ride).

	in a pattern, finding all the words that begin with a given sound, dragging letters to build a word, selecting only items in particular colour and shape, etc	tapping on blue objects among colourful objects, tapping on a particular letter or number among other letters/numbers	
Feedback	Feedback is specific, meaningful, constructive and age appropriate (i.e. app provides positive feedback when child makes an error and in this way motivates the child to improve, e.g. by repeating the instruction or by demonstrating how to perform an action using visual help, such as arrows showing the direction of the tracing in a tracing shape activity, or index finger pointing to the correct element on the screen). Feedback relates directly to the activity/task and supports the learning goal, e.g. "Good job counting all the ducks!", "This is letter 'a', well spotted!", "Oh dear, this is not a toothbrush – have another go and look for a toothbrush".	Feedback either (a) includes motivational message ("Well done!", "Good job!", etc.) presented via audio or onscreen, (b) comes as points, badges or stars together with an audio message (e.g. "Well done, you've earned a star!"), or (c) comes as visual age appropriate signal of the reason for the reward (e.g. a correctly selected object is highlighted or shaken), but it is not specific, meaningful or constructive, i.e. it does not specifically relate to the reason for feedback (e.g. "That wasn't right, try again"), or app does not demonstrate how to perform an action (e.g. no arrows showing the direction of the tracing in a tracing shape activity)	Feedback is either (a) limited to correctness of child's responses, e.g. "Correct!", "That's right!", (b) non-specific (e.g. cheering, beeping) or (c) comes as points, badges or stars but is not accompanied by an audio message, and is not an age-appropriate signal of the reason for the award (e.g. confetti on the screen).
Social interactions	During use, app involves "social" interactions with characters onscreen (e.g. a character asks to repeat after him/her, asks questions or gives instructions). The character must be present onscreen when it is communicating with the child and it must	App either (a) involves some "~~s~~ocial" interactions with characters onscreen that are not related to the learning material, or (c) involves "social" interactions with characters onscreen that are related to learning but the character is	App does not involve "social" interactions with characters onscreen.

	"look" directly at the child and be animated (i.e. move its mouth or gesture)	rarely present on the screen during instructions, or it is not animated	
Opportunities for exploration	App is semi-structured and gives child the opportunity for exploratory use, e.g. the order of activities/games is fixed, but within the activity child can move freely across the screens and try different interactions in his/her preferred order, or app provides a significant free play space but comes with frequent fixed questions or challenges within the play.	App is either (a) mostly structured and does not give child many opportunities for exploratory use, e.g. child can choose which activity/game to play first but interactions in the activities/games are fixed, without the opportunity for the child to choose what to do and in which order, or (b) app provides mostly free play with only occasional fixed questions or challenges within the play	App is either (a) fully structured and does not give child any opportunity for exploratory use, activities are framed and come in a fixed order, e.g. a set of games being introduced one after another in a fixed order, with fixed interactions in them, or (b) app provides only free play and no fixed questions or challenges within the play
Storyline	The content is created to be on either one overall storyline that connects all activity goals (e.g. character goes on an adventure with dinosaurs) or number of mini storylines and routines (e.g. storylines can connect a set of activities such as character goes on a submarine or treasure hunt)	The content is not created to be on an overall storyline (or there are no multiple storylines connecting sets of activities) but the app may follow a routine, or some individual activities may follow a routine (e.g., character is brushing teeth, taking bath, getting dressed).	Challenges in the app are not combined into an overall storyline or individual storylines (e.g., characters talk about their hobbies), and the app does not encourage the child to engage in routines
Quality of language	App always contains age-appropriate and child-directed language; speech is clear, its pace is slow or moderate and easy to follow. Sentences are not overly complex and not too long. Language is comprehensible	App sometimes doesn't contain age-appropriate language and/or sentences are sometimes overly complex, speech is unclear or its pace is too fast and not easy to follow.	App does not contain language, contains very limited language, or the language is age-inappropriate and not child-directed, speech is unclear, its pace is fast and not easy to follow, sentences are overly complex.
Adjustable content	Content is usually adapted according to child's performance, i.e. (a) if child gives	Content is not automatically adapted to child's performance, but app enables	Content is not adapted to child's performance (i.e. app never simplifies the

	a wrong answer (or several wrong answers), the app might provide item that is similar to the one missed, simplify the skill, and/or (b) if child's performance is very good, the app provides higher level of difficulty	child/caregiver to manually set an age/stage appropriate level of difficulty (e.g. app asks about child's age, child can choose to read the story or being read to, child can choose small vs large letters or tracing vs no tracing)	content if child struggles with a task and never makes the content more challenging if child is doing very well), and the app does not enable child/caregiver to manually set an age/stage appropriate level of difficulty.
Suitability of design	The design is simple and consistent, the pictures and letters are clearly visible, operating buttons are arranged in a clear way, the app does not include unnecessary advertisement, additional in-app purchases and loads quickly. App is also easy to use and is always responsive to touch interactions.	The design is generally quite simple and consistent but minor problems may occur: (a) the pictures and letters are not clearly visible, (b) operating buttons are not arranged in a clear way, (c) the app includes some unnecessary advertisement, (d) takes a while to load activities, (e) has some additional in-app purchases, (e) is not easy to use or (f) not always responsive to touch interactions.	The design is overly complicated and not consistent, the images are not clear, app includes advertisement, content is very restricted without additional in-app purchases, takes a while to load activities, is difficult to use or is often unresponsive to touch interactions.

Appendix B. Coding criteria for quantifying the app features

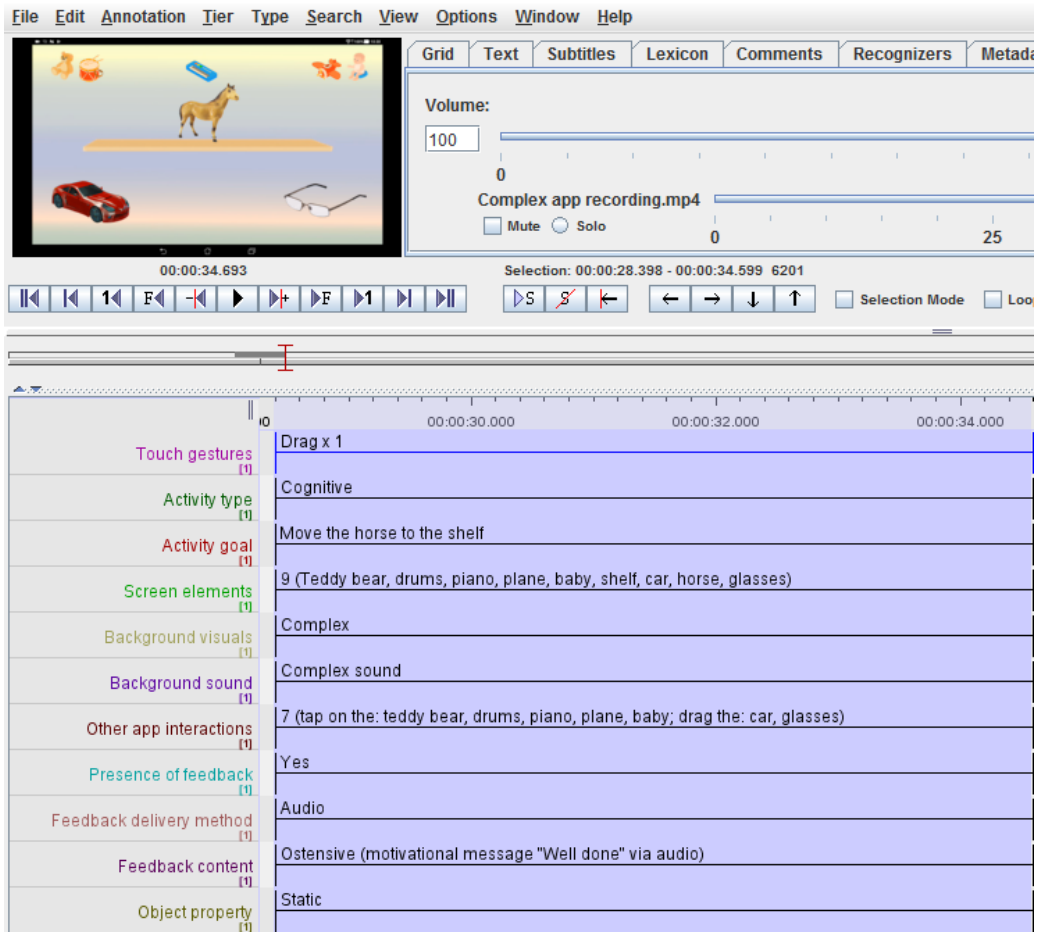

Figure B1. Example of screen coding in ELAN with app features and coding options. An experimental app developed in the lab was used as an example here.

Coding instructions:

For the coding, use ELAN software. Open the ELAN template available on the project OSF site (https://osf.io/atg78/?view_only=6a9a71afe39e453ea20c732b14324524) and the video recording of app use[3] that you want to code. Coding should start when the app finished loading and when the first opportunity for an interaction is presented on the screen or when the first audio or onscreen instruction is given. Using ELAN, select each screen of the app use and prepare it for the annotations. Ideally, each screen should contain a single activity (e.g. tap on a box to open it, avoid obstacles in a race in order to get to the finish line, drag the pieces to make a puzzle). If the background visual, background sound, screen elements or additional interactions available on the screen change during an activity, it is recommended to split the activity into different screen selections, so that each screen selection can capture the different features available on the screen. Code each screen of the app for the 11 coding categories (see Figure B1 for an example of coding). Use coding options and examples in Table B to correctly capture all the features on the screen. Note that depending on the level of detail that you are interested in, you can either code only the frequencies of the features available on each screen, or you can additionally list all the app elements or interactions. For example, for the Screen elements, you can either code only the number of screen elements on the screen, or name all the elements for future reference. Similarly, for the Other app interactions, you can either code only the number of other app interactions on the screen, or you can list all the possible interactions (see Figure B1). Once you have completed the coding, you can export the file to Excel to enable automatic calculation of the numbers and frequencies of the features. Use calculating instructions in Table B1 to calculate the numbers and frequencies required for the analysis. The table includes the calculation method

[3] Record the screen when you use the app on the tablet or on the phone.

that was used in the paper (Calculations after the coding), as well as the alternative ways of calculating the app features (e.g. frequencies vs proportions), depending on the preferences of the researcher and on what is considered more informative for the planned analyses.

Table B1. Coding criteria: app features, coding options on each screen with examples, and guidelines for calculations after the coding

App feature	Coding options on each screen	Examples	Calculations after the coding
Touch gestures			
Touch gestures	Classify the touch gesture(s) on the screen as either of the four listed below. Make sure that you also note the frequency of the gestures if more than one gesture leads to an activity on the screen (e.g. drag the piece *12 (12 drag gestures) to make a puzzle (activity)) Tap Swipe Drag Trace		Calculate the total frequency of tap, swipe, drag, and trace gestures during app use. Alternatively: Calculate the proportion of each type of touch gesture during app use.
Active learning			
Activity type	Classify the activity on the screen as either of the two: Stimulus-reaction activity (action needing basic cognitive involvement)	Avoid moving objects in an arcade-style spaceship game or obstacles in a car race	Calculate the total frequency of stimulus-reaction activities and cognitive activities during app use Alternatively: calculate the proportion of cognitive activities during app use: Proportion of cognitive activities = total frequency of cognitive activities / total frequency of all activities

	Cognitive activity (action involving active cognition and mental attention)	Select correct answer, complete the pattern, write a letter Note: Sometimes a single gesture constitutes an activity (e.g. tap on the box (gesture) to select correct answer (activity)); sometimes a chain of gestures is needed to complete an activity (e.g. drag the piece *12 (12 drag gestures) to make a puzzle (activity)). For coding the Activity type feature, count **only** the activities, not the gestures leading to the activities (see: Touch gesture).	
Activity goal	Name the activity goal on the screen	Select correct answer, complete the pattern, write a letter	Calculate the number of all unique activity goals during app use Note: If a given activity occurs more than once during app use (e.g. the user is asked to write letter 'k' 5 times during app use) and hence leads to the same goal (write letter 'k'), it should be counted as activity goal only once. This allows capturing the variety of different activities, rather than just their frequency
Complexity of the learning environment			
Screen elements	Calculate the number of elements giving an opportunity for an interaction displayed on the screen	Screen elements are any of the following: 1. Objects on the screen that give an opportunity for an interaction (i.e. elements that can be swiped, dragged, tapped or traced).	Calculate the mean number of screen elements across all screens during app use

		2. Objects on the screen that cannot be interacted with but are animated. 3. Game characters (e.g. Peppa Pig, Horrid Henry), regardless of whether they are animated, interactive or not. 4. Essential objects, that are neither animated nor interactive per se, but contain other interactive elements, e.g. a tree that contains interactive apples that user has to pick up 5. Essential objects, that are neither animated nor interactive per se, but enable other objects interact with them, e.g. a chair onto which user can drag to a character to make them sit down Note: All other objects should be treated as part of the background, e.g. non-interactive (and not animated) farm animals and plants that a user passes on his way while driving a train.	
Background visual	Classify the background visual on the screen as either of the two: Simple background visual Complex background visual	 No background or plain colour in the background E.g. on a farm, in the room, in the school, colourful background	Calculate the total frequency of simple and complex background visual during app use Alternatively: Calculate the proportion of simple and complex background during app use

Background sound	Classify the background sound on the screen as either of the four: No sound Simple sound Music Complex sound	 No background sound is played E.g. ping, jingle, whistle, crowd cheering, animal sound, vehicle sound Sound events that last for the whole selected screen duration Two or more sounds played simultaneously, e.g. music & jingle, train sound & animal sound	Calculate the total frequency of no sound, simple sound, music and complex sound during app use. Alternatively: Calculate the proportion of each sound during app use
Other app interactions	Calculate the number of all other app interactions (gestures) available on a screen	Other app interactions are all interactions (gestures) that are available on a screen at the same time as the target app gesture. For example, a target app gesture might be to trace a letter, but at the same time there might be three other buttons available on the screen: a "go back" button, a "move to the next screen" button and a "choose another letter" button. In that case, there are three other app interactions available on a screen.	Calculate the mean number of other app interactions across all screens during app use
Feedback			
Proportion of feedback during app use	If there is an opportunity for feedback on the screen, specify whether the app provided feedback ('Yes) or not ('No)	This app feature is specified based only on the activities that give an opportunity for feedback, e.g. user has to select the correct answer, drag all the elements to create a picture or find all the words that begin with a given letter. If there is no	Calculate the total frequency of feedback and no feedback during app use, and then calculate the proportion of feedback during app use.

		opportunity for feedback during app use, e.g. user can move freely on a screen during app use and explore various screen elements, but app provides no specific tasks or activities, it is coded as NA rather than 'No'.	
Feedback delivery method	Classify the feedback delivery method on the screen as either of the three: Feedback delivered via audio during app use Feedback delivered onscreen Feedback delivered simultaneously via audio and onscreen	 Motivational message such as "Great job!" delivered after an interaction Points or badges appearing on the screen after an interaction Audio message "Well done" & points or visual prizes appearing on the screen simultaneously	Calculate the total frequency of feedback delivered audio, onscreen, and audio & onscreen simultaneously during app use. Alternatively: Calculate the proportion of each type of feedback during app use.

Feedback content	Classify the feedback content on the screen as either of the two: Ostensive/referential feedback Non-specific feedback	 Feedback is categorised as ostensive if it includes either (a) motivational message or an age-appropriate indication of the reason for the award delivered via audio (e.g. “Well done”, “Correct”) and/or (b) visual cues to indicate the correct answer (e.g. the box with the correct answer is shaken or highlighted). Feedback that cannot be categorised as ostensive, e.g. crowd cheering, confetti on the screen Note: If feedback is delivered simultaneously in two different ways, e.g. motivational message via audio (ostensive) & visual points onscreen (non-specific), and thus can be classified as both ostensive and non-specific, code it as ostensive.	Calculate the total frequency of ostensive and non-specific feedback during app use. Alternatively: Calculate the proportion of ostensive feedback during app use.
App design sophistication			

Object property	Classify the object property on the screen as either of the four: Static screen Static movement Animation Mixed	 Static screen with no animation or moving elements Screen elements are in motion on the screen, but they are static per se, e.g. a static image of a character moves across the screen All screen elements, or majority of them, are animated Some screen elements are animated, and some are static, e.g. there are four boxes with different activities to choose from on the screen but only one of them is animated and the remaining ones are static.	Calculate the total frequency of static screen, static movement, animation and mixed during app use. Alternatively: Calculate the proportion of each object property during app use.