Tests d'utilisabilité : comparaison entre une méthode avec modération en présentiel et une méthode automatisée à distance, étude de cas appliquée au site e-rara.ch

Raphaël Rey
rey.rapha@gmail.com
Haute Ecole de Gestion, Genève

Valérie Meystre
valerie_meystre@hotmail.com
Haute Ecole de Gestion, Genève

Résumé

Face à la multitude des méthodes de test d'utilisabilité existantes, il peut être difficile de choisir la plus adaptée. Cette étude se propose de comparer deux d'entre elles en les appliquant au site e-rara.ch : la première avec modérateur et think-aloud, et la seconde automatisée à distance. A l'aide du même test construit avec le logiciel Loop11, le but est de déterminer si l'une de ces deux méthodes permet de mieux détecter certains types de problèmes. L'analyse des données de quatre-vingt-quatre participants à distance et de sept autres avec modérateur nous a permis d'identifier vingt-cinq problèmes d'utilisabilité. Nous avons constaté que le test à distance se prête bien à la comparaison de deux variantes d'un même site et permet de remonter des problèmes marginaux, tandis que le test avec modérateur relève mieux le ressenti des participants, les doutes et les erreurs de compréhension de certains éléments.

Zusammenfassung

So zahlreich sind die Methoden, einen Usability-Tests durchzuführen, dass es schwierig ist, die am Besten geeignete auszuwählen. Diese Studie vergleicht zwei dieser Methoden, indem sie die Website e-rara.ch analysiert : die erste ist moderiert und benutzt think-aloud ; die zweite ist unmoderiert und automatisiert. Durch den gleichen Test, der mit Loop11 Software aufgebaut worden ist, haben wir geprüft, ob eine dieser Methoden besser hilft, bestimmte Arten von Problemen zu erkennen. Durch die Datenanalyse von vierundachtzig Teilnehmer des unmoderierten Tests und sieben Teilnehmer des moderierten Tests identifizierten wir fünfundzwanzig Usability-Probleme. Wir haben festgestellt, dass der automatisierte Test besonders gut ist, um zwei Versionen einer gleichen Website zu vergleichen und um Randprobleme zu entdecken. Im Vergleich ist der moderierte Test besser, um zu verstehen, was die Teilnehmer fühlen, warum sie zweifeln oder sich irren.

Abstract

There is a lot of different usability test methods and it can be hard to choose the most relevant. This study propose to compare two of those methods : one using a moderator and think-aloud and a remote method. We used the same test, built with the Loop11 software, including five tasks to perform in the e-rara.ch website and a questionnaire. We wanted to see if one of those methods could detect some kind of usability problems better than the other, and what kind of problems. We analysed the data of eighty-four remote participants and seven moderated participants and found twenty-five usability problems. They show that remote test is better at comparing two website versions and identifying little problems, but that the moderated test allows us to see the participant's feelings and doubts and if some elements are difficult for him/her to understand. Choosing one usability test method depends on the data type we're searching for.

Mots-clés

test d'utilisabilité, test à distance, test automatisé, test modéré, étude comparative, bibliothèque numérique, e-rara.ch

1. Introduction

Il existe une grande diversité dans les méthodes existantes pour analyser l'utilisabilité des sites web. Les tests peuvent s'effectuer à distance ou en présentiel, avec ou sans modérateur, en recourant à un nombre de participants très variable allant de quatre ou cinq à plusieurs centaines, etc. Quand il s'agit de se lancer dans une telle démarche, il n'est donc pas aisé de choisir la méthode qui apportera les résultats les plus pertinents. L'article de Fernandez, Insfran et Abrahão (2011) illustre parfaitement cette richesse en dressant un panorama détaillé sur la base d'une sélection de 206 articles tirés d'un corpus initial de 2703 publications appartenant à ce domaine spécifique.

Pour effectuer des choix méthodologiques pertinents au sein de ce foisonnement, il existe certes des manuels qui fournissent des conseils pratiques comme celui de Barnum, *Usability testing essentials* (2011) où ce dernier décrit les avantages et inconvénients de la plupart des modes de test. Ces paroles d'experts sont précieuses, mais elles gagnent à être complétées et confirmées par des études de cas et notamment des études comparatives qui à l'épreuve des faits et à travers une démarche rigoureuse mettent en évidence les forces et faiblesses des différentes méthodes analysées. Le projet dont cet article se fait l'écho s'inscrit dans cette démarche.

Il se base, en effet, sur une série de tests appliqués au site e-rara.ch. Cette plateforme offre au visiteur la possibilité de consulter des imprimés anciens numérisés qui appartiennent aux collections d'un grand nombre de bibliothèques suisses partenaires. Pour évaluer l'utilisabilité du portail, nous avons d'une part recouru à un dispositif automatisé à distance et d'autre part effectué des sessions de test modérées en présentiel avec think-aloud. Le tableau suivant résume les principales différences entre les deux formes de test :

Méthode 1	**Méthode 2**
A distance	En présentiel
Automatisée	Modérée
Asynchrone	Synchrone
Sans think aloud	Avec think-aloud
Analyse quantitative et qualitative	Analyse quantitative

Notre objectif dans cette étude est de comparer ces deux méthodes sous les aspects : de l'efficacité et de l'efficience. Concernant l'efficacité, l'enjeu sera de déterminer quelle approche permet de repérer le plus de problèmes et si l'une d'elles permet de repérer particulièrement facilement certains types d'entre eux, ou au contraire a tendance à ignorer d'autres. La question

de la fiabilité des résultats sera également abordée ainsi que celle de l'efficience qui s'intéressera à la question du coût investi, notamment en temps.

2. Etat de l'art

Les tests d'utilisabilité peuvent se mener de multiples manières différentes comme en témoigne l'abondante littérature les concernant. Mentionnons, par exemple, l'article de Bastien (2010) qui dresse un panorama général ou encore le travail imposant de Fernandez, Insfran et Abrahão qui anaylsent l'état de l'art sur le sujet à partir d'un ensemble de 206 publications sur un total initial de 2703.

Plusieurs manuels exposent les différentent méthodes et en présentent le fonctionnement, mais aussi leurs avantages et inconvénients respectifs : le livre de Barnum, *Usability testing essentials : ready, set... test !* (2011), constitue à ce titre une excellente introduction et offre un aperçu assez complet de la question.

2.1. Le think-aloud, un outil aux mutiples visages

Le mode le plus traditionnel pour réaliser un test d'utilisabilité passe par le recours à un modérateur. Le participant se voit proposer une série de tâches à réaliser sur un site web et est convié à commenter à haute voix ses actions et les difficultés qu'il rencontre : il s'agit là du think-aloud dont il sera spécifiquement question dans ce chapitre.

Le manuel de Barnum et notamment son chapitre 7 définissent en détail comment utiliser cette méthode pour récolter des données (2011, p 199-237). Comme l'auteur l'explique dans son premier chapitre à la page 19, le but est de comprendre pourquoi l'utilisateur accomplit telle ou telle action : « Not only do you see the actions users take, but you also benefit from hearing *why* users are taking an action and *what* they think about the process - good and bad. When users think out loud, you don't have to guess what they're thinking. They tell you ». Pour une réflexion approfondie sur cette question de l'introspection et de la verbalisation de la pensée, le lecteur pourra se référer à l'article de Nielsen et al. (2002, en particulier les pages 106-107).

Afin de mieux comprendre le type de propos que l'on recueille via cette méthode, l'article de Cooke (2010) est d'un grand intérêt. A travers une comparaison entre les données d'un dispositif d'eye tracking et les propos recueillis, elle arrive à la conclusion que 58% des énoncés relèvent de la lecture et 19% de la procédure. Quant aux silences et aux paroles de remplissage des participants, ils ne traduisent pas une absence d'action. Au contraire, ils signalent souvent une difficulté ou un besoin d'analyse. Le dispositif d'eye tracking permet de repérer ce que regarde l'utilisateur dans ces moments-là et de comprendre dans une certaine mesure ses intentions. Il en ressort que le think-aloud ne donne accès qu'à une portion assez restreinte de l'activité cérébrale des utilisateurs.

Le risque principal de la pensée à voix haute est que le modérateur biaise les résultats par ses interventions ou son attitude. Barnum consacre à cette problématique les pages 207 à 218 de son ouvrage déjà mentionné (2011). Le langage corporel peut trahir des appréciations qui ensuite influencent les choix du participant. Quant aux compliments, il se doivent d'être équilibrés quel que soit le résultat de la tâche. Enfin, il faut poser des questions adéquates qui ne vont pas induire certains comportements.

Une étude de Nørgaard et Hornbæc (2006) démontre à travers l'analyse de plusieurs sessions de test avec think-aloud que les modérateurs recherchent en réalité dans leur manière d'intéragir avec les participants la confirmation de problèmes qu'ils ont déjà repérés.

Afin d'éviter ce type de biais, le plus efficace est d'intervenir le moins possible en tant que modérateur et de se contenter d'inviter le participant à verbaliser ses pensées (à ce propos voir le modèle proposé par Ericsson et Simon,1993).

Des alternatives avec davantage d'interactions ont toutefois été proposées. En 2000, Boren et Ramey constatent un décalage entre les recommandations du modèle d'Ericsson et Simon et la pratique des professionnels. Ces auteurs proposent dans un article un cadre théorique plus souple. Par exemple, si un participant estime avoir terminé une tâche et que ce n'est pas le cas, ne serait-il pas préférable de le relancer pour qu'il poursuive plutôt que de perdre une partie de ce qui aurait pu être testé? Le biais que cela induit est-il compensé par les données supplémentaires récoltées? (voir Boren et Ramey 2000, p. 273-274). Une étude de Khramer et Ummelen (2004) montre que les résultats concernant le repérage des problèmes d'utilisabilité sont à peu près équivalents, mais le confort et la performance des participants sont meilleurs avec une modération plus active. Ils ont moins le sentiment d'être « perdus » (p. 116). Zhao et McDonald (2010) arrivent à des conclusions similaires, mais soulignent le nombre plus importants d'énoncés produits grâce aux échanges avec le modérateurs sans que cela ne permettent d'améliorer la détection d'éléments problématiques. Greiner observe de même sauf pour le dernier point (2012, p. 35).

En 2014, une nouvelle étude de Zhao et McDonald auxquels vient encore s'ajouter Edwards tente une nouvelle expérience pour définir un protocole plus efficace de think-aloud. Cette fois-ci, les auteurs semblent avoir obtenu des résultats convaincants en donnant des instructions assez précises sur le type de propos qu'ils souhaitent recueillir pendant les tests. Ils ont ainsi obtenu non seulement davantage d'énoncés comme dans leur étude précédente, mais aussi davantage de contenus explicatifs. Les problèmes d'utilisabilité supplémentaires détectés par ce biais sont toutefois en général de peu d'importance.

Un mode complètement différent de recourir à la pensée à voix haute est le « retrospective think-aloud » où la personne réalise dans un premier temps le test, puis s'exprime à son propos, éventuellement en visonnant un vidéo de ses interactions avec le site. L'étude de De Jong, Schellens et van den Haak (2003) arrive à la conclusion qu'en terme de repérage des problèmes d'utilisabilité, les deux méthodes sont équivalentes. La détection s'effectue toutefois de manière différente : avec la pratique simultanée du think-aloud, les problème sont constatés surtout grâce à l'observation, tandis que si la récolte des propos se fait après la réalisation des tâches, ces derniers deviennent la source la plus importante (p. 345).

Les auteurs ont également observé une baisse des performances des participants dans la réalisation de leurs tâches lorsque ceux-ci doivent verbaliser leurs actions en prarallèle. Par conséquent, il y a un risque d'influence plus important si la tâche proposée est particulièrement complexe (p. 350).

Ces trois auteurs ont réalisé trois autres études sur des thématiques proches (2004, 2007 et 2009) et arrivent à des conclusions similaires. Ils notent toutefois l'influence du type de site testé sur l'efficacité des méthodes (voir par exemple De Jong, Schellens et van den Haak 2009, p. 201). Un article de McDonald, Zhao et Edwards (2013) estiment que ces deux

approches sont complémentaires. L'usage du think-aloud pendant la réalisation des tâches permet de repérer davantage de problèmes, mais les données récoltées après coup viennent utilement confirmer ces éléments et les compléter en donnant des explications sur les difficultés rencontrées.

Même si les méthodes sont perfectibles, en particulier pour le confort des participants, l'influence du type de modération sur l'efficacité des tests semble assez négligeable, si on exclut évidemment les réels faux-pas qui biaisent les tests sans aucun profit.

2.2. Comparaison entre test en laboratoire et test à distance synchrone

Comme alternative à cette méthode traditionnelle en laboratoire, il est possible de procéder au même type de test à distance cette fois-ci. Dans ce cas, le modérateur et l'internaute se trouvent dans des lieux séparés, mais gardent un contact vocal. On parle dans ce cas de méthode à distance synchrone.

Selon Barnum, les résultats sont très similaires : « moderated, also called synchronous, remote testing is very much like lab testing » (2011, p. 42). Plusieurs études viennent confirmer ce fait : voir à ce propos Andreasen et al. (2007, p. 1413), Hartson et al. (1996, p. 234), Thompson et al. (2004, p. 136), Selvaraj (2004, p. 32). Un article de Chalil Madathil et Greenstein (2011, p. 2233) arrive à la même conclusion, mais introduit une nouvelle méthode qui semble légèrement meilleure : le laboratoire virtuel. Lors du déroulement du test, un laboratoire s'affiche à l'écran avec un navigateur partagé et deux avatars qui peuvent interagir : le participant et le modérateur. Cette méthode aurait permis de détecter davantage de problèmes de faible gravité. Ce résultat serait à confirmer.

Au niveau de l'efficience, les données montrent que les tests à distance prennent en général plus de temps (Thompson et al. 2004, p. 234 et Andreasen 2007, p. 1412), même si Slevaraj affirme que la différence qu'elle observe n'est pas suffisamment significative (2004 p. 30).

L'article de Dray et Siegel (2004, p. 12-14) mentionne trois avantages :

- Réduction des coût : toutefois, même si cela diminue par exemple les frais de déplacement ou facilite le recrutement, le gain reste faible à ce niveau.
- Liberté de l'interface : l'utilisateur peut choisir son interface et est en principe familier avec cette dernière.
- Gain de temps : il ne faut pas sous-estimer le temps nécessaire pour mettre en place le dispositif qui, à distance, peut prendre un certain temps.

Slevaraj (2004, p. 19) reprend une partie de ces éléments en ajoutant la facilité pour élargir le recrutement et mieux cibler les utilisateurs par rapport à la population cible du site. De plus, le modérateur risque moins d'influencer le déroulement, s'il est physiquement absent. Même si cela ne semble pas avoir d'impact sur les résultats, Andreasen et al. ont observé que celui-ci pouvait avoir un effet intimidant comme en témoigne ce participant au test à distance dont les auteurs rapportent les propos : « I liked this test method better than the traditional method where the test leader looks over your shoulder. » (2007, p. 1413).

En contrepartie, les données sont en général plus pauvres. Le visage de la personne qui effectue le test n'est pas toujours filmé, de plus il est assez difficile d'interpréter le langage

paraverbal et non-verbal quand on n'est pas physiquement présent dans la salle (Dray et Siegel, 2004, p. 14-15). Au niveau de la satisfaction des participants, Slevaraj constate que ceux qui ont participé à un test à distance préfèrent massivement cette méthode, alors que ceux qui ont pris part à la méthode en présentiel sont beaucoup plus partagés dans leur engouement pour l'un ou l'autre dispositif (2004, p. 32-34).

2.3. Comparaison entre test synchrone et test asynchrone

Les différences au niveau de l'efficacité et de l'efficience sont plus marquées losrqu'on oppose test synchrone et test asynchrone. A noter que les études à ce propos sont assez rares, comme le font observer Rodriguez et Resnick (2010, p. 760). En 2002, un article de Tullis et al. fait le point sur la question. Il en ressort que les deux méthodes arrivent à des résultats similaires concernant le repérage des problèmes d'utilisabilité. Les plus sérieux sont toujours détectés et les utilisateurs se comportent de manière très semblable en rencontrant les mêmes difficultés dans l'un et l'autre contexte. La population plus importante des participants du test automatisé augmente la probabilité de découvrir des problèmes mineurs et constitue une source de commentaires sur les sites à la fois plus abondante et plus variée que dans une petite série de tests modérés. Ces derniers sont toutefois plus à même de découvrir certains types de problèmes, lorsque justement une observation directe est nécessaire comme dans le cas d'un scrolling excessif. Au final, l'étude conseille de combiner dans la mesure du possible les deux méthodes, même si l'une ou l'autre est suffisante s'il s'agit uniquement de repérer les problèmes les plus importants.

L'article d'Andreasen et al. déjà mentionné plus haut (2007) compare une méthode en laboratoire, une à distance modérée et deux autres automatisées (la première avec des experts et la seconde des utilisateurs ordinaires). Les conclusions sont assez défavorables pour les méthodes asynchrones qui ont permis de découvrir moins de problèmes d'utilisabilité et ont demandé plus de temps aux participants.

A noter toutefois que cette étude a effectué exactement le même nombre de tests pour chacune des méthodes (6 soit 24 au total), alors que l'intérêt des méthodes asynchrones est d'offrir la possibilité de recourir à des analyses quantitatives avec de nombreux utilisateurs comme le rappelle Bastien dans son article (2010, p. e20). Selon ce même auteur (p. e21), une des raisons de ces résultats est peut-être à chercher dans le design des tâches ou le type de site web analysé.

Schmidt a réalisé en 2013 une étude dans laquelle elle a comparé une méthode à distance automatisée et une autre en présence d'un modérateur. Les conclusions sont similaires : si les problèmes les plus sérieux sont détectés dans les deux cas, la présence d'un observateur direct donne de meilleurs résultats pour tous les autres éléments posant des difficultés. Les commentaires laissés par les participants du test automatisé sont extrêmement précieux et révèlent une part non négligeable des problèmes.

Ces résultats défavorables pour la forme asynchrone doivent toutefois être relativisé puisque le site cible de ne proposait pas d'url différentes lors du passage d'une page à l'autre ce qui rendait très difficile de suivre le parcours des internautes dans le test à distance. Les données utilisables étaient donc très incomplètes.

Notre étude se place dans la continuité de ces travaux, et, sur la base d'une nouvelle expérience, tente de confirmer ou d'infirmer ces conclusions.

3. Méthodologie

3.1. Descriptif des tests réalisés

Dans le but de limiter autant que possible les différences entre les tests, nous avons demandé aux participants de réaliser exactement les mêmes tâches dans les deux cas. Ce choix est discutable dans la mesure où en vue d'une efficacité maximale, il faudrait sans doute adapter le design de celles-ci à chacune des deux méthodes, ce qui compromettrait ensuite drastiquement les possibilités de comparaison. Par conséquent, les résultats de la comparaison des deux méthodes sont en partie liés aux tâches telles que nous les avons définies et ne peuvent donc pas être généralisés sans précaution.

La trame du test se compose d'un questionnaire préalable récoltant des informations personnelles comme l'âge et l'expertise dans l'utilisation d'Internet, des bibliothèques numériques et des livres anciens.

Arrivent ensuite cinq tâches :

1. vérifier si la Bibliothèque des Pasteurs de Neuchâtel propose une partie de ses collections sur e-rara.ch ;
2. trouver la liste des flux rss ;
3. repérer l'ouvrage le plus ancien présent sur e-rara.ch et dont l'auteur est Erasme (moteur de recherche, index, facettes) ;
4. en partant de la notice de l'*Histoire naturelle des oiseaux* de Buffon, déterminer quel genre d'oiseau est le griffon pour cet auteur (navigation dans un livre) ;
5. dans le même ouvrage, trouver une illustration d'un faucon sort (navigation dans un livre, utilisation des vignettes).

La tâche 3 présente une différence importante selon que le test est réalisé en français ou en allemand. En effet, le système ne connaît qu'Erasmus en latin, forme qu'utilisent également les germanophones. Par contre, entrer « Erasme » dans le moteur de recherche ne retourne aucun résultat pertinent.

Dans la tâche 4, ce sont les francophones qui sont avantagés puisque le livre est en français et que dans l'énoncé nous avons traduit « griffon » par « Greif ». Ces différences linguistiques permettent de simuler une forme de test A/B avec deux versions d'un même site. Quelle méthode est la plus à même de mesurer l'impact de ces divergences ?

Une fois l'étape centrale achevée, le participant est invité à s'exprimer sur des problèmes d'utilisabilité qu'il aurait pu constater, puis arrive un test SUS quelque peu remanié et surtout raccourci sur lequel nous ne nous étendrons pas dans cet article (1)

Nous avons débuté cette étude par le test automatisé à distance. Ce choix se justifie dans la mesure où les données recueillies via la méthode avec think-aloud sont davantage qualitatives. En effet, les résultats auraient été biaisés, si nous avions à l'esprit les propos des participants du test modéré lors de l'analyse des données du test à distance, certes plus abondantes, mais plus pauvres étant donné l'absence de think-aloud.

Ce dernier a été réalisé à l'aide du logiciel Loop11 qui permet notamment d'enregistrer des flux de clics, des heatmaps(2) et de gérer des questionnaires. Nous avons recruté la grande majorité

des participants grâce à la liste de diffusion Swiss-lib. Au total, 160 personnes ont démarré le test et 84 l'ont terminé avec une légère majorité pour les francophones (46 contre 38). Dans l'analyse des résultats, nous avons également pris en compte toutes les tâches effectuées dans leur intégralité, même lorsque la totalité du test n'a pas été réalisée par la suite.

Le test modéré a également été réalisé avec Loop11, mais avec en plus un enregistrement audio des commentaires des utilisateurs. Le modérateur n'a joué qu'un rôle minimal en invitant uniquement le participant à s'exprimer lorsque celle-ci cessait d'expliquer les raisons de ses actions. Comme nous l'avons vu dans notre état de l'art, le protocole du think-aloud ne possède qu'assez peu d'influence sur les résultats. Nous avons donc choisi le mode le plus simple et qui nous rapprochait le plus du test automatisé, donc sans modération. En tout, 7 personnes ont pris part au test modéré : deux germanophones et cinq francophones. Toutes avaient des affinités avec l'histoire (de par leur formation ou leur activité professionnelle).

3.2. Mode de comparaison des deux méthodes de test

A travers les questionnaires avant et après les tâches, notre objectif était de segmenter la population des participants au test automatisé en décelant des corrélations entre certains comportements et des données relatives à la personne. Par exemple, les utilisateurs plus âgés rencontreraient spécifiquement tel ou tel problème. Cette indication serait une aide ensuite pour proposer des améliorations adaptées.

Pour comparer les méthodes, nous sommes partis des problèmes qu'elles ont permis de détecter et avons regardé si un problème était clairement identifié, soupçonné ou totalement invisible. Nous avons également pris en compte la gravité des problèmes observés et la probabilité de leur observation (en fonction du nombre de participants au test). De plus, nous avons comparé l'efficience du repérage de ces problèmes en fonction de la nécessité d'une manipulation particulière des données : soit le problème pouvait être détecté via un traitement standard des données (observation directe ou commentaire de la part d'un participant pour le test modéré et analyse des questionnaires ou des flux de clics pour le test automatisé), soit le problème requérait une méthode plus complexe (par exemple le recoupement de plusieurs types de données).

Afin de systématiser cette analyse, nous avons créé une grille à l'aide de laquelle nous avons pour chacun des problèmes évalué l'une et l'autre méthode. En attribuant un score aux différents critères, il a donc été possible de comparer quantitativement les deux tests. Notre étude est toutefois également qualitative, puisque nous avons également exploité nos fiches pour catégoriser les éléments qui posaient des difficultés aux utilisateurs afin de mieux comprendre les forces et faiblesses de chacune des deux approches.

4. Résultats

4.1. Tests d'utilisabilités

Les analyses de chacune des cinq tâches nous ont permis de repérer vingt-cinq problèmes d'utilisabilité au total, toutes méthodes confondues. Il serait ici peu pertinent de les énumérer tous et nous nous contenterons de citer quelques exemples particulièrement intéressants pour notre propos.

Certaines fonctionnalités se sont montrées défectueuses : après une recherche, l'option de tri « relevance » s'affiche toujours automatiquement quand on modifie l'ordre (« croissant », « décroissant »).

Moins graves, certains éléments, comme la fenêtre de recherche, n'étaient pas toujours suffisamment visibles pour être utilisés autant qu'ils l'auraient mérité.

D'autres éléments se sont révélés peu ergonomiques à l'usage : par exemple, l'affichage des résultats manque de clareté. En effet, il n'y a pas d'étiquette pour indiquer qu'il s'agit d'un auteur, d'un titre ou d'autre chose. Le contexte permet souvent de trancher, mais pour une lecture rapide, ce n'est pas pratique.

Pareillement, la navigation dans les miniatures n'est pas aisée et le cadre rouge qui devrait servir de guide est très peu visible. Par conséquent, de nombreux utilisateurs n'ont pas pu se repérer efficacement et ont fini par se perdre dans les pages à consulter.

Pour finir, mentionnons une « fonctionnalité » manquante qui a occasionné beaucoup d'erreurs : l'absence de solution pour repérer la « bonne » variante orthographique d'un nom d'auteur (renvois, index alphabétique clair, etc.).

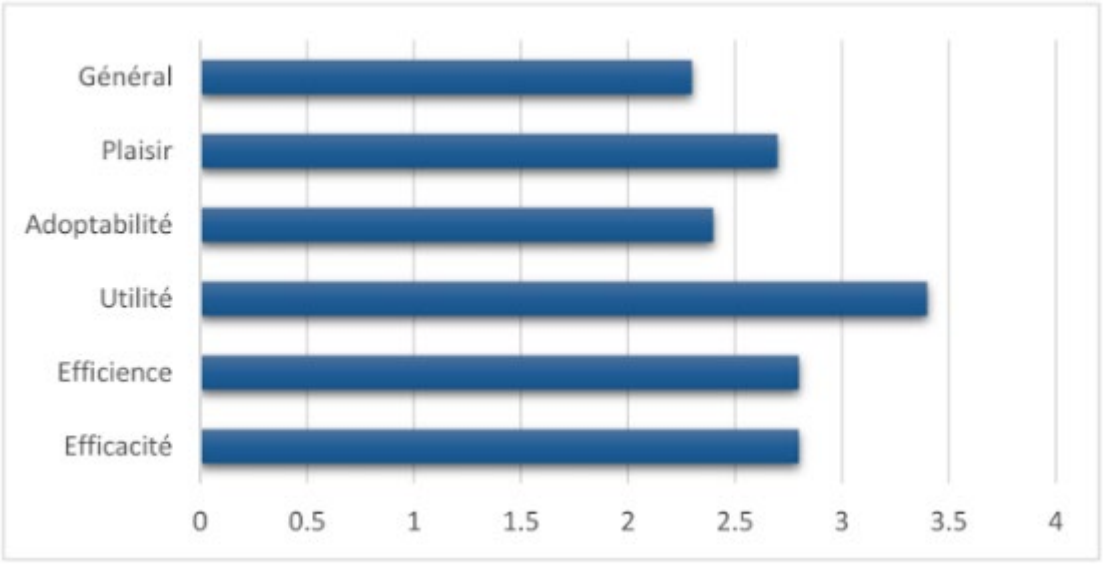
Nous avons également constaté de grosses différences entre les tests francophones et germanophones et avons donc pu analyser efficacement la capacité de chacune des deux méthodes à évaluer l'utilisabilité de deux versions d'un même site. Un exemple des plus emblématique est la tâche 3 (recherche d'un livre d'Erasme). Ainsi, 22% des francophones du test à distance ont réussi la tâche contre 61% des germanophones. 35% des erreurs ont été débuté avec une recherche « Erasme » au lieu du nom latin. Ce dernier étant le nom utilisé en allemand, aucun des participants germanophones n'a commis cette erreur. Ces observations se retrouvent dans le test en présentiel, mais sans qu'une analyse statistique soit pertinente : 2 francophones sur 5 ont effectués une recherche avec « Erasme ». La section suivante portera sur le comparatif des deux méthodes utilisées dans cette étude.

4.2. Evaluation et comparaison des deux méthodes

Les informations personnelles récoltées avant et après les tâches se sont révélées difficiles à exploiter. En effet, malgré nos efforts, nous n'avons pas pu déceler de corrélation significative entre plusieurs éléments (questions ou problèmes détectés). Par conséquent, ces informations ont au final été de peu d'utilité pour le test automatisé, si ce n'est pour pouvoir prouver la diversité de la population qui a pris part au test.

Lorsqu'un modérateur est présent, l'enjeu est un peu différent. Ces questions permettent de mieux cerner le participant et aident à comprendre les problèmes qu'il rencontre, sans pour autant (du moins dans notre étude) contribuer directement à la détection de problèmes.

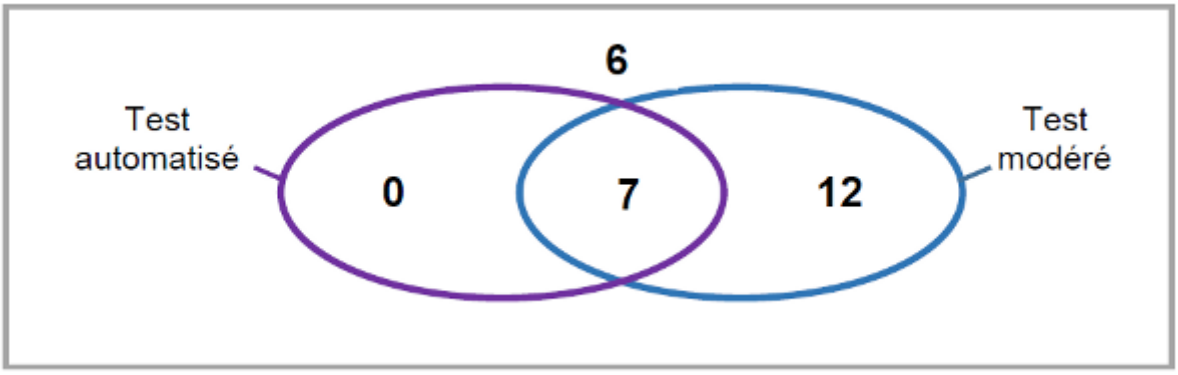
Concernant le questionnaire post-test (similaire au SUS, mais avec uniquement six questions), le résultat du test automatisé montre qu'e-rara.ch obtient une note très moyenne de 64/100. Si ces questions avaient été déjà posées lors d'états antérieurs du site, il aurait alors été possible de suivre l'évolution du site en parallèle avec celui de son score. De plus, avant même d'entrer dans l'analyse de détail des tâches, ces informations suffisent à donner un diagnostic préalable du site. Le diagramme ci-dessous, dont les entrées de gauche constituent les critères évalués par les questions, livre un premier aperçu sur la manière dont les utilisateurs appréhendent le site.



Résultats du questionnaire post-test

Dans le test modéré, les calculs quantitatifs ne sont que peu pertinents et ces questions ont surtout permis d'ouvrir la discussion et de revenir sur certains éléments du site qui ont marqué les utilisateurs. Par exemple, un participant a observé une certaine surcharge dans les contenus des pages et estimait que tous les éléments n'étaient pas utiles au même titre. Cela ne permet pas encore de définir un problème précis, mais donne malgré tout une piste intéressante.

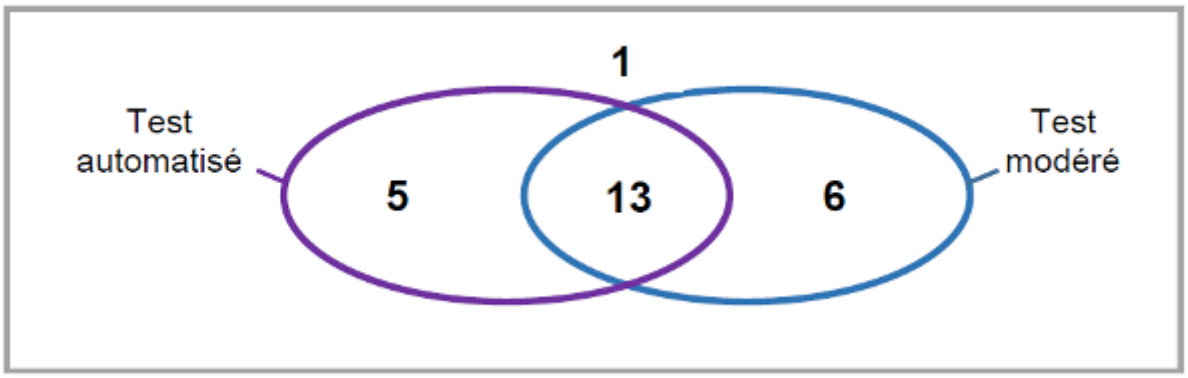
En comparant les problèmes d'utilisabilités détectés par l'une et l'autre méthodes, nous avons pu mettre en évidence que le test automatisé révélait principalement des problèmes majeurs ou entraînant des erreurs dans la résolution des tâches (7 problèmes sur 25 relevés, dont aucun n'est passé inaperçu lors du test modéré).



Nombre de problèmes repérés (sans question ouverte pour le test automatisé)

Pour tout ce qui ressort d'incompréhensions de certains éléments de l'interface, d'un manque de visibilité ou d'un mauvais d'affichage et qui n'est pas directement décelable dans le flux de clic, le test modéré est clairement plus performant (19 problèmes observés, dont 12 ignorés dans l'autre test). Ce phénomène est lié d'une part au think-aloud et d'autre part à l'observation directe des interactions, qui nous apportent des éléments indispensables pour réellement comprendre le sens des actions des participants.

La question ouverte à la fin du test (propositions d'amélioration pour le site) a permis de recueillir de précieuses informations lors du test automatisé et de détecter 18 problèmes, dont 5 passés inaperçus durant le test modéré. Cette seule question est donc à l'origine de l'identification de 11 éléments, ce qui rétablit quelque peu la balance entre les deux méthodes, puisque le test modéré ne révèle alors que 6 problèmes que le test automatisé ne détecte pas.



Nombre de problèmes repérés par méthode

Cela montre bien l'importance primordiale de cette question dans le cadre du test automatisé, même si seul 40% (36 participants sur 86) y ont répondu. Dans le cas du test modéré, cette importance est nettement moindre, puisque les difficultés des participants ont été verbalisées et observées au fur et à mesure du déroulement du test. Dans notre travail, un seul problème supplémentaire a pu être détecté de cette manière.

En revanche, le nombre de participants plus important du test automatisé nous a aidé à mieux évaluer la fréquence des problèmes détectés, le taux d'échec causé par ces problèmes et l'effort nécessaire pour leur apporter une solution, point important pour décider si une modification importante est nécessaire ou si elle ne concerne finalement que peu de participants. Une faiblesse du test modéré réside dans le fait qu'il est difficile de déterminer s'il vaut la peine d'entrer dans une telle action, pouvant demander beaucoup de temps et de moyens, si elle ne concerne que 2 participants sur 7.

Par exemple, pour la tâche 3, nous avons constaté un taux de succès de 61% chez les germanophones et de seulement 22% parmi les francophones. La seule différence significative provenait du nom d'un auteur à rechercher : « Erasme » dans l'énoncé de la tâche en français et « Erasmus » dans le test germanophone. Comme le système ne reconnaît que la forme « Erasmus » (forme latine également utilisée en allemand), les participants au test francophone se voyaient proposer une tâche bien plus ardue. On peut donc affirmer que l'absence de renvois entre les diverses formes des noms d'auteur ou du moins d'un réel index alphabétique de ces derniers occasionne jusqu'à près de 40% d'échec. Ce chiffre suffit à montrer l'importance de trouver une solution à ce problème. Une telle évaluation n'aurait évidemment pas été possible avec le nombre restreint de participants à un test modéré.

Concernant le critère de l'efficacité, nos résultats n'indiquent pas une nette prééminence d'une des méthodes sur l'autre. Si on additionne le score de nos grilles relatives à chaque problème le test modéré obtient un résultat d'un peu moins de 10% supérieur.

Certains problèmes sont toutefois plus facilement détectés avec une méthode en particulier. Les phénomènes qui ne se manifestent que rarement ont peu de chance d'être mis en évidence avec le test modéré et son nombre réduit de participants. Ces faits apparaissent plus

facilement dans un test à distance auquel prennent part bien plus d'utilisateurs. En revanche, le fil de clics du test à distance ne permet pas de percevoir le ressenti des usagers. Ainsi, si un problème d'utilisabilité ne provoque pas d'erreur, mais génère un désagrément, il faut que ce dernier soit exprimé dans la question ouverte en fin de test, faute de quoi il passera totalement inaperçu. Dans un test modéré, en revanche, le participant fera immédiatement part de sa frustration et de ses interrogations.

Par contre, lorsqu'il s'agit de comparer deux versions d'un même site, le test quantitatif est clairement plus performant que le test qualitatif. En effet, les statistiques permettent aisément de chiffrer l'impact de la différence entre les deux variantes, en fonction de la langue dans notre cas. Un test modéré pourrait seulement constater la présence de la difficulté sans en évaluer l'importance. De plus, la fiabilité des résultats est souvent plus faible, notamment si par exemple le fait considéré n'a été observé qu'avec un ou deux participants.

Au niveau de l'efficience, nos résultats ont montré que les deux méthodes supposent un investissement conséquent. Outre la conception du test en lui-même, l'analyse des résultats du test à distance demande un temps considérable. En effet, si le logiciel Loop11 donne différents graphiques, il est nécessaire de suivre le flux de clics de chaque participant pour bien comprendre le cheminement et les erreurs rencontrées. Quant au test modéré, il demande de planifier les rendez-vous avec les différents participants et de prévoir suffisamment de temps pour la passation de chaque test. Or, nous avons pu constater que ceux qui ont passé le test en présentiel avaient parfois tendance à s'acharner sur les différentes tâches et à vouloir absolument les mener jusqu'au bout, là où le participant à distance aurait depuis longtemps passé à la question suivante.

Au final, si l'on additionnait le temps passé par tous les acteurs (organisateurs et participants) dans la réalisation des tests, il ne fait pas de doute que le test modéré est bien plus léger. En terme de coûts réels, à l'exemple de notre étude, il est souvent possible de trouver des participants à peu de frais, mais dans le cas contraire un test automatisé avec une visée quantitative deviendrait rapidement extrêmement cher.

5. Conclusion

Même si le projet dont cet article se fait l'écho consiste en une étude de cas, plusieurs des résultats décrits précédemment peuvent dans une certaine mesure se généraliser à d'autres contextes.

Nous n'avons en effet pas opposé les deux méthodes uniquement de manière quantitative, mais défini des types de problèmes d'utilisabilité que l'une ou l'autre d'entre elles permettait de repérer plus aisément ou au contraire avait tendance à ignorer. Ces constatations restent valides quel que soit le site analysé. Naturellement, certains types de faiblesses ont pu nous échapper ou n'être tout simplement pas présent sur e-rara.ch. Par conséquent, d'autres études auraient encore tout leur sens pour confirmer et compléter nos conclusions.

Au vu des résultats décrits précédemment, le scénario idéal serait donc de recourir successivement aux deux méthodes et en cela nous rejoignons l'étude de Tullis (2002) et ce que nous en disions dans notre état de l'art. Les tests modérés permettent de repérer la plupart des problèmes et les tests automatisés à distance viendraient enrichir la liste et chiffrer de manière quantitative l'impact des éléments détectés.

Dans le cas où les ressources à disposition exigeraient de se limiter à l'une des deux méthodes, nous estimons en général plus avantageux (notamment en termes d'efficience) de recourir à un test modéré. Par contre, dans le cas où il s'agirait de comparer deux versions d'un même site, un test non modéré à distance avec de nombreux participants est largement plus intéressant. En effet, l'analyse quantitative permet de déterminer très aisément quelle variante rencontre le plus de succès ou d'échec.

Le choix dépend aussi du type de données que l'on souhaite recueillir. Si au niveau de la détection des problèmes les deux méthodes se valent environ, les caractéristiques des résultats produits sont très différentes. D'un côté, le test automatisé repose sur une part d'interprétation de l'analyste qui s'efforce de comprendre ce qui a posé des difficultés à l'utilisateur, par exemple en révisant un flux de clics. En contrepartie, des chiffres sont fournis sur la gravité des problèmes en indiquant notamment le nombre de personnes touchées, ce qui n'est pas sans intérêt pour prendre des décisions qui ont un certain coût en vue d'apporter des modifications à un site.

Quant au test modéré, il décèle les problèmes de manière explicite grâce au think-aloud, mais repose uniquement sur un nombre restreint de témoignages. Difficile donc de décider s'il s'agit de cas isolés ou d'éléments réellement importants.

L'analyste ne doit donc pas uniquement se poser la question de savoir quelle méthode permettra de détecter le plus de problèmes d'utilisabilité, mais aussi quelle forme de résultats répondra de la manière la plus adéquate aux besoins des propriétaires du site concerné.

BIBLIOGRAPHIE

ANDREASEN, Morten Sieker, NIELSEN, Henrik Villemann, SCHRØDER, Simon Ormholt et STAGE, Jan, 2007. What Happened to Remote Usability Testing? : An Empirical Study of Three Methods. In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* [en ligne]. New York, NY, USA : ACM. 2007. pp. 1405–1414. [Consulté le 14 mars 2014]. CHI '07. Disponible à l'adresse : http://doi.acm.org/10.1145/1240624.1240838

BARNUM, Carol M. (éd.), 2011. *Usability testing essentials: ready, set... test !*. Amsterdam : Elsevier. ISBN 9780123750921.

BASTIEN, J. M. Christian, 2010. Usability testing: a review of some methodological and technical aspects of the method. *International Journal of Medical Informatics*. avril 2010. Vol. 79, n° 4, pp. e18-e23. DOI 10.1016/j.ijmedinf.2008.12.004.

BOREN, T. et RAMEY, J., 2000. Thinking aloud : reconciling theory and practice. *IEEE Transactions on Professional Communication*. septembre 2000. Vol. 43, n° 3, pp. 261-278. DOI 10.1109/47.867942.

CHALIL MADATHIL, Kapil et GREENSTEIN, Joel S., 2011. Synchronous Remote Usability Testing : A New Approach Facilitated by Virtual Worlds. In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* [en ligne]. New York, NY, USA : ACM. 2011. pp. 2225–2234. [Consulté le 15 octobre 2014]. CHI '11. ISBN 978-1-4503-0228-9. Disponible à l'adresse : http://doi.acm.org/10.1145/1978942.1979267

COOKE, Lynne, 2010. Assessing Concurrent Think-Aloud Protocol as a Usability Test Method: A Technical Communication Approach. *IEEE Transactions on Professional Communication*. septembre 2010. Vol. 53, n° 3, pp. 202-215. DOI 10.1109/TPC.2010.2052859.

DE JONG, M. D. T., SCHELLENS, P. J. et VAN DEN HAAK, M. J., 2004. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues : a methodological comparison. *Interacting with Computers*. 2004. Vol. 16, n° 6, pp. 1153-1170.

DE JONG, Menno D. T., SCHELLENS, Peter Jan et VAN DEN HAAK, Maaike J., 2003. Retrospective vs. concurrent think-aloud protocols : testing the usability of an online library catalogue. *Behaviour and Information Technology*. 2003. Vol. 22, n° 5, pp. 339-351.

DRAY, Susan et SIEGEL, David, 2004. Remote Possibilities? : International Usability Testing at a Distance. *interactions*. mars 2004. Vol. 11, n° 2, pp. 10–17. DOI 10.1145/971258.971264.

ERICSSON, Karl Anders, 1993. *Protocol analysis : verbal reports as data*. Rev. ed. Cambridge Mass. [etc.] : The MIT Press. A Bradford book. ISBN 0262050471.

FERNANDEZ, Adrian, INSFRAN, Emilio et ABRAHÃO, Silvia, 2011. Usability evaluation methods for the web : A systematic mapping study. *Information and Software Technology*. août 2011. Vol. 53, n° 8, pp. 789-817. DOI 10.1016/j.infsof.2011.02.007.

GREINER, Katie, 2012. *A Comparison of two concurrent think-aloud protocols : Categories and relevancy of utterances* [en ligne]. Thesis. [Consulté le 14 mars 2014]. Disponible à l'adresse : https://ritdml.rit.edu/handle/1850/15950

HARTSON, H. Rex, CASTILLO, José C., KELSO, John et NEALE, Wayne C., 1996. Remote Evaluation : The Network As an Extension of the Usability Laboratory. In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* [en ligne]. New York, NY, USA : ACM. 1996. pp. 228–235. [Consulté le 20 mars 2014]. CHI '96. Disponible à l'adresse : http://doi.acm.org/10.1145/238386.238511

KRAHMER, E. et UMMELEN, N., 2004. Thinking about thinking aloud : a comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication*. juin 2004. Vol. 47, n° 2, pp. 105-117. DOI 10.1109/TPC.2004.828205.

MCDONALD, Sharon, ZHAO, Tingting et EDWARDS, Helen M., 2013. Dual Verbal Elicitation : The Complementary Use of Concurrent and Retrospective Reporting Within a Usability Test. *International Journal of Human - Computer Interaction* [en ligne]. 2013. Vol. 29, n° 10. [Consulté le 16 octobre 2014]. Disponible à l'adresse : http://search.proquest.com/docview/1421973064/2BDB0E0BDC074EE2PQ/1?accountid=15920

MEYSTRE, Valérie et REY, Raphaël, 2014. *Tests d'utilisabilité : comparaison de deux méthodes appliquées au site e-rara.ch* [en ligne]. Haute école de gestion de Genève. [Consulté le 25 octobre 2014]. Disponible à l'adresse : http://doc.rero.ch/record/209599

NIELSEN, Janni, CLEMMENSEN, Torkil et YSSING, Carsten, 2002. Getting access to what goes on in people's heads? reflections on the think-aloud technique. In : *Proceedings of the second Nordic conference on Human-computer interaction* [en ligne]. ACM. 2002. pp. 101–110. [Consulté le 15 octobre 2014]. Disponible à l'adresse : http://dl.acm.org/citation.cfm?id=572033

NØRGAARD, Mie et HORNB\A EK, Kasper, 2006. What Do Usability Evaluators Do in Practice? An Explorative Study of Think-aloud Testing. In : *Proceedings of the 6th Conference on Designing Interactive Systems* [en ligne]. New York, NY, USA : ACM. 2006. pp. 209–218. [Consulté le 15 octobre 2014]. DIS '06. ISBN 1-59593-367-0. Disponible à l'adresse : http://doi.acm.org/10.1145/1142405.1142439

OLMSTED-HAWALA, Erica L., MURPHY, Elizabeth D., HAWALA, Sam et ASHENFELTER, Kathleen T., 2010. Think-aloud Protocols : A Comparison of Three Think-aloud Protocols for Use in Testing Data-dissemination Web Sites for Usability. In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* [en ligne]. New York, NY, USA : ACM. 2010. pp. 2381–2390. [Consulté le 16 octobre 2014]. CHI '10. ISBN 978-1-60558-929-9. Disponible à l'adresse : http://doi.acm.org/10.1145/1753326.1753685

RODRIGUEZ, Ania et RESNICK, Marc L., 2010. Head to Head : Remote Usability Testing Takes on Live Usability Testing in the HFES Ultimate Fighting Challenge. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 1 septembre 2010. Vol. 54, n° 11, pp. 759-762. DOI 10.1177/154193121005401103.

SCHMIDT, Eveline, 2013. Remote oder In-person Usability-Test ? [en ligne]. 2013. [Consulté le 3 janvier 2014]. Disponible à l'adresse : http://doc.rero.ch/record/208871?ln=fr

SELVARAJ, Prakaash, 2004. *Comparative study of synchronous remote and traditional in-lab usability evaluation methodes* [en ligne]. Master thesis, industrial and systems engineering. Blacksburg : Virginia Polytechnic Institute and State University. [Consulté le 16 mars 2014]. Disponible à l'adresse : http://vtechworks.lib.vt.edu/bitstream/handle/10919/9939/Thesis_Prakaash_Selvaraj.pdf?sequence=2

THOMPSON, Katherine E., ROZANSKI, Evelyn P. et HAAKE, Anne R., 2004. Here, There, Anywhere : Remote Usability Testing That Works. In : *Proceedings of the 5th Conference on Information Technology Education* [en ligne]. New York, NY, USA : ACM. 2004. pp. 132–137. [Consulté le 14 mars 2014]. CITC5 '04. Disponible à l'adresse : http://doi.acm.org/10.1145/1029533.1029567

TULLIS, Tom, FLEISCHMAN, Stan, MCNULTY, Michelle, CIANCHETTE, Carrie et BERGEL, Marguerite, 2002. An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. In : *Usability Professionals Association Conference* [en ligne]. Boston. 2002. [Consulté le 14 mars 2014]. Disponible à l'adresse : http://home.comcast.net/%7Etomtullis/publications/RemoteVsLab.pdf

VAN DEN HAAK, Maaike J., DE JONG, Menno D. T. et SCHELLENS, Peter Jan, 2007. Evaluation of an informational Web site : three variants of the think-aloud method compared. *Technical Communication*. 2007. Vol. 54, n° 1, pp. 58-71.

VAN DEN HAAK, Maaike J., DE JONG, Menno D. T. et SCHELLENS, Peter Jan, 2009. Evaluating municipal websites : A methodological comparison of three think-aloud variants. *Government Information Quarterly*. janvier 2009. Vol. 26, n° 1, pp. 193-202. DOI 10.1016/j.giq.2007.11.003.

ZHAO, Tingting, MCDONALD, Sharon et EDWARDS, Helen M., 2014. The impact of two different think-aloud instructions in a usability test : a case of just following orders? *Behavior & Information Technology*. 2014. Vol. 33, n° 2, pp. 163-183. DOI http://dx.doi.org/10.1080/0144929X.2012.708786

ZHAO, Tingting et MCDONALD, Sharon, 2010. Keep Talking : An Analysis of Participant Utterances Gathered Using Two Concurrent Think-aloud Methods. In : *Proceedings of the 6th Nordic Conference on Human-Computer Interaction : Extending Boundaries* [en ligne]. New York, NY, USA : ACM. 2010. pp. 581–590. [Consulté le 16 octobre 2014]. NordiCHI '10. ISBN 978-1-60558-934-3. Disponible à l'adresse : http://doi.acm.org/10.1145/1868914.1868979

NOTES

(1) A ce propos, voir notre rapport de recherche à l'adresse http://doc.rero.ch/record/209599?ln=fr, p. 14-17.

(2) Carte représentant les points les plus cliqués pour chaque page du test.