Psychologie française 69 (2024) 269–286

Disponible en ligne sur www.sciencedirect.com

ScienceDirect

et également disponible sur www.em-consulte.com

Article original

Adaptation française et propriétés psychométriques de l'échelle d'utilisabilité perçue des sites web « Design-oriented evaluation of perceived usability (DEEP) »

French adaptation and psychometric properties of the "Design-Oriented Evaluation of Perceived Usability (DEEP)" scale

G. Gronier[a,*], E. Lazure[b], I. Dussouet[b]

[a] *Luxembourg Institute of Science and Technology, 5, avenue Des Hauts Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg*

[b] *Onisep, Direction de l'offre de service et de la relation à l'usager, 12, mail Barthélemy-Thimonnier, CS 10450 Lognes, 77437 Marne-la-Vallée cedex 2, France*

INFO ARTICLE

Historique de l'article :
Reçu le 30 septembre 2022
Accepté le 16 octobre 2023

Mots clés :
Adaptation française
Analyses psychométriques
Utilisabilité
Questionnaire
Site web

RÉSUMÉ

Introduction. – Face à l'utilisation omniprésence et souvent incontournable de sites web, il est nécessaire que ceux-ci présentent la meilleure utilisabilité possible pour satisfaire au mieux les utilisateurs. Le questionnaire Design-Oriented Evaluation of Perceived Usability (DEEP), comprenant 19 items, permet d'évaluer la facilité d'utilisation perçue sur la base de six critères génotypiques : contenu, structure et architecture de l'information, navigation, effort cognitif, cohérence de la mise en page, et guidage visuel. À l'issue de la passation du DEEP, les concepteurs sont alors en mesure d'identifier les principaux facteurs du site web qui posent problème.

Objectif. – Le but de cette étude est de proposer une adaptation transculturelle en langue française du DEEP, et d'en évaluer les qualités psychométriques.

Méthode. – Quatre cent soixante-dix utilisateurs ont complété une enquête en ligne afin de donner leur avis sur le site web Onisep.fr. Cette enquête comprenait notamment les items traduits en

* Auteur correspondant.
Adresses e-mail : guillaume.gronier@list.lu (G. Gronier), elise.lazure1@gmail.com (E. Lazure), isabelle.dussouet@onisep.fr (I. Dussuet).

https://doi.org/10.1016/j.psfr.2023.10.003

français du DEEP (F-DEEP), ainsi que le questionnaire Net Promoter Score. Une analyse factorielle en composantes principales (ACP) avec rotation oblimin, une analyse de la cohérence interne, plusieurs analyses factorielles confirmatoires (AFC) ainsi qu'une étude de la validité convergente entre le F-DEEP et le NPS ont été menées.
Résultat. – L'ACP suggère une structure à 6 facteurs, qui correspond en partie aux 6 facteurs du questionnaire originale. La cohérence interne est très satisfaisante, avec un coefficient alpha de Cronbach de 0,944 et un oméga de McDonalds de 0,945. L'AFC qui obtient les indices d'ajustement les plus satisfaisants correspond au modèle de premier-ordre comprenant les 6 facteurs du F-DEEP. Toutes les corrélations calculées entre les facteurs du F-DEEP et le NPS sont positives et hautement significatives.
Conclusion. – Le F-DEEP présente de très bonnes qualités psychométriques, ce qui lui permet de pouvoir être utilisé par les chercheurs et les professionnels en psychologie ergonomique des interactions humain-machine, pour l'évaluation des sites web auprès des publics francophones. Ce questionnaire peut également être appliqué à l'évaluation de l'utilisabilité des applications mobiles.

Keywords:
French adaptation
Psychometric analysis
Usability
Questionnaire
Website

ABSTRACT

Introduction. – With the ubiquitous and often unavoidable use of websites, it is necessary that they present the best possible usability to satisfy users. The Design-Oriented Evaluation of Perceived Usability (DEEP) questionnaire, comprising 19 items, assesses perceived usability on the basis of six genotypic criteria: content, structure and information architecture, navigation, cognitive effort, layout consistency, and visual guidance. After completing the DEEP, designers are then able to identify the main factors of the website that cause problems.
Objective. – The aim of this study was to propose a cross-cultural adaptation of the DEEP in French and to evaluate its psychometric qualities.
Method. – Four hundred and seventy users completed an online survey to give their opinion on the Onisep.fr website. The survey included the French translation of the DEEP (F-DEEP) and the Net Promoter Score questionnaire. A principal component analysis (PCA) with oblimin rotation, an internal consistency analysis, several confirmatory factor analyses (CFA) and a convergent validity study between the F-DEEP and the NPS were conducted.
Results. – The PCA suggests a 6-factor structure, which partly corresponds to the 6 factors of the original questionnaire. The internal consistency is very satisfactory, with a Cronbach's alpha coefficient of 0.944 and a McDonalds omega of 0.945. The CFA with the most satisfactory fit indices corresponds to the first-order model including the 6 factors of the F-DEEP. All the correlations calculated between the F-DEEP factors and the NPS are positive and highly significant.
Conclusion. – The F-DEEP has very good psychometric qualities, which means that it can be used by researchers and professionals in the field of ergonomic psychology of human-computer interaction to evaluate websites for French-speaking audiences. This questionnaire can also be applied to the evaluation of the usability of mobile applications.

1. Introduction

L'utilisabilité des systèmes interactifs, tels que les logiciels, les applications mobiles et les sites web, représente depuis les années 1990 (Nielsen, 1993) un champ de recherche et d'application en ergonomie (Brangier, Dufresne & Hammes-Adelé, 2009), en psychologie ergonomique (Dubois et Bobillier-Chaumon, 2009) et en interactions humain-machine (IHM). Même s'il est difficile de trouver une définition unique de la notion d'utilisabilité (Barcenilla et Bastien, 2009) l'une des plus usuelles est celle de la norme ISO 9241-11 (2018), qui caractérise l'utilisabilité comme le « degré selon lequel un produit peut être utilisé, par des utilisateurs identifiés, pour atteindre des buts définis avec efficacité, efficience et satisfaction, dans un contexte d'utilisation spécifié ».

La norme ISO/IEC 25010:2011 (2011), qui s'intéresse à la qualité des systèmes, complète cette définition en précisant que l'utilisabilité comprend six principaux facteurs :

- l'aptitude à la reconnaissance : il s'agit de la capacité d'un utilisateur à reconnaître si un produit ou un système est adapté à ses besoins ;
- la capacité d'apprentissage : il s'agit de la facilité avec laquelle un utilisateur peut apprendre à utiliser un produit ou un système ;
- l'opérabilité : il s'agit de savoir si un produit ou un système possède des caractéristiques qui en facilitent l'usage et le contrôle ;
- la protection contre les erreurs de l'utilisateur : il s'agit de la façon dont un système protège les utilisateurs contre les erreurs qu'ils peuvent commettre ;
- l'esthétique de l'interface utilisateur : elle concerne le caractère agréable de l'interface utilisateur ;
- l'accessibilité : il s'agit de la capacité d'un produit ou d'un système à être utilisé avec le plus grand nombre de caractéristiques et de capacités possible.

Sur la base des normes ISO relatives à l'utilisabilité, Abran, Khelifi & Suryn (2003) propose un modèle de l'utilisabilité renforcé, comprenant l'efficacité, l'efficience, la satisfaction, la facilité d'apprentissage mais aussi la sécurité comme facteur garantissant la prévention des accès non autorisés (piratage, virus, etc.).

Dans une démarche similaire, Quesenbery (2003) relève cinq dimensions qui caractérisent l'utilisabilité : l'efficacité, l'efficience, l'engagement, la tolérance à l'erreur et la facilité d'apprentissage. Dans ce contexte, l'engagement doit être compris comme le degré selon lequel le ton et le style de l'interface rendent le produit agréable ou satisfaisant à utiliser.

Plus récemment, plusieurs études ont cherché à instancier le concept d'utilisabilité aux sites web, afin de mieux répondre aux spécificités de ce type de système. Les enjeux sont importants car les sites web sont probablement l'une des technologies les plus fréquemment utilisés par tous les profils d'utilisateurs. Ainsi, Flavián, Guinalíu & Gurrea (2006) considèrent que l'utilisabilité d'un site web doit favoriser : (1) la facilité de compréhension de la structure d'un site web, de ses fonctions, de son interface et des contenus que l'utilisateur peut observer ; (2) la simplicité d'utilisation du site web dans ses phases initiales ; (3) la rapidité avec laquelle les utilisateurs peuvent trouver ce qu'ils recherchent ; (4) la facilité perçue de la navigation sur le site en termes de temps requis et d'action nécessaire pour obtenir les résultats souhaités ; et (5) la capacité de l'utilisateur à contrôler ce qu'il fait et où il se trouve à tout moment. Belanche, Casaló & Guinaliu (2012) soulignent quant à eux que l'utilisabilité d'un site web commercial (e-commerce) aura un impact sur les attitudes de ses utilisateurs et sur le taux de conversion. Elle renforcera la fidélité, l'intention de revisiter le site, et par conséquent son acceptation (Amar Raju, Roy & Mandal, 2018).

Évaluer l'utilisabilité est dès lors un enjeu essentiel pour garantir la qualité ergonomique des systèmes interactifs, dans le cadre d'un processus d'amélioration itératif et continu. Plusieurs méthodes d'évaluation de l'utilisabilité existent (Fernandez, Insfran & Abrahão, 2011), dont celles basées sur des modèles (comme par exemple, le *Cognitive Task Analysis* ou GOMS), les méthodes d'inspection (comme par exemple, l'évaluation heuristique, l'inspection ergonomique ou le *Cognitive Walkthrough*), les tests

utilisateurs (Bastien, 2010) et les méthodes d'enquêtes (comme par exemple, les entretiens ou les Focus Group).

Parmi les méthodes d'enquêtes, les questionnaires standardisés (Sauro et Lewis, 2012) sont un des outils privilégiés d'évaluation car ils sont rapides à faire passer et permettent d'obtenir un score de l'utilisabilité. Ce score pourra être comparé au score d'autres systèmes, au score des différentes versions d'un même système, ou aux scores obtenus par différents profils d'utilisateurs dans une comparaison sociodémographique. La norme ISO 16982:2002 (2002) référence ainsi les questionnaires parmi l'une des dix principales méthodes de la mesure de l'utilisabilité, et les définit comme des « méthodes d'évaluation indirecte qui recueillent, au moyen de questionnaires prédéfinis, les opinions des utilisateurs sur l'interface ». Maguire (2001) référence aussi les questionnaires parmi les méthodes de conception centrée sur l'humain.

Parmi les questionnaires qui mesurent l'utilisabilité des sites web, beaucoup sont dédiés à des sites particuliers : le Website Evaluation Questionnaire (WEQ) (Elling, Lentz, de Jong & van den Bergh, 2012) s'intéresse à l'utilisabilité perçue des sites gouvernementaux ; le WebQual (Loiacono, Watson & Goodhue, 2007) et le Perceived Web Quality (PWQ) (Aladwani et Palvia, 2002) sont consacrés à l'évaluation des sites de commerce ; le questionnaire Academic Social Networking Sites (ASNS) (Koranteng, Ham, Wiafe & Matzat, 2021) est spécialisé dans l'évaluation des sites de réseaux sociaux universitaires. Le *Website Analysis and MeasureMent Inventory* (WAMMI) (Kirakowski et Cierlik, 1998) s'adresse quant à lui à tous les types de sites web. Il différencie l'attractivité, la contrôlabilité, l'efficacité, l'utilité et la facilité d'apprentissage d'un site à l'aide de 60 items. Les auteurs soulignent que compléter ce nombre important d'items n'est pas une tâche que l'utilisateur peut s'acquitter de son plein gré, sans compensation ou recrutement spécifique, car elle demande du temps et un effort soutenu.

D'autres questionnaires de mesure de l'utilisabilité s'appliquent à des systèmes plus génériques que les sites web. Le Purdue Usability Testing Questionnaire (PUTQ) (Lin, Choong & Salvendy, 1997) comprend 100 items et mesure 8 aspects d'une interface humain-machine : la compatibilité, la consistance, la flexibilité, la facilité d'apprentissage, les actions minimales, la charge de mémoire minimale, la limite perceptive et le guidage. Tout comme le WAMMI, le grand nombre d'items du PUTQ ne permet pas de l'administrer à des utilisateurs qui le compléteraient de façon spontanée, sans un protocole expérimental prédéfini. Le questionnaire Usefulness, Satisfaction, and Ease of use (USE) (Lund, 2001), comprenant 30 items, mesure quant à lui 4 principales dimensions d'une interface, comprenant l'utilité, la facilité d'utilisation, la facilité d'apprentissage et la satisfaction.

D'autres questionnaires plus répandus, comme le *System Usability Scale* (SUS) (Brooke, 1996), le Computer System Usability Questionnaire (CSUQ) (Lewis, 1995) ou l'Usability Metric for User Experience (UMUX) (Finstad, 2010), ne permettent pas d'avoir une vue détaillée sur les problèmes d'utilisabilité qui peuvent être relevés par les utilisateurs. En effet, le SUS et l'UMUX offrent un score unique d'utilisabilité, compris entre 0 et 100, qui peut être interprété en termes de qualité de convivialité. (Bangor, Kortum & Miller, 2009) ont ainsi calculé qu'un score supérieur à 86/100 signifiait que le système évalué était jugé excellent ; à l'inverse, un système était jugé mauvais en dessous de 39/100. Néanmoins, ce score ne permet pas d'établir l'origine des problèmes d'utilisabilité. Le CSUQ propose quant à lui une évaluation de trois principaux aspects d'un système : son utilité, la qualité de l'information et la qualité de l'interface. Si cette distinction permet de poser un diagnostic plus précis sur l'utilisabilité d'un système que le SUS ou l'UMUX, ce questionnaire s'applique toutefois davantage aux outils informatiques en entreprise qu'aux sites web. Plusieurs items s'intéressent en effet à l'efficacité au travail, comme par exemple l'item 3 « Je peux faire mon travail efficacement en utilisant ce système » (Gronier et Johannsen, 2022).

1.1. Le questionnaire Design-Oriented Evaluation of Perceived Usability (DEEP)

Le questionnaire Design-Oriented Evaluation of Perceived Usability (DEEP) (Yang, Linder & Bolchini, 2012) a été construit pour évaluer l'utilisabilité de tous les types de sites web, en proposant une structure qui permet de relever les éléments d'un site qui posent un problème aux utilisateurs. C'est en ce sens que ce questionnaire est orienté vers la conception (*Design-Oriented*). Les auteurs introduisent à ce titre la notion d'alignement évaluation-conception. Il s'agit de la capacité d'un instrument d'utilisabilité à conduire progressivement l'activité d'évaluation vers des exigences de reconception.

Ainsi, plus le *feedback* de l'évaluation est générique et ne permet pas de recommander des exigences de conception exploitables, plus l'alignement évaluation-conception est faible.

L'objectif du DEEP est alors de rapprocher la perception d'un problème d'utilisabilité à sa cause. En effet, un problème d'utilisabilité qui est un obstacle à une utilisation efficace, efficiente et satisfaisante d'un système, comporte invariablement deux composantes : la manifestation perçue de l'obstacle sur l'expérience réelle de l'utilisateur, appelée le phénotype d'utilisabilité, et le défaut réel de conception du système qui cause le problème, appelé le génotype d'utilisabilité (Lavery, Cockton & Atkinson, 1997). La plupart des questionnaires d'utilisabilité restent cantonnés au niveau du phénotype. Le DEEP cherche à identifier le génotype afin de réduire l'écart entre l'utilisabilité perçue et les conseils pour améliorer la conception du site web.

Le questionnaire DEEP contient 19 items sous la forme de phrases affirmatives, pour chacun desquels l'utilisateur est invité à se positionner sur une échelle de Likert allant de 1 « Pas du tout d'accord » à 5 « Tout à fait d'accord ». Une option non applicable offre la possibilité de ne pas répondre aux items qui ne correspondent pas au système évalué.

Les 19 items sont répartis en 6 dimensions : 1. le contenu perçu à 4 items ; 2. la structure et architecture de l'information à 3 items ; 3. la navigation perçue à 3 items ; 4. l'effort cognitif perçu à 3 items ; 5 la cohérence de la mise en page perçue à 3 items ; 6. le guidage visuel perçu à 3 items.

La conception du DEEP s'est appuyée sur plusieurs questionnaires d'utilisabilité dont le Purdue Usability Testing Questionnaire (PUTQ) (Lin et al., 1997), le *Website Analysis and MeasureMent Inventory* (WAMMI) (Kirakowski et Cierlik, 1998) et le USE (Lund, 2001). Les critères d'inspection des sites web MILE+ (Bolchini et Garzotto, 2008) ont également été utilisés. Plusieurs itérations basées sur des analyses psychométriques ont permis de ne retenir que les 19 items les plus pertinents pour l'évaluation des sites web.

Les auteurs du DEEP concluent leur article sur un ensemble de recommandations qui permettra de corriger certains aspects du site en fonction du score obtenu pour chaque item. Par exemple, si la moyenne des réponses à l'item « La formulation du texte était claire » est inférieure à 3, alors il est conseillé de rechercher des moyens de simplifier le style et la langue du texte. Des liens sont proposés vers des références en ligne, mais ces liens n'étaient plus accessibles au moment où cette étude a été menée.

2. Objectifs

Cette étude propose de tester la validité d'une adaptation française du questionnaire DEEP, nommée F-DEEP. Le premier objectif de l'étude est d'adapter en français le DEEP selon la méthodologie proposée par Vallerand (1989) et Gana, Boudouda, Ben Youssef, Calcagni & Broc (2021). Le second objectif est de vérifier les propriétés psychométriques du F-DEEP.

3. Méthodologie

3.1. Adaptation française du questionnaire DEEP

Nous avons tout d'abord demandé et obtenu l'accord des auteurs du DEEP (Yang et al., 2012) pour procéder à son adaptation française.

Nous avons ensuite procédé à la traduction en français du DEEP en suivant une approche par comité, comme le recommande Vallerand (1989). Dans ce cadre, trois chercheurs en ergonomie des interactions humain-machine (IHM), bilingues de langue natale française, ont tout d'abord procédé individuellement à une traduction des items. Ensuite, chaque traduction a été confrontée aux deux autres, puis discutée de manière à ce qu'une traduction consensuelle soit trouvée pour chaque item. En accord avec Gana et al. (2021), il était demandé aux traducteurs de chercher la meilleure approximation possible du sens original, et de ne pas procéder obligatoirement à une traduction mot à mot. L'objectif était ainsi d'ajuster la signification de l'item source, en anglais, aux spécificités culturelles et linguistiques de la langue française.

Une contre-traduction, du français vers l'anglais, a ensuite été réalisée par deux autres chercheurs bilingues, mais de langue natale anglaise. Cette dernière étape a permis d'approuver la traduction française.

La traduction des items est présentée dans le Tableau 1.

3.2. Analyse des propriétés psychométriques du F-DEEP

3.2.1. Cas d'étude

Afin d'analyser les qualités psychométriques du F-DEEP dans un contexte naturel (versus expérimental) et auprès d'utilisateurs réels d'un site web (versus des utilisateurs interrogés uniquement pour la validation du questionnaire), nous nous sommes associés à l'Office national d'information sur les enseignements et les professions (ONISEP). L'Onisep est un opérateur de l'État qui relève du ministère de l'Éducation nationale, de la Jeunesse et des Sports, et du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Éditeur public, l'Onisep produit et diffuse toute l'information sur les formations et les métiers. Il propose aussi des services aux élèves, aux parents et aux équipes éducatives.

Le site web onisep.fr, qui a servi de cas d'étude, présente des actualités nationales et régionales sur les formations et les métiers, la géolocalisation des lieux de formation et d'information, les structures dédiées à l'information ou à la scolarisation des élèves en situation de handicap. Il propose des informations ciblées « collège », « lycée, CFA », « après le bac », « parents », « équipes éducatives », « handicap » ainsi que des espaces thématiques consacrés à l'enseignement professionnel, les formations du sport et leurs débouchés, ou encore les études scientifiques.

3.2.2. Procédures de diffusion de l'enquête

L'objectif de l'étude était de connaître la satisfaction des visiteurs du site Onisep.fr et d'identifier des potentiels points de blocage en termes d'utilisabilité. Une enquête a été élaborée avec l'outil Sphinx. Un lien vers l'enquête a été publié durant 2 semaines en bas de 6700 pages, sous la forme d'un visuel cliquable invitant les visiteurs à donner leur avis sur le site. En cliquant sur le visuel, les participants étaient redirigés vers une nouvelle page contenant l'enquête.

Les données ont été collectées de façon anonyme et il n'y avait pas de critères d'inclusion. Aucune compensation n'était proposée. L'enquête contenait plusieurs questions relatives au profil du répondant (genre, âge, profession, degré et cursus de scolarisation), ainsi que les questionnaires NPS, un item unique de mesure de l'utilisabilité, et le F-DEEP.

3.2.3. Participants

L'échantillon de notre étude était composé de 470 participants dont 274 femmes (58 %), 113 hommes (24 %), 64 personnes qui n'ont pas souhaité exprimer leur genre (14 %) et 19 personnes qui se considéraient comme faisant partie d'une autre catégorie sexuelle (autre) (4 %). La taille de l'échantillon est supérieure à la recommandation de 10 participants par item suggérée par Nunnally (1978), c'est-à-dire dans notre cas 190 participants. Elle est aussi proche des 500 répondants nécessaires selon Gana et al. (2021) pour la validation des qualités psychométriques d'une échelle traduite.

Parmi les répondants, 7 avaient moins de 12 ans (1 %), 292 avaient entre 12 et 17 ans (62 %), 75 avaient entre 18 et 34 ans (16 %), 80 avaient entre 35 et 64 ans (17 %) et 16 avaient plus de 65 ans (3 %). 314 des répondants étaient des élèves (67 %), 42 des étudiants (9 %), 82 des parents d'élèves (17 %) et 32 des membres des équipes éducatives (7 %) (Tableau 2).

3.2.4. Instruments de mesure de la validité

Deux instruments complémentaires au F-DEEP ont été diffusés dans le cadre de la même enquête et ont permis de mesurer une validité nomologique. De manière générale, la validité d'un questionnaire est établie en analysant les relations entre les scores de ce questionnaire et les scores à des construits qui lui sont proches et reliés (Gana et al., 2021). La validité nomologique consiste à comparer les scores à des questionnaires qui ne mesurent pas directement le même concept, mais qui s'intéressent à des concepts proches ou qui peuvent être sémantiquement liés. Dans ce cadre, les scores au F-DEEP ont été

Tableau 1
Traduction en français des items du DEEP dans une approche par comité.

	Items originaux en anglais	Traduction des items en français
Contenu perçu		
1	*The wording of the text was clear*	La formulation du texte était claire
2	*The content (including text, pictures, audios, and videos, etc.) was easy to understand*	Le contenu (texte, images, sons, vidéos, etc.) était facile à comprendre
3	*The text was useful*	Le texte était utile
4	*The text was relevant*	Le texte était pertinent
Structure perçue et architecture de l'information		
5	*I could quickly get to know the structure of the website by skimming its home page*	Je me suis rapidement familiarisée avec la structure du site en parcourant sa page d'accueil
6	*The organization of the website was clear*	Le site était bien organisé
7	*Under each section of the website, the web pages were well organized*	Dans chaque section du site, les pages étaient bien organisées
Navigation perçue		
8	*It was easy to find the information I needed on the website*	J'ai facilement trouvé l'information que je cherchais sur le site
9	*This website helped me find what I was looking for*	Ce site m'a permis de trouver ce que je cherchais
10	*I got what I expected when I clicked on things on this website*	J'ai trouvé ce que je cherchais en cliquant sur les éléments du site
Effort cognitif perçu		
11	*Using this website was effortless*	L'utilisation de ce site ne m'a pas posé de difficultés
12	*Using this website made me feel tired*	J'ai trouvé l'utilisation de ce site éprouvante
13	*I learned to use this website quickly*	J'ai appris à utiliser ce site rapidement
Cohérence de la mise en page perçue		
14	*The layout of the pages throughout the website was consistent*	La mise en page était cohérente sur l'ensemble du site
15	*I noticed abrupt changes in the layout of the pages*	J'ai remarqué des changements surprenants dans la mise en page
16	*The layout under each section of the website was consistent*	La mise en page dans chaque section du site était cohérente
Guidage visuel perçu		
17	*The colors helped me to distinguish different sections of the website*	Les couleurs m'ont aidé à distinguer les différentes sections du site
18	*The highlighted areas of a page helped me locate the information I needed*	Les parties en évidence m'ont aidé à trouver l'information que je cherchais
19	*I got to know the content of a page by skimming the highlighted areas*	Je me suis familiarisé·e avec le contenu d'une page en parcourant les parties mises en évidence

Tableau 2
Données sociodémographiques de l'échantillon.

Variables	(n = 470)	%
Genre		
Homme	113	24
Femme	274	58
Ne souhaite pas répondre	64	14
Autre	19	4
Groupes d'âge (ans)		
< 12	7	1
12–17	292	62
18–34	75	16
35–64	80	17
≥ 65	16	3
Profil		
Élève	314	67
Étudiant	42	9
Parent d'élève	82	17
Équipe éducative	32	7

comparés au score d'un item unique mesurant le concept d'utilisabilité (Christophersen et Konradt, 2011) et au Net Promoter Score (NPS) (Reichheld, 2003).

Ces deux instruments sont présentés dans les sections suivantes.

3.2.4.1. L'item unique de Christophersen et Konradt (2011). L'item unique proposé par Christophersen et Konradt (2011) correspond à l'affirmation « De manière générale, je suis satisfait·e de ce site web », face à laquelle les utilisateurs sont invités à se positionner sur une échelle de Likert. Pour notre étude, les possibilités de réponse allaient de 1 « Pas du tout d'accord » à 10 « Tout à fait d'accord ».

Cette échelle a tout d'abord été choisie afin de réduire au maximum le nombre de questions contenu dans l'enquête générale menée auprès des utilisateurs, tout en conservant une validité scientifique de mesure de l'utilisabilité. Christophersen et Konradt (2011) ont en effet démontré que cet item unique disposait d'une bonne fiabilité, validité de contenu, validité prédictive et une sensibilité de mesure satisfaisante pour la mesure de l'utilisabilité.

Cette échelle nous a semblé également intéressante car elle a été conçue et appliquée pour mesurer l'utilisabilité des sites de commerce. En termes d'utilisabilité, les sites commerciaux nécessitent ainsi que soit prises en compte non seulement la facilité d'utilisation, mais aussi la confiance accordée au site, l'appréciation de son esthétique, et la loyauté perçue en termes d'intention d'usage régulier ou d'intention d'achat (Christophersen et Konradt, 2011). Ces caractéristiques nous semblaient correspondre à notre cas d'étude, puisque le site de l'Onisep doit répondre à des critères de 1. facilité d'utilisation car il s'adresse à un large public parfois peu à l'aise avec les technologies ; 2. confiance envers les informations fournies car il doit présenter de manière exhaustive toutes les formations et métiers qui existent en France ; 3. esthétique en tant que composante de la satisfaction et de facilité d'utilisation perçue (Tractinsky et al., 2000) ; 4. loyauté puisque le site doit apporter une valeur suffisamment importante à ses utilisateurs pour qu'ils aient envie de le consulter à nouveau et le promouvoir autour d'eux (Casalo, Flavian & Guinaliu, 2008).

3.2.4.2. Le Net Promoter Scale. Le Net Promoter Score (NPS) (Reichheld, 2003) est une échelle qui mesure l'intention de promouvoir une entreprise, un produit ou un service. Il est composé d'un seul item correspondant à la question « Quelle est la probabilité que vous recommandiez [cette entreprise/ce produit/ce service] à un ami ou un collègue ? ». Les participants sont invités à y répondre sur une échelle de Likert allant de 0 « Pas du tout probable » à 10 « Extrêmement probable ». En fonction du score qu'ils ont coché, les répondants sont classés dans l'une des trois catégories suivantes : détracteurs pour un score compris entre 0 et 6 ; passifs pour un score de 7 ou 8 ; promoteurs pour un score de 9 ou 10.

Bien que quelque fois critiquée en raison, notamment, de l'absence d'indicateurs sur les pistes d'améliorations possibles du produit évalué (Fisher et Kordupleski, 2019), le NPS reste toutefois une échelle très utilisée en ergonomie, et possède une corrélation élevée avec les échelles de mesure de l'utilisabilité. Par exemple, Borsci, Buckle & Walne(2020) ont mesuré une corrélation positive significative entre l'UMUX-LITE et le NPS ($r = 0{,}455$, $n = 116$, $p < 0{,}001$). D'autres recherches ont également combiné le NPS avec une autre échelle d'utilisabilité comme le *System Usability Scale* (SUS) (Brooke, 1996) afin de mesurer l'utilisabilité globale d'un système (Ismail, Nalawati & Putra, 2021 ; Pradini, Kriswibowo & Ramdani, 2019 ; Sasmito, Zulfiqar & Nishom, 2019).

Pour les besoins de notre étude, le NPS a été traduit par la question « Recommanderiez-vous ce site internet autour de vous pour s'informer sur l'orientation ? ». Les participants étaient invités à se positionner sur une échelle allant de 0 « Je ne recommande absolument pas » à 10 « Je recommande absolument ».

3.2.5. Analyses statistiques

Toutes les analyses statistiques ont été réalisées à l'aide des logiciels SPSS 26 et de JASP 0.16.2.

Nous avons suivi le cadre méthodologique dédié à l'évaluation psychométrique des adaptations transculturelles des échelles en psychologie proposée par Gronier (2022). Les analyses exploratoires ont été réalisées à l'aide d'une analyse en composantes principales (ACP) avec rotation oblique oblimin et d'une analyse factorielle exploratoire (EFA) ave rotation oblique oblimin. La cohérence interne a été mesurée avec les alphas de Cronbach et l'oméga de McDonald (Hayes et Coutts, 2020). Selon Nunnally

(1978), le coefficient minimum acceptable ne doit pas être inférieur à 0,70. Pour les statistiques descriptives et afin d'évaluer un éventuel effet de l'âge et du sexe, nous avons effectué le test *t* de Student et des ANOVA à mesures répétées.

Nous avons également testé le modèle structurel avec une analyse factorielle confirmatoire (AFC). Pour cette analyse, la qualité de l'ajustement a été testée avec le test du Chi^2 et un Chi^2 normalisé, une statistique d'ajustement dérivée moins dépendante de la taille de l'échantillon. Le Chi^2 normalisé est calculé en divisant l'indice du Chi^2 par le degré de liberté. En outre, selon Schweizer (2010), nous avons choisi quatre autres indices d'ajustement pour l'analyse : le résidu quadratique moyen normalisé (SRMR), l'erreur d'approximation quadratique moyenne (RMSEA), l'indice d'ajustement comparatif (CFI) et l'index de Tucker-Lewis (TLI). On suppose que le modèle peut être considéré comme satisfaisant lorsque le Chi^2/df est inférieur à 3 ; le SRMR et le RMSEA respectivement inférieurs à 0,06 et 0,08 ; et le CFI et le TLI respectivement supérieurs à 0,90 et 0,95 (Hu et Bentler, 1999 ; Steiger, 2007 ; Tabachnick et Fidell, 2019).

4. Résultats

4.1. Analyse exploratoire

4.1.1. Analyse factorielle en composantes principales

Une analyse en composantes principales (ACP) a été réalisée pour tester la validité de construction du questionnaire F-DEEP. L'objectif de l'ACP est de vérifier si la structure factorielle de notre traduction est similaire à celle de l'échelle originale (Yang et al., 2012).

Tout d'abord, nous obtenons un indice de Kaiser-Meyer-Olkin (KMO) de 0,942, avec un test de sphéricité de Bartlett hautement significatif ($p < 0{,}000$). Ceci nous permet de nous assurer que les items du F-DEEP sont fortement corrélés entre eux et de poursuivre l'analyse en composantes principales. Nous avons ensuite procédé à une analyse par composantes principales en nous appuyant sur la méthode d'analyse parallèle proposée par Horn (1965). L'analyse parallèle est basée sur la comparaison des valeurs propres des données réelles avec celles de données simulées. Avec cette analyse, l'accent est mis sur le nombre de facteurs obtenus à partir des données réelles dont la valeur propre est supérieure à celle des données simulées, ce qui permet de déterminer le nombre de facteurs.

L'ACP par rotation oblique oblimin suggère une structure du F-DEEP à 6 facteurs. En accord avec Nunnally (1978), seules les valeurs supérieures à 0,3 sont à retenir pour l'analyse (Tableau 3).

On observe plusieurs similitudes entre la structure factorielle du DEEP (Yang et al., 2012) et celle du F-DEEP. Tout d'abord, le facteur 2 regroupe les items 8, 9 et 10, qui correspondent pour le DEEP à l'utilisabilité perçue relative à la navigation. Le facteur 3 est composé des items 14, 15 et 16 relatifs pour le DEEP à la cohérence de la mise en page perçue. Le facteur 4 regroupe les items 17, 18 et 19 qui correspondent pour le DEEP au guidage visuel perçu. Les facteurs 5 et 6 regroupent les items 1, 3 et 4, qui pour le DEEP concernent le contenu perçu. Néanmoins, l'ACP du F-DEEP distinguent d'une part avec le facteur 5 les items qui concernent plus directement l'intérêt perçu du texte (« Le texte était utile », « Le texte était pertinent »), et d'autre part avec le facteur 6, les items qui traitent du contenu de façon plus générale (« La formulation du texte était claire », « Le contenu (texte, images, sons, vidéos, etc.) était facile à comprendre »). Ainsi, nous nous proposons de nommer le facteur 5 « Utilité du texte perçue », et le facteur 6 « Contenu perçu ».

Pour finir, le facteur 1 regroupe les items 5, 6, 11, 12 et 13, qui font partie de 2 dimensions distinctes dans le DEEP : la structure perçue et l'architecture de l'information pour les items 5 et 6, et l'effort cognitif perçu pour les items 11, 12 et 13. Pour le F-DEEP, nous nous proposons de nommer le facteur 1 « Facilité d'utilisation perçue », termes qui nous semble regrouper de façon suffisamment générique les deux dimensions du DEEP.

Ces similitudes et différences entre le DEEP et le F-DEEP seront débattues dans la section Discussion.

4.1.2. Analyse factorielle exploratoire

Puisque le modèle du F-DEEP ne correspond pas exactement à celui du DEEP, nous avons procédé à une analyse factorielle exploratoire (AFE) afin d'identifier les éventuels facteurs latents de notre traduction. L'objectif principal de l'AFE est en effet d'extraire les variables ou facteurs latents sous-

Tableau 3
Charges factorielles de la version française du DEEP issues de l'ACP par rotation oblique oblimin (les charges supérieures à 0,3 sont surlignées en gras).

	Facteur 1 (effort d'utilisation)	Facteur 2 (navigation)	Facteur 3 (cohérence de la mise en page)	Facteur 4 (guidage visuel)	Facteur 5 (utilité du texte)	Facteur 6 (contenu perçu)
Item 1					0,404	0,558
Item 2						0,461
Item 3					0,881	
Item 4					0,840	
Item 5	0,714					
Item 6	0,578					
Item 7						0,401
Item 8		0,733				
Item 9		0,910				
Item 10		0,783				
Item 11	0,754					
Item 12	0,755					
Item 13	0,761					
Item 14			0,502			
Item 15			0,935			
Item 16			0,693			
Item 17				0,921		
Item 18				0,694		
Item 19				0,623		

Tableau 4
Structure factorielle du F-DEEP issue de l'AFE par rotation oblique oblimin.

	Facteur 1	Facteur 2	Facteur 3	Facteur 4	Facteur 5	Facteur 6
Item 1		0,430				
Item 2		0,333				
Item 3		0,865				
Item 4		0,762				
Item 5					0,492	
Item 6					0,657	
Item 7					0,478	
Item 8	0,650					
Item 9	0,847					
Item 10	0,681					
Item 11						0,503
Item 12						0,835
Item 13						0,447
Item 14			0,527			
Item 15			0,716			
Item 16			0,677			
Item 17				0,630		
Item 18				0,879		
Item 19				0,586		

jacents d'une mesure en explorant la relation entre les variables observées (Roberson, Elliott, Chang & Hill, 2014).

Une analyse factorielle exploratoire avec rotation oblique oblimin a ainsi été réalisée. L'indice KMO est de 0,940 et le test de Bartlett suffisamment satisfaisants ($Chi^2 = 5835{,}125$; $p < 0{,}001$) pour conduire l'analyse (Cerny et Kaiser, 1977-XXX). L'analyse factorielle, calculée sur la base de la méthode d'analyse parallèle (Horn, 1965), permet d'identifier un modèle en six facteurs (Tableau 4).

La distribution des items dans les différents facteurs correspond au modèle du DEEP. Autrement dit, les 6 facteurs de l'EFA correspondent aux 6 dimensions du DEEP. Soulignons toutefois que l'item 2 « Le contenu (texte, images, sons, vidéos, etc.) était facile à comprendre » possède une saturation inférieure à 0,30.

Tableau 5
Statistiques descriptives et coefficients de fidélité pour les dimensions du F-DEEP.

Dimensions du DEEP	M	ET	Alpha de Cronbach	Oméga de McDonald
Contenu perçu	3,977	0,695	0,830	0,834
Structure perçue et architecture de l'information	3,578	0,907	0,849	0,851
Navigation perçue	3,409	0,982	0,863	0,865
Effort cognitif perçu	3,702	0,931	0,868	0,870
Cohérence de la mise en page perçue	3,669	0,777	0,823	0,826
Guidage visuel perçu	3,456	0,859	0,802	0,811

M : moyenne ; ET : écart-type.

Tableau 6
Statistiques descriptives et coefficients de fidélité pour les items de chaque dimension du F-DEEP si un item donné avait été supprimé.

Dimensions du DEEP	Item	M	ET	Alpha de Cronbach	Oméga de McDonald
Contenu perçu	1	3,996	0,773	0,804	0,807
	2	4,047	0,857	0,809	0,811
	3	3,976	0,910	0,763	0,766
	4	3,878	0,871	0,759	0,763
Structure perçue et architecture de l'information	5	3,582	1,101	0,829	0,832
	6	3,558	1,032	0,724	0,726
	7	3,587	0,970	0,810	0,814
Navigation perçue	8	3,316	1,152	0,832	0,835
	9	3,439	1,109	0,803	0,805
	10	3,472	1,047	0,789	0,792
Effort cognitif perçu	11	3,709	1,066	0,827	0,830
	12	3,670	1,093	0,787	0,789
	13	3,729	0,979	0,825	0,829
Cohérence de la mise en page perçue	14	3,717	0,925	0,758	0,762
	15	3,594	0,930	0,790	0,793
	16	3,705	0,837	0,720	0,723
Guidage visuel perçu	17	3,322	1,061	0,807	0,813
	18	3,467	1,036	0,652	0,656
	19	3,576	0,931	0,725	0,728

M : moyenne ; ET : écart-type.

4.1.3. Cohérence interne

La cohérence interne du F-DEEP a été testée à l'aide de l'alpha de Cronbach et de l'oméga de McDonald. L'alpha de Cronbach est de 0,944, tandis que l'oméga global de McDonald est de 0,945. Ces scores sont tous les deux supérieurs au seuil de 0,70 suggéré par Nunnally (1978) et attestent d'une fidélité très satisfaisante. Aucune suppression d'items ne permet d'améliorer la cohérence interne. Ces résultats sont proches de ceux obtenus par Yang et al. (2012), qui avaient calculé un alpha de Cronbach pour l'ensemble des items du DEEP de 0,954.

Nous avons ensuite calculé l'alpha et l'oméga des items pour chacune des 6 dimensions du F-DEEP (Tableau 5), ainsi que l'alpha et l'oméga pour les items de chaque dimension si un item donné avait été supprimé (Tableau 6). Les deux tableaux reprennent également les analyses descriptives relatives aux scores pour chaque dimension et pour chaque item.

4.1.4. Différences sociodémographiques

Nous nous sommes intéressés aux différences qui pouvaient exister entre le score total au F-DEEP et les caractéristiques sociodémographiques de notre échantillon.

Nous avons ainsi observé une différence significative selon le genre ($F(3, 466) = 5{,}666$, $p < 0{,}001$), selon la tranche d'âge ($F(7{,}462) = 3{,}128$, $p = 0{,}003$) et selon le profil du visiteur ($F(3, 466) = 2{,}936$, $p = 0{,}033$).

Tableau 7
Valeurs des indices d'ajustement pour les 4 modèles testés.

Modèle	Chi^2	*p*	ddl	Chi^2/ddl	TLI	CFI	SRMR	RMSEA
Modèle de premier-ordre	466,954	< 0,001	137	3,408	0,922	0,937	0,042	0,076
Modèle de second-ordre	534,827	< 0,001	146	3,663	0,913	0,926	0,049	0,080
Modèle unifactoriel	1252,128	< 0,001	152	8,238	0,764	0,791	0,071	0,131
Structure factorielle du DEEP	571,673	< 0,001	146	3,916	0,905	0,919	0,055	0,083

TLI : Index de Tucker-Lewis ; CFI : Indice d'ajustement comparatif ; SRMR : Résidu quadratique moyen normalisé ; RMSEA : Erreur d'approximation quadratique moyenne.

Tableau 8
Coefficients de corrélations de Pearson entre les 6 dimensions du F-DEEP (utilisation, navigation, guidage, cohérence, texte, contenu) et son score global (Global), et les scores à l'item unique (satisfaction) de Christophersen et Konradt (2011) et le Net Promoter Score (NPS) de Reichheld (2003).

	F-DEEP						
	Utilisation	Navigation	Guidage	Cohérence	Texte	Contenu	Global
Satisfaction	0,597[a]	0,613[a]	0,514[a]	0,496[a]	0,485[a]	0,518[a]	0,671[a]
NPS	0,575[a]	0,610[a]	0,508[a]	0,533[a]	0,551[a]	0,514[a]	0,676[a]

[a] La corrélation est significative au seuil de 0,001.

4.2. Analyse confirmatoire

4.2.1. Analyse factorielle confirmatoire

Plusieurs analyses factorielles confirmatoires (AFC) ont été conduites afin de comparer différents modèles du F-DEEP. Nous avons tout d'abord testé un modèle de premier-ordre, comprenant les 6 facteurs issus de l'ACP, puis une structure hiérarchique de second-ordre comprenant le F-DEEP comme facteur commun et regroupant les 6 facteurs de premier-ordre. Nous avons également testé un modèle unifactoriel, ainsi qu'un modèle reprenant l'organisation hiérarchique originale du DEEP.

Les résultats sont présentés dans le Tableau 7.

Le modèle le plus satisfaisant est celui de premier-ordre comprenant les 6 facteurs du F-DEEP (utilisation, navigation, guidage, cohérence, texte, contenu). Tous les indices sont en dessous des seuils prérequis, excepté le TLI (0,937) dont le seuil habituel est situé à 0,95 (Hu et Bentler, 1999). Néanmoins, cet index est considéré comme l'un des plus pénalisants, et plusieurs auteurs considèrent qu'un TLI supérieur à 0,90 reste satisfaisant (Wang et Wang, 2019).

L'analyse factorielle confirmatoire de premier-ordre est représentée dans la Fig. 1.

4.2.2. Validités nomologiques

La validité nomologique a été mesurée en comparant les scores des différentes dimensions du F-DEEP ainsi que son score global, avec l'item unique de Christophersen et Konradt (2011) et le Net Promoter Score (NPS) (Reichheld, 2003). Pour cela, un coefficient de corrélation de Pearson a été calculé. Les résultats sont présentés dans le Tableau 8.

Les résultats montrent une corrélation significative pour toutes les dimensions du F-DEEP et pour l'item unique de satisfaction, ainsi que pour le NPS. Par exemple, on observe un coefficient de corrélation de 0,671 ($p < 0{,}001$) entre le score global au F-DEEP et la satisfaction, et de 0,676 ($p < 0{,}001$) pour le NPS.

5. Discussion

Cette étude avait pour objectif de proposer une adaptation française du questionnaire de mesure de l'utilisabilité des sites web « Design-Oriented Evaluation of Perceived Usability (DEEP) », définie par Yang et al. (2012), et de vérifier si ses qualités psychométriques étaient suffisamment satisfaisantes pour qu'elle soit utilisée dans le cadre des recherches en psychologie ergonomique sur les interactions humain-machine ou par les professionnels qui travaillent sur l'évaluation de l'expérience

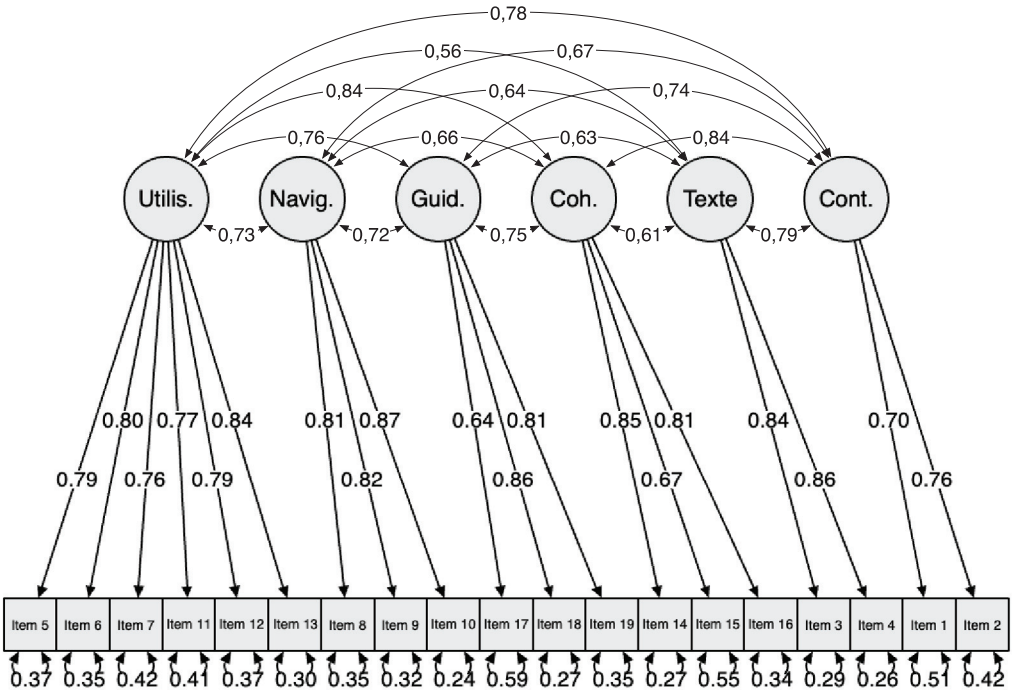

Fig. 1. Analyse factorielle confirmatoire de premier-ordre du F-DEEP ($n = 470$) (Utilis. : facilité d'utilisation perçue ; Navig. : navigation perçue ; Guid. : guidage visuel perçu ; Coh. : cohérence de la mise en page perçue ; Texte : utilité du texte perçu ; Cont. : contenu perçu).

utilisateur. Les mesures psychométriques ont ainsi montré que le F-DEEP présentait des valeurs aux indices d'ajustement très satisfaisantes selon un modèle de premier-ordre reprenant les 6 dimensions extraites de l'analyse factorielle en composantes principales. En nous inspirant des intitulés des dimensions du DEEP lorsque la correspondance entre les items était évidente entre le DEEP et le F-DEEP, nous avons nommé les 6 dimensions : facilité d'utilisation perçue, navigation perçue, guidage visuel perçu, cohérence de la mise en page perçue, utilité du texte perçue et contenu perçu.

Aussi, suite aux résultats que nous avons obtenus, plusieurs aspects sont-ils discutés dans cette section.

5.1. Similitudes et différences entre le DEEP et le F-DEEP

Nous avons déjà souligné, dans la section « 4.1.1. Analyse factorielle en composantes principales », que certaines propriétés psychométriques du F-DEEP étaient proches du DEEP, notamment en ce qui concerne sa structure factorielle. En effet, nous avons retrouvé les mêmes regroupements d'items pour les dimensions « Navigation perçue », « Guidage visuel perçu » et « Cohérence de la mise en page perçue ».

En revanche, les items relatifs aux dimensions « Structure perçue et architecture de l'information » et « Effort cognitif perçu » du DEEP font partie d'un seul et même facteur, que nous avons nommé « Facilité d'utilisation perçue », pour le F-DEEP. Ce résultat peut s'expliquer par les liens étroits qui existent entre l'architecture de l'information d'un site web et les efforts cognitifs nécessaires pour naviguer dans ce site. En effet, une architecture de l'information et une structure homogènes constituent un critère important de satisfaction chez l'utilisateur (Scapin et Bastien, 1997 ; Tan et Wei, 2006). À l'inverse, une structure incohérente augmente chez l'utilisateur le temps nécessaire à trouver l'information pertinente et la charge de travail mental. Par conséquent, plus l'organisation des pages d'un site web

est inconsistance, plus l'utilisateur devra mobiliser de ressources cognitives pour naviguer dans le site et rencontrera des difficultés d'interaction. Cette corrélation peut ainsi expliquer que les items comme « Le site était bien organisé », relatifs à la structure perçue et à l'architecture de l'information, fassent partie du même facteur que les items de type « J'ai appris à utiliser ce site rapidement », relatifs à l'effort cognitif perçu.

Par ailleurs, la dimension originale du DEEP nommée « Contenu perçu » et regroupant les items 1, 2, 3 et 4, a fait l'objet de deux facteurs différents pour le F-DEEP. Le premier facteur comprenant les items 1 et 2 a été intitulé « Contenu perçu », et le second facteur regroupant les items 3 et 4 a été nommé « Utilité du texte perçue ». La différenciation entre ces deux facteurs lors de l'ACP peut s'expliquer par notre cas d'étude, le site Onisep.fr. En effet, le site de l'Onisep a pour vocation d'informer sur les formations, les métiers et les secteurs professionnels. Il cherche à guider les élèves et étudiants dans leurs choix de parcours de formation et de projet professionnel, en fournissant une grande diversité de ressources en ligne : fiches métiers, quiz, vidéos, documentations sur les structures éducatives, référentiels de compétences, etc. Dans ce cadre, le contenu des informations fournies est l'élément central du site web, avec une grande majorité d'informations textuelles. Dès lors, il est probable que les utilisateurs portent une attention toute particulière à la compréhension du contenu d'une part, et à la pertinence des informations fournies d'autre part (intérêt et utilité des informations, adéquation par rapport aux attentes, aide vis-à-vis du projet professionnel envisagé, etc.). Par conséquent, cela pourrait expliquer que les items « La formulation du texte était claire » et « Le contenu (texte, images, sons, vidéos, etc.) était facile à comprendre » relatifs à la clarté du contenu, se démarquent des items « le texte était utile » et « Le texte était pertinent » qui sont quant à eux relatifs à la pertinence des informations fournies.

5.2. Réduire le nombre d'items

Tout comme les concepteurs du DEEP, nous pensons qu'il est important de réduire le plus possible le nombre d'items d'un questionnaire afin de diminuer l'ennui lors de sa complétion chez les répondants. Dans ce cadre, si la structure factorielle du F-DEEP se vérifiait dans d'autres études, il serait envisageable de réduire le nombre d'items de la dimension « facilité d'utilisation perçue », qui regroupe actuellement 6 items. Une analyse de la fidélité après suppression d'items, à l'aide de l'alpha de Cronbach et de l'oméga de McDonald, pourrait être menée afin de ne conserver que les items qui favorisent le plus la cohérence interne.

Concernant les autres facteurs, et en accord avec Yang et al. (2012), il nous semble difficile de réduire à moins de 3 le nombre d'items par dimension, au risque de fragiliser sa fiabilité.

5.3. Limites de l'étude et perspectives

Malgré des résultats encourageants, cette étude comporte certaines limites qui pourront être traitées lors de recherches complémentaires. La première limite concerne le site web qui a servi à valider l'adaptation française du DEEP. Comme nous l'avons souligné dans la section « 3.2.1. Cas d'étude », Onisep.fr propose des informations relatives à l'orientation professionnelle. Le site s'adresse ainsi avant tout à une population jeune composée de collégiens et de lycéens de moins de 18 ans, représentant 63 % des répondants (voir Tableau 1). Ce profil particulier a probablement un impact sur les résultats obtenus. En effet la génération Z, caractérisée par les personnes nées en pleine ère du numérique entre 1997 et 2010 (qui sont donc âgées entre 12 et 25 ans au moment de la diffusion de notre enquête), présente des attentes différentes des précédentes générations vis-à-vis des sites web : par exemple, elle préfère les médias digitaux (images, schémas, vidéos, podcasts. . .) aux contenus textuels traditionnels (Szymkowiak, Melović, Dabić, Jeganathan & Kundi, 2021) ; elle passe peu de temps à chercher l'information qu'elle souhaite, car elle a tendance à être versatile et possède une capacité réduite à prêter une attention soutenue (Ding, Guan & Yu, 2017) ; elle est très sensible à l'attractivité et l'esthétisme des services numériques en ligne (Windasari, Kusumawati, Larasati & Amelia, 2022). Ainsi, ces caractéristiques pourraient expliquer des jugements différents de ceux obtenus avec la population interrogée par Yang et al. (2012) pour la validation du DEEP.

La deuxième limite concerne l'absence de réplication de l'étude afin de procéder à une contre-validation. Gana et al. (2021) recommandent en effet de réaliser une nouvelle étude sur un large échantillon (N > 500) représentatif de la population cible, et de s'assurer que les nouveaux résultats, en termes de structure interne, fiabilité, validité des scores, etc., soient identiques ou proches de ceux obtenus précédemment. Même si cette dernière étape du processus d'adaptation transculturelle des échelles de mesure proposé par Gana et al. (2021) est très peu souvent appliquée par les chercheurs, elle nous semble intéressante dans la mesure où elle garantit que la traduction proposée est suffisamment robuste pour être appliquée à différents contextes, tout en garantissant des résultats psychométriques similaires. Concernant le F-DEEP, une étude menée sur le site web d'une société de transport public est en cours de diffusion, mais n'a pas obtenu un nombre de participants suffisants au moment de la rédaction de cet article.

La troisième limite concerne les instruments utilisés pour l'analyse de la validité du F-DEEP. En l'absence de comparaison avec d'autres questionnaires de mesure de l'utilisabilité suffisamment robustes en termes de qualités psychométriques, nous avons uniquement procédé à une validité nomologique, c'est-à-dire qui s'intéresse au concept global de la mesure de la satisfaction des utilisateurs. En revanche, nous n'avons pas pu procéder à une analyse de la validité concourante, qui aurait permis de s'assurer que le F-DEEP mesurait bien le concept d'utilisabilité en tant que tel. En effet, les qualités psychométriques de l'item unique de Christophersen et Konradt (2011) ont été largement remises en cause par Cairns (2013), qui souligne notamment que cet item n'a pas été validé du point de vue de son construit, en étant par exemple comparé à des données objectives d'utilisabilité comme les temps d'exécution des tâches, le nombre d'erreurs commises ou les retours en arrière lors de la navigation. De plus, cet item présente l'inconvénient de ne pas être une échelle d'utilisabilité fréquemment utilisée dans le domaine de la recherche ou en contexte professionnel. L'utilisation du NPS pour l'analyse de la validité nomologique peut aussi être critiquée car cette échelle est d'une part décriée en raison de sa structure à un seul item et du mode de calcul de son score, et d'autre part parce qu'elle porte sur l'intention de promouvoir une entreprise ou un produit et non pas sur l'évaluation de l'utilisabilité. Une comparaison des scores entre le F-DEEP et un autre questionnaire d'utilisabilité plus robuste nous semblerait alors pertinente pour compléter cette étude. Nous pensons que le *System Usability Scale* (SUS), traduit et validé en français par Gronier et Baudet (2021), serait un bon candidat pour mener une analyse de la validité concourante. Le SUS est en effet un questionnaire largement utilisé pour l'évaluation de l'utilisabilité de tout type de système (Gronier et Baudet, 2021), et a été construit pour aider les professionnels dans leurs pratiques, tout comme l'a été le DEEP. Le F-DEEP pourrait être également comparé à une échelle de mesure de l'expérience utilisateur (UX), comme l'AttrakDiff (Hassenzahl, Burmester & Koller, 2003) dont une version française a été validée par Lallemand, Koenig, Gronier & Martin (2015).

Pour finir, la quatrième limite concerne l'absence de situation contrôlée lors de la complétion du F-DEEP. Les répondants étaient en effet en situation réelle et naturelle de navigation sur le site Onisep.fr. Leur parcours ou le temps passé sur le site n'étaient pas enregistrés. Par conséquent, il n'était pas possible de déterminer si le répondant était un visiteur assidu, qui avait parcouru de nombreuses pages, ou s'il était un visiteur « en coup de vent ». Il n'était pas non plus possible de déterminer si les participants avaient atteint leur objectif ou non. Ces biais méthodologiques s'inscrivent dans le courant de recherche appelé « The turn to the wild » (que l'on peut traduire par « Retour à la nature »), décrit par Crabtree, Chamberlain, Grinter, Jones, Rodden & Rogers (2013) comme une nouvelle manière de comprendre et de façonner les interventions des nouvelles technologies dans la vie quotidienne. Dans le cadre de ce courant, les chercheurs ont quitté les laboratoires d'utilisabilité et d'expérimentation pour aller sur le terrain, réalisant des études *in situ* sur les technologies de l'information et de la communication. Selon Crabtree et al. (2013), l'une des raisons de cette approche contemporaine est la reconnaissance qu'une grande partie de la technologie est désormais intégrée et utilisée dans la vie de tous les jours. C'est dans ce cadre particulier que cette recherche sur la validation psychométrique du F-DEEP a été menée.

Pour le DEEP, Yang et al. (2012) avaient utilisé un environnement de tests en ligne qui fournissait un accès contrôlé à un site web et permettait de présenter des tâches précises à réaliser. Chaque tâche était vérifiée à l'aide d'un questionnaire à choix multiples, dont une seule réponse sur trois correspondait à l'objectif de la tâche à réaliser. Ce dispositif permettait ainsi aux chercheurs de s'assurer que les

participants prêtaient une attention soutenue au site web sur lequel ils naviguaient. Un protocole similaire pourrait ainsi être mis en place pour valider le F-DEEP lors d'une étude complémentaire.

Déclaration de liens d'intérêts

Les auteurs déclarent ne pas avoir de liens d'intérêts.

Remerciements

Les auteurs s'associent pour remercier Isabelle Dussouet, Directrice de l'offre de service et de la relation à l'usager à l'Onisep, pour son soutien et sa collaboration tout au long de cette étude. Laurence Johannsen, Juliette Decroix et Lindsey Stokes sont également remerciées pour leur contribution à la traduction du DEEP.

Références

Abran, A., Khelifi, A., Suryn, W., & Seffah, A. (2003). Usability meanings and interpretations in ISO standards. *Software Quality Journal*, *11*(4), 325–338.
Aladwani, A. M., & Palvia, P. C. (2002). Developing and validating an instrument for measuring user-perceived web quality. *Information and Management*, *39*(6), 467–476. https://doi.org/10.1016/S0378-7206(01)00113-6
Amar Raju, G., Roy, S., & Mandal, S. (2018). Determinants of website usability: Empirical evidence from tourism sector in India. *Global Business Review*, *19*(6), 1640–1662. https://doi.org/10.1177/0972150918794976
Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123. https://uxpajournal.org/determining-what-individual-sus-scores-mean-adding-an-adjective-rating-scale/
Barcenilla, J., & Bastien, J. M. C. (2009). L'acceptabilité des nouvelles technologies : quelles relations avec l'ergonomie, l'utilisabilité et l'expérience utilisateur. *Le Travail Humain*, *72*(4), 311–331.
Bastien, J. M. C. (2010). Usability testing: a review of some methodological and technical aspects of the method. *International Journal of Medical Informatics*, *79*(4), e18–e23. https://doi.org/10.1016/j.ijmedinf.2008.12.004
Belanche, D., Casaló, L. V., & Guinalíu, M. (2012). Website usability, consumer satisfaction and the intention to use a website: the moderating effect of perceived risk. *Journal of Retailing and Consumer Services*, *19*(1), 124–132. https://doi.org/10.1016/j.jretconser.2011.11.001
Bolchini, D., & Garzotto, F. (2008). Quality and potential for adoption of usability evaluation methods: an empirical study on MiLE+. *Journal of Web Engineering*, *7*(4), 299–317. http://dl.acm.org/citation.cfm?id=2011277.2011281
Borsci, S., Buckle, P., & Walne, S. (2020). Is the LITE version of the usability metric for user experience (UMUX-LITE) a reliable tool to support rapid assessment of new healthcare technology? *Applied Ergonomics*, *84*(November 2019), 1–5. https://doi.org/10.1016/j.apergo.2019.103007
Brangier, E., Dufresne, A., & Hammes-Adelé, S. (2009). Approche symbiotique de la relation humain-technologie : perspectives pour l'ergonomie informatique. *Le Travail Humain*, *74*(4), 333–353.
Brooke. (1996). SUS: A 'quick and dirty' usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). London: Taylor & Francis. http://hell.meiert.org/core/pdf/sus.pdf
Cairns, P. (2013). A commentary on short questionnaires for assessing usability. *Interacting with Computers*, *25*(4), 312–316. https://doi.org/10.1093/iwc/iwt019
Casalo, L., Flavian, C., & Guinaliu, M. (2008). The role of perceived usability, reputation, satisfaction and consumer familiarity on the website loyalty formation process. *Computers in Human Behavior*, *24*(2), 325–345. https://doi.org/10.1016/j.chb.2007.01.017
Christophersen, T., & Konradt, U. (2011). Reliability, validity, and sensitivity of a single-item measure of online store usability. *International Journal of Human Computer Studies*, *69*(4), 269–280. https://doi.org/10.1016/j.ijhcs.2010.10.005
Crabtree, A., Chamberlain, A., Grinter, R. E., Jones, M., Rodden, T., & Rogers, Y. (2013). Introduction to the special issue of "The Turn to The Wild". *ACM Transactions on Computer-Human Interaction*, *20*(3), 0–3. https://doi.org/10.1145/2491500.2491501
Ding, D., Guan, C., & Yu, Y. (2017). Game-based learning in tertiary education: a new learning experience for the generation Z. *International Journal of Information and Education Technology*, *7*(2), 148–152. https://doi.org/10.18178/ijiet.2017.7.2.857
Dubois, M., & Bobillier-Chaumon, M. E. (2009). L'acceptabilité des technologies : bilans et nouvelles perspectives. *Le Travail Humain*, *72*(4), 305–310. https://doi.org/10.3917/th.724.0305
Elling, S., Lentz, L., de Jong, M., & van den Bergh, H. (2012). Measuring the quality of governmental websites in a controlled versus an online setting with the "Website Evaluation Questionnaire". *Government Information Quarterly*, *29*(3), 383–393. https://doi.org/10.1016/j.giq.2011.11.004
Fernandez, A., Insfran, E., & Abrahão, S. (2011). Usability evaluation methods for the web: a systematic mapping study. *Information and Software Technology*, *53*(8), 789–817. https://doi.org/10.1016/j.infsof.2011.02.007
Finstad, K. (2010). The usability metric for user experience. *Interacting with. Computers*, *22*(5), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004
Fisher, N. I., & Kordupleski, R. E. (2019). Good and bad market research: a critical review of Net Promoter Score. *Applied Stochastic Models in Business and Industry*, *35*(1), 138–151. https://doi.org/10.1002/asmb.2417
Flavián, C., Guinalíu, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & Management*, *43*(1), 1–14. https://doi.org/10.1016/j.im.2005.01.002

Gana, K., Boudouda, N. E., Ben Youssef, S., Calcagni, N., & Broc, G. (2021). Adaptation transculturelle de tests et échelles de mesure psychologiques : guide pratique basé sur les Recommandations de la Commission internationale des tests et les standards de pratique du testing de l'APA. *Pratiques Psychologiques*, *27*(3), 223–240. https://doi.org/10.1016/j.prps.2021.02.001
Gronier, G. (2022). Psychometric analyses in the transcultural adaptation of psychological scales. In S. Misciagna (Ed.), *Psychometrics – new insights (pp. 1–13)*. IntechOpen [https://doi.org/10.1039/C7RA00172J%0A. https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics%0Ahttps://doi.org/10.1016/j.colsurfa.2011.12.014].
Gronier, G., & Baudet, A. (2021). Psychometric evaluation of the F-SUS: Creation and validation of the French version of the System Usability Scale. *International Journal of Human-Computer Interaction*, *37*(16), 1571–1582. https://doi.org/10.1080/10447318.2021.1898828
Gronier, G., Johannsen, L. (2022). Proposition d'une adaptation française et premières validations de l'échelle d'utilisabilité Computer System Usability Questionnaire (F-CSUQ). 33e Conference Internationale Francophone Sur l'Interaction Homme-Machine: Interaction Humain and IA, IHM 2022 – Actes de La Conference. https://doi.org/10.1145/3500866.3516379.
Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwillus, & J. Ziegler (Eds.), *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187–196). Stuttgart: B.G. Teubner. https://doi.org/10.1007/978-3-322-80058-9_19
Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But. . .. *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrica*, *30*(2), 179–185.
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118
Ismail, I. E., Nalawati, R. E., & Putra, A. (2021). *System usability scale and net promoter score on donation application of toddlers equipment*. pp. 170–174. Proceedings – 2021 4th International Conference on Computer and Informatics Engineering: IT-Based Digital Industrial Innovation for the Welfare of Society, IC2IE. https://doi.org/10.1109/IC2IE53219.2021.9649186
ISO 16982:2002. (2002). Méthodes d'utilisabilité pour la conception centrée sur l'opérateur humain.
ISO/IEC 25010:2011. (2011). Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models.
Kirakowski, J., & Cierlik, B. (1998). Measuring the usability of web sites. *Proceedings of the Human Factors and Ergonomics Society*, *1*, 424–428. https://doi.org/10.1177/154193129804200405
Koranteng, F. N., Ham, J., Wiafe, I., & Matzat, U. (2021). The role of usability, aesthetics, usefulness and primary task support in predicting the perceived credibility of academic social networking sites. *Behaviour and Information Technology*, 1–16. https://doi.org/10.1080/0144929X.2021.2009570
Lallemand, C., Koenig, V., Gronier, G., & Martin, R. (2015). Création et validation d'une version française du questionnaire AttrakDiff pour l'évaluation de l'expérience utilisateur des systèmes interactifs. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *65*(5), 239–252. https://doi.org/10.1016/j.erap.2015.08.002
Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour and Information Technology*, *16*(4–5), 246–266. https://doi.org/10.1080/014492997119824
Lin, H., Choong, Y.-Y., & Salvendy, G. (1997). A proposed index of usability: a method for comparing the relative usability of different software systems. *Behaviour & Information Technology*, *16*(4), 267–277. https://doi.org/10.1080/014492997119833
Loiacono, E. T., Watson, R. T., & Goodhue, D. L. (2007). WebQual: an instrument for consumer evaluation of web sites. *International Journal of Electronic Commerce*, *11*(3), 51–87.
Lund, A. M. (2001). Measuring usability with the USE questionnaire. *Usability Interface*, *8*(2), 3–6. https://doi.org/10.1177/1078087402250360
Maguire, M. (2001). Methods to support human-centred design. *International Journal of Human-Computer Studies*, *55*(4), 587–634. https://doi.org/10.1006/ijhc.2001.0503
Nielsen, J. (1993). *Usability Engineering*. San Francisco: Morgan Kaufmann Publishers Inc.
Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
Pradini, R. S., Kriswibowo, R., & Ramdani, F. (2019). *Usability evaluation on the SIPR website uses the system usability scale and net promoter score*. pp. 280–284. Proceedings of 2019 4th International Conference on Sustainable Information Engineering and Technology, SIET, Lombok, Indonesia. https://doi.org/10.1109/SIET48054.2019.8986098
Quesenbery, W. (2003). Dimensions of usability. In M. Albers, & B. Mazur (Eds.), *Content and complexity: information design in technical communication* (pp. 81–102). New-York : Erlbaum.
Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*(12), 46–54.
Roberson, R. B., Elliott, T. R., Chang, J. E., & Hill, J. N. (2014). Exploratory factor analysis in rehabilitation psychology: A content analysis. *Rehabilitation Psychology*, *59*(4), 429–438. https://doi.org/10.1037/a0037899
Sasmito, G. W., Zulfiqar, L. O. M., & Nishom, M. (2019). *Usability testing based on system usability scale and net promoter score*. pp. 540–545. 2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI, Yogyakarta, Indonesia. https://doi.org/10.1109/ISRITI48646.2019.9034666
Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience*. Waltham : Elsevier. https://doi.org/10.1016/B978-0-12-384968-7.00013-8
Scapin, D., & Bastien, J. M. C. (1997). Ergonomic criteria for evaluating the ergonomic quality of interactive systems. *Behaviour & Information Technology*, *16*(4), 220–231. https://doi.org/10.1080/014492997119806
Schweizer, K. (2010). Some guidelines concerning the modeling of traits and abilities in test construction. *European Journal of Psychological Assessment*, *26*(1), 1–2. https://doi.org/10.1027/1015-5759/a000001
Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*(5), 893–898. https://doi.org/10.1016/j.paid.2006.09.017
Szymkowiak, A., Melović, B., Dabić, M., Jeganathan, K., & Kundi, G. S. (2021). Information technology and Gen Z: the role of teachers, the internet, and technology in the education of young people. *Technology in Society*, *65*(March) https://doi.org/10.1016/j.techsoc.2021.101565
Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). London : Pearson.

Tan, G. W., & Wei, K. K. (2006). An empirical study of Web browsing behaviour: towards an effective website design. *Electronic Commerce Research and Applications*, *5*(4), 261–271. https://doi.org/10.1016/j.elerap.2006.04.007
Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127–145. https://doi.org/10.1016/S0953-5438(00)00031-X
Vallerand, R. J. (1989). Vers une méthodologie de validation trans-culturelle de questionnaires psychologiques : Implications pour la recherche en langue française. *Psychologie Canadienne*, *30*(4), 662–680. https://doi.org/10.1037/h0079856
Wang, J., & Wang, X. (2019). *Structural equation modeling*. Hoboken : John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119422730.ch2
Windasari, N. A., Kusumawati, N., Larasati, N., & Amelia, R. P. (2022). Digital-only banking experience: Insights from gen Y and gen Z. *Journal of Innovation and Knowledge*, *7*(2), 100170. https://doi.org/10.1016/j.jik.2022.100170
Yang, T., Linder, J., & Bolchini, D. (2012). DEEP: design-oriented evaluation of perceived usability. *International Journal of Human-Computer Interaction*, *28*(5), 308–346. https://doi.org/10.1080/10447318.2011.586320
ISO 9241-11:2018. (2018). Ergonomie de l'interaction homme-système. Partie 11 : Utilisabilité - Définitions et concepts.