The Development of a Checklist To Assess Tablet Games on Usability for Language Therapy

Name	L. Reinhoudt
Student number	4196619
Course	Afstudeeronderzoek
Status	Final research report
Date	July 1st, 2015
Research institution	University of applied sciences Utrecht (HU)
Supervisors HU	Prof. dr. P.A.M. Gerrits & Drs. I. Singer
Supervisor UU	R.L.M. Zwitserlood MA
Journal intended	Child language teaching and therapy
Requirements journal	max. 6000 words; reference style: SAGE Harvard
Reference style	Vancouver
Reporting guidelines	COREQ (qualitative part) & COSMIN (quantitative part)
Word count	3712
Word count abstract	300

Master Clinical Health Sciences (KGW), Clinical Language, Speech and Hearing Sciences, Utrecht University

Introduction

About seven percent of children between two and six years old have specific language impairment (SLI), a form of developmental language impairment in which children demonstrate unexpected difficulties with the acquisition of spoken language.(1,2) Children with SLI are at risk for social and behavioural problems and educational difficulties.(2,3) Speech language therapy is effective in treating SLI.(4) There are no specific guidelines about which type of intervention to offer children with SLI.(4) The speech-language therapist (SLT) is responsible for the choice of appropriate methods and materials.

With the first appearance of tablet computers and the rapidly growing market for tablets since then, an increasing number of SLTs have started to use tablet games (TGs) in therapy to train the language skills of children with SLI. SLTs report that TGs are a valuable addition to traditional treatment methods because most children are highly motivated to work with TGs.(5) Although the TGs are not specifically developed for language therapy, SLTs report that there are many opportunities to incorporate TGs in language therapy to achieve language goals.(5) Specific studies on the effectiveness of TGs in language therapy are lacking. The results from some scoping studies suggest that the use of TGs may enhance engagement and motivation of children(6,7), improve the language skills of children with special educational needs(8) and increase the reading skills of typically developing children(9). Although these are small scoping studies without any statistical testing or expert opinions and not in all cases specifically focussed on the use of TGs in language therapy, the results imply a positive trend.

The importance of carefully choosing appropriate TGs is highlighted in literature(6), but because most TGs are not specifically developed for language therapy, it is not easy to select usable TGs. Some Dutch sources document TGs for language stimulation, such as www.logo-apps.nl and www.praatapps.nl. On these websites, the evaluation of the TGs is based on the clinical experience of one or more SLTs. There are also Dutch and English checklists which can be used to assess the usability of a TG for language therapy, for example the Dutch “Checklist for apps”(10), which is based on the checklist of Klarowska(11), and the checklist of Tomarakos(12). These sources use their own set of criteria and there is no information about the validity and reliability of these checklists. This causes a great disparity in reviews and confusion about the usability of TGs specifically for language therapy. Because of the importance of making well-founded choices when using TGs in language therapy, SLTs need a valid and reliable instrument to assess the usability of a TG for language therapy.

Because of the lack of such an instrument, this study aims to determine which characteristics make a TG useful for language therapy and to develop a valid and reliable Dutch checklist to assess the usability of TGs for language therapy.

Research questions

1. Which characteristics of a tablet game determine whether the game is useful, according to SLTs, for language therapy for children between two and six years old with SLI?

2. How can these characteristics be questioned with a checklist?

3. What are the clinimetric properties of this checklist?

Methods

Design

The study has a mixed methods design with a sequential approach.(13,14) First, a checklist to assess the usability of TGs for language therapy was developed based on qualitative data collected in focus groups. Second, the clinimetric properties of this checklist were examined in the quantitative part of the study.

Participants

The participants were 41 Dutch SLTs who were recruited by advertisements on social media. The participants had knowledge about the study based on the information letter and informed consent form. Ten of the 41 SLTs participated in the qualitative part of the study. The remaining 31 SLTs participated in the quantitative part of the study.

The inclusion criteria were as follows: 1) working with children with SLI (aged two to six years old); 2) possession of an Apple iPad and 3) current use of the tablet in their therapy sessions with children.

Tablet game selection

The required sample size for reliability and validity assessment is 50 subjects.(15) Therefore, 50 TGs were selected by stratified sampling with three strata, including 1) rating of the TG according to an existing judgment on a review website; 2) purpose of the TG; 3) age category of the TG. Table 1 shows the results of this selection. Because few games with a low rating are present on the review websites, the category of low quality games is smaller. Appendix I provides a list of all TGs that were studied.

Table 1: Characteristics of tablet games

Data collection

Qualitative data were collected from two focus groups, each consisting of 90 minutes with five participants. Focus groups are a useful method for exploring the SLTs' knowledge and experiences with TGs in language therapy because the group interaction and discussion provides information about what the SLTs think about TGs and also how and why they think that way.(14,16) The focus groups were conducted by the first author (LR) and were organized in Utrecht and Nijmegen in February 2015. The focus group in Utrecht took place at the University of Applied Sciences Utrecht (HU) and the focus group in Nijmegen was held at KC Stijntje Buys, the workplace of the first author. The first author is a master's student in clinical health sciences at the University of Utrecht. In addition to her part-time academic training, the first author works as an SLT in an institution for people with intellectual disabilities and is a

member of the editorial board of www.praatapps.nl. The focus groups consisted of a structured discussion about important characteristics of a TG for language therapy and a practical part in which the participants assessed selected TGs and discussed their usefulness for language therapy. Adding this practical part to the focus group made it possible to gather rich data about the characteristics that make a TG useful for language therapy.(16) The discussions were structured by using a pilot-tested topic list which was developed based on a systematic literature search. Through the use of this topic list and avoidance of any substantial participation in the discussion, the influence of the first author on the data was limited.(14)The focus groups were audio recorded and the first author took field notes during the meetings.

Based on the themes that emerged from the qualitative data, items for the checklist were formulated. To collect data about the importance and weighing of these items, all participants rated the checklist items on the degree of importance using a 5-point scale from unimportant [1] to very important [5].

To examine the inter-rater reliability and ensure the validity of the checklist, data were collected by having the 31 SLTs who did not participate in the focus groups assess the 50 selected TGs using the checklist. Each TG was assessed three times. The TGs were randomly assigned to the SLTs and each TG was assessed by a different set of SLTs. The SLTs used the selected TGs in at least one therapy session and filled in the online checklist. This resulted in 150 completed checklists.

Data analysis

Content validity and checklist construction. The focus groups were audio recorded, transcribed and analysed anonymously. The transcripts were analysed by open coding, axial coding and selective coding, using NVivo software.(14) The coding process was peer reviewed by a second researcher to improve the reliability of the analysis. Interpretations were compared and disagreements were discussed until consensus was reached. Emerging themes were considered as the categories of the checklist. Subthemes were expressed as dichotomous checklist items. The total score for the checklist is the percentage of questions answered with “yes”. Based on this percentage a star score is assigned to the TG, see Table 2.

Table 2: Calculation of star scores

The checklist items were pretested by the ten SLTs who participated in the focus groups to obtain information about the content validity. They gave written feedback about the comprehensibility and relevance of the items by reading the questions and filling in a checklist for a TG. The items were revised in response to the feedback from this pretest.

Weighing items. The quantitative data about the importance of the items were analysed using descriptive statistics. Based on the mean value of importance the weighing of

each item was determined. When any items had a mean rating <3 (unimportant), they were removed from the checklist.

Inter-rater reliability. Inter-rater reliability was assessed by comparing the opinions of three independent raters about the same TGs. First, the inter-rater reliability per item was calculated using Fleiss' kappa.(17) Items with poor agreement (Fleiss' kappa < 0.20) were removed from the checklist. After the per-item analysis, the inter-rater reliability of the entire checklist was analysed using an intraclass correlation coefficient (ICC). Although the ICC is primarily designed for use with interval/ratio data, it can also be applied to these data on the ordinal scale because the intervals between the star scores (1-5) are assumed to be equivalent(17). Because each TG was assessed by a different set of raters, ICC model 1 (one-way random) is used. A value of ICC >0.60 was considered sufficiently reliable. The significance level was set at < 0.05.(17)

Construct validity. Construct validity evidence involves the degree to which the content of the checklist matches the content domain associated with the construct.(15) Usually, convergent validity is used to determine the construct validity of a measurement instrument. Because there is no comparable instrument regarding the usability of TGs for language therapy, principal component analysis (PCA) was used. PCA examines the construct validity by exploring the underlying constructs of the checklist items. Knowledge about these constructs makes it possible to determine whether the checklist content fits the underlying constructs.(17) Because the PCA is based on a correlation matrix, only the items with fair reliability were included in the PCA. Because of the binary data, a tetrachoric correlation matrix is used. The number of factors retained was determined by Velicer's minimum average partial (MAP) test, which is based on the average partial correlations between the variables after successively removing the effect of the factors. The number of factors that minimizes the average partial correlations was retained.(18)

Ethical issues

The study was conducted according to the principles of the Declaration of Helsinki.(19) The study is approved by the Medical Ethical Screening Committee of the HU Faculty of Health Care. All participating SLTs signed to indicate informed consent.

Results

Themes emerging from focus groups

The coding and analysis of the qualitative data were performed in Dutch, the original language of the data. To provide a representation of the data and the checklist for this research report, the themes, quotations and checklist items were translated into English by the first author in cooperation with a professional translator.(20) For use in practise, a checklist must not only be translated, but must also be adapted culturally with specific research conducted on the clinimetric properties.(21) See appendix II for the original themes in Dutch.

Analysis of the focus groups resulted in seven main themes: user-friendliness, attractiveness, influence on the attention of the child, adaptability, suitability to therapy goals, costs and risks. Each main theme consisted of several subthemes, of which 21 were identified. See table 3 for an overview of all themes, subthemes, descriptions and examples of quotations from the qualitative data in which the themes emerged.

Table 3: Themes and sub-themes emerged from the analysis of the focus groups, translated into English

Construction of the checklist

The aforementioned main themes were made into the categories of the checklist. The 21 subthemes were transformed into one or more dichotomous questions, which constitute the checklist items. The first draft of the checklist had 28 items. These items were pretested by the focus group participants. The items “the game fits the goals and methods of the SLT” and “the content of the game is acceptable for use in language therapy” were not understandable or relevant according to the test panel and were removed. The items “the game is easy to use” and “the game is attractive for use in language therapy” were too broad to answer and were also removed. The items “the game is customizable for the child and adjustments can be shared with parents” and “the game provides specific feedback and offers the child the opportunity to correct mistakes” were ambiguously worded and were both divided into two items. After revision, the checklist consisted of 27 items that the test panel found relevant and understandable.

Weighing items

One item on the draft checklist had a mean importance rating lower than three and was removed, see Table 4. Therefore, the final checklist used to collect data for the quantitative part of the study contained 26 items (see Appendix III and IV). Because the mean importance values of the remaining items were similar, it was not necessary to use a weighing rule when calculating the total score for the checklist.

Table 4: Mean rating of importance checklist items

Inter-rater reliability

Results of the per-item analysis of inter-rater reliability showed poor agreement between the three independent raters (Fleiss' kappa < .20) for fourteen items, see Table 5. To improve the reliability of the checklist, the unreliable items were removed. After removing the items, the entire checklist (star score) had a fair inter-rater reliability of ICC = 0.392.

Table 5: Agreement between three raters per item (Fleiss' kappa)

Construct validity

Since Velicer's MAP test yielded a minimum of 0.1 with three factors, a three-factor solution was extracted. The twelve reliable items from the 26 original were included. The three-factor solution, accounting for 60% of the total variance, is presented in table 6. All items loaded >0.40 on one of the three factors. Item 9, 21 and 26 cross-loaded and were assigned to the factor with the highest loading. The first factor consisted of the items for adaptation of the game to the child's needs, sharing adjustments, switching off music and the usability of a game for different language therapy goals. This factor was labelled "suitability for therapy". The second factor contained the items involving sounds and voices, feedback, attractive rewards and clarity about how the game works. These factor seem to define whether a TG is attractive for a child and therefore usable in language therapy; and was defined as "attractiveness for the child". The third factor, "practical characteristics," consisted of items involving language, availability of a demo version and the absence of pop-up advertisements. The item regarding the availability of a demo version loaded negatively. This may be caused by interpretation problems. Free demo versions of a TG generally do not provide all the capabilities of the game, and perhaps the SLTs therefore reported these TGs as less usable for language therapy. The proportions of the explained variance were as follows: for "suitability for therapy," 0.22; for "attractiveness for the child," 0.20; and for "practical characteristics" 0.17.

Table 6 Factor structure of the checklist

Discussion

More and more SLTs are using TGs during language therapy. Because of the large number of TGs and the fact that the TGs are not specially developed for language therapy, it is difficult for the SLTs to select appropriate TGs. Therefore, the aims of this study are to identify characteristics that make a TG useful for language therapy and to develop a valid and reliable checklist to measure these characteristics.

A strength of the current study is the mixed methods design in which qualitative methods are used to determine important characteristics of TGs according to SLTs, supplemented with quantitative methods to evaluate the clinimetric properties of the checklist developed. The content and face validity of the checklist were verified by incorporation of the focus groups and pretesting of the checklist by SLTs.

User-friendliness, attractiveness, influence on the attention of the child, adaptability, suitability to therapy goals, costs and risks appear to be main themes that influence the usability of a TG for language therapy. To some extent, these themes overlap with criteria found in previous literature, such as the checklist of Tomarakos.(12) However, Tomarakos uses only three categories and questions multiple characteristics per item.(12) The list of criteria found at www.praatapps.nl contains some of the same themes but does not use subthemes or items to make the assessment more specific. The Dutch checklist for apps(10) is partially consistent with the new checklist, but also contains themes which were not mentioned during the focus groups.

The inter-rater reliability analysis of the checklist shows disappointing results for the agreement among three independent raters. After removing a large number of unreliable items (50%), only a fair inter-rater reliability was reached. This revision necessitated an undesirable loss of content. Analysis of the unreliable items suggests that the low inter-rater reliability of the checklist is caused by items that are relatively subjective, for example “The game’s content gives ample opportunity to use speech therapy techniques”. These subjective items question opinions of the SLTs which can differ based on interpretation of the item or different opinions about how to use the TG in therapy. Items that question more factual information about the TGs had higher agreement. Methodological limitations might also be a reason for the low inter-rater reliability of the checklist. The inter-rater reliability is measured by comparing dichotomous data regarding three independent raters. The dichotomous answer scale of the checklist is very practical for the users, but causes methodological limitations such as the use of Fleiss’ kappa to calculate agreement on item level, and a rapidly decrease in reliability when the answer of one rater differs from the others.

A PCA was conducted to obtain information about the construct validity of the checklist. The PCA shows a three-factor solution that demonstrates the importance of taking into account

multiple dimensions when assessing the usability of a TG for language therapy. The PCA indicates that “suitability for therapy”, “attractiveness for the child” and “practical characteristics” are the underlying constructs which must be measured when assessing a TG for usability in language therapy. A limitation of the PCA is that only the twelve reliable items from the checklist were included, while other underlying constructs might have been neglected. Looking at the qualitative data and the items removed, the theme “effect on the attention of the child” seems to be insufficiently represented in the PCA.

To improve the reliability and validity of the checklist, a recommendation for further research is to revise the checklist items for better clarity. Explanations and examples should be used to clarify the items and additional feedback sessions with users are recommended to rule out any terms or items in the checklist that might have multiple interpretations. Appendix V contains specific recommendations for revision. The answer scale of the checklist must be changed. An answer scale on an interval or ratio scale and statistical methods appropriate to these measurement levels should be used to obtain more detailed information about the reliability of the checklist. After these revisions, it would be possible to perform a more complete analysis of construct validity. A PCA with more items would likely result in more underlying constructs to be measured when assessing the usability of TGs for language therapy.

Conclusions and clinical implications

This study presents a checklist to assess the usability of TGs for language therapy consisting of 26 dichotomous questions, based on seven main themes and 21 subthemes derived from qualitative data. The checklist appears to be complete and has good content validity and face validity.

Because deletion of the unreliable items causes an undesirable loss of data, revision and further research are needed to improve the reliability of the checklist to make it a useful instrument for SLTs. After revision, the instrument can be used by SLTs to assess the usability of TGs for language therapy.

References

1. Tomblin JB, Records NL, Buckwalter P, Zhang X, Smith E. Prevalence of Specific Language Impairment in Kindergarten Children. J Speech, Lang Hear Res. 2014;40:1245–60.

2. Leonard L. Specific Language Impairment Across Languages. Child Dev Perspect. 2014;1(8):1–5.

3. St Clair MC, Pickles A, Durkin K, Conti-Ramsden G. A longitudinal study of behavioral, emotional and social difficulties in individuals with a history of specific language impairment (SLI). J Commun Disord. 2011;44(2):186–99.

4. Law J, Garrett Z, Nye C. Speech and language therapy interventions for children with primary speech and language delay or disorder (Review). Cochrane Database Syst Rev. 2008; 3:1-79.

5. Noordzij DS. Onderzoeksverslag Gebruik iPad in de Logopedische behandeling [Bachelor thesis]. Zwolle; University of applied sciences Windesheim; 2012.

6. Sidock J. Critical Review: Is the integration of mobile device apps into speech and language therapy effective clinical practice? [Master thesis]. London, Ontario; University of Western Ontario: School of Communication Sciences and Disorders; 2011.

7. Kucirkova N. iPads in early education: separating assumptions and evidence. Front Psychol. 2014 Jan;5:715.

8. Fernández-López Á, Rodríguez-Fórtiz MJ, Rodríguez-Almendros ML, Martínez-Segura MJ. Mobile learning technology based on iOS devices to support students with special education needs. Comput Educ. 2013;61:77–90.

9. Van de Ven M, De Leeuw L, Van Weerdenburg M. Gamen-der-wijs leren lezen met Letterprins [Master thesis]. Nijmegen; Radboud University; 2014.

10. Eijl L, Heusen A, Lolkema L. Checklist voor apps: Applicaties beoordelen op geschiktheid voor taaltherapie bij kinderen [Bachelor thesis]. Utrecht; University of Applied Sciences Utrecht; 2010.

11. Klarowska B. The Evaluation Rubric for Speech Therapy Apps [Internet]. www.virtualspeechcenter.com. Available from: http://www.virtualspeechcenter.com/articles/100.aspx

12. Tomarakos D. App Review Checklist & Rating Chart [Internet]. www.speechgadget.com. 2012. Available from: http://3.bp.blogspot.com/-LVtvaUvtg4s/T-odu0YpGoI/AAAAAAAAAuw/Xtcud0kao3E/s1600/app+review+checklist.jpg

13. Creswell JW, Fetters M, Ivankiva N. Designing A Mixed Methods Study In Primary Care. Ann Fam Med. 2004;2(1):7–12.

14. Boeije H. Analysis In Qualitative Research. London: Sage Publications Ltd; 2009.

15. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine, A practical Guide. New York: Cambridge University Press; 2011.

16. Kitzinger J. Qualitative research. Introducing focus groups. BMJ. 1995;311:299–302.

17. Portney LG, Watkins MP. Foundations of Clinical Research, Applications to Practice. New Jersey: Pearson Education; 2009.

18. O'Connor BP. SPSS and SAS programs for determining the number of components using parallel analysis and velicer's MAP test. Behav Res Methods Instrum Comput. 2000;32(3):396–402.

19. WMA Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects [Internet]. [cited 2014 Nov 16]. Available from: http://www.wma.net/en/30publications/10policies/b3/.

20. Van Nes F, Abma T, Jonsson H, Deeg D. Language differences in qualitative research: is meaning lost in translation? Eur J Ageing. 2010;7:313–6.

21. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine. 2000;25(24):3186–91.

Tables

Table 1: Characteristics of tabletgames

	Age category 2-4 years			
	Education	Therapy	Entertainment	Total
Low rating	2	0	3	5
Moderate rating	3	0	7	10
High rating	6	0	3	9
	11	0	13	**24**
	Age category 4-6 years			
	Education	Therapy	Entertainment	Total
Low rating	4	0	4	8
Moderate rating	4	4	1	9
High rating	7	1	1	9
	15	5	6	**26**

Table 2: Calculation of star scores

Percentage "yes"	**Stars**
0 – 20 %	1
20 – 40 %	2
40 – 60 %	3
60 – 80 %	4
80 – 100%	5

Table 3: Themes and sub-themes derived from analysis of the focus groups, translated into English

Main theme	Description	Subtheme(s)	Examples of quotes
User-friendliness	A TG is defined as user-friendly when the navigation and operation is clear, the language used was Dutch and the pace of the TG is adequate. The aim of the game must be immediately clear and the TG must load quickly	Navigation Operation Language Clarity Pace Recording results	*"It should look simple and be clear what a child has to do."* *"A game has to load quickly so it is immediately ready for use."*
Attractiveness	A TG must be attractive for the child and for the SLT. Attractiveness also depends on adequate feedback and attractive rewards for the child.	Quality images Quality sounds and spoken language Feedback Reward Content of the game Game elements Surveyability	*"Recognizable images and bright colours are very important."* *"A short, nice reward is important to keep the child interested and motivated."*
Influence on the attention of the child	A game which provides a lot of opportunities for interaction between the child and the SLT appears to be more useful in language therapy.	Purposeful therapy Motivation	*"Children can be so focused on the game that they are not interested in interacting with me anymore."* *"It is important to keep the therapy goal in mind when playing."*
Adaptability	A TG is more usable for language therapy when it is possible to make adjustments to the game. Customizing the TG should be easy and not time consuming.	Required time investment	*"It is very nice when you can customize the game, with specific images and words appropriate to the goals of the child."* *"The possibility of making adjustments must be in balance with the time needed to make a TG ready for use."*
Suitability to therapy goals	A TG becomes more usable for language therapy when it is suitable for different language goals.	Multifunctionality Suitable as homework assignment	*"You can do a lot of different exercises with this TG, which makes it attractive and usable and very suitable to recommend to parents as a homework assignment."* *"it is nice when you work toward different goals with the same TG."*
Costs	It appears to be important that a TG not cost more than five euros, because otherwise the threshold for purchasing the game is too high. It is also important that a trial version be available free of charge.	Free demo version	*"I want to check the possibilities in a free version, and after that I will decide whether I want to buy the TG."* *"When a TG costs, for example, 10 euros, and I cannot try it first, I don't purchase it."*
Risks	The presence of advertisements makes a TG less usable for language therapy. It is also important that a TG prevent accidental in-app purchases.	Advertising In-app purchase	*"Children are distracted by the pop-up ads, they tap on it and then suddenly another website appears."* *"I don't recommend TGs for homework when they have the possibility of in-app purchases."*

Table 4: Mean rating of importance checklist items

	Item	Mean rating importance[a]
1	The game holds the child's attention	4,05
2	The game offers enough possibilities for interaction between the child and the speech therapist	3,32
3	Children like to play the game	4,22
4	The game can be adapted to the child's needs: images can be added , the speaker's voice can be recorded, speed and level can be adjusted	4,34
5	The game's content (such as adjustments, images, scores, etc.) can be shared with parents	4,59
6	The game is ready to use	4,39
7	The game is clearly structured	4,02
8	The game uses plain, recognisable images	3,63
9	The voices and sounds used are clear and distinct	3,68
10	Any background music can be switched off	3,29
11	The game gives practical feedback, clearly indicating if the child has made any mistakes	4,12
12	The game gives the child the opportunity to correct any mistakes	3,61
13	The game offers attractive rewards	3,46
14	The game's content gives ample opportunity to use speech therapy techniques (slowing down the pace of speech, expanding, rephrasing, etc.)	3,76
15	The game offers a variety of themes and play tasks	4,68
16	There is a motivating element of play in the game	4,22
17	The game can be used to pursue various speech therapy objectives in language therapy	4,22
18	The game can be recommended to parents as a homework task	3,80
19	The game is easy to control	3,90
20	The game is either in Dutch or no language is used	2,90[b]
21	It is immediately apparent how the game works	3,90
22	The speed of the game is good	4,54
23	It is possible to mark the progress made by several children	4,10
24	The game has a demo version free of charge	3,32
25	The game costs no more than € 5,00	3,44
26	The game is free from advertising pop-ups	4,27
27	The game ensures that it is impossible for the child to accidentally buy something during a game	4,00

[a] 5 point Likert scale: 1: very unimportant, 2: unimportant, 3: not important not unimportant, 4: important, 5: very important
[b] Items with a mean rating of importance <3,00 were considered unimportant and were removed from the checklist

Table 5: Agreement between three raters per items (Fleiss' kappa)

	Item	Fleiss' kappa[a]
1	The game holds the child's attention	0,05
2	The game offers enough possibilities for interaction between the child and the speech therapist	0,04
3	Children like to play the game	0,15
4	The game can be adapted to the child's needs: images can be added , the speaker's voice can be recorded, speed and level can be adjusted	0,55
5	The game's content (such as adjustments, images, scores, etc.) can be shared with parents	0,40
6	The game is ready to use	0,14
7	The game is clearly structured	0,08
8	The game uses plain, recognisable images	0,02
9	The voices and sounds used are clear and distinct	0,22
10	Any background music can be switched off	0,20
11	The game gives practical feedback, clearly indicating if the child has made any mistakes	0,30
12	The game gives the child the opportunity to correct any mistakes	0,39
13	The game offers attractive rewards	0,23
14	The game's content gives ample opportunity to use speech therapy techniques (slowing down the pace of speech, expanding, rephrasing, etc.)	0,07
15	The game offers a variety of themes and play tasks	0,01
16	There is a motivating element of play in the game	0,19
17	The game can be used to pursue various speech therapy objectives in language therapy	0,22
18	The game can be recommended to parents as a homework task	0,01
19	The game is easy to control	0,00
20	The game is either in Dutch or no language is used	0,37
21	It is immediately apparent how the game works	0,33
22	The speed of the game is good	-0,05
23	It is possible to mark the progress made by several children [b]	0,68
24	The game has a demo version free of charge	0,29
25	The game costs no more than € 5,00	0,03
26	The game is free from advertising pop-ups	0,33
27	The game ensures that it is impossible for the child to accidentally buy something during a game	0,05

[a] Fleiss Kappa ≥ 0.20 = fair agreement
[b] Item removed because of mean rating of importance < 3

Table 6 Factor structure of the checklist

	Items [a]	Factor loadings [b]		
		Factor 1 Suitability for therapy	Factor 2 Attractiveness for the child	Factor 3 Practical characteristics
5	The game's content (such as adjustments, images, scores, etc.) can be shared with parents	0.94		
4	The game can be adapted to the child's needs: images can be added, the speaker's voice can be recorded, speed and level can be adjusted	0.79		
17	The game can be used to pursue various language therapy goals in language therapy	0.78		
10	Any background music can be switched off	0.46		
11	The game gives practical feedback, clearly indicating if the child has made any mistakes		0.90	
13	The game offers attractive rewards		0.78	
12	The game gives the child the opportunity to correct any mistakes		0.75	
21	It is immediately apparent how the game works		0.56	0.47
9	The voices and sounds used are clear and distinct		0.54	0.46
20	The game is either in Dutch or no language is used			0.78
26	The game is free from advertising pop-ups	0.47		0.73
24	The game has a demo version free of charge			-0.60

[a] Only items with a fair reliability (Fleiss kappa ≥0.20) were included

[b] Factor loadings <0.40 not reported

Samenvatting

De ontwikkeling van een betrouwbare en valide checklist om de bruikbaarheid van tabletgames voor taaltherapie te beoordelen

INLEIDING: Steeds meer logopedisten gebruiken tabletgames in de therapie van kinderen met taalontwikkelingsstoornissen. Omdat de meeste tabletgames niet speciaal voor taaltherapie zijn ontwikkeld, moeten logopedisten zelf beoordelen welke tabletgames bruikbaar zijn.
DOELEN: De doelen van de studie waren het onderzoeken welke kenmerken van een tabletgame de bruikbaarheid voor taaltherapie bepalen en het ontwikkelen van een betrouwbare en valide checklist waarmee deze kenmerken kunnen worden beoordeeld.
METHODE: De studie had een mixed methods design. De deelnemers waren 41 logopedisten die werken met kinderen en al tabletgames gebruiken tijdens taaltherapie. De checklist werd opgesteld naar aanleiding van de kwalitatieve data van 2 focusgroepbijeenkomsten over kenmerken die een tabletgame bruikbaar maken voor taaltherapie. In het kwantitatieve gedeelte van de studie werd de interbeoordelaarsbetrouwbaarheid bepaald door de ingevulde checklists van drie onafhankelijke logopedisten over 50 tabletgames te vergelijken. De constructvaliditeit werd bepaald met behulp van factoranalyse.
RESULTATEN: Er is een checklist ontwikkeld die bestaat uit 26 dichotome items, gebaseerd op zeven hoofdthema's en 21 subthema's die naar voren kwamen uit de kwalitatieve data. Veertien items hadden een zwakke overeenkomst tussen 3 onafhankelijke beoordelaars (Fleiss Kappa <0.20). Na het verwijderen van de onbetrouwbare items had de checklist een redelijke interbeoordelaarsbetrouwbaarheid (ICC 0,392). De factoranalyse resulteerde in drie onderliggende theoretische constructen: "geschiktheid voor therapie", "aantrekkelijkheid voor kinderen" en "praktische kenmerken".
CONCLUSIE: De checklist met 26 items heeft een goede inhouds- en indruksvaliditeit. Revisie van de checklist is nodig om de betrouwbaarheid van de checklist te verbeteren zodat deze bruikbaar wordt voor gebruik in de praktijk.
AANBEVELINGEN: De items van de checklist moeten worden verduidelijkt om verschillen in interpretatie te voorkomen. De antwoordschaal moet worden herzien zodat logopedisten hun antwoord kunnen nuanceren en er beter geschikte statistische methoden gebruikt kunnen worden om de betrouwbaarheid van de checklist te onderzoeken.

TREFWOORDEN: tablet games, taaltherapie, checklist, TOS.

Abstract

The Development of a Checklist To Assess Tablet Games on Usability for Language Therapy

BACKGROUND: More and more speech-language therapists (SLTs) are using tablet games (TGs) in the treatment of children with specific language impairment. Because most TGs are not specifically developed for language therapy, SLTs must select usable TGs themselves.
AIMS: The aims of the study were to determine which characteristics make a TG useful for language therapy and to develop a valid and reliable Dutch checklist to assess the usability of TGs for language therapy.
METHOD: The study had a mixed methods design. The participants consisted of 41 SLTs who work with children and use TGs in language therapy. The checklist was constructed based on qualitative data from two focus groups regarding the characteristics that make a TG useful for language therapy. In the quantitative part, inter-rater reliability was assessed by comparing the assessments from three independent raters of 50 TGs. Construct validity was determined by principal component analysis (PCA).
RESULTS: The study resulted in a checklist consisting of 26 dichotomous items, based on seven main themes and 21 subthemes derived from the qualitative data. Fourteen items showed poor agreement (Fleiss' kappa <0.20). After removing the unreliable items, the checklist had a fair inter-rater reliability (ICC 0.392). The PCA resulted in a three-factor solution. "Suitability for therapy", "attractiveness for the child" and "practical characteristics" constitute the underlying theoretical constructs of the checklist.
CONCLUSION: The checklist has good content validity and face validity. Revision is needed to improve the reliability and validity of the checklist to make it a useful instrument for SLTs.
RECOMMENDATIONS: The checklist items must be clarified to avoid differences in interpretation, and the answer scale must be changed to facilitate more nuanced ratings and to enable the use of other statistical methods to obtain information about the reliability of the revised checklist.

KEYWORDS: tablet games, language therapy, checklist, SLI

Appendix I Characteristics studied tablet games

Name	Quality rating on existing review platform	Purpose of the game	Age category (years)
Absurd free	Moderate	Therapy	4-6
Animal puzzles for toddlers	Moderate	Entertainment	2-4
App voor kinderen	Low	Education	2-4
App voor peuters	High	Education	2-4
Associations for kids	Moderate	Education	4-6
Bas op de kinderboerderij	Moderate	Education	2-4
Bitsboard	High	Education	4-6
Bobo zoeken	Low	Education	4-6
Brandweermannetjes	Moderate	Education	2-4
Build it up	Low	Education	2-4
De boerderij	Moderate	Entertainment	2-4
Dish puzzle	Low	Education	2-4
Doodle buddy	Low	Entertainment	4-6
Dr. Panda's restaurant	High	Entertainment	2-4
Electro voor iPad	Low	Education	4-6
Fiete	High	Entertainment	2-4
Geluiden van het leven lite	High	Education	4-6
iSequences lite	High	Education	4-6
Jop gaat eten	High	Education	4-6
Juf Jannie kinderboerderij	High	Education	4-6
Juf Jannie seizoenen	High	Education	4-6
Kenny	Low	Education	4-6
Kikker viert feest	Moderate	Entertainment	2-4
Lego duplo zoo	Low	Education	4-6
LEGO® Juniors Create & Cruise	Low	Entertainment	4-6
Lexico cognitie	Moderate	Education	4-6
Logo art Oops	Moderate	Therapy	4-6
Making sequences	Moderate	Education	4-6
Match it up 2	High	Education	2-4
Mijn lichaamsdelen ontdekken	High	Education	4-6
My playhome lite	High	Entertainment	4-6
My scene	High	Education	2-4
Nijntje apps	Moderate	Entertainment	2-4
Pepi bath lite	High	Education	2-4
Photomatch lite	Low	Entertainment	2-4
Put it away	Low	Education	4-6
Slaap lekker	High	Entertainment	2-4
Sort it out Therapy	High	Education	2-4
Story cubes	Moderate	Education	4-6
Suus & Luuk sinterklaas	High	Entertainment	4-6
Talking Tom	Moderate	Entertainment	2-4
Tap it too	Moderate	Entertainment	2-4
Timo en het toverstokje	Low	Entertainment	4-6
Tiny tap	Moderate	Education	2-4

Toca doctor HD lite	Low	Education	4-6
Video touch	High	Entertainment	2-4
What's diff 3	Moderate	Education	4-6
Zacht zijn de wolken	Moderate	Education	4-6
Zoek en vind lite	Moderate	Education	4-6
Zoekspel Gonnie & vriendjes	Moderate	Entertainment	2-4

Appendix II : Themes and sub-themes emerged from the analysis of the focus groups in Dutch (original language)

Hoofdthema	**Sub-thema('s)**
Gebruiksvriendelijkheid	Navigatie
	Besturing
	Taal
	Duidelijkheid
	Tempo
	Registratie resultaten
Aantrekkelijkheid	Kwaliteit afbeeldingen
	Kwaliteit geluiden en gesproken taal
	Feedback
	Beloning
	Inhoud van het spel
	Spelelement
	Overzichtelijkheid
Aandacht	Doelgericht werken
	Motivatie
Aanpasbaarheid	Tijdsinvestering die nodig is voor aanpassingen
Doel van de game	Multifunctionaliteit
	Geschikt als huiswerkopdracht
Kosten	Gratis demo versie beschikbaar
Risico's	Pop-up reclames
	In-app aankopen

Appendix III Checklist in Dutch (original language)

	Item	Score	
		JA	NEE
1	De game houdt de aandacht van het kind vast [b]		
2	De game biedt voldoende mogelijkheden voor interactie tussen het kind en de logopedist [b]		
3	Kinderen spelen de game graag [b]		
4	De game is aan te passen per kind, je kunt bijvoorbeeld eigen afbeeldingen invoegen, zelf woorden inspreken, het tempo of het niveau aanpassen		
5	De inhoud van de game (zoals aanpassingen, afbeeldingen, prestaties etc.) kan worden gedeeld met ouders		
6	De game is snel klaar voor gebruik [b]		
7	De game ziet er overzichtelijk uit [b]		
8	De game maakt gebruik van duidelijke, herkenbare afbeeldingen [b]		
9	De gebruikte stemmen en geluiden zijn helder en duidelijk		
10	Achtergrondmuziek kan worden uitgeschakeld		
11	De game voorziet in gerichte feedback, de game geeft aan of een kind iets goed of fout heeft gedaan		
12	De game biedt het kind de kans om gemaakte fouten te herstellen		
13	De game maakt gebruik van aantrekkelijke beloning		
14	De inhoud van de game biedt voldoende kansen om logopedische technieken toe te passen (vertraagd spreektempo, expanderen, refraseren, etc.) [b]		
15	De game biedt variatie in thema en spel [b]		
16	Er zit een motiverend spelelement in de game [b]		
17	De game is in te zetten voor verschillende logopedische doelen in taaltherapie		
18	De game is geschikt om te adviseren aan ouders als huiswerkopdracht [b]		
19	De game is gemakkelijk te bedienen [b]		
20	De game is in het Nederlands of er wordt geen taal gebruikt		
21	Het is meteen duidelijk hoe de game werkt		
22	Het tempo van de game is goed [b]		
23	Het is mogelijk om voortgang van meerdere kinderen te registreren [a]		
24	De game heeft een gratis demoversie		
25	De game kost niet meer dan 5 euro [b]		
26	De game is vrij van reclame pop-ups		
27	De game voorkomt dat een kind per ongeluk aankopen kan doen in een game [b]		
	Somscore [c]		
	Percentage “ja” [d]		
	Sterscore [e]		

[a] Item verwijderd i.v.m. weging belangrijkheid

[b] Interbeoordelaarsbetrouwbaarheid onvoldoende (Fleiss’ Kappa <0.20)

[c] aantal items beantwoord met “ja”

[d] somscore / 26 * 100

[e] 0-20% “ja” = 1 ster; 20-40% “ja” = 2 sterren; 40-60% “ja” = 3 sterren; 60-80% “ja” = 4 sterren; 80-100% “ja” = 5 sterren

Appendix IV Checklist in English (translation only for research report, not for use in practice)

	Item	Score	
		YES	NO
1	The game holds the child's attention [b]		
2	The game offers enough possibilities for interaction between the child and the speech therapist [b]		
3	Children like to play the game [b]		
4	The game can be adapted to the child's needs: images can be added , the speaker's voice can be recorded, speed and level can be adjusted		
5	The game's content (such as adjustments, images, scores, etc.) can be shared with parents		
6	The game is ready to use [b]		
7	The game is clearly structured [b]		
8	The game uses plain, recognisable images [b]		
9	The voices and sounds used are clear and distinct		
10	Any background music can be switched off		
11	The game gives practical feedback, clearly indicating if the child has made any mistakes		
12	The game gives the child the opportunity to correct any mistakes		
13	The game offers attractive rewards		
14	The game's content gives ample opportunity to use speech therapy techniques (slowing down the pace of speech, expanding, rephrasing, etc.) [b]		
15	The game offers a variety of themes and play tasks [b]		
16	There is a motivating element of play in the game [b]		
17	The game can be used to pursue various speech therapy objectives in language therapy		
18	The game can be recommended to parents as a homework task [b]		
19	The game is easy to control [b]		
20	The game is either in Dutch or no language is used		
21	It is immediately apparent how the game works		
22	The speed of the game is good [b]		
23	It is possible to mark the progress made by several children [a]		
24	The game has a demo version free of charge		
25	The game costs no more than € 5,00 [b]		
26	The game is free from advertising pop-ups		
27	The game ensures that it is impossible for the child to accidentally buy something during a game [b]		
	Total score [c]		
	Percentage "yes" [d]		
	Starscore [e]		

[a] Item removed because of mean rating importance

[b] Poor agreement (Fleiss' Kappa <0.20)

[c] Items answered with "yes"

[d] Total score/ 26 * 100

[e] 0-20% "yes" = 1 star; 20-40% "yes" = 2 stars; 40-60% "yes" = 3 stars; 60-80% "yes" = 4 stars; 80-100% "yes" = 5 stars

Appendix V Recommendations for revision

General recommendations

In general, it seems advisable to clarify the items with narrative descriptions and examples. To improve the items, it would be useful to conduct a user consultation with SLTs to study the interpretation of the items and the cause of the poor agreement. User consultation also improves the feasibility of an instrument and facilitates implementation in the future.

Answer scale

The answer scale of the checklist is in need of revision. An answer scale based on an interval or ratio scale, for example a 5 point Likert scale or a score between 1 and 10, seems to be more appropriate because this would facilitate a more nuanced rating and provide opportunities to use other statistical methods to assess reliability. If an interval or ratio scale were used, it would be possible to assess the inter-rater reliability of the items using ICC. ICC measures inter-rater reliability, whereas kappa measures agreement between raters. ICC is a more appropriate method to measure the inter-rater reliability of the checklist items.

Recommendations for specific items

1 *The game holds the child's attention.*
The item had poor agreement (Fleiss' kappa <0.20). Discussion with a group of SLTs about the topic of "attention" and the formulation of the item would be useful to improve the understandability and therefore the reliability of the item. A narrative description to define "holding attention" may be necessary to avoid differences in interpretation.

2 *The game offers enough possibilities for interaction between the child and the speech therapist.*
The item had poor agreement (Fleiss' kappa <0.20). Discussion about the formulation of the item seems to be necessary. A narrative description to explain "possibilities for interaction" and defining "enough" might also improve the reliability.

3 *Children like to play the game.*
The item had poor agreement (Fleiss' kappa <0.20). A narrative description to explain which characteristics an SLT should observe to know that the child likes the game is a possible way of improving this item. Perhaps reformulating the item and, especially, asking for specific characteristics (for example whether the child asks to play the game again) would help make the item more objective.

4 *The game can be adapted to the child's needs: images can be added, the speaker's voice can be recorded, speed and level can be adjusted.*

No revision needed.

5 *The game's content (such as adjustments, images, scores, etc.) can be shared with parents.*

No revision needed.

6 *The game is ready to use.*

The item had poor agreement (Fleiss' kappa <0.20). A narrative description to define "ready to use" is needed. Perhaps it would be better to ask for more directly observable characteristics, for example "the game loads in … seconds" or "any introductory stories can be skipped". It is advisable to discuss this with users.

7 *The game is clearly structured.*

The item had poor agreement (Fleiss' kappa <0.20). Discussion about the formulation of the item seems to be necessary. A narrative description to define "clearly structured" might also help improve reliability.

8 *The game uses plain, recognisable images.*

The item had poor agreement (Fleiss' kappa <0.20). Discussion with users about the definition of "plain and recognizable" is recommended. The poor agreement on this item may have been caused by differences in taste. Different SLTs might simply prefer different types of images. If this is the case, it would be possible to improve the item by defining, in cooperation with users, which type of image is the best for language therapy.

9 *The voices and sounds used are clear and distinct.*

No revision needed.

10 *Any background music can be switched off.*

No revision needed.

11 *The game gives practical feedback, clearly indicating if the child has made any mistakes.*

No revision needed.

12 *The game gives the child the opportunity to correct any mistakes.*

No revision needed.

13 *The game offers attractive rewards.*

The item had poor agreement (Fleiss' kappa <0.20). Defining "attractive rewards" with a narrative description would possibly improve the reliability of the item. In this case as well, the tastes of users can play a role in rating the item. Not every SLT likes the same types of rewards. Asking for more objective characteristics may yield better results, for example by simply asking whether the tablet game provides any rewards.

14 *The game's content gives ample opportunity to use speech therapy techniques (slowing down the pace of speech, expanding, rephrasing, etc.).*

The item had poor agreement (Fleiss' kappa <0.20). For this item, it would likely be useful to provide examples of how to use speech therapy techniques when using a tablet game.

15 *The game offers a variety of themes and play tasks.*
The item had poor agreement (Fleiss' kappa <0.20). Examples of themes and play tasks might improve the reliability of this item.

16 *There is a motivating element of play in the game.*
The item had poor agreement (Fleiss' kappa <0.20). A narrative description of a "motivating element of play" would be useful. It is possible that not all users understand what is meant by this term.

17 *The game can be used to pursue various speech therapy objectives in language therapy.*
No revision needed.

18 *The game can be recommended to parents as a homework task.*
The item had poor agreement (Fleiss' kappa <0.20). In the focus groups, SLTs mentioned that a disadvantage of working with games was that they cannot recommend the exercise to parents as homework. The poor agreement might indicate that it is not clear enough which characteristics of a game make it useful as a homework task. It would be useful to discuss this further with users of the checklist.

19 *The game is easy to control.*
The item had poor agreement (Fleiss' kappa <0.20). The aim of this item was to determine how the game is controlled and whether this is easy enough for children. Working with some games requires advanced fine motor skills. The item can be improved by questioning more objective characteristics, for example the size of the buttons or difficulty of specific actions.

20 *The game is either in Dutch or no language is used.*
No revision needed.

21 *It is immediately apparent how the game works.*
No revision needed.

22 *The speed of the game is good.*
The item had poor agreement (Fleiss' kappa <0.20). The item can possibly be improved by defining "good speed". Because this is difficult to define, more objective characteristics about speed can be questioned, for example "the game offers the opportunity to think about an answer" or "the game offers the opportunity to talk about a task before a reward follows or a new task starts".

23 *It is possible to mark the progress made by several children.*

This item was removed from the checklist because it was rated as unimportant during the study for weighting the items. This mean rating does not agree with the qualitative data. In the focus groups, the recording of progress was defined as an important characteristic. Because of this difference, it is advisable to discuss this theme again with a group of SLTs.

24 *The game has a demo version free of charge.*
No revision needed.

25 *The game costs no more than € 5.00.*
The item had poor agreement (Fleiss' kappa <0.20). This item is an objective item, and therefore it is notable that the agreement is low. A possible explanation is that the SLTs were confused by different versions of the tablet game. In most cases, a tablet game has a free demo version and a paid full version. To avoid this confusion, the item should be reworded to clarify that it refers to the price of the full version.

26 *The game is free from advertising pop-ups.*
No revision needed.

27 *The game ensures that it is impossible for the child to accidentally buy something during a game.*
The item had poor agreement (Fleiss' kappa <0.20). This item is an objective item. An explanation of in-app purchases might make the item more clear and improve its reliability.

Representation constructs

In the PCA of the current study, only the twelve reliable items were included. This results in a loss of data and therefore some underlying theoretical constructs may be neglected. Looking at the qualitative data and the items not included in PCA, the topic of "attention" seems to be insufficiently represented.

After revision of the checklist, a new PCA with more items can be conducted to explore whether there are additional underlying constructs to be measured when assessing the usability of tablet games for language therapy.