Effects of Different Types of Correctness Feedback on Children's Performance with a Mobile Math App

Ahmed Sabbir Arif[1,3], Cristina Sylla[2], Ali Mazalek[3]

[1]University of California, Merced
Merced, California, USA
asarif@ucmerced.edu

[2]Universidade do Minho
Braga, Portugal
sylla@engagelab.org

[3]Ryerson University
Toronto, Ontario, Canada
mazalek@ryerson.ca

***Abstract*—This paper presents results of an exploratory study that examined the effects of different types of correctness feedback on children's actual and perceived performance with a math app. In the study, forty-five grade-2 students solved easy, moderate, and hard drill questions with a math app augmented with textual, icon, and emoticon correctness feedback. Results suggested that, for the most part, neither the feedback type nor the difficulty level affect children's actual and perceived performance with the app.**

Keywords—children; mathematics; correctness feedback; visual feedback; graphical feedback; education; mobile apps; games.

I. INTRODUCTION

Mobile devices are becoming increasingly popular among children. A recent survey revealed that about 72% of children aged eight and under in the U.S. have access to mobile devices, particularly tablets, which they are using for various purposes, including for learning, playing games, watching videos, taking pictures, and gaining access to social networks [1]. A different survey [2] found out that about 95% of all apps on the Apple App Store are targeted at children aged from three to thirteen, among which math apps are the most popular.

Many researchers have emphasized that special care must be taken when designing apps for children since they are a special user-group that have different needs, desires, and expectations than adults [3], [4]. Many have argued that child users need to be continuously updated on the current state of a user interface, as it enhances their performance by making them aware of the input and interactions necessary to perform a task [5]–[7]. Directive and facilitative visual feedback that are typically composed of hints, suggestions, and/or tutorials (e.g., [8]–[10], etc.) could also motivate children to learn and understand new concepts and skills [6], [11], [12].

However, the most common type of feedback in mobile apps is correctness feedback that simply informs the user whether an input is correct or incorrect [9], [13]. Unlike directive and facilitative feedback, most apps cannot function without correctness feedback; because without it, it is often impossible to determine if an input is valid or not. Nevertheless, not many studies have investigated this specific type of feedback.

To address this, first, we informally surveyed the most popular math apps for children. Similar to several prior investigations [9], [13], results revealed that correctness feedback was the most used feedback type in mobile apps. We also identified three different types of correctness feedback that were commonly used, namely textual, icon, and emoticon. We tested the effects of these three feedback types on children's performance with a math app in a pilot study. Results suggested that feedback type does not affect children's actual and perceived performance with the math app. We then further validated the findings in a cross-sectional study.

Blair [9] conducted an informal survey to study the types of feedback used in math apps aimed at preschool children. Results revealed that 87% of all apps provided correctness feedback that only inform the user of whether an answer is correct or incorrect. The remaining 13% apps provided hints and tutorials to facilitate learning. Blair, however, focused only on the evaluation of the apps [14], thus did not investigate the effects of different types of correctness feedback on children's performance or preference.

Masood and Hoda [15] designed and developed an app to help 5 to 6 year-olds learn and practice early numeracy addition and subtraction. It used a combination of auditory and graphics-heavy icon-based correctness feedback. However, they did not evaluate the effectiveness of the app or the feedback methods. Zhang et al. [16], in contrast, evaluated the effectiveness of three existing apps that used different scaffolding strategies to support learning of decimals and multiplication. These apps also used a mixture of auditory and icon-based feedback. They conducted an exploratory study in an inclusive grade-4 class, where about half of the students were either at-risk or had disabilities. Results showed that using the apps reduces the achievement gap between average and struggling students. Yet, similar to Blair et al. [14], they did not explore the effects of different types of correctness feedback on children's performance or preference.

Correctness feedback is not exclusive to math apps. Walker [13] identified four different types of feedback in apps, regardless of the domain or the target audience. Two of them were correctness feedback, either with or without re-entry, and the other two were facilitative feedback, composed of hints, suggestions, or tutorials. Based on this finding, Buckler and Peterson [17] evaluated six commercial apps aimed at helping adults with special needs to

This work was supported in part by the Ontario Ministry of Research and Innovation (MRI), the Canada Research Chairs (CRC) program, the Canada Foundation for Innovation (CFI), NSERC, the Portuguese Foundation for Science and Technology (FCT), and the European Regional Development Fund (ERDF) through the Operational Programme for Competitiveness and Internationalisation – COMPETE 2020, within the Research Centre on Child Studies of the University of Minho (CIEC) POCI-01-0145 FEDER-007562 with the Postdoctoral Grant: SFRH/BPD/111891/2015.

perform activities of daily living, such as telling time, counting money, etc. Although they did not elaborate on the exact types of feedback used in these apps, their respective ratings suggest that they all provided correctness feedback almost exclusively.

Sandvik et al. [18] investigated the effectiveness of two apps intended for improving kindergarten students' language skills. Like many popular math apps, they used a mixture of auditory and icon-based correctness feedback, but did not explore their effects on children's performance or preference.

Some researchers have studied the effects of seductive details in educational apps. Seductive details are *"appealing elements that are inserted alongside educational content with the intent that children's interest in these elements will make the educational content more compelling and memorable"* [19]. Although the use of seductive details are more common in directive and facilitative feedback [20], some have also studied their effects on correctness feedback. Fisch [19] suggested caution in using seductive details, such as colorful graphics, animations, and sound effects, since it could make children more interested in the appealing elements rather than the intended educational content. Yet, this suggestion was based on prior findings in the field of educational psychology that suggested that seductive details result in poorer retention of information and transfer of learning [21]–[23], and not on a user study. The design of correctness feedback could also be of interest to color researchers who attempt to identify the relationships between colors and psychological functioning [24]. However, this is outside the scope of this work.

TABLE I. DIFFERENT TYPES OF CORRECTNESS FEEDBACK IN POPULAR MATH APPS FOR CHILDREN[a]

Correctness Feedback	**Apps** (%)	**Animation** (%)	**Sound Effects** (%)
Text	23.1	7.7	23.1
Icon	61.5	46.2	61.5
Emoticon	15.4	15.4	15.4

[a] The surveyed apps are Kids Math Games, Math Teacher for Children, Math for Kids, Puzzles Math Games for Kids!, Toddler Math Plus, Kids Math, Math Puppy, Kindergarten Math Class, Math Challenge–Brain Workout, Math Kid, Todo Maths, AB Math, and Monkey Math School Sunshine.

II. AN INFORMAL SURVEY

Since not much work has explored the types of correctness feedback used in math apps, we conducted an informal survey to study the most downloaded math apps aimed at kindergarten to elementary aged children (5–13 years old). We picked this age range since it is the most targeted age range for app developers [2]. Six of these apps were for the Google Android OS, five were for the Apple iOS, and two were cross-platform. All thirteen apps required children to solve drill questions involving the four basic arithmetic operations, i.e., addition, subtraction, multiplication, and division. Most of these apps (77%) were proprietary, that is either paid or required in-app purchases to activate all features. The remaining 23% were free.

Results of the survey (TABLE I) revealed that roughly 62% of the apps used graphics-heavy icons in correctness feedback, i.e., colorful check marks and balloons. About 23% used textual feedback that displayed messages, such as "Outstanding!" and "Correct". The remaining 15% used emoticons, such as happy and sad faces. About 70% of all apps also used animations, such as fade-in/out and fly-in/out effects that took from 1 to 5 seconds (average 2.5) to complete. Besides, all apps used sound effects, such as cheers and claps, and some apps also provided facilitative feedback containing hints, suggestions, and tutorials. But, auditory and facilitative feedback are outside the scope of this work.

III. MOTIVATION

Since textual, icon, and emoticon correctness feedback are commonly used in apps for children (TABLE I, [9]), identifying how they influence children's actual and perceived performance could inform practices, impacting a large number of apps, hence the users. It could also provide an understanding of these feedback types' roles in more complex, hybrid feedback that combines text, different kinds of icons, and/or emoticons with sound and visual effects. Sated simply, if the basic feedback types do not affect performance, but some of the hybrid feedback types do, then one can assume that factors other than text, icons, and emotions are contributing towards the influence.

IV. PILOT STUDY

We conducted a pilot study to explore whether different types of correctness feedback, i.e., textual, icon, and emoticon, affect children's performance with a math app.

The pilot used multiple LG Optimus L7 II P710 smartphones, 121.5×66.6×9.7 mm, 118 grams. The devices ran on Android OS Jelly Bean 4.1.2 at 480×800. We developed a custom app for the pilot using the default Android SDK. It generated drill questions at run-time involving the addition, subtraction, multiplication, and division of whole numbers between 0 and 10, e.g., "5 + 10 = ?" and "10 – 2 = ?". It displayed one question at a time, prompted children to enter the answer using a keypad, and then displayed textual, icon, or emoticon feedback (Fig. 1). It kept scores for all responses in a scorecard. The app did not provide any auditory feedback to eliminate a potential confounding factor.

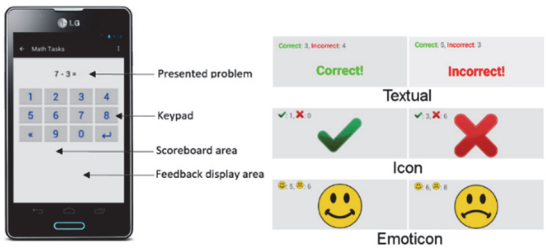

Fig. 1. The device(s), the custom app, and the three feedback types used in the pilot study.

Twenty-one grade-2 students participated in the pilot study. We collected consents from the school, the children, and their guardians. Their average age was 7.3 years (SD = 0.5). Eight of them were male and thirteen were female. They all used tablets for various class activities, thus, were familiar with touchscreens.

The pilot used a within-subjects design with three blocks for the three independent variables: textual, icon, and emoticon. The dependent variable (metric) was *Attempts per Operation* (APO) that measured the average number of attempts per correct answer. In each block, they solved 12 drill questions. The blocks were counterbalanced to eliminate the effect of learning.

During the study, we first demonstrated the app to the children, and allowed them to practice with it. We then started the pilot. Error correction was forced—children had to enter the correct answer to see the next question (Fig. 2). Upon completion of the pilot, all children rated their perceived impact of the examined feedback types on their math skills and performance using a children-friendly five-point Likert scale [25]. In summary, the design was: 21 children × 3 blocks × 12 drills = 756 drills, excluding practice questions.

Fig. 2. Two children participating in the pilot study.

An ANOVA failed to identify a significant effect of feedback type on APO ($F_{2,20} = 1.78$, ns). On average children took 1.05 (SD = 0.07), 1.1 (SD = 0.14), and 1.07 (SD = 0.11) attempts per questions with textual, icon, and emoticon correctness feedback, respectively. We used a Friedman test to analyze the subjective data by converting the five-point scale to a three-point scale using linear transformation. The test failed to identify significance with respect to perceived impact on math skills ($\chi^2 = 0.67$, ns, $df = 2$). About 52.4%, 47.6% 52.4% children felt that textual, icon, and emoticon feedback impacted their performance.

This suggests that the type of correctness feedback does not affect children's actual and perceived performance with a math app. We conducted a cross-sectional user study to investigate this further. Unlike the pilot, it used three difficulty levels to explore if children find some feedback types more rewarding when solving difficult math problems. It also made several procedural changes to increase its validity (e.g., used tablets instead of smartphones since children were more familiar with the device, recorded more representative metrics, etc.). We discuss these in more detail in the following sections.

V. User Study

The purpose of this user study was to test the following null hypothesis, which we assumed it would fail to reject.

H₀) *The examined correctness feedback types do not influence children's actual and perceived performance with a math app, even when solving problems with varying difficulty levels.*

A. Apparatus

We used multiple Apple iPad 3 Wi-Fi tablets, 241.2×185.7×9.4 mm, 652 g, running on Apple iOS 9.2 at 1536×2048. We used a custom app, developed with HTML5 and JavaScript, for the study. It presented children with drill questions for addition and subtraction. The app displayed one problem at a time, and asked children to enter the answer using a keypad. Upon each entry, the app displayed either textual, icon-based, or emoticon-based correctness feedback (Fig. 3). No animation was used and no auditory feedback was provided to eliminate any potential confounds. The app maintained a scorecard of all answers. It processed all interactions on the client side but recorded all data in a PHP database. The Apple iPad tablets were placed on a table using commercial cases (Fig. 4).

B. Difficulty Levels

The math problems used in the study were selected from a popular math workbook for grade-2 students [26]. We categorized the problems into easy, moderate, and hard difficulty levels in consultation with three experienced math teachers from the school. We then consulted with a fourth teacher from a different school to make sure that the selected problems and the difficulty levels were appropriate. Examples of easy, moderate, and hard problems are "2 + 2 + 4 = ?", "40 + ? = 53", and "345 – 234 = ?", respectively. We used the three difficulty levels to find out if children's performance with and preference for the math app changes when they are solving problems with different difficulty levels. This is based on a prior work that suggested that rewarding children with visually appealing and entertaining feedback could carry some level of motivational value, encouraging them to engage in the drill and practice to see the feedback [19].

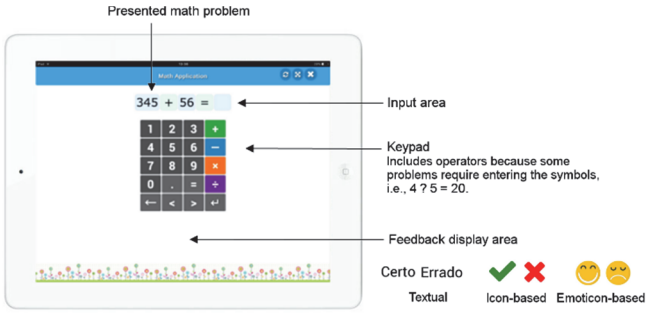

Fig. 3. The device(s), the custom app, and the three feedback types used in the final user study.

C. Participants

Three grade-2 classes from the same school, each consisting of 15 students, in total 45 students, voluntarily participated in the final study. None of them participated in the pilot study. We collected consents from the school, the children, and their guardians for the study. There were 8 females and 7 males in the first class, 6 females and 9 males in the second class, and 8 females and 7 males in the third class. Average age for the three classes were 7.1 (SD = 0.3), 7 (SD = 0), and 7.1 (SD = 0.3) years, respectively. All participants were familiar with touchscreens since they all used tablets for various class activities. This enabled us to attain permission from the school to conduct the study in these classes.

Fig. 4. Two children interacting with the custom app during the user study.

We recruited the three classes because their respective math teachers confirmed (based on the children's previous test scores) that, collectively, the classes were roughly the same in terms of competence. Besides, the children used the same textbooks and

participated in the same activities at school. We did not conduct a formal test to assess competence since it is difficult to achieve in a single test. Nevertheless, we acknowledge this as a limitation of the study, since we cannot claim beyond a reasonable doubt that the classes were, in fact, similar in competence.

D. Design

During the study, each class (group) used the three different feedback types with alternating difficulty levels in three sessions that expanded over three days. Hence, the independent variables were feedback type × difficulty level and the dependent variables (or metrics) were Preparation Time and Success Rate (see below). TABLE II illustrates the design.

TABLE II. THE DESIGN OF THE USER STUDY

Group	Session 1 (Day 1)	Session 2 (Day 2)	Session 3 (Day 3)
1	Emoticon × Easy	Textual × Moderate	Icon × Hard
2	Icon × Moderate	Emoticon × Hard	Textual × Easy
3	Textual × Hard	Icon × Easy	Emoticon × Moderate

Initially, we wanted to test all feedback types and all difficulty levels with each group, but had to deviate from the plan due to practical reasons. Such a design would have required either nine groups or nine sessions to ensure a reasonable sample size for each condition, increasing the length of the study. Neither the teachers nor the guardians were comfortable with these design alternatives, since they would have affected children's regular class activities. We also decided against recruiting students from other schools, because we could not guarantee that they had a similar level of competence as our participants. Hence, we settled on the above design that assured 15 children per condition, which we believe is a reasonable sample size for a study involving ~7-year-olds.

One limitation of this design is, children in different classes started with a different difficulty level. For example, children in the first class started with hard questions paired with emoticon feedback, whereas children in the second class started with moderate questions paired with icon feedback. However, we scheduled the sessions on different days to reduce any potential effects of skill transference.

E. Procedure

We started each session with a demonstration of the app and the respective feedback type. We then asked children to interact with the app in a practice block that included three drill questions. The actual session started after the practice, where all children solved the same five drill questions for addition and subtraction. Hence, there were five drill questions per difficulty level that were repeated across group (not sessions). We instructed children to be careful in solving the problems, but assured them that it was alright to make mistakes since the app did not allow repetitive attempts. The app provided children with textual, icon, or emoticon correctness feedback. We also encouraged children to correct all input errors as they notice them, however we did not enforce this. The three sessions were carried out on three consecutive days, during the class hours (Fig. 4). The day after the completion of the study, children were asked to participate in a brief interview session, where they were asked to pick the feedback type(s) that had the most impact on their math skills, attitude towards math, and their preference of the math app. This session was conducted in private to avoid any bias due to mutual influence.

F. Metrics

The app recorded the following metrics for each child.

Preparation Time is the average time (seconds) children took before entering an answer. This is an estimation of the compound time for processing the previous feedback and solving the current drill question. This was calculated from the moment a feedback was displayed to the moment a digit was entered. We are calling this an "estimation" because some children may start inputting while still solving the problem.

Success Rate is simply the average percentage of correct answers entered in a session.

VI. RESULTS

To analyze the effects of feedback type, we filtered the data for each difficulty level and then ran an ANOVA on the dependent variables. Similarly, to analyze the effects of difficulty level, we filtered the data for each feedback type and then ran an ANOVA on the dependent variables. TABLE III displays the results.

TABLE III. RESULTS OF THE STUDY.THE BOLD VALUES SIGNIFY STATISTICAL SIGNIFICANCE AND THE VALUES INSIDE THE BRACKETS SIGNIFY STANDARD ERROR. THE PERCENTAGES DO NOT ALWAYS ADD UP TO 100% SINCE CHILDREN COULD PICK MULTIPLE FEEDBACK TYPES AS THEIR ANSWERS

	Actual Performance						**Perceived Performance**						**Preference**		
	Preparation Time (Seconds)			**Success Rate** (%)			**Math Skills** (%) *"The examined feedback type improved my math skills"*			**Attitude** (%) *"The examined feedback type made me like math more"*			**Overall Rating** (%) *"I would like to keep using the examined feedback type"*		
Difficulty	***Text***	***Icon***	***Emoticon***	***Text***	***Icon***	***Emoticon***	***Text***	***Icon***	***Emoticon***	***Text***	***Icon***	***Emoticon***	***Text***	***Icon***	***Emoticon***
Easy	14.45 (6.4)	18.14 (12.1)	31.55 (9.7)	95.89 (5.7)	75.01 (7.2)	81.08 (6.8)	53.33	40.0	20.0	6.67	33.33	20.0	53.30	**86.67**	80.0
Moderate	42.44 (6.3)	33.59 (12.1)	27.96 (9.4)	69.73 (5.5	**87.50 (7.2)**	45.21 (6.9)	46.67	40.0	46.67	26.66	33.33	46.66	66.67	80.0	80.0
Hard	41.45 (6.4)	**96.63 (12.2)**	**81.46 (9.3)**	41.09 (5.7)	56.33 (7.2)	56.01 (6.8)	46.67	20.0	20.0	46.66	33.33	46.66	73.33	80.0	100

A. Preparation Time

An ANOVA failed to identify a significant effect of feedback type on Preparation Time for easy ($F_{2,42} = 2.96$, ns) or moderate ($F_{2,42} = 1.34$, ns) difficulty level. However, a significant effect was identified for hard difficulty level ($F_{2,42} = 3.92$, $p < .05$). A Tukey-Kramer test recognized three distinct groups in hard—textual, icon, and emoticon.

An ANOVA also identified a significant effect of difficulty level on Preparation Time for textual ($F_{2,42} = 6.15$, $p < .0001$), icon ($F_{2,42} = 11.76$, $p < .0001$), and emoticon ($F_{2,42} = 10.26$, $p < .0005$) feedback. A Tukey-Kramer test identified two distinct groups in all feedback methods—easy-moderate and hard.

B. Success Rate

An ANOVA failed to identify a significant effect of feedback type on Success Rate for easy ($F_{2,42} = 2.71$, ns) or hard ($F_{2,42} = 2.01$, ns) difficulty level. But a significant effect was identified for moderate ($F_{2,42} = 8.93$, $p < .001$). A Tukey-Kramer test identified two distinct groups in moderate—textual-icon and emoticon.

Unsurprisingly, a significant effect of difficulty level was identified on Success Rate for textual ($F_{2,42} = 23.37$, $p < .00001$), icon ($F_{2,42} = 4.71$, $p < .05$), and emoticon ($F_{2,42} = 7.25$, $p < .005$) feedback. A Tukey-Kramer test identified two distinct groups in all feedback types—easy and hard.

VII. User Feedback

We used a Kruskal Wallis test to compare the non-parametric data from the interview.

A. Perceived Impact on Math Skills

A Kruskal Wallis test failed to identify significance in regard to children's perceived impact on skills for the three feedback types with easy ($\chi^2 = 1.6957$, ns, $df = 2$), moderate ($\chi^2 = 0.3913$, ns, $df = 2$), and hard ($\chi^2 = 1.1739$, ns, $df = 2$) difficulty level. There was also no significant effect of difficulty level for textual ($\chi^2 = 0.3913$, ns, $df = 2$), icon ($\chi^2 = 1.1739$, ns, $df = 2$), or emot-icon ($\chi^2 = 2.0870$, ns, $df = 2$) feedback.

B. Attitude Towards Math

A Kruskal Wallis test failed to identify significance in regard to children's attitude towards math after using the feedback types with easy ($\chi^2 = 1.5652$, ns, $df = 2$), moderate ($\chi^2 = 1.5652$, ns, $df = 2$), and hard ($\chi^2 = 0.1304$, ns, $df = 2$) difficulty level. There was also no significant effect of difficulty level for textual ($\chi^2 = 3.6522$, ns, $df = 2$), icon ($\chi^2 = 0.1304$, ns, $df = 2$), or emoticon ($\chi^2 = 2.0870$, ns, $df = 2$) feedback.

C. Overall Rating

A Kruskal Wallis test identified significance in regard to children's preference of the three feedback methods for easy ($\chi^2 = 8.9243$, $p < .05$, $df = 2$), but not for moderate ($\chi^2 = 3.5782$, ns, $df = 2$) or hard ($\chi^2 = 2.0303$, ns, $df = 2$) difficulty levels. A Tukey-Kramer test revealed that substantially more children liked icon than textual for easy problems.

However, a Kruskal Wallis test failed to identify significance regarding children's preference for textual ($\chi^2 = 4.6245$, ns, $df = 2$), icon ($\chi^2 = 0.4597$, ns, $df = 2$), and emoticon ($\chi^2 = 2.8500$, ns, $df = 2$) for the three difficulty levels.

VIII. Discussion

As expected, results showed that children took significantly more Preparation Time and made significantly more mistakes in solving the hard problems. An ANOVA failed to find a significant effect of feedback type on Preparation Time for easy and moderate problems. However, a significant effect was identified for hard problems. A Tukey-Kramer test revealed out that children took significantly more time for hard problems with icons. This is also noticeable in TABLE III, where one can see that both icon and emoticon feedback took roughly twice as much time as textual feedback. This could be because of a phenomenon known as the "seductive details effect" [20] that suggests, visually attractive feedback often divert children's attention away from the task at hand, causing them to focus on the feedback instead [21]. It is also possible that children took relatively more time to "process" the graphics, increasing the overall cognitive load [27], affecting their performance for the questions that are already difficult to solve. Yet, further studies are necessary to fully investigate these possibilities.

Results failed to find a definite relationship between feedback type and Success Rate. During the study, children yielded mostly comparable Success Rates with all feedback types. Although, a significant effect of feedback type was identified for moderate problems (noticeable in TABLE III), the qualitative data suggest that it was an outlier.

A. Qualitative Data

Analysis failed to identify significance regarding children's perceived impact on math skills for the examined feedback types. Analysis also failed to identify significance regarding children's attitude towards math after using the different feedback types. Most children felt that the feedback types had no or comparable impacts on their math skills and their attitude towards math. This suggests that different types of correctness feedback do not affect children's perceived performance with a math app.

There was a significance regarding children's preference of the feedback types for easy problems—most children preferred icon and emoticon feedback than textual feedback. This suggests that children generally prefer attractive visual feedback, but are also aware that they could affect their performance when solving challenging problems. Note that children did take relatively more time to solve hard drill questions with these feedback types.

Although not the focus of our work, we asked children if the feedback types influenced their willingness to use the math app. Almost all children responded that the feedback types did not affect their willingness to use the app. A Kruskal Wallis test also failed to identify a significant effect regarding this and the three difficulty levels. This suggests that correctness feedback does not affect children's impression of the app.

B. Implications

Results of this investigation suggest that, for the most part, different types of correctness feedback do not affect children's actual and perceived performance with a math app, regardless of

the difficulty level of the tasks. We hope that these findings will encourage app developers to reconsider using graphics-heavy correctness feedback in apps, since it not only increases the production time and cost but also slows down the interactions due to the increased processing and cognitive demand. However, we caution that our recommendations are in the context of correctness feedback. Using graphics/animations in directive and facilitative feedback could make some children more interested in learning by stimulating their senses [7], [28].

IX. Conclusion

We first presented results of an informal survey that revealed that the most downloaded math apps targeted at children used three types of correctness feedback: textual, icon, and emoticon, typically augmented with various animations and sound effects. We evaluated these feedback types in a pilot study that suggested that they do not affect children's actual and perceived performance with a math app. We extended our investigation in a cross-sectional study where 45 grade-2 students solved easy, moderate, and hard drill questions with a math app augmented with textual, icon, and emoticon correctness feedback. Results suggested that these correctness feedback, for the most part, do not influence children's actual and perceived performance with a math app. We hope that work will inspire further research in the area, and encourage app developers to reconsider the use of graphics-heavy correctness feedback in math apps targeted at children.

X. Future Work

We discussed several limitations of the study, particularly the absence of a pre-test to evaluate children's competence and the use of a mixed design that failed to properly counterbalance the conditions. In the future, we will conduct more controlled studies to address these limitations. We will also extend our work to other feedback types, educational apps, and games.

Acknowledgments

We thank all the students and their guardians, teachers, and staff of Colégio Teresiano de Braga for their support. We also thank TLCI 2 – Soluções Integradas de Telecomunicações, S.A. for lending us the smartphones for the pilot study.

References

[1] V. Rideout, "Zero to eight: children's media use in America 2013," *Pridobljeno*, pp. 1–31, 2013.

[2] C. Shuler, Z. Levine, and J. Ree, "iLearn II: an analysis of the education category of Apple's app store," in *The Joan Ganz Cooney Center at Sesame Workshop, New York, NY, USA*, 2012.

[3] A. Druin, "The role of children in the design of new technology," *Behav. Inf. Technol.*, vol. 21, no. 1, pp. 1–25, Jan. 2002.

[4] A. S. Arif and C. Sylla, "A comparative evaluation of touch and pen gestures for adult and child users," in *Proceedings of the 12th International Conference on Interaction Design and Children - IDC '13*, 2013, pp. 392–395.

[5] L. Anthony, Q. Brown, J. Nias, and B. Tate, "Examining the need for visual feedback during gesture interaction on mobile touchscreen devices for kids," in *Proceedings of the 12th International Conference on Interaction Design and Children - IDC '13*, 2013, pp. 157–164.

[6] D. H. Clements, "Effective use of computers with young children," in *Mathematics in the Early Years*, J. V. Copley, Ed. Reston, VA, USA, 1999, pp. 119–128.

[7] A. Druin, *Mobile Technology for Children: Designing for Interaction and Learning*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009.

[8] A. S. Arif, C. Sylla, and A. Mazalek, "Learning new words and spelling with autocorrections," in *Proceedings of the 2016 ACM on Interactive Surfaces and Spaces - ISS '16*, 2016, pp. 409–414.

[9] K. P. Blair, "Learning in Critter Corral: evaluating three kinds of feedback in a preschool math app," in *Proceedings of the 12th International Conference on Interaction Design and Children - IDC '13*, 2013, pp. 372–375.

[10] K. Highfield and K. Goodwin, "Apps for mathematics learning: a review of 'educational' apps from the iTunes app store," in *The 36th Annual Conference of Mathematics Education Research Group of Australasia*, 2013, no. 2009, pp. 378–385.

[11] C. Finegan and N. J. Austin, "Developmentally appropriate technology for young children," *Inf. Technol. Child. Educ. Annu.*, vol. 1, pp. 87–102, 2002.

[12] S. W. Haugland, "The effect of computer software on preschool children's developmental gains," *J. Comput. Child. Educ.*, vol. 3, no. 1 (January 1992), pp. 15–30, 1992.

[13] H. Walker, "Evaluating the effectiveness of apps for mobile devices," *J. Spec. Educ. Technol.*, vol. 26, no. 4, pp. 59–63, 2011.

[14] K. P. Blair, J. Pfaffman, M. Cutumisu, N. Hallinen, and D. Schwartz, "Testing the effectiveness of iPad math game: lessons learned from running a multi-classroom Study," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '15*, 2015, vol. 2, pp. 727–734.

[15] Z. Masood and R. Hoda, "Math tutor: an interactive Android-based numeracy application for primary education," in *Proceedings of the Fifteenth Australasian User Interface Conference - Volume 150 (AUIC '14)*, 2014, pp. 3–10.

[16] M. Zhang, R. P. Trussell, B. Gallegos, and R. R. Asam, "Using math apps for improving student learning: an exploratory study in an inclusive fourth grade classroom," *TechTrends*, vol. 59, no. 2, pp. 32–39, 2015.

[17] M. Malone and M. Peterson, "Is there an app for that? Developing an evaluation rubric for apps for use with adults with special needs," *J. BSN Honor. Res.*, vol. 5, no. 1, pp. 19–32, 2012.

[18] M. Sandvik, O. Smørdal, and S. Østerud, "Exploring iPads in practitioners' repertoires for language learning and literacy practices in kindergarten," *Nord. J. Digit. Lit.*, vol. 2012, no. 3, pp. 204–220, 2012.

[19] S. M. Fisch, "Making educational computer games 'educational,'" in *Proceeding of the 2005 conference on Interaction design and children - IDC '05*, 2005, no. 1, pp. 56–61.

[20] H. R. Schugar, C. A. Smith, and J. T. Schugar, "Teaching with interactive picture e-books in grades K-6," *Read. Teach.*, vol. 66, no. 8, pp. 615–624, May 2013.

[21] S. F. Harp and R. E. Mayer, "How seductive details do their damage: a theory of cognitive interest in science learning.," *J. Educ. Psychol.*, vol. 90, no. 3, pp. 414–434, 1998.

[22] R. Garner, M. G. Gillingham, and C. . S. White, "Effects of 'seductive details' on macroprocessing and microprocessing in adults and children," *Cogn. Instr.*, vol. 6, no. 1, pp. 41–57, Mar. 1989.

[23] G. D. Rey, "A review of research and a meta-analysis of the seductive detail effect," *Educ. Res. Rev.*, vol. 7, no. 3, pp. 216–237, Dec. 2012.

[24] A. J. Elliot, "Color and psychological functioning: a review of theoretical and empirical work," *Front. Psychol.*, vol. 6, no. APR, pp. 1–8, Apr. 2015.

[25] J. C. Read and S. MacFarlane, "Endurability, engagement and expectations: measuring children's fun," *Interact. Des. Child.*, vol. 2, pp. 1–23, 2002.

[26] A. Rodrigues and L. Azevedo, *Matemática - Ensino Básico 2 ° Ano*. Oporto, Portugal: Areal Editores, 2011.

[27] R. E. Mayer and R. Moreno, "Nine ways to reduce cognitive load in multimedia learning," *Educ. Psychol.*, vol. 38, no. 1, pp. 43–52, Mar. 2003.

[28] L. Lazaris, "Designing websites for kids: trends and best practices," *Smashing Magazine*, 2009. [Online]. Available: https://www.smashingmagazine.com/2009/11/designing-websites-for-kids-trends-and-best-practices/. [Accessed: 29-Dec-2016].