International Journal of Medical Informatics 129 (2019) 366–373

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Eye-tracking retrospective think-aloud as a novel approach for a usability evaluation

Hwayoung Cho[a,*], Dakota Powell[b], Adrienne Pichon[b], Lisa M. Kuhns[c,d], Robert Garofalo[c,d], Rebecca Schnall[b]

[a] College of Nursing, University of Florida, Gainesville, FL, United States
[b] School of Nursing, Columbia University, New York, NY, United States
[c] Division of Adolescent Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, United States
[d] Department of Pediatrics, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

ARTICLE INFO

Keywords:
Eye movement measurements
Eye movements
Eye-tracking
Mobile applications
Mobile health
Information technology
Health IT
Usability evaluation

ABSTRACT

Objective: To report on the use of an eye-tracking retrospective think-aloud for usability evaluation and to describe its application in assessing the usability of a mobile health app.

Materials and Methods: We used an eye-tracking retrospective think-aloud to evaluate the usability of an HIV prevention mobile app among 20 young men (15–18 years) in New York City, NY; Birmingham, AL; and Chicago, IL. Task performance metrics, critical errors, a task completion rate per participant, and a task completion rate per task, were measured. Eye-tracking metrics including fixation, saccades, time to first fixation, time spent, and revisits were measured and compared among participants with/without a critical error.

Results: Using task performance analysis, we identified 19 critical errors on four activities, and of those, two activities had a task completion rate of less than 78%. To better understand these usability issues, we thoroughly analyzed participants' corresponding eye movements and verbal comments using an in-depth problem analysis. In areas of interest created for the activity with critical usability problems, there were significant differences in time spent ($p = 0.008$), revisits ($p = 0.004$), and total numbers of fixations ($p = 0.007$) by participants with/without a critical error. The overall mean score of perceived usability rated by the Health IT Usability Evaluation Scale was 4.64 ($SD = 0.33$), reflecting strong usability of the app.

Discussion and Conclusion: An eye-tracking retrospective think-aloud enabled us to identify critical usability problems as well as gain an in-depth understanding of the usability issues related to interactions between end-users and the app. Findings from this study highlight the utility of an eye-tracking retrospective think-aloud in consumer health usability evaluation research.

1. Introduction

With the rapid expansion of mobile technology in healthcare [1], it is crucial to ensure that mobile health (mHealth) technologies are usable [2]. Usability is a measure of the quality of an end-user's experience when interacting with the technology [3]. Usability factors are closely linked to the success or failure of the technology as usability is related to the quality in use of the technology [4]. The 'quality in use' is the capability of the software product to enable specified users to achieve specified goals with effectiveness, productivity, safety and satisfaction in specified contexts of use [5,6]. To ensure quality in use of the technology, it is important to assess its usability during system development, which helps ensure that the system meets the needs of end-users [2,7,8].

In order to successfully achieve the goals of the system, it is critical to choose the most appropriate evaluation techniques which best meet the study aims during the system development process [9]. Usability evaluation methods are broadly classified as expert-based usability testing methods such as a heuristic evaluation and a cognitive walkthrough and end-user-based usability testing methods such as a think-aloud protocol, field observation, interview, focus group, and questionnaire [9–11]. With a particular focus on usability testing with intended end-users in this paper, traditional usability testing most commonly uses a think-aloud protocol [10,12]. Think-aloud protocols are used to identify the cognitive behavior of performing tasks while using technology and determine how that information is used to facilitate

* Corresponding author at: University of Florida College of Nursing, 1225 Center Drive, PO Box 100197, Gaineville, FL 32610-0197, United States
E-mail address: hcho@ufl.edu (H. Cho).

https://doi.org/10.1016/j.ijmedinf.2019.07.010
Received 27 February 2019; Received in revised form 9 July 2019; Accepted 11 July 2019

problem resolution [10,13]. Think-aloud protocols are generally categorized into concurrent and retrospective protocols. In a concurrent think-aloud protocol, users are asked to think and talk aloud at the same time while performing cognitive tasks; in a retrospective think-aloud protocol, users are asked to recall what they were thinking during a prior experience. Both concurrent and retrospective think-aloud protocols are popular approaches since they provide comprehensive insights into the problems that end-users encounter in their interaction with the system [14]. However, there are several limitations of the think-aloud protocol. The qualitative information provided by end-users are unstructured, and there are often gaps of silence where the end-users are thinking but not verbalizing, and as a result, some data collection is limited at those time [15]. Specific to adolescents, studies report that this age group is less likely to articulate their thought processes during a think-aloud protocol [16,17]. Findings from our past work suggest that a traditional think-aloud protocol to assess the usability of technology with adolescents may not provide sufficient information to identify usability problems [15,18].

To address this gap, eye-tracking technology can be used to assess usability of new technologies by illuminating the decision-making through the examination of eye movement patterns [19–21]. Eye-tracking is the process of measuring the point of gaze and/or the motion of an eye relative to the head, which has the potential to improve usability assessments by providing valuable ocular data. However, there is a paucity of research on how the eye-tracking method can be applied in usability testing of mHealth technology as a single rigorous usability evaluation method by achieving its full potential [22]. Prior use of eye-tracking has not standardized the use of this data making interpretation of eye-tracking data difficult [20–23]. The purpose of this paper is to report a novel methodological approach of an eye-tracking retrospective think-aloud for usability evaluation, and to describe its application in assessing the usability of a mHealth app.

1.1. Study context

This study was conducted as part of a larger study to adapt a group-based theory-driven, manualized HIV prevention curriculum for diverse sexual minority adolescents [24]. We adapted an evidence-based, group-level, face-to-face HIV prevention curriculum onto a mobile platform using an iterative design process [25–27]. The mobile app, the Male Youth Pursuing Education, Empowerment & Prevention around Sexuality (MyPEEPS App), delivers HIV prevention information through 21 activities which are comprised of: didactic content, graphical reports, videos, and true/false and multiple-choice quizzes. Upon completing each activity, users are rewarded with a stylized trophy, which is used to promote continued use of the app. A combination of usability evaluation techniques including usability experts as well as intended end-users is recommended [28,29]; therefore, we assessed the usability of the MyPEEPS App from both expert and end-user perspectives [30]. In this paper, we focused on the end-user usability testing utilizing an eye-tracking retrospective think-aloud.

2. Methods

We conducted an eye-tracking retrospective think-aloud to evaluate the usability of the MyPEEPS App. The Institutional Review Board of Columbia University Medical Center served as the central IRB (#AAAQ6500) for this study and approved all research activities.

2.1. Sample

Participants were recruited using flyers, posting on social media, and direct outreach at local community-based organizations in New York City, NY; Birmingham, AL; and Chicago, IL. Our sample was comprised of 20 young men since 95% of usability issues are identified with 20 end-users [31]. Inclusion criteria were: 1) 13 to 18 years of age; 2) self-identified as male; 3) male sex assigned at birth; 4) understand and read English; 5) living within the metropolitan area of one of the three cities listed above; 6) ownership of a smartphone; 7) sexual interest in men; and 8) self-reported HIV-negative or unknown status. Participants who wore bifocal/progressive glasses or who experienced eye surgery (e.g., corneal, cataract, intraocular implants) were excluded from participation since these types of glasses or eye conditions affect the precision of the gaze estimation while collecting participants' eye movements [32].

2.2. Procedures

We explained the purpose of the study and study procedures to the participants who were then asked to sign an informed consent (18 years old) / assent (13–17 years old) form. Participants were asked to sit down at a desk. The eye tracker (i.e., Tobii X2-30) was calibrated with a nine-point system where the participant watched a circle move across the screen and paused at each of nine fixed points. With the moving calibration test, the measurement accuracy was provided within 0.5 degrees providing an error of less than 0.5 cm between measured and intended gaze points [32]. The resolution of the computer monitor was set to 1920*1080 pixels.

First, participants were provided with use case scenarios of the MyPEEPS App and asked to complete the tasks using the app on an iOS simulator utilizing a Windows desktop computer. The first half of participants were provided with use case scenario, version 1; the remaining half of the participants were provided with use case scenario, version 2. Two versions of use case scenarios were used in order to capture representative tasks of the app (e.g., comics, animated videos, true/false questions, and multiple-choice quizzes). Activities which were necessary to navigate the app (e.g., log-in/out, set-up of profile) and those activities which were difficult for the first ten participants to complete were included in use case scenario version 2. The tasks associated with each of the use case scenarios are presented in Table 1. iMotions software was used to record participant's eye movements and the computer screen while performing each task [33], which allows researchers to present app screen recordings and synchronized eye-tracking data simultaneously.

Participants were allowed to ask questions before starting the app testing, but once testing began, we encouraged participants to complete all tasks by themselves. Participants were instructed not to turn to the researcher for assistance because a shift in visual focus increases the risk of losing eye-tracking data [22]. If participants had trouble and were unable to proceed, they were instructed to say 'HELP'.

Following use of the app, participants were asked to describe their experience dealing with errors and their perception of their overall performance. Then participants viewed the recordings of their use of the app which depicted their eye movements overlaid on the app screen on a computer. Participants were asked to think-aloud and verbalize their thoughts about the tasks they completed and the difficulties they encountered while using the app. Participant's verbal comments were audio-recorded.

Following the testing of the app, participants were asked to rate usability of the MyPEEPS App using the Health Information Technology Usability Evaluation Scale (Health-ITUES). [34] Participants were compensated $40-50, depending on the geographic site, for their time.

2.3. Data collection

Eye-tracking data were collected using Tobii X2-30 [35], which has a sampling rate of 30 Hz (i.e., 30 gaze points were collected per second for each eye), and saved into iMotions software [33]. Table 2 lists the task performance metrics collected to capture usability problems by examining how capable participants were at using the MyPEEPS App on given tasks (i.e., a task completion rate was calculated in two ways: by participant and by task) [4,36], and the eye-tracking metrics collected

Table 1
Task included in use case scenarios.

Task	Use case scenario - version	
Log-in to the MyPEEPS App	I	II
Collect the trophy from activity #1 Set Up MyPEEPS Profile	I	II
[Activity: Set-up]		
Introduction to the app explaining what the user is to expect. User inputs name, telephone number, e-mail address, and how they prefer to get notifications.		
Collect the trophy from activity #2 BottomLine	I	II
[Activity: Select from options]		
Users are asked the farthest they will go with a one-time hookup in a number of sexual scenarios and given a selection of responses about what they will and won't do and how they will do it.		
Collect the trophy from activity #3 Underwear Personality Quiz	I	
[Activity: Sliders]		
Users complete a personality quiz and are introduced to the avatars that they will be seeing in the app. Avatars' personality traits and identities are shared with 'gossip'.		
Collect the trophy from activity #4 My Bulls-I	I	
[Activity: Text input] Users are asked to think about their important identity traits and create a list of their top five identity traits after seeing an example of the activity done by one of the app avatars, P.		
Collect the trophy from activity #5 P's On-Again Off-Again BottomLine	I	
[Activity: Video, select from options]		
Video of a text conversation between two avatars, P and Nico, about P's new relationship and P ignoring his BottomLine. Users are asked to complete questions about why P should be concerned about his BottomLine with a new partner. There are two videos with two sets of questions.		
Collect the trophy from activity #6 Sexy Settings	I	
[Activity: Select from options]		
Users are presented with a setting in which sex could be taking place and are given one potential threat to a BottomLine and asked to select another potential threat for the given setting.		
Collect the trophy from activity #7 Goin' Downhill Fast	I	
[Activity: Click through information, select from options]		
Users are presented with information about drugs and alcohol and how they can affect a BottomLine. Resources for additional information about drugs/alcohol are provided. After reading through the information, users complete a set of questions about drugs/alcohol's potential impact on their BottomLine.		
Collect the trophy from activity #8 Step Up, Step Back	I	
[Activity: Select from options]		
Users are introduced to identity traits that may identify them as a VIP (privileged)/Non-VIP (non-privileged) and then asked a series of identity-related questions. An avatar representing the user moves back and forth in a line for a night club, relative to the avatars in the app, as questions are answered.		
Collect the trophy from activity #9 HIV True/False	I	
[Activity: True/False button answer]		
Users complete a series of True/False questions related to HIV, with information following a correct answer.		
Collect the trophy from activity #10 Checking In On Your BottomLine	I	
[Activity: Select from options]		
Users are given the opportunity to review and make changes to their BottomLine, taking into consideration any information that they may have learned from completing the activities prior to this check-in.		
Collect the trophy from activity #13 Well Hung	I	
[Activity: Drag and drop]		
Users are introduced to the association of HIV transmission risk with different sexual behaviors categorized into no risk, low, medium, and high risk. Users complete an activity dragging and dropping a given sexual activity onto the risk category associated with the sex act.		
Collect the trophy from activity #15 Checking In On Your BottomLine Again		II
[Activity: Select from options]		
Users are again given the opportunity to review and make changes to their BottomLine, taking into consideration any information that they may have learned from completing the activities prior to this check-in.		
Collect the trophy from activity #17 4 Ways To Manage Stigma		II
[Activity: Click through, select from options] Users are presented with four stigma management strategies, then a scene for each of the four app avatars and asked to answer which strategy each character is using in the scene.		
Collect the trophy from activity #18 Rubber Mishap	I	
[Activity: Shaking select from options]		
Users are asked to complete a series of questions relating to condom usage as the screen shakes to mimic being under the influence of drugs/alcohol.		
Collect the trophy from activity #19 Get a Clue!		II
[Activity: Shake device situation builder]		
Jumbled scenarios are created using either a shake of the phone or press of a button. Users answer from given options how they would act in the scenario, keeping the BottomLine and communication strategies in mind.		
Collect the trophy from activity #20 Last Time Checking In On Your BottomLine		II
[Activity: Select from options]		
Users are again given the opportunity to review and make changes to their BottomLine, taking into consideration any information that they may have learned from completing the activities prior to this check-in.		
Collect the trophy from activity #21 BottomLine Overview		II
[Activity: View list of changes]		
Users are presented with a list of their BottomLine selections since the initial activity and subsequent check-ins.		
View settings	I	II
Log Out	I	II

for an in-depth analysis of usability problems.

All survey data were collected electronically using Qualtrics® survey software [39]. Demographics and mobile technology use was assessed through (our research team-designed) questions on age, race, ethnicity, frequency of using mobile devices or laptop/desktop to access the Internet, and duration of using mobile apps on a smartphone. Data on perceived usability were collected using the Health-ITUES [34], a customizable questionnaire with a four-factor structure: system impact, perceived usefulness, perceived ease of use, and user control, and it has been validated for use with mHealth technology [40]. The Health-ITUES consists of 20 items rated on a five-point Likert scale from strongly disagree (1) to strongly agree (5). A higher scale value indicates higher perceived usability of the technology. Table 3 lists the 20 items on the Health-ITUES and how they were customized for this study.

2.4. Data analysis

Data analysis was based on the iMotions video-recordings of user sessions synchronized with eye movements, and transcriptions of participants' verbal comments from the audio-recordings collected during the think-aloud. Two research team members reviewed the transcripts to identify common usability concerns, then a third reviewer consulted in instances of discrepancy. STATA SE 14 was used for analysis of descriptive statistics [41].

Data analysis focused on: 1) *task performance analysis* of task performance metrics, and 2) *problem analysis* of eye-tracking metrics and participants' verbal comments. Since the average task completion rate in the literature (i.e., an analysis of nearly 1200 usability tasks) is 78% [42], any task with less than 78% of a task completion rate was identified as a problem. In the problem analysis, the eye-tracking metrics including time to first fixation, time spent, revisits, and total numbers of fixations were compared among participants with/without a critical error using a two-sample t-test. Level of significant was set as alpha less than 0.05.

3. Results

3.1. Sample

The mean age of study participants was 17.4 years (SD = 0.88;

Table 2
Task performance and eye-tracking metrics.

Task performance metrics	
Critical error	Number of critical errors (e.g., if a participant said 'HELP' during the app testing, it was considered a critical error in this study).
Task completion rate per participant	Percentage of tasks that were completed without a critical error by a participant.
Task completion rate per task	Percentage of participants who completed a given task without a critical error.
Eye-tracking metrics	
Fixation	Moments when the eyes are relatively stationary, indicating the moments when the brain is processing information received by the eyes. The fixation generally ranges from 100 to 300 milliseconds. Longer fixations on a specific area reflect a participant's difficulty with information processing [22, 37].
Saccades	Rapid eye movement from one target to another between two consecutive fixations [38].
Time to first fixation	Amount of time it took a participant to look at a specific area from stimulus onset [37].
Time spent	Amount of time that a participant spent looking at a specific area.
Revisit	Number of times that a participant repeatedly viewed a specific area.

range = 15–18 years of age). 45% (N = 9) of participants self-identified as White, 20% (N = 4) as African American, 10% (N = 2) as Asian, and 45% (N = 9) of participants self-identified as Hispanic. 85% of participants (N = 17) reported using Internet almost constantly every day. The majority of participants (85%) reported using mobile devices as opposed to using laptop/desktop (15%) to access the Internet. The mean duration of participants' use of mobile apps on a smartphone per day was 9.40 h (SD = 5.52).

3.2. Eye-Tracking retrospective think-aloud

The visit took between 2 and 2.5 h. Before watching the recordings displaying their eye movements, participants described their experience dealing with errors and their perception of their task performance. More than half of participants who had difficulty completing tasks (e.g., participants who said 'HELP' during the app testing) stated, '*Everything was okay*', '*It was pretty easy*', or '*I didn't have any difficulties*' until they viewed their eye movements on an app screen page where they encountered the difficulty.

3.2.1. Task performance analysis

3.2.1.1. Critical error. A total of 19 critical errors were identified across four activities: *#2 BottomLine, #5 P's On-Again Off-Again BottomLine, #8 Step Up, Step Back, and #13 Well Hung*. The number of critical errors for the activities is presented in Table 4.

Table 4
Critical error.

Activity	Critical errors
#2 BottomLine	6
#5 P's On-Again Off-Again BottomLine	1
#8 Step Up, Step Back	1
#13 Well Hung	11
Total number of critical errors	19

3.2.1.2. Task completion rate per participant. The percentage of tasks that were completed without a critical error by a participant ranged from 79% to 100%. Six participants successfully completed tasks without any critical error.

3.2.1.3. Task completion rate per task. The percentage of participants who completed each task without a critical error ranged from 45% to 100%. Two tasks had a task completion rate less than 78% [42]; in our study, the tasks related to the activities *#2 BottomLine* 70% and *#13 Well Hung* 45%.

3.2.1.4. Summary of task performance analysis. There were two activities with critical errors, which were identified through task performance analysis: *#5 P's On-Again Off-Again BottomLine* and *#8 Step Up, Step Back*. These two activities were reported by participants as a user error (e.g., they closed the app screen by mistake while they were

Table 3
Health-ITUES (customized for this study).

System Impact	
1	MyPEEPS is a positive addition to my sexual health.
2	MyPEEPS helps me make safe decisions when it comes to sex and relationships.
3	MyPEEPS gives me the information and skills I need to avoid situations that make me uncomfortable and that put my sexual health at risk from HIV or other STIs.
Perceived Usefulness	
4	Using MyPEEPS makes it easier to make safer decisions about my sexual health.
5	Using MyPEEPS allows me to make safer decisions about my sexual health more quickly.
6	Using MyPEEPS makes me more likely to make safer decisions about my sexual health.
7	MyPEEPS is useful for making safer decisions about my sexual health.
8	I think MyPEEPS presents a more open-minded process for learning about my sexual health.
9	I am satisfied with MyPEEPS for making safer decisions about my sexual health.
10	I make safer decisions about my sexual health in a timely manner because of MyPEEPS.
11	Using MyPEEPS lowers my risk of getting HIV.
12	I am able to find the information I need about sexual health and HIV whenever I use MyPEEPS.
Perceived Ease of Use	
13	I am comfortable with my ability to use MyPEEPS.
14	Learning to operate MyPEEPS is easy for me.
15	I have the skills to use MyPEEPS.
16	I find MyPEEPS easy to use.
17	I remember how to log on to and use MyPEEPS.
User Control	
18	MyPEEPS gives error messages that clearly tell me how to fix problems.
19	Whenever I make a mistake using MyPEEPS, I recover easily and quickly.
20	The information (such as on-line help, on-screen messages and other documentation) provided with MyPEEPS is clear.

reading contents), and reviewed/determined as non-usability-related problems by two research team members and were excluded from problem analysis. There were also two activities with critical errors, which were identified with a task completion rate less than 78% via the task performance analysis: *#2 BottomLine* and *#13 Well Hung*. These two activities were thoroughly reviewed and analyzed using eye-tracking data and participants' verbal comments, and included in the problem analysis.

3.2.2. Problem analysis

3.2.2.1. Problem 1. #2 BottomLine; navigating the map after completing the prior activity #1. Task description: Within the MyPEEPS App, a total of 21 activities are displayed along a virtual 'Map'. One activity at a time on a smartphone screen is shown in consecutive order on the Map. User begins each activity by clicking the activity's number in a circle or name in a box. Upon completing each activity, the user is taken back to the Map showing the activity's number and name the user just completed. In order to navigate to the next activity, the user needs to scroll or swipe to the left on the Map.

Problem description: Participants were confused about moving forward to the second activity #2 *BottomLine* on the Map after completing the very first activity *#1 Set Up MyPEEPS Profile* since they expected to view the next activity by default.

> Quotations: *"This is the part where I was confused. I didn't understand that I should move to the side. I didn't know. I kept clicking number one because I thought that was where I had to go and then it would just take me back… there should be instructions or something or like a hint… like arrows."* [UMP07]
> *"I am trying to figure out how to… I think that would be really helpful if it just went automatically over. I meant I want to see the next one automatically right after I completed the previous one. Otherwise, I cannot remember if I did or not."* [UMP03]

Gaze plots: Based on participants' fixations and saccades, gaze plots depicting fixation sequences were generated in conjunction with *Problem 1*. The gaze plots were compared among participants with/without a critical error. The number of fixations on *Problem 1* ranged from 19 (without a critical error) to 200+ (with a critical error). A sample of gaze plots with/without a critical error is presented in Fig. 1(1) and (2).

3.2.2.2. Problem 2. #13 well hung; Drag/drop response option. Task description: User is introduced to the association of HIV transmission risk with different sexual behaviors categorized into 'no risk', 'low', 'medium', and 'high risk'. The user completes an activity (i.e., quizzes) dragging and dropping a given sexual activity onto the four risk categories associated with the sex act. For instance, user drags a card labeled with the sexual activity down to the corresponding risk level, then selects the 'Next' button to continue in the activity. In order to see the 'Next' button, the user needs to scroll down.

Problem description: Participants were confused about the drag/drop response option on the quizzes. Several participants tried to figure it out in a way of either of dragging the sexual activity card down to the risk category or dragging the risk category up to the sexual activity card since there was no feedback on whether their selected response was correct or incorrect unless they clicked the 'Next' button.

> Quotations: *"Didn't I have to drag something? That's what was confusing. I felt it should have just been a click. Then, even I didn't know there was next button at the bottom. I couldn't move forward."* [UMP005]
> *"I didn't know what to do. I figured it out but I didn't know if I had to click it or drag it. I don't know… I was expecting it to be like an empty line. I was expecting it just to be a line, empty, and then I would drag the answers into the clips. I was expecting the clothing line clips to be empty because you see how there are four and it says high, medium, low, or no risk… I was expecting it to be like a spectrum and I would drag the answers into the line depending where they fell."* [UMP07]

Heat maps: Heat maps are static aggregations of gaze fixations revealing the distribution of visual attention, which represent where participants concentrated their gaze and how long they gazed at a given point in different colors [22,43]. Red areas on a heat map reflect a high number of gaze fixations, while yellow and green areas indicate fewer gaze fixations. The heat maps were compared among app pages with/without a critical error. For instance, a participant who successfully completed this task without difficulty would see the first given sexual activity card, drag the card down to a correct (risk level) answer 'lower', and then click the 'Next' button. While several participants had these difficulties on the first page of the quizzes, no one had the difficulties on the remaining pages. For the reason, the heat maps for every page within the activity *#13 Well Hung* were compared with the first page. Fig. 2(1) depicts a heat map of the first page without a critical error, while Fig. 2(2) depicts that of the page with a critical error.

Areas of interest: Areas of interest refer to specific areas in the

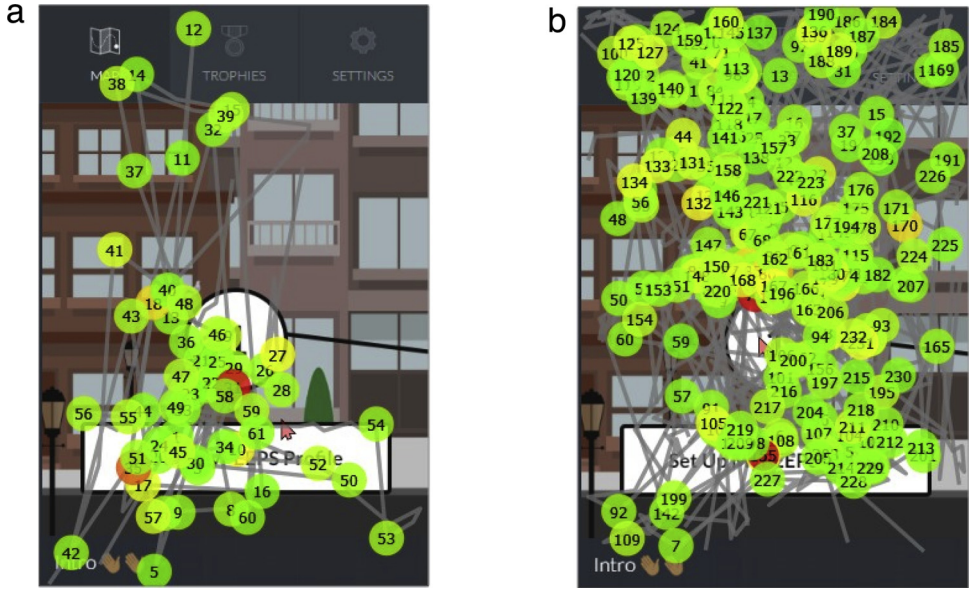

Fig. 1. (1) Gaze plot without a critical error. (2) Gaze plot with a critical error.

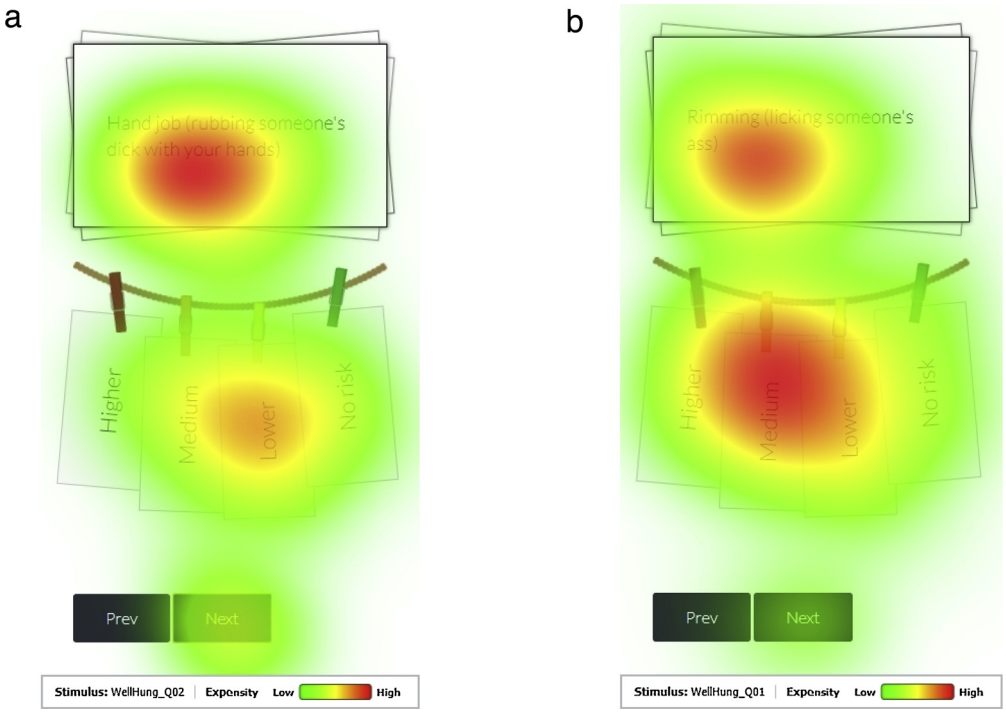

Fig. 2. (1) Heat map without a critical error. (2) Heat map with a critical error.

interface that are of interest to researchers [20]. Given that a participant with a critical error on the first page of quizzes did not experience any critical error on the remaining quiz pages, a total of eight areas of interest on the first page of the quizzes were created to compare time to first fixation, time spent, revisit, and total number of fixations among participants with/without a critical error. On an area of interest, 'lower' (i.e., the card corresponding to the correct answer for the first quiz), there were significant differences in time spent ($p = 0.008$), revisits ($p = 0.004$), and total number of fixations ($p = 0.007$) by participants with/without a critical error.

3.2.3. Perceived usability

Perceived usability was rated using the Health-ITUES (Table 5) [34]. The overall Health-ITUES score was the mean of all the items with each item weighted equally, and in this study it was 4.64 ($SD = 0.33$) with a range of 3.80–5.00, reflecting strong usability of the app.

4. Discussion

In this study we successfully used an eye-tracking retrospective think-aloud to conduct a comprehensive usability evaluation of a mHealth app with adolescents. We identified two critical usability problems through participants' eye movements and verbal comments. Eye-tracking data and qualitative data were integrated to provide a more holistic understanding of the usability issues with the mHealth app. Our methodological approach is briefly compared with a traditional stand-alone usability testing method, a think-aloud only used in literature in Table 6 [11,44].

Table 5
Health-ITUES scores.

Health-ITUES Construct	Mean (SD)	Median (range)
System Impact	4.80 (0.38)	5.00 (3.67-5.00)
Perceived Usefulness	4.63 (0.40)	4.78 (3.56-5.00)
Perceived Ease of Use	4.76 (0.33)	4.90 (4.00-5.00)
User Control	4.28 (0.74)	4.33 (2.00-5.00)
Overall Health-ITUES Score	4.64 (0.33)	4.75 (3.80-5.00)

*Rating score from 1- worst to 5-best (20 items)

Our findings demonstrate the usefulness of an eye-tracking retrospective think-aloud for usability evaluations. While a concurrent think-aloud is the predominant data collection method for traditional usability testing [45], it suffers from some notable shortcomings such as distractions of end-users' attention or negative effects on their natural task performance when incorporated with other techniques/technologies [22,46–49]. In contrast, an eye-tracking retrospective think-aloud allowed participants to avoid interferences during the usability evaluation [22,48,49], which is an especially important strength of our methodological approach.

Findings from this study showed that use of an eye-tracking retrospective think-aloud for a usability evaluation allows participants to share their real-time experience using the app and stimulates verbal expression of their experience using the app which is more difficult to achieve using traditional stand-alone usability testing [14,15]. Our study participants had difficulty successfully completing some tasks but could only explain these issues at a descriptive level until the participants reviewed a recording of their task performance depicting their eye movements overlaid on the app screen on a computer. For example, we asked participants to describe their experience dealing with errors and their perception of their task performance right after they had completed tasks. Despite the difficulties our participants encountered, more than half of the participants briefly commented, *'Everything was okay'*, *'It was pretty easy'*, or *'I didn't have any difficulties'*. On the other hand, while watching the screen-recording presenting participants' unusual eye movements, the participants expressed difficulties during the app testing. This suggests that traditional stand-alone usability testing among youth may underestimate problems, reflecting social desirability among some adolescents. By showing the screen-recordings with the eye-tracking data, we were able to explore participants' reason (s) for their eye movements and their challenges using the app. A previous study on the usability of mHealth apps among adolescents reported difficulty in capturing the adolescents' verbalizations in a think-aloud protocol [15,18,50]. Other existing evidence also suggests that in think-aloud protocols, some adolescents did not clearly discuss their difficulties finding a solution [46]. Therefore, our findings suggest

Table 6
Comparison with traditional stand-alone usability testing method.

	Think-aloud only (mostly concurrent)[11 44]	Eye-tracking retrospective think-aloud	
Benefit	Direct insights into end-users' thoughts and strategies during the task performance	Deep insights into end-users' behavior related to the identified usability problems Holistic understandings of the usability issues	Objective eye movements of the identified usability problems
Measurement	Time for task performance	Time for task performance; critical error	Eye fixation; saccades; time to first fixation; time spent; revisit
Needed users	3+	20+	
Required users' skills during testing (particularly for adolescents)	High (unnatural/distracting/strenuous) to think and talk aloud at the same time)	Low	Low
Required equipment	Low (audio-recorder)	Low (audio-recorder)	High (eye-tracking device and software)
Required time for data collection	Medium	High	Low
Required time for data analysis	Medium	Medium	High
Required expertise	Medium	High	High

that it can be beneficial to show eye-tracking data during a retrospective think-aloud to elicit rich comments as well as usability problem-related comments from adolescents.

The application of a comprehensive usability evaluation method, an eye-tracking retrospective think-aloud, enabled us to gain a better understanding of usability issues of a mobile app. In our study, the eye-tracking data was illustrated using gaze plots and aggregated heat maps in addition to areas of interest. In gaze plots tracing participants' eye movements by representing the sequence of fixation and saccades in the form of a scan path, the eye-tracking data were presented with circles and lines. By comparing the gaze plots among participants with/without a critical error, we identified a specific app page within each activity. Heat maps aggregating the fixations revealed which parts were most frequently looked at using colors. In the heat maps created for our study, the areas of each page of quizzes and its correct answer were displayed as a red color indicating participants' high visual attention if there were no critical usability problems. Also, the eight areas of interest - a sexual activity card, four risk categories, 'Prev' button, 'Next' button, and 'Back' button - created to compare eye-tracking metrics among participants with/without a critical error, showed significant differences in time spent, revisits, and number of fixations. Findings from our study demonstrate that eye-tracking data indicating differences between participants with/without a critical error can help capture usability problems where end-users cannot recognize the problems right away, monitor the specific areas where they encountered difficulties, and further make inferences about their actual cognitive processes by researchers. Our work highlights that the use of eye-tracking data can provide researchers a rich representation and an in-depth understanding of the end-users' experience participating in usability testing.

The eye-tracking retrospective think-aloud approach was time-consuming. For example, upon completing the tasks employing a use case scenario, participants were encouraged to think aloud retrospectively and asked to verbalize their thoughts about the tasks while watching a recording of their use of the app that depicted their eye movements overlaid on the app screen. The process took a significant amount of time (i.e., between 2 and 2.5 h). Moreover, our approach using eye-tracking technology required additional time and researcher's technical and extensive analysis skills. The eye tracker (i.e., Tobii X2-30) and software (i.e., iMotions) is very costly, which may limit others ability to access this technology. With the benefits from the method of an eye-tracking retrospective think-aloud, however, the use of the eye-tracking technology (device and software) is highly recommended for a usability evaluation.

The Health-ITUES was used as a measure of usability, which has been validated for use with mHealth technology [40]. Although several usability problems were identified in this study, the overall Health-ITUES mean score (i.e., mean of all 20 items; a higher score indicates higher perceived usability of the technology) was as high as 4.64 (5-best). Nearly 95% of teens in the US ages 13–17 own or have access to a smartphone [51]. Given that 85% of our participants reported using mobile devices constantly every day, and the mean duration of their use of mobile apps per day was 9.40 h, the high usability score of the MyPEEPS App may be because our participants could easily resolve problems while using the app, and/or quickly learn how to use the new app as they are largely heavy smartphone users. In our study, participants who had difficulty on the first page of quizzes no longer had difficulty on the remaining pages within the same activity. End-users who perceive the mHealth app to be useful may be more likely to show an improvement in its impact in their everyday lives [52], which is another strength of our study.

4.1. Limitations

The generalizability of the results may be limited by the study sample who live in the metropolitan areas of New York City, NY; Chicago, IL; and Birmingham, AL. Results may differ in other groups who live in rural areas. We employed an iOS simulator on the computer in order for the mobile app to be used in the same manner as on a smartphone, therefore there may be differences in end-users' experience when interacting with the app on a computer. Additionally, it was time-consuming to collect and analyze data through a retrospective think-aloud with an eye-tracking technique, as compared to a traditional stand-alone usability testing method.

5. Conclusions

In this paper, we presented a methodological approach of an eye-tracking retrospective think-aloud and its application in evaluating the usability of an HIV prevention mobile app intended for diverse sexual minority young men. Our approach enabled us to identify critical usability problems as well as gain an in-depth understanding of the usability issues related to interactions between end-users and the MyPEEPS App. Findings from this study highlight the utility of an eye-tracking retrospective think-aloud to enhance end-user usability testing of a mHealth app. Our methodological approach may encourage other researchers who design/develop mHealth apps for adolescents to conduct comprehensive usability evaluations in a collaborative manner utilizing an eye-tracking technique with high-skilled usability experts in future research.

Declaration of Competing Interest

The authors declare that they have no conflicts of interest in the research.

Acknowledgements

This research was supported by the National Institute of Minority and Health Disparities of the National Institutes of Health (NIH) under award number U01MD11279 (MPI: RS and RG). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] U.S. Food and Drug Administration, Mobile Medical Applications. Secondary Mobile Medical Applications, (2018) https://www.fda.gov/MedicalDevices/DigitalHealth/MobileMedicalApplications/default.htm.
[2] W. Brown 3rd, P.Y. Yen, M. Rojas, R. Schnall, Assessment of the Health IT Usability Evaluation Model (Health-ITUEM) for evaluating mobile health (mHealth) technology, J. Biomed. Inform. 46 (6) (2013) 1080–1087, https://doi.org/10.1016/j.jbi.2013.08.001 [published Online First: Epub Date]|.
[3] A. Abran, A. Khelifi, W. Suryn, A. Seffah, Usability Meanings and Interpretations in ISO Standards, Softw. Qual. J. 11 (4) (2003) 325–338, https://doi.org/10.1023/a:1025869312943 [published Online First: Epub Date]|.
[4] J. Nielsen, Usability Engineering, Elsevier, 1994.
[5] ISO/IEC 9126, Software Engineerting- Product Quality, (2001).
[6] ISO 9241-11, Ergonomic Requirements for Office Work With Visual Display Terminals (VDTs) –Part 11: Guidance on Usability, (1998).
[7] R. Louho, M. Kallioja, P. Oittinen, Factors affecting the use of hybrid media applications, Graphic arts in Finland 35 (3) (2006) 11–21.
[8] P. Ziemba, J. Wątróbski, A. Karczmarczyk, J. Jankowski, W. Wolski, Integrated approach to e-commerce websites evaluation with the use of surveys and eye tracking based experiments. 2017, Federated Conference on Computer Science and Information Systems (FedCSIS) (2017) 1019–1030, https://doi.org/10.15439/2017F320 [published Online First: Epub Date]|.
[9] H. Cho, P.-Y. Yen, D. Dowding, J.A. Merrill, R. Schnall, A multi-level usability evaluation of mobile health applications: a case study, J. Biomed. Inform. 86 (2018) 79–89, https://doi.org/10.1016/j.jbi.2018.08.012 [published Online First: Epub Date]|.
[10] M.W.M. Jaspers, A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence, Int. J. Med. Inform. 78 (5) (2009) 340–353, https://doi.org/10.1016/j.ijmedinf.2008.10.002 [published Online First: Epub Date]|.
[11] A. Holzinger, Usability Engineering Methods For Software Developers, (2005).
[12] M.J. Van Den Haak, M.D.T. De Jong, P.J. Schellens, Retrospective vs. Concurrent think-aloud protocols: testing the usability of an online library catalogue, Behav. Inf. Technol. 22 (5) (2003) 339–351, https://doi.org/10.1080/0044929031000 [published Online First: Epub Date]|.
[13] P.-Y. Yen, S. Bakken, A comparison of usability evaluation methods: heuristic evaluation versus end-user think-aloud protocol–an example from a web-based communication tool for nurse scheduling, AMIA Annual Symposium Proceedings, (2009), p. 714.
[14] Haak MJvd, Jong MDTd, Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols, IEEE International Professional Communication Conference (2003), https://doi.org/10.1109/IPCC.2003.1245501 [published Online First: Epub Date]|.
[15] L. Cooke, E. Cuddihy, Using eye tracking to address limitations in think-aloud protocol. IPCC 2005, Proceedings. International Professional Communication Conference, 2005 (2005) 653–658, https://doi.org/10.1109/IPCC.2005.1494236 [published Online First: Epub Date]|.
[16] A. Donker, P. Markopoulos, A comparison of think-aloud, Questionnaires and Interviews for Testing Usability with Children (2002) 305–316.
[17] G.H. Seng, The effects of think-aloud in a collaborative environment to improve comprehension of L2 texts, The reading matrix 7 (2) (2007).
[18] B. Sheehan, Y. Lee, M. Rodriguez, V. Tiase, R. Schnall, A comparison of usability factors of four mobile devices for accessing healthcare information by adolescents, Appl. Clin. Inform. 3 (4) (2012) 356–366, https://doi.org/10.4338/aci-2012-06-ra-0021 [published Online First: Epub Date]|.
[19] L. Cooke, Is eye tracking the next step in usability testing? 2006, IEEE International Professional Communication Conference (2006) 236–242, https://doi.org/10.1109/IPCC.2006.320355 [published Online First: Epub Date]|.
[20] R.J. Jacob, K.S. Karn, Eye Tracking in Human-computer Interaction and Usability Research: Ready to Deliver the Promises. The Mind's Eye, Elsevier, 2003, pp. 573–605.
[21] L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, Eye tracking and online search: lessons learned and challenges ahead, J. Am. Soc. Inf. Sci. Technol. 59 (7) (2008) 1041–1052, https://doi.org/10.1002/asi.20794 [published Online First: Epub Date]|.
[22] O. Asan, Y. Yang, Using eye trackers for usability evaluation of health information technology: a systematic literature review, JMIR Hum. Factors 2 (1) (2015) e5, https://doi.org/10.2196/humanfactors.4062 [published Online First: Epub Date]|.
[23] E.M. Kok, H. Jarodzka, Before your very eyes: the value and limitations of eye tracking in medical education, Med. Educ. 51 (1) (2017) 114–122, https://doi.org/10.1111/medu.13066 [published Online First: Epub Date]|.
[24] M.A. Hidalgo, L.M. Kuhns, A.L. Hotton, A.K. Johnson, B. Mustanski, R. Garofalo, The MyPEEPS randomized controlled trial: a pilot of preliminary efficacy, feasibility, and acceptability of a group-level, HIV risk reduction intervention for young men who have sex with men, Arch. Sex. Behav. 44 (2) (2015) 475–485, https://doi.org/10.1007/s10508-014-0347-6 [published Online First: Epub Date]|.
[25] R. Schnall, L. Kuhns, M. Hidalgo, et al., Development of MyPEEPS mobile: a behavioral health intervention for young men, Stud. Health Technol. Inform. 250 (2018) 31.
[26] R. Schnall, L. Kuhns, K. Bullock, et al., Participatory End-user feedback to update MyPEEPS: a theory-driven evidence based intervention for YMSM, American Public Health Association 2018 Annual Meeting & Expo, (2018).
[27] R.K.L. Schnall, M. Hidalgo, D. Powel, J. Thai, C. Pearson, S. Hirshfield, J. Bruce, M. Ignacio, A. Radix, U. Belkind, R. Garofalo, Adaptation of a group-based, HIV risk reduction intervention to a mobile app for young sexual minority men, Aids Educ. Prev. 30 (6) (2018).
[28] C. Rusu, S. Roncagliolo, V. Rusu, C. Collazos, A Methodology to Establish Usability Heuristics, (2011).
[29] A. Solano, C.A. Collazos, C. Rusu, H.M. Fardoun, Combinations of methods for collaborative evaluation of the usability of interactive software systems, Advances in Human-Computer Interaction 2016 (2016).
[30] H. Cho, D. Powell, A. Pichon, et al., A mobile health intervention for HIV prevention among racially and ethnically diverse young men: usability evaluation, JMIR Mhealth Uhealth 6 (9) (2018) e11450, , https://doi.org/10.2196/11450 [published Online First: Epub Date]|.
[31] L. Faulkner, Beyond the five-user assumption: benefits of increased sample sizes in usability testing, Behav. Res. Methods Instrum. Comput. 35 (3) (2003) 379–383.
[32] Tobii Technology I. Tobii Technology, Stockholm, Sweden. Secondary Tobii Technology, Stockholm, Sweden, (2016) http://www.tobiipro.com/product-listing/tobii-pro-x2-60/.
[33] iMotions Biometric Research Platform 6.0, iMotions A/S, Copenhagen, Denmark. [program], (2016).
[34] P.Y. Yen, D. Wantland, S. Bakken, AMIA Symposium 2010Development of a Customizable Health IT Usability Evaluation Scale. AMIA … Annual Symposium Proceedings.2010, Development of a Customizable Health IT Usability Evaluation Scale. AMIA … Annual Symposium Proceedings. (2010) 917–921.
[35] iMotions. Tobii X2-30. Secondary Tobii X2-30. https://imotions.com/tobii-x2-30/.
[36] A.L. Russ, J.J. Saleem, Ten factors to consider when developing usability scenarios and tasks for health information technology, J. Biomed. Inform. 78 (2018) 123–133, https://doi.org/10.1016/j.jbi.2018.01.001 [published Online First: Epub Date]|.
[37] A. Poole, L. Ball, Eye tracking in human-computer interaction and usability research: current Status and future prospects, in: C. Ghaoui (Ed.), Encyclopedia of Human Computer Interaction: IGI Global, 2005.
[38] M.-L. Lai, M.-J. Tsai, F.-Y. Yang, et al., A review of using eye-tracking technology in exploring learning from 2000 to 2012, Educ. Res. Rev. 10 (2013) 90–115, https://doi.org/10.1016/j.edurev.2013.10.001 [published Online First: Epub Date]|.
[39] Qualtrics, Provo, Utah, USA [program], (2005).
[40] R. Schnall, H. Cho, J. Liu, Health information technology usability evaluation scale (Health-ITUES) for usability assessment of mobile health technology: validation study, JMIR Mhealth Uhealth 6 (1) (2018) e4, https://doi.org/10.2196/mhealth.8851 [published Online First: Epub Date]|.
[41] StataCorp, Stata Statistical Software: Release 14, StataCorp LP, College Station, TX, 2015.
[42] Jeff Sauro, What Is A Good Task-Completion Rate? Secondary What Is A Good Task-Completion Rate? (2011) https://measuringu.com/task-completion/.
[43] O. Špakov, D. Miniotas, Visualization of eye gaze data using heat maps, Electron. Electr. Eng. (2007) 55–58.
[44] A. Fernandez, E. Insfran, S. Abrahão, Usability evaluation methods for the web: a systematic mapping study, Inf. Softw. Technol. 53 (8) (2011) 789–817, https://doi.org/10.1016/j.infsof.2011.02.007 [published Online First: Epub Date]|.
[45] J. Nielsen, T. Clemmensen, C. Yssing, Getting access to what goes on in people's heads?: Reflections on the think-aloud technique, Proceedings of the Second Nordic Conference on Human-Computer Interaction (2002) 101–110.
[46] J.L. Branch, Investigating the information-seeking processes of adolescents: the value of using think alouds and think afters, Libr. Inf. Sci. Res. 22 (4) (2000) 371–392.
[47] J. Preece, Y. Rogers, H. Sharp, Interaction Design: Beyond Human-computer Interaction, John Wiley & Sons, 2015.
[48] J.B. Bavelas, L. Coates, T. Johnson, Listener responses as a collaborative process: the role of gaze, J. Commun. 52 (3) (2002) 566–580, https://doi.org/10.1111/j.1460-2466.2002.tb02562.x [published Online First: Epub Date]|.
[49] S. Elling, L. Lentz, Md. Jong, Retrospective think-aloud method: using eye movements as an extra cue for participants' verbalizations, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2011) 1161–1170.
[50] K.S. Karn, S. Ellis, C. Juliano, The Hunt for Usability: Tracking Eye Movements. CHI' 99 Extended Abstracts on Human Factors in Computing Systems, ACM, Pittsburgh, Pennsylvania, 1999 173-73.
[51] PewResearchCenter, Teens, Social Media & Technology 2018, (2018).
[52] L. Casaló, C. Flavián, M. Guinalíu, The role of perceived usability, reputation, satisfaction and consumer familiarity on the website loyalty formation process, Comput. Human Behav. 24 (2) (2008) 325–345, https://doi.org/10.1016/j.chb.2007.01.017 [published Online First: Epub Date]|.