Mesure de la sensibilité et de la signification de la version française du System Usability Scale

Measuring the sensitivity and significance of the French version of the System Usability Scale

Guillaume Gronier
Luxembourg Institute of Science and Technology

ABSTRACT

The System Usability Scale (SUS) is the most widely cited and most widely used of the User Experience Scales (UX). To date, there is no scientifically validated translation into French of this scale. This article therefore proposes to use a first French version of the SUS, the F-SUS, to measure its sensitivity and meaning. The sensitivity of the F-SUS was studied by applying it to 11 different interactive systems (websites, mobile application, expert systems), with 439 users. The scores of these systems were compared in order to verify that they were sufficiently differentiated. The measure of significance was to replicate the Bangor et al (2008) study, asking F-SUS respondents to qualify the system they were assessing using 7 adjectives. These adjectives were then positioned on the overall F-SUS score out of 100.

Pourtant la plus citée et la plus utilisée parmi les échelles de mesure de l'expérience utilisateur (UX), l'échelle System Usability Scale (SUS) ne possède pas à ce jour de traduction validée scientifiquement en français. Cet article propose ainsi de s'appuyer sur une première version française du SUS, le F-SUS, pour mesurer sa sensibilité et sa signification. La sensibilité du F-SUS a été étudiée en l'appliquant à 11 systèmes interactifs différents (sites web, application mobile, systèmes experts), auprès de 439 utilisateurs. Les scores de ces systèmes ont été comparés afin de vérifier qu'ils se différenciaient suffisamment. La mesure de la signification a consisté à répliquer l'étude de Bangor et al. (2008), en demandant aux répondants du F-SUS de qualifier le système qu'ils évaluaient à l'aide de 7 adjectifs qualificatifs. Ces adjectifs ont ensuite été positionnés par rapport au score global, sur 100, du F-SUS.

CCS CONCEPTS

• **Human-centered computing**; • **Human computer interaction (HCI)**; • **HCI design and evaluation methods**; • **Usability testing**;

KEYWORDS

Echelle d'utilisabilité, Usability scale, SUS, questionnaire, translation, validation

IHM '20.21, April 13–16, 2021, Virtual Event, France

ACM ISBN 978-1-4503-8362-2/21/04…$15.00
https://doi.org/10.1145/3450522.3451241

ACM Reference Format:
Guillaume Gronier. 2021. Mesure de la sensibilité et de la signification de la version française du System Usability Scale: Measuring the sensitivity and significance of the French version of the System Usability Scale. In *32e Conférence Francophone sur l'Interaction Homme-Machine (IHM '20.21), April 13–16, 2021, Virtual Event, France.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3450522.3451241

1 INTRODUCTION

Parmi les méthodes qui permettent d'évaluer l'utilisabilité, les questionnaires dits «standardisés »sont souvent employés pour capturer la satisfaction de l'utilisateur vis-à-vis d'un produit, d'un service ou d'un système [34]. Le caractère «standardisé »de ces questionnaires les distinguent des questionnaires «faits maison »[20, 41] en raison de la validation scientifique qui leur a été appliquée. En tant que mesure de l'utilisabilité, ces questionnaires «recueillent l'avis des utilisateurs sur la facilité d'utilisation perçue d'un système et la satisfaction liée à l'interaction »[26]. La norme ISO/TR 16982:2002(F) définit également les questionnaires comme une méthode «d'évaluation indirecte qui recueillent, au moyen de questionnaires prédéfinis, les opinions des utilisateurs sur l'interface »[22].

De nombreux questionnaires standardisés existent dans le domaine de l'utilisabilité, ou plus généralement de l'expérience utilisateur (UX). Lallemand et Gronier [26] en listent une vingtaine, couvrant différentes facettes de l'évaluation d'un système, comme par exemple sa facilité d'utilisation, les émotions qu'il génère, ou son esthétique. L'un des premiers questionnaires à avoir été créé est le System Usability Scale (SUS), conçu à son origine comme une échelle «Quick and Dirty », c'est-à-dire facile et rapide à faire passer [13]. Aujourd'hui encore, le SUS est très largement utilisé et s'impose pour l'évaluation de très nombreux types de systèmes: application mobile [1, 7], site web [21], systèmes experts [43, 46], serious games [44], e-learning [37], etc. Il a également fait l'objet de nombreuses publications, qui se sont non seulement intéressées à sa validité scientifique [11, 31–33], mais qui ont également approfondi la signification de ses scores [3, 4, 25].

Malgré ce succès, le SUS est peu étudié en France, et ne connait aucune traduction officielle même s'il est quelquefois utilisé dans les recherches en Interaction Homme-Machine (IHM) [29]. Cette recherche se donne alors pour objectif d'étudier deux composantes essentielles du SUS dans sa version française (que nous nommerons le F-SUS), dont la validation de la traduction a fait l'objet d'une autre recherche en cours de publication au moment de la rédaction de cet article [18]. Les deux composantes relatées dans cet article sont d'une part la sensibilité du F-SUS, c'est-à-dire sa capacité à discriminer des systèmes différents lorsqu'ils sont évalués, et d'autre part la signification du score du F-SUS, c'est-à-dire l'interprétation qu'il peut être donné à la note globale obtenue sur 100.

2 LE SYSTEM USABILITY SCALE (SUS)

2.1 Présentation du SUS

Créé en 1986 par Brooke dans le cadre d'un programme en ingénierie sur l'utilisabilité des systèmes, le questionnaire SUS s'est rapidement diffusé pour l'évaluation de l'utilisabilité des systèmes interactifs [13, 14]. Le SUS a été conçu comme un questionnaire «quick and dirty », rapide à faire passer, de manière à compléter la passation de tests utilisateurs en laboratoire par une mesure subjective de l'utilisabilité perçue, tout en garantissant une passation rapide et non contraignante pour les utilisateurs interrogés.

Le SUS comprend dix items, formulés sous la forme de phrases affirmatives, sur chacune desquelles l'utilisateur est invité à se positionner, en exprimant son accord ou son désaccord, à l'aide d'une échelle de Lickert à 5 points (1 = Pas du tout d'accord ; 5 = Tout à fait d'accord). Si l'utilisateur ne sait pas comment se positionner par rapport à un item, il est invité à y répondre malgré tout en cochant le centre de l'échelle (score 3).

Les 10 items de la version définitive du SUS ont été sélectionnés à partir d'une liste préalable de 50 items, rédigés de manière à couvrir les trois principaux concepts de l'utilisabilité selon la norme ISO 9241-11 [23]: l'efficacité, l'efficience et la satisfaction. Les 50 items ont d'abord été soumis à un panel de 20 utilisateurs pour l'évaluation de deux systèmes interactifs. Puis seuls les 10 items obtenant des réponses les plus extrêmes ont été retenus [13].

2.2 Mode de calcul du score du SUS

Le score global du SUS est calculé de manière à tenir compte des items inversés (items pairs: 2, 4, 6, 8 et 10) et à obtenir un score total compris entre 0 et 100. Pour cela, le calcul se fait en 3 étapes :

1. Il faut tout d'abord soustraire un point au score coché par l'utilisateur pour les items 1, 3, 5, 7 et 9 (items impairs, non inversés).
2. Ensuite, pour les items 2, 4, 6, 8 et 10 (items pairs, inversés), il faut calculer 5 moins le score coché par l'utilisateur.
3. Les 10 nouveaux scores ainsi recalculés sont additionnés et multipliés par 2,5.

Brooke [14] explique que ce mode de calcul a été défini afin de répondre à des exigences marketings, plutôt que scientifiques. Au moment de la création du SUS, Brooke et son équipe ont considéré que les chefs de projet, les chefs de produit et les ingénieurs étaient plus susceptibles de comprendre une échelle qui allait de 0 à 100, plutôt qu'une échelle allant de 10 à 50 (50 étant la note maximale qui aurait été obtenue en utilisant le mode de calcul habituel pour les échelles avec items inversés). Brooke indique également que l'obtention d'une note sur 100 facilite la compréhension du score et la comparaison entre différents systèmes, puisque les différences entre plusieurs scores sont perçues comme plus importantes que si la note était sur 50.

2.3 Signification du score du SUS

Bangor, Kortum et Miller [4] ont cherché à donner du sens aux scores du SUS calculés sur 100. Les auteurs relatent que la signification du score était toujours un problème lorsqu'il fallait reporter le résultat d'une étude à un chef de projet ou une équipe de conception. Aussi, un programme pilote a-t-il été lancé pour déterminer si des adjectifs pouvaient être associés à des intervalles de scores du SUS afin de donner une note plus absolue. Une échelle d'évaluation à 7 adjectifs a alors été utilisée, en complément du SUS. 212 participants étaient invités à compléter le SUS, puis à répondre à la question: «Dans l'ensemble, je qualifierais la convivialité de ce produit de. . . ». Les participants devaient ainsi se positionnaient sur l'un des 7 adjectifs suivants: La pire qu'on puisse imaginer ; Horrible ; Mauvaise ; Acceptable ; Bonne ; Excellente ; La meilleure qu'on puisse imaginer. Un peu plus tard, les auteurs ont répliqué la même étude auprès d'un échantillon plus important de participants (959 résultats exploitables) [3]. Tous les adjectifs ont obtenu des scores significativement différents, sauf «La pire qu'on puisse imaginer »et «Horrible ». Par conséquent, Bangor, Kortum et Miller [3] ont retenus 6 adjectifs, repris dans la figure 1

2.4 Traductions du SUS

Suite aux nombreuses évaluations pour lesquelles le SUS a été mobilisé, plusieurs traductions ont été réalisées. Le SUS a ainsi été traduit dans une démarche scientifique en indonésien [42], en portugais [35], en polonais [10], en arabe [2], en slave [9], en grecque [24] et en perse [16]. Une version allemande existe également [39], même si elle n'a pas été publiée. En France, si le SUS est utilisé en évaluation des systèmes (voir par exemple [29]), il n'a jamais fait l'objet ni d'une traduction officielle ni d'une validation scientifique.

Relevons pour finir que le SUS a été transcrit dans le langage des signes américain [6], mais aussi en langage pictural [5].

Le table 1 reprend les traductions publiées du SUS, accompagnées du score de fidélité (alpha de Cronbach) mesuré.

3 PROBLÉMATIQUE ET MÉTHODOLOGIE

3.1 Problématique

Il nous semble désormais grand temps d'offrir à la communauté des chercheurs et des professionnels francophones, qui travaillent dans les domaines de l'utilisabilité et de l'expérience utilisateur, une traduction validée du SUS, et de s'assurer de sa robustesse psychométrique. Étonnamment, peu de recherches portent sur la traduction en français d'échelles dans ces deux domaines. A notre connaissance, seule le questionnaire de mesure de l'expérience utilisateur AttrakDiff [19] a fait l'objet d'une traduction dans le cadre rigoureux d'une démarche scientifique [27].

Aussi, cette recherche vise-t-elle d'une part à diffuser une version française du SUS, baptisée F-SUS, et d'autre part à approfondir les qualités psychométriques de cette traduction en s'intéressant à sa sensibilité et à la signification de son score. La démarche de traduction du SUS a fait l'objet d'une autre recherche en cours de publication, au moment où le présent article a été rédigé [18]. Nous n'en proposons ici qu'une synthèse reprise dans la partie méthodologie. Cela nous permet ainsi de concentrer cet article sur les résultats approfondis obtenus pour la mesure de la sensibilité et de la signification du F-SUS.

3.2 Recherche préliminaire: traduction de la version française du SUS

Pour assurer la traduction du French-System Usability Scale (F-SUS), nous nous sommes inspirés de la démarche proposée par Vallerand

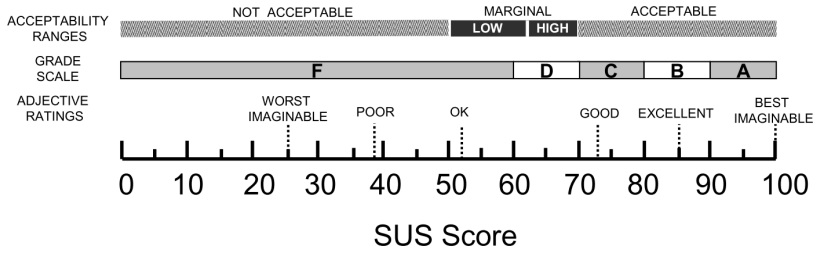

Figure 1: Signification des scores du SUS, d'après l'étude de Bangor, Kortum et Miller [3].

Table 1: Traductions publiées du SUS et leur score de fidélité (alpha de Cronbach)

Traduction du SUS	Alpha de Cronbach	Référence
SUS original (version anglaise)	,91	[32]
Indonésien	,841	[42]
Portugais	Non calculé	[35]
Polonais	,805	[10]
Arabe	,82	[2]
Slave	,81	[9]
Grecque	,777	[24]
Perse	,79	[16]
Langage des signes américain	,7	[6]
Langage pictural	Non calculé	[5]

Table 2: Items originaux et traduction du SUS

Items originaux	Traduction en français par comité
1. I think that I would like to use this system frequently.	Je voudrais utiliser ce système fréquemment.
2. I found this system unnecessarily complex.	Ce système est inutilement complexe.
3. I thought the system was easy to use.	Ce système est facile à utiliser.
4. I think that I would need the support of a technical person to be able to use this system.	J'aurais besoin du soutien d'un technicien pour être capable d'utiliser ce système.
5. I found the various functions in this system were well integrated.	Les différentes fonctionnalités de ce système sont bien intégrées.
6. I thought there was too much inconsistency in this system.	Il y a trop d'incohérences dans ce système.
7. I would imagine that most people would learn to use this system very quickly.	La plupart des gens apprendront à utiliser ce système très rapidement.
8. I found this system very cumbersome to use.	Ce système est très lourd à utiliser.
9. I felt very confident using the system.	Je me suis senti très en confiance en utilisant ce système.
10. I needed to learn a lot of things before I could get going with this system.	J'ai eu besoin d'apprendre beaucoup de choses avant de pouvoir utiliser ce système.

[45]. Cet auteur propose une méthodologie de validation transculturelle de questionnaires psychologiques. Cette méthodologie comprend sept étapes qui permettent la traduction et la validation de questionnaires anglophones vers le français. Parmi ces sept étapes, deux sont fondamentales et sont reprises ici: la préparation d'une version expérimentale et l'évaluation psychométrique.

3.2.1 Préparation d'une traduction du SUS. Cette première étape consiste à préparer une version expérimentale du questionnaire original dans la langue cible (ici en langue française). Pour se faire, nous avons opté pour une traduction de type comité. Dans ce cadre, trois chercheurs bilingues français-anglais, de nationalité française, ont été sollicités pour proposer une traduction des 10 items du SUS. Dans un premier temps, chaque traducteur a procédé à une traduction individuelle. Puis dans un second temps, les chercheurs présentaient chacun leur traduction, et une discussion de groupe était engagée sur le contenu de leur traduction. A l'issue de la session par comité, une principale version du SUS a été retenue (table 2).

3.2.2 Evaluation psychométrique du F-SUS. Cette étape a consisté à procéder à une évaluation psychométrique du F-SUS. Plusieurs analyses statistiques ont été réalisées, afin de mesurer la fidélité, la structure factorielle, la sensibilité et la validité de contenu. La

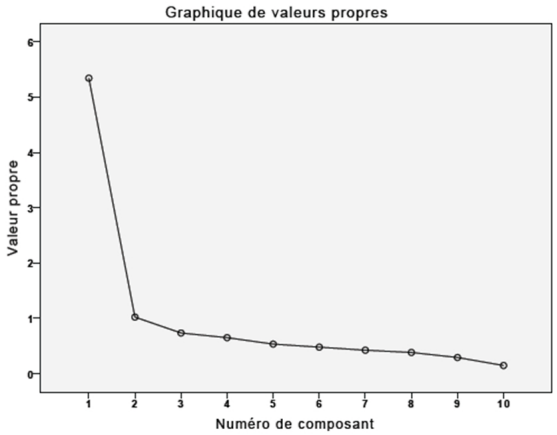

Figure 2: Test des éboulis pour le F-SUS

taille minimale recommandée de l'échantillon pour les tests psychométriques, et plus spécifiquement pour l'analyse factorielle, est d'avoir au moins cinq participants par item, ce qui, pour la SUS à 10 items, correspond à un minimum de 50 participants [36].

Nous avons administré le F-SUS à 88 étudiants bénévoles (78 femmes, 10 hommes ; âge moyen = 19,66 ans ; écart-type = 3,85 ; âge minimum = 17 ; âge maximum = 42) en sciences humaines et sociales, à qui il était demandé de penser aùn système qu'ils connaissaient bien, ou d'afficher aḷ'ećran, sur leur ordinateur ou leur smartphone par exemple, un systeme aèv́aluer. Pour cette évaluation psychométrique, nous ne nous sommes volontairement pas intéressés au système évalué. En effet, nous avons considéré que le système importait peu, mais qu'il était primordial que ce soit toujours le même système qui soit évalué pour tous les items du F-SUS. Les utilisateurs ont été sensibilisés sur ce point précis.

Mesure de la fidélité. Le coefficient alpha de Cronbach mesuré est de ,899. Cela atteste d'une fidélité suffisante, avec un seuil supérieur à ,70, tel que recommandé par Landauer [28]. Ce score est parmi les plus élevés qui aient été mesurés au cours des différentes traductions du SUS [31]. Il est également proche de ceux observés lors des analyses psychométriques du SUS, avec un alpha de ,91 calculé par Lewis et Sauro [32], ou de ,911 trouvé par Bangor et al. [4].

Analyses factorielles. Une analyse factorielle en composantes principales (ACP) a été réalisée pour tester la validité de construit du F-SUS. L'intérêt d'une ACP est de vérifier si la structure factorielle d'une traduction est similaire à la structure du questionnaire original [13]. Le test des éboulis de Cattell [15] (aussi appelé Eigenvalues) suggère une structure à 2 facteurs (figure 2). Si ce résultat ne converge pas vers une structure à facteur unique tel qu'a été conçu le SUS par Brooke [13], et confirmée par Bangor et al. [4], il rejoint celui obtenu par Lewis et Sauro [32], puis vérifié par Borsci, Federici et Lauriola [11].

Le table 3 présente la rotation Varimax à 2 facteurs pour les 10 items du SUS. L'ACP à deux facteurs montrent que les items 1, 2, 3, 5, 6, 7, 8 et 9 sont alignés avec le premier facteur, et les items 4 et 10 sont alignés avec le second facteur. Ces résultats rejoignent les analyses factorielles menées par Lewis et Sauro [32]. Les auteurs ont

Table 3: Rotation Varimax à 2 facteurs pour les 10 items du F-SUS

Item	Facteur 1	Facteur 2
1	**.728**	.098
2	**.757**	.280
3	**.675**	.275
4	.344	**.682**
5	**.678**	.400
6	**.852**	.143
7	**.710**	.303
8	**.847**	.250
9	**.735**	.049
10	.054	**.878**

alors intitulé l'échelle identifiée par le premier facteur «Utilisabilité », et la seconde échelle identifiée par le second facteur (avec les items 4 et 10) «Apprentissage ».

3.3 Etude de la sensibilité du F-SUS

Afin de mesurer la sensibilité du F-SUS, c'est-à-dire la capacité du F-SUS à obtenir des scores d'utilisabilité différents pour des systèmes différents, et identiques pour des systèmes identiques, nous avons appliqué le F-SUS à 11 systèmes. Ces systèmes ont été choisis par des groupes d'étudiants en L3, dans le cadre d'une formation à l'Expérience utilisateur des systèmes interactifs, et devaient être les plus diversifiés possibles: jeux vidéos sur mobile, applications mobiles de réseaux sociaux, site web administratif, site web d'un fournisseur d'accès à Internet, application mobile d'un service de rencontres, etc. Les groupes d'étudiants étaient donc totalement libres du choix de leur système, même si une validation par l'enseignant était nécessaire afin de s'assurer qu'il s'agissait bien de systèmes interactifs (et non pas de produit), et de nature variée. Chacun de ces systèmes est décrit dans le table 4

3.4 Etude de la signification du score du F-SUS

Pour étudier la signification du score du F-SUS, nous avons repris les travaux de [3, 4], qui ont appliqué un adjectif au score du SUS (voir figure 1). Pour notre recherche, les 7 adjectifs ont été traduits en français (table 5), tout comme l'affirmation à laquelle les utilisateurs devaient répondre («Dans l'ensemble, vous qualifieriez la convivialité de ce système de. . . »). Notons à ce sujet que nous avons repris au plus près la phrase affirmative employée par Bangor, Kortum et Miller [4] («Overall, I would rate the user-friendliness of this product as »), qui justifiaient le choix des termes par trois principaux arguments :

1. Le terme «dans l'ensemble »(«Overall ») était utilisé afin de recueillir l'expérience cumulative de l'utilisateur vis-à-vis du système évalué [38].
2. Le terme «convivialité »(«User-friendless ») était utilisé parce qu'il est l'un des synonymes les plus connus d'utilisabilité, et que les participants étaient susceptibles de le comprendre immédiatement.
3. En revanche, les auteurs avaient préféré utiliser le terme «produit »(«product ») plutôt que «système », car ils avaient

Table 4: Description des systèmes étudiés (les liens ont été testés le 13/10/2020)

Système	Description	URL
Archero	Jeu video sur mobile	https://www.habby.fun/
Crunchyroll	Jeu video sur mobile	https://www.crunchyroll.com/
Deliveroo	Site web d'un service de livraison de repas à domicile	https://deliveroo.fr/fr/
Discord	Application mobile d'un réseau social	https://discord.com/
Free	Site web d'un fournisseur d'accès à Internet	https://www.free.fr/
Impôts.gouv	Site web de déclaration d'impôts	https://www.impots.gouv.fr/
Messenger	Application mobile d'un réseau social	https://messenger.com/
Superprof	Site web d'un service de cours particuliers	https://www.superprof.fr
Tinder	Application mobile d'un service de rencontres	https://tinder.com
TV Time	Application mobile de gestion de films et de séries	https://www.tvtime.com/
Youtube	Application mobile d'un service de diffusion de vidéos	http://youtube.fr/

Table 5: Traduction des adjectifs appliqués à la suite du F-SUS

Adjectif original	Traduction	Codage numérique
Worst imaginable	Pire qu'on puisse imaginer	1
Awful	Horrible	2
Poor	Mauvaise	3
OK	Acceptable	4
Good	Bonne	5
Excellent	Excellente	6
Best imaginable	Meilleure qu'on puisse imaginer	7

Table 6: Population des répondants en fonction des systèmes évalués

Système	N	Nbre de femmes	Nbre d'hommes	Age moyen
Archero	20	4	16	20,9
Crunchyroll	19	6	13	19,9
Deliveroo	38	18	20	24,2
Discord	29	8	21	22,9
Free	35	14	21	22,6
Impôts.gouv	34	15	19	26,2
Messenger	50	21	29	23,0
Superprof	117	46	71	21,6
Tinder	39	20	18	25,1
TV Time	30	9	21	23,5
Youtube	29	9	20	22,2
Total	**439**	**170**	**269**	**22,9**

adapté le SUS afin qu'il soit mieux adapté à leur objet d'étude. En ce qui nous concerne, nous avons souhaité conserver le terme «système », afin de respecter la formulation originale du SUS [13].

3.5 Populations et déroulement des évaluations

439 utilisateurs ont répondu au F-SUS, avec des répartitions différentes pour chacun des 11 systèmes évalués (table 6). A noter que tous les utilisateurs ont eu accès à la même version du F-SUS (table 2), et que nous ne l'avons pas adaptée au type de système: par exemple, si l'utilisateur devait donner son avis sur une application mobile, les items gardaient bien la formule «Je voudrais utiliser ce système fréquemment », et non pas «Je voudrais utiliser cette application mobile fréquemment ». Ainsi, toutes les réponses ont pu être comparées les unes aux autres.

Les utilisateurs ont été contactés via une annonce sur différents réseaux sociaux (Facebook, groupes Messenger, LinkedIn, Groupe Discord, Slack), ou ont été interrogés en face-à-face sur le campus du Pôle Lorrain de Gestion de l'Université de Lorraine à Nancy. Il ne leur était pas demandé de connaître le système qu'il devrait évaluer. Puisque chaque groupe d'étudiants, en charge d'un des 11 systèmes, avait la responsabilité de recruter eux-mêmes leurs

Table 7: Résultats au F-SUS (de 0 à 100) en fonction des différents systèmes testés

Système	Score au F-SUS	Ecart-type	Médiane	Score minimum	Score maximum
Archero	78,13	15,93	85	30	95
Crunchyroll	71,71	12,91	72,5	45	93
Deliveroo	72,92	12,54	75	48	100
Discord	77,84	14,48	80	45	100
Free	66,36	14,43	67,5	40	95
Impôts.gouv	60,51	15,07	60	25	90
Messenger	75,40	13,74	75	50	100
Superprof	77,57	12,98	77,5	43	100
Tinder	67,89	16,70	66,25	45	100
TV Time	83,58	11,61	86,25	58	100
Youtube	86,47	8,95	87,5	68	100

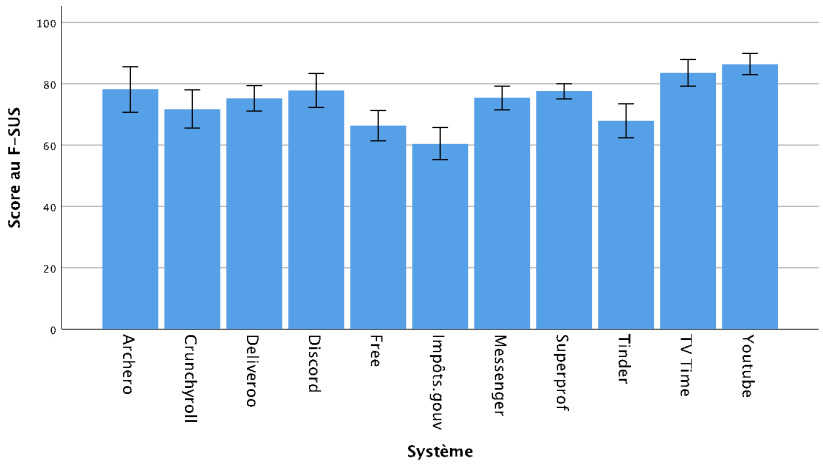

Figure 3: Score moyen (de 0 à 100) au F-SUS pour chacun des systèmes évalués (barres d'erreur: intervalle de confiance à 95%)

utilisateurs, il n'a pas été organisé de répartition particulières des répondants. Cela explique également les différences de population entre les systèmes (table 6).

Avant de compléter le F-SUS, chaque utilisateur était invité à réaliser plusieurs tâches avec un des 11 systèmes afin d'en prendre connaissance. Ces tâches s'apparentaient à des scénarios d'usage utilisés dans le cadre de tests utilisateurs. Par exemple, pour le système Youtube, les utilisateurs devaient chercher une vidéo particulière, la lire, et l'enregistrer dans leurs favoris.

Ensuite, un lien vers un questionnaire en ligne, sous Google Form, était transmis à l'utilisateur. Dans ce questionnaire, ce dernier devait indiquer son âge, son sexe, puis répondre aux dix items du F-SUS, et se positionner pour finir sur un des 7 adjectifs pour l'étude de la signification.

4 RÉSULTATS

4.1 Sensibilité du F-SUS

Pour rappel, la sensibilité du F-SUS a été mesurée en comparant les évaluations de 11 systèmes couvrant différents types de systèmes différents, auprès d'un total de 439 utilisateurs. Les scores au F-SUS pour chacun des systèmes sont présentés dans le table 7 et la figure 3

Une analyse de variance univariée (ANOVA) montre que les scores au F-SUS varient de façon significative selon le système évalué (α = ,05, p < ,001). Nous avons également appliqué le test d'additivité de Tukey, qui permet d'évaluer si deux éléments mesurent sensiblement la même chose. Pour notre étude, cela consiste à observer si certains systèmes sont évalués de la même façon. Les résultats, présentés dans le table 8, mettent en évidence 5 sous-ensembles, autrement dit 5 principales catégories de systèmes, que nous pouvons décrire comme suit :

- Catégorie 1: le site web Impôts.gouv.
- Catégorie 2: le site web Free.
- Catégorie 3: l'application mobile Tinder et le jeu vidéo sur mobile Crunchyroll.
- Catégorie 4 :, le site web Deliveroo et l'application mobile Messenger.
- Catégorie 5: Le site web Superprof, l'application mobile de réseau social Discord, le jeu sur mobile Archero, et les applications mobiles TV Time et Youtube.

Table 8: Résultats au test d'additivité de Tukey pour les différents systèmes testés

		Sous-ensemble pour alpha = 0.05				
Système	N	1	2	3	4	5
Impôts.gouv	34	60,51				
Free	35	66,36	66,36			
Tinder	38	67,89	67,89	67,89		
Crunchyroll	19		71,71	71,71		
Deliveroo	38		75,20	75,20	75,20	
Messenger	50		75,40	75,40	75,40	
Superprof	115			77,57	77,57	77,57
Discord	29			77,84	77,84	77,84
Archero	20			78,13	78,13	78,13
TV Time	30				83,58	83,58
Youtube	29					86,47
Sig.		,531	,225	,097	,331	,246

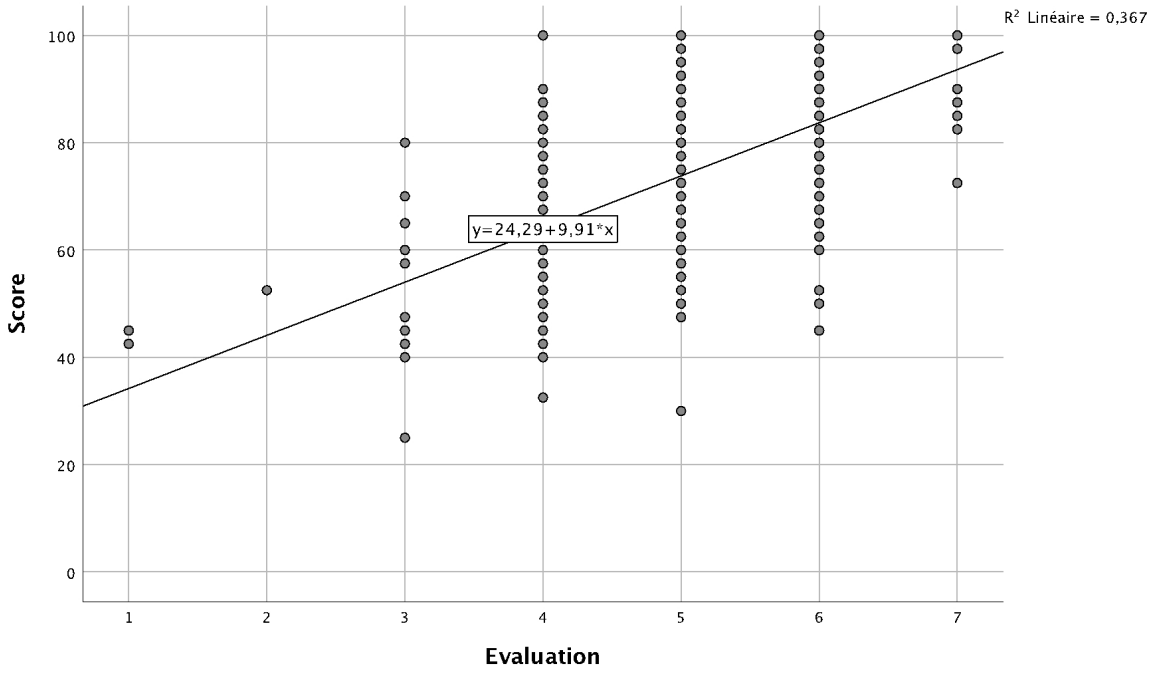

Figure 4: Représentation du lien entre les scores du F-SUS (en ordonnée, de 0 à 100) et l'adjectif (noté de 1, «Pire qu'on puisse imaginer »à 7 «Meilleure qu'on puisse imaginer »)

4.2 Signification du score du F-SUS

Afin d'évaluer la signification du score du F-SUS, nous avons tout d'abord calculé un coefficient de corrélation entre les scores obtenus et l'adjectif sélectionné par l'utilisateur. Le coefficient de corrélation de Pearson obtenu est r = ,606 significatif à ,000. Il existe donc un lien très fort entre la notation attribuée au système évalué par le F-SUS, et la qualification donnée par l'adjectif.

La représentation graphique de la régression linéaire calculée à partir du coefficient de corrélation illustre la dispersion des scores du F-SUS par rapport à chaque adjectif, ainsi que le lien progressif entre les deux formes d'évaluations (figure 4).

Le table 9 et la figure 5 présentent un résumé des résultats statistiques de l'association des scores au F-SUS avec les 7 adjectifs. A noter qu'il y a eu très peu d'évaluations de systèmes considérés comme «Pire qu'on puisse imaginer »(2 évaluations) et «Horrible »(1 évaluation).

Pour finir, nous avons mis en correspondance le score au F-SUS ainsi que l'adjectif associé à l'évaluation de chaque système de notre étude (figure 6).

Table 9: Résultats statistiques pour chaque adjectif

Adjectif	N	Moyenne	Ecart-type
Pire qu'on puisse imaginer	2	43,75	1,25
Horrible	1	52,50	0
Mauvaise	13	55,96	15,30
Acceptable	84	60,86	12,78
Bonne	183	75,11	11,63
Excellente	139	83,96	11,36
Meilleure qu'on puisse imaginer	14	88,75	7,83

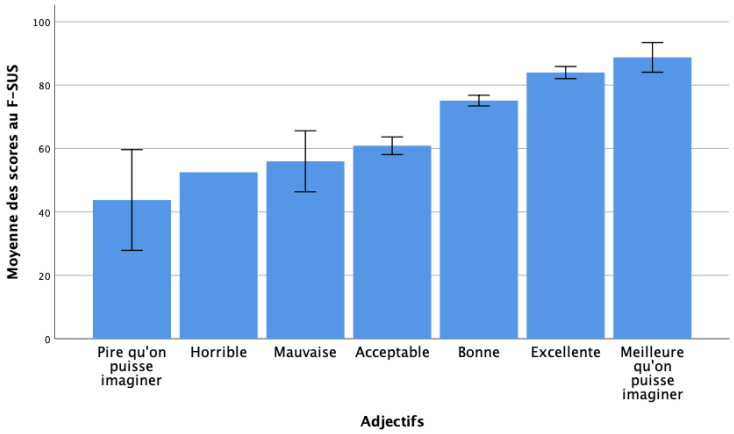

Figure 5: Moyenne des scores au F-SUS pour chaque adjectif (barres d'erreur: intervalle de confiance à 95%)

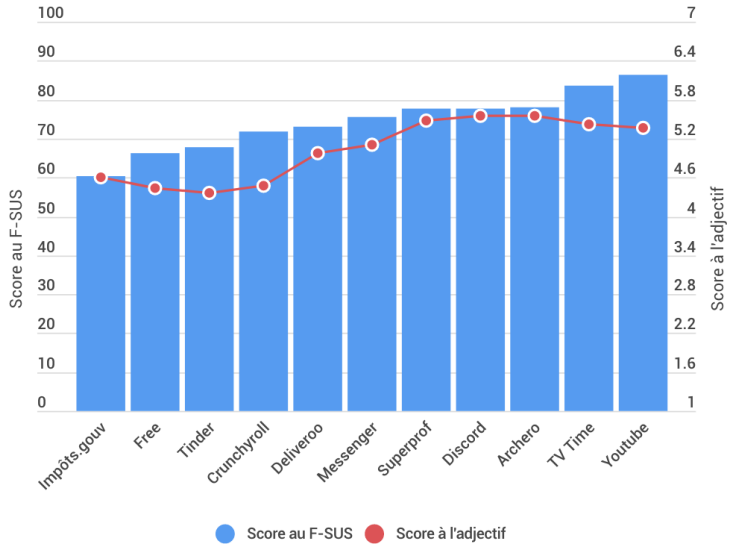

Figure 6: Moyennes des scores au F-SUS et de l'adjectif associé (Pire qu'on puisse imaginer = 1 ; Horrible = 2 ; Mauvaise = 3 ; Acceptable = 4 ; Bonne = 5 ; Excellente = 6 ; Meilleure qu'on puisse imaginer = 7) pour chaque système, classé par ordre croissant de score au F-SUS

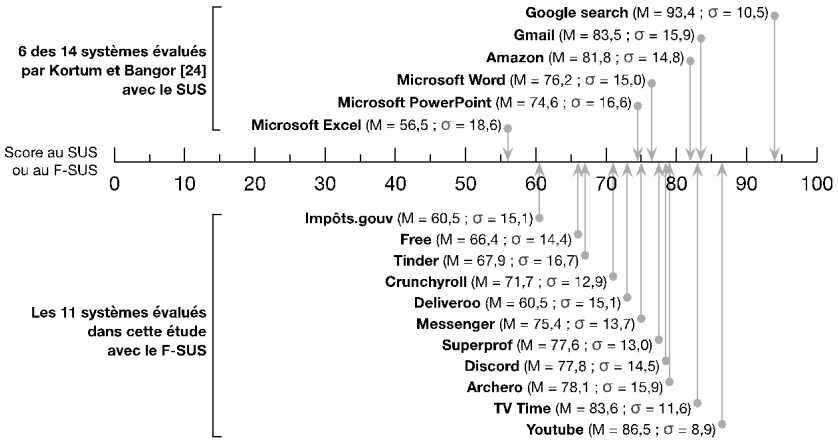

Figure 7: Positionnement des scores (avec moyennes et écart-types) au SUS pour 6 des 14 systèmes du quotidien évalués par Kortum et Bangor [25], et des scores au F-SUS pour les 11 systèmes évalués dans cette étude

5 DISCUSSION

5.1 Sensibilité du F-SUS

La sensibilité du F-SUS a été mesurée en comparant les scores obtenus pour l'évaluation de 11 systèmes interactifs de différents types (site web, application mobile, jeux vidéos, etc.), auprès de 439 d'utilisateurs. Les résultats ont montré que les scores au F-SUS différaient de façon significative selon le système. Ces résultats rejoignent ceux obtenus par Bangor et al. [4], qui avaient toutefois préféré comparer différents types de systèmes entre eux, plutôt que les systèmes eux-mêmes.

A titre d'illustration, nos 11 systèmes ont été positionnés par rapport à 6 des 14 principaux systèmes du quotidien évalués par Kortum et Bangor [25] (figure 7). Même s'il est difficile d'établir une analyse comparative, nous pouvons toutefois observer que nos systèmes se répartissent de façon cohérente selon leur score au F-SUS, avec une supériorité des outils Google (Gmail, Google search et Youtube) en termes d'utilisabilité.

Les catégorisations créées par le test d'additivité de Tukey (table 8) nous semblent intéressantes d'être discutées dans cette partie. En effet, ces catégories de systèmes, différenciées en fonction des scores obtenus au F-SUS, font apparaître des types de systèmes particuliers, et mettent par là-même en évidence la capacité du F-SUS à différencier ces types de systèmes. Ainsi, le site impôts.gouv, qui permet d'accéder à l'ensemble des services en ligne proposés par la Direction générale des finances publiques (DGFiP), et en particulier de déclarer et payer ses impôts, se distingue de tous les autres systèmes de notre étude et obtient également le score le plus faible, avec une moyenne de 60,51. Cela peut tout d'abord s'expliquer par la nature-même du site impôts.gouv, dont l'utilité et l'usage sont étroitement liés à la réalisation de tâches généralement perçues comme peu attrayantes. De plus, de tous les autres systèmes de notre étude, le site impôts.gouv est celui dont les utilisateurs ne sont pas des clients effectifs ou potentiels, mais des citoyens à qui il doit être rendu un service public. La part commerciale, et de ce fait inévitablement attractive de ce site, est par conséquent plus faible que pour les autres systèmes.

La seconde catégorie obtenue par le test d'additivité est le site Free, en tant que vitrine des offres proposées par Free pour accéder à Internet. Par rapport aux 10 autres systèmes de notre étude, le site Free est le seul à ne proposer aucun service, mais uniquement des informations et des offres à vocation commerciale. Il nous semble alors que c'est pour cela qu'il se distingue des autres systèmes évalués.

Les catégories suivantes, 3, 4 et 5, regroupent indistinctement des applications mobiles de réseaux sociaux (Discord, Messanger), et d'autres types d'applications ou de site web dont les finalités sont très variées: jeux (Archero et Crunchyroll), enseignements particuliers (Superprof), service de livraison à domicile (Deliveroo), application mobile de vidéos (Yotube), application mobile de gestion de films et de séries (TV Time). Si ces regroupements hétérogènes ne permettent pas d'identifier clairement différentes catégories de systèmes, comme l'ont fait Bangor et al. [4], ils montrent néanmoins que l'évaluation par le F-SUS peut être appliquée à de nombreux types de systèmes, et que les scores obtenus varient d'un système à l'autre. Ces résultats sont en faveur d'une sensibilité forte du F-SUS, qui permettra donc de différencier la qualité de l'utilisabilité entre plusieurs systèmes.

5.2 Signification du score du F-SUS

Pour mesurer la signification du F-SUS, nous nous sommes appuyés sur les travaux de Bangor et al. [3, 4] et nous avons repris les 7 adjectifs qualificatifs proposés par les auteurs (table 4). Nos résultats ont montré une corrélation significative (r = ,606) entre le score au F-SUS et l'adjectif sélectionné par l'utilisateur. Néanmoins, ces résultats ne sont pas aussi saillants que ceux obtenus par Bangor et al., qui avaient calculé un coefficient de corrélation r = ,806 dans le cadre de leur première étude [4], et un coefficient de corrélation r = ,822 lors de leur seconde étude [3].

Nous pensons que la différence de résultats entre notre étude et celles des auteurs peut tout d'abord s'expliquer par la traduction en français des adjectifs originaux. En effet, nous pouvons constater un aplatissement des scores pour les adjectifs négatifs. Autrement dit,

Table 10: Moyennes des scores au F-SUS et au SUS pour chacun des adjectifs

Adjectif traduit	Adjectif original	Score obtenu au F-SUS	Score obtenu au SUS par Bangor et al. [4] (première étude)	Score obtenu au SUS par Bangor et al. [3] (seconde étude)
Pire qu'on puisse imaginer	Worst imaginable	43,75	25	12,5
Horrible	Awful	52,50	NA	20,3
Mauvaise	Poor	55,96	39,17	35,7
Acceptable	OK	60,86	52.01	50,9
Bonne	Good	75,11	72.75	71,4
Excellente	Excellent	83,96	85,58	85,5
Meilleure qu'on puisse imaginer	Best imaginable	88,75	100	90,9

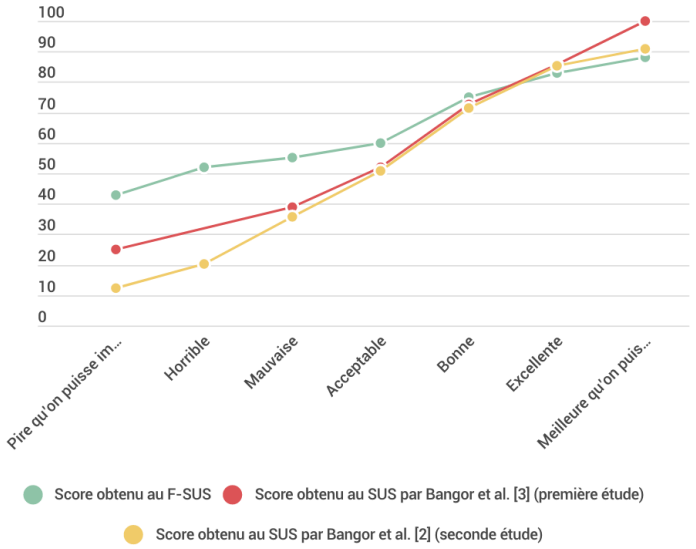

Figure 8: Moyenne des scores au F-SUS et au SUS pour chaque adjectif

les adjectifs à valeur négative (Pire qu'on puisse imaginer, Horrible, Mauvaise, Acceptable) couvrent une plage de scores plus réduite que ceux obtenus par Bangor et al. au cours de leurs deux études. Ainsi, l'amplitude des scores du F-SUS entre les adjectifs «Pire qu'on puisse imaginer »(M = 43,75) et «Acceptable »(M = 60,86), est égale à 17,11 points. En revanche, cette amplitude est de 27,01 points pour la première étude de Bangor et al., et de 38,4 points pour la seconde (table 10).

Il semble ainsi que pour notre étude en français, les adjectifs négatifs n'aient peut-être pas les mêmes nuances linguistiques qu'en anglais, et que pour les utilisateurs francophones, un système qui soit par exemple le «pire qu'on puisse imaginer », est proche d'un système jugé «horrible ». Aussi pourrions-nous utiliser et tester d'autres adjectifs, et demander aux utilisateurs d'en établir un classement afin de s'assurer qu'ils expriment bien des intensités qualificatives différentes. A titre d'exemple, ces adjectifs pourraient être :

- «La pire jamais vécue », au lieu de «La pire qu'on puisse imaginer »;
- «Épouvantable »au lieu de «Horrible »;
- «Déplorable »au lieu de «Mauvaise ».

Les résultats divergents entre ceux de notre étude et ceux de Bangor et al. pourraient également s'expliquer par ce qu'on pourrait appeler une *différence culturelle de jugement*, avec une exigence vis-à-vis de la qualité des systèmes plus élevée chez les francophones, et donc un jugement plus sévère. En effet, on peut observer que la moyenne des scores pour les 4 adjectifs à valeur négative (Pire qu'on puisse imaginer, Horrible, Mauvaise, Acceptable) correspond à des scores au F-SUS plus élevés que ceux obtenus par Bangor et al. [3, 4] avec une population anglophone. Cela signifie que les utilisateurs francophones jugent les systèmes plutôt négativement à l'aide des adjectifs, mais n'attribuent pas pour autant de faibles notes aux items du F-SUS.

Cette différence culturelle de jugement est illustrée par la figure 8. On peut ainsi constater que plus la valeur des adjectifs est négative (vers la gauche du graphique), plus l'écart avec les résultats obtenus auprès d'une population anglophone et une population francophone est important.

Toutefois, les hypothèses exprimées ci-dessus pour expliquer les différences de jugement n'excluent pas les effets de population: il

est ainsi tout à fait possible que le profil très jeune (voir table 6) des utilisateurs interrogés ait un impact sur la notation au F-SUS et l'attribution d'un adjectif. Peut-être que les utilisateurs de moins de 25 ans sont alors plus exigeants vis-à-vis de l'utilisabilité des systèmes, que ne le seraient des utilisateurs plus âgés ? C'est en partie ce que conclut Bilgihan [8], dans une étude sur les attentes de la génération Y vis-à-vis des sites de e-commerce: les jeunes utilisateurs sont non seulement sensibles aux qualités pragmatiques des interfaces (facilité d'utilisation, facilité d'apprentissage, utilité, etc.), mais attendent aussi beaucoup des qualités hédoniques (esthétique, facilité d'identification, partage de valeurs, etc.). Cette hypothèse pourrait être vérifiée dans une étude complémentaire, où le même protocole que l'étude présentée dans cet article serait appliqué à des profils d'utilisateurs significativement différent en terme d'âge, mais aussi en terme d'expérience avec les technologies.

Pour finir, notons que les études de Bangor et al. [3, 4] ont été publiées en 2008 et 2009, alors que notre recherche a été menée en 2020. Il est ainsi probable que les utilisateurs interrogés par Bangor n'avaient pas les mêmes exigences par rapport aux systèmes interactifs que ceux interrogés de nos jours. Avec l'omniprésence des systèmes interactifs et la surexposition de ces systèmes par rapport à 2008, la comparaison entre les deux études doit être menée avec prudence.

6 CONCLUSION ET PERSPECTIVES DE RECHERCHE

Cette recherche visait à mesurer les capacités de la version française du System Usability Scale (baptisée F-SUS) pour deux éléments psychométriques fondamentaux: la sensibilité, en tant que composante de fiabilité, et la signification du score global. Des études complémentaires, dont certaines sont déjà en cours, pourront être menées afin de poursuivre la diffusion SUS auprès de la communauté francophone des chercheurs et des professionnels.

6.1 Mesurer la sensibilité sur d'autres variables

En ce qui concerne la sensibilité, nous avons observé que le F-SUS permettait de recueillir des scores d'utilisabilité différents, pour des systèmes différents, et ce de façon significative. Cela nous conforte dans l'idée que le System Usability Scale reste, après plus de 30 ans d'existence, une échelle disposant d'une bonne finesse discriminative.

Pour poursuivre ces analyses psychométriques, nous envisageons d'étudier la sensibilité du F-SUS vis-à-vis d'autres variables que le système lui-même. Le sexe et l'âge sont généralement des paramètres discriminants qui sont étudiés. Nous pensons également que la compétence perçue de l'utilisateur vis-à-vis des technologies constitue également une variable intéressante. Une échelle de mesure de l'aisance technologique, comme celle proposée par Brangier, Dufresne et Hammes-Adelé [12], pourrait être utilisée.

6.2 Comparaison avec d'autres échelles

Il nous semble également important de comparer les résultats du F-SUS avec d'autres échelles francophones de mesure de l'utilisabilité ou de l'UX. Malheureusement, comme nous l'avons souligné, il existe à ce jour que l'échelle AttrakDiff [19] qui a été traduite en 2015 par Lallemand et al. [27]. L'attrakDiff présentant quatre dimensions distinctes relatives à l'expérience utilisateur (qualité hédonique identité, qualité hédonique stimulation, qualité pragmatique et attractivité globale), dont une seule peut être clairement apparentée à la mesure de l'utilisabilité (qualité pragmatique), nous envisageons de comparer les résultats du F-SUS avec ceux obtenus par la dimension pragmatique de l'AttrakDiff.

6.3 Validation du F-SUS en version «positive »

Les recherches de Lewis et Sauro [33, 40] ont montré qu'une version positive du SUS, c'est-à-dire n'incluant pas d'items inversés, pouvaient éviter des erreurs d'inattention et d'incompréhension chez les utilisateurs. De plus, une version positive pouvait éviter aux chercheurs de faire des erreurs de codage. Il nous semble ainsi intéressant de poursuivre ces études en s'attachant à valider une version positive du F-SUS, et de comparer les résultats que nous obtiendrons avec ceux de Lewis et Sauro.

6.4 Validation de la version française de l'UMUX et UMUX-LITE

Pour finir, en tant que version «allégée »du SUS, les échelles UMUX [17] et UMUX-LITE [30] pourront être également traduites et validées en version française. Ces deux questionnaires nous semblent intéressants car ils permettent de conserver un score global sur 100, tout comme le SUS, à partir d'un nombre réduit d'items (4 pour l'UMUX, 2 pour l'UMUX-LITE) et sans items inversés pour l'UMUX-LITE. Ces questionnaires correspondent ainsi aux nouvelles formes de passations à distance, qui requiert des protocoles allégés avec le moins de sources d'erreurs possible de complétion pour les utilisateurs.

6.5 Approfondissement de la signification du F-SUS

Plusieurs critiques peuvent être formulées vis-à-vis des études sur la signification menées par Bangor et al. [3, 4]. Deux d'entre elles nous semblent particulièrement importantes. La première critique est que les adjectifs ont été choisis par les auteurs sur une base arbitraire: aucune étape dans le protocole expérimental n'explique comment ces adjectifs ont été sélectionnés ou définis. Or, il existe beaucoup d'autres adjectifs, plus nuancés, qui correspondraient peut-être mieux à la qualification de l'utilisabilité d'un système. Pour répondre à cette remarque, nous avons débuté une nouvelle recherche dans laquelle nous interrogeons des experts en UX au cours de focus groups impliquant 3 à 4 experts. Dans un premier temps, les experts étaient invités à définir individuellement 27 adjectifs selon 3 critères de valence (négative, neutre et positive), et selon 3 niveaux d'intensité (forte, moyenne, faible). Les adjectifs devaient tous permettre de compléter la phrase «Dans son ensemble, je qualifierais l'utilisabilité de ce système de... ». Ensuite, les adjectifs définis individuellement étaient réunis, triés (les doublons étaient fusionnés) puis discutés entre les experts. Dans une dernière étape de priorisation, chaque expert étaient invités à se positionner sur 5 adjectifs par valence, ou autrement dit sélectionner 15 adjectifs dans la liste. Les adjectifs recueillant plus de deux votes ont été ceux que nous avons retenus, soit un total de 16 adjectifs.

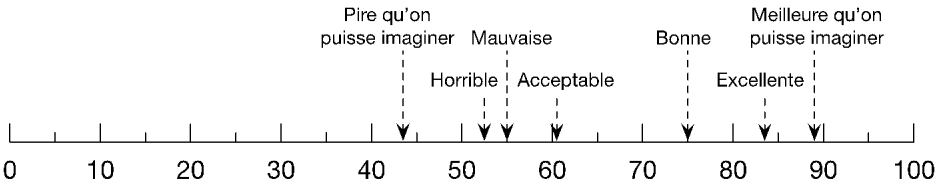

Figure 9: Echelle de signification des scores du F-SUS

La seconde critique que nous pouvons faire à l'égard des études de Bangor et al. est que les adjectifs ont été présentés aux utilisateurs sous une forme similaire à une échelle de Licket à 7 points. En effet, les 7 adjectifs ont été présentés par ordre croissant de valence (de «pire qu'on puisse imaginer »à «meilleure qu'on puisse imaginer ») ; les utilisateurs devaient alors cocher l'une des 7 cases. Or, dans cette configuration, ce n'est pas le choix d'un adjectif arbitraire qui est réalisé, mais un positionnement sur une seconde échelle, après celle du SUS. Pour remédier à ce biais, nous proposons de présenter aux utilisateurs les 16 adjectifs, définis dans le paragraphe précédent, dans un ordre aléatoire. Les utilisateurs devront également sélectionner un adjectif dans une liste déroulante, et non plus le choisir dans échelle semi-ordinale.

6.6 Conclusion provisoire sur la signification

Le calcul de la signification est un élément important de communication du score du SUS auprès des commanditaires ou des clients [14]. Nous avons notamment souligné que nos résultats se différenciaient de ceux obtenus par Bangor et al. [3, 4], en ce qui concerne plus particulièrement la correspondance des scores du SUS aux adjectifs qualificatifs à valeur négative. Si nous avons proposé quelques pistes pour approfondir cet axe de recherche, nous pouvons toutefois établir une échelle de signification, qui aidera à interpréter les scores du F-SUS (figure 9). Cette échelle devrait permettre d'identifier rapidement si l'utilisabilité d'un système est acceptable ou non.

7 REMERCIEMENTS

Les auteurs remercient les étudiants de la promotion 2019-2020 en L3 MIAGE de l'Université de Nancy, pour la qualité de leur travail en groupe en cours d'Interaction Homme-Machine.

Cet article s'appuie également en partie sur les travaux réalisés avec Alexandre Baudet, chercheur en Psychologie au Luxembourg Institute of Science and Technology.

REFERENCES

[1] . Adinda, P.P. and Suzianti, A. Redesign of user interface for E-government application using usability testing method. *ICCIP*, (2018), 145–149.

[2] . AlGhannam, B.A., Albustan, S.A., Al-Hassan, A.A., and Albustan, L.A. Towards a Standard Arabic System Usability Scale: Psychometric Evaluation using Communication Disorder App. *International Journal of Human-Computer Interaction 34*, 9 (2018), 799–804.

[3] . Bangor, A., Kortum, P., and Miller, J. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies 4*, 3 (2009), 114–123.

[4] . Bangor, A., Kortum, P.T., and Miller, J.T. An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction 24*, 6 (2008), 574–594.

[5] . Baumgartner, J., Frei, N., Kleinke, M., Sauer, J., and Sonderegger, A. Pictorial System Usability Scale (P-SUS): Developing an instrument for measuring perceived usability. *CHI 2019, May 4–9, 2019, Glasgow, Scotland, UK*, (2019), 1–11.

[6] . Berke, L., Huenerfauth, M., and Patel, K. Design and psychometric evaluation of American sign language translations of usability questionnaires. *ASSETS'17, Oct. 29–Nov. 1, 2017, Baltimore, MD, USA*, (2017), 175–184.

[7] . Beul-Leusmann, S., Samsel, C., Wiederhold, M., Krempels, K.H., Jakobs, E.M., and Ziefle, M. Usability evaluation of mobile passenger information systems. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2014, 217–228.

[8] . Bilgihan, A. Gen y customer loyalty in online shopping: An integrated model of trust, user experience and branding. *Computers in Human Behavior 61*, (2016), 103–113.

[9] . Blažica, B. and Lewis, J.R. A Slovene Translation of the System Usability Scale: The SUS-SI. *International Journal of Human-Computer Interaction 31*, 2 (2015), 112–117.

[10] . Borkowska, A. and Jach, K. Pre-testing of polish translation of system usability scale (SUS). In *Advances in Intelligent Systems and Computing*. 2017, 143–153.

[11] . Borsci, S., Federici, S., and Lauriola, M. On the dimensionality of the System Usability Scale: A test of alternative measurement models. *Cognitive Processing 10*, 3 (2009), 193–197.

[12] . Brangier, E., Dufresne, A., and Hammes-Adelé, S. Approche symbiotique de la relation humain-technologie: perspectives pour l'ergonomie informatique. *Le Travail Humain 74*, 4 (2009), 333–353.

[13] . Brooke. SUS: A "quick and dirty" usability scale. In P.W. Jordan, B. Thomas, B.A. Weerdmeester and I.. McClelland, eds., *Usability evaluation in industry*. London: Taylor & Francis, 1996, 189–194.

[14] . Brooke, J. SUS: A Retrospective. *Journal of Usability Studies 8*, 2 (2013), 29–40.

[15] . Cattell, R.B. The Scree Test For The Number Of Factors. *Multivariate Behavioral Research 1*, 2 (1966), 245–276.

[16] . Dianat, I., Ghanbari, Z., and AsghariJafarabadi, M. Psychometric properties of the persian language version of the system usability scale. *Health promotion perspectives 4*, 1 (2014), 82–829.

[17] . Finstad, K. The usability metric for user experience. *Interacting with Computers 22*, 5 (2010), 323–327.

[18] . Gronier, G. and Baudet, A. Psychometric evaluation of the F-SUS: Creation and validation of the French version of the System Usability Scale. *International Journal of Human-Computer Interaction*, .

[19] . Hassenzahl, M., Burmester, M., and Koller, F. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwillus and J. Ziegler, eds., *Mensch & Computer 2003: Interaktion in Bewegung*. 2003, 187–196.

[20] . Hornbaek, K. Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies 64*, 2 (2006), 79–102.

[21] . Hussain, A., Mkpojiogu, E.O.C., and Hussain, Z. Usability evaluation of a web-based health awareness portal on smartphone devices using ISO 9241-11 model. *Jurnal Teknologi 77*, 4 (2015), 1–5.

[22] . ISO 16982:2002. Méthodes d'utilisabilité pour la conception centrée sur l'opérateur humain. 2002.

[23] . ISO 9241-11:1998. Exigences ergonomiques pour travail de bureaux avec terminaux à écrans de visualisation. Partie 11: Lignes directrices concernant l'utilisabilité. 1998.

[24] . Katsanos, C., Tselios, N., and Xenos, M. Perceived usability evaluation of learning management systems: A first step towards standardization of the system usability scale in Greek. *Proceedings of the 2012 16th Panhellenic Conference on Informatics, PCI 2012, Piraeus, Greece: IEEE CPS, 5-7 Oct., 2012*, (2012), 302–307.

[25] . Kortum, P.T. and Bangor, A. Usability Ratings for Everyday Products Measured With the System Usability Scale. *International Journal of Human-Computer Interaction 29*, 2 (2013), 67–76.

[26] . Lallemand, C. and Gronier, G. Méthodes de Design UX. 30 méthodes fondamentales pour concevoir des expériences optimales. Paris, 2018.

[27] . Lallemand, C., Koenig, V., Gronier, G., and Martin, R. 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 (2015), 239–252.

[28] . Landauer, T.K. Chapter 9 - Behavioral Research Methods in Human-Computer Interaction. In *Handbook of Human-Computer Interaction (Second Edition)*. 1997, 203–227.

[29] . Larue, V. Apport du System Usability Scale à l'activité d'ergonomie d'évaluation. *IHM 2009, 13-16 Octobre 2009, Grenoble, France*, (2009), 155–161.

[30] . Lewis, J., Utesch, B., and Maher, D. UMUX-LITE: when there's no time for the SUS. *Proceedings of the SIGCHI Conference on Human Computer Interaction*, (2013), 2099–2102.

[31] . Lewis, J.R. The System Usability Scale: Past, Present, and Future. *International Journal of Human-Computer Interaction 34*, 7 (2018), 577–590.

[32] . Lewis, J.R. and Sauro, J. The Factor Structure of the System Usability Scale. Human Centered Design, First International Conference, HCD 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, (2009), 26–31.

[33] . Lewis, J.R. and Sauro, J. Revisiting the Factor Structure of the System Usability Scale. *Journal of Usability Studies 12*, 4 (2017), 183–192.

[34] . Maguire, M. Methods to support human-centred design. *International Journal of Human-Computer Studies 55*, 4 (2001), 587–634.

[35] . Martins, A.I., Rosa, A.F., Queirós, A., Silva, A., and Rocha, N.P. European Portuguese Validation of the System Usability Scale (SUS). *Procedia Computer Science 67*, Dsai (2015), 293–300.

[36] . Nunnally, J.C. An Overview of Psychological Measurement. In B.B. Wolman, ed., *Clinical Diagnosis of Mental Disorders*. New York, 1978, 97–146.

[37] . Peruri, A., Borchert, O., Cox, K., Hokanson, G., and Slator, B.M. Using the system usability scale in a classification learning environment. In *Advances in Intelligent Systems and Computing*. 2017, 167–176.

[38] . Roto, V., Law, E., Vermeeren, A., and Hoonhout, J. *User Experience White Paper. Bringing clarity to the concept of user experience*. 2010.

[39] . Rummel, B. System Usability Scale – jetzt auch auf Deutsch. 2015. System Usability Scale – jetzt auch auf Deutsch.

[40] . Sauro, J. and Lewis, J.R. When designing usability questionnaires, does it hurt to be positive? *Conference on Human Factors in Computing Systems - Proceedings*, (2011), 2215–2223.

[41] . Sauro, J. and Lewis, J.R. *Quantifying the User Experience*. Elsevier, 2012.

[42] . Sharfina, Z. and Santoso, H.B. An Indonesian adaptation of the System Usability Scale (SUS). *2016 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2016*, (2016), 145–148.

[43] . South, H., Taylor, M., Dogan, H., and Jiang, N. Digitising a medical clerking system with multimodal interaction support. *ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction 2017-Janua*, (2017), 238–242.

[44] . Tolentino, G.P., Battaglini, C., Pereira, A.C.V., De Oliveria, R.J., and De Paula, M.G.M. Usability of serious games for health. *Proceedings - 2011 3rd International Conferenceon Games and Virtual Worlds for Serious Applications, VS-Games 2011*, (2011), 172–175.

[45] . Vallerand, R.J. Vers une méthodologie de validation trans-culturelle de questionnaires psychologiques: Implications pour la recherche en langue française. *Psychologie Canadienne 30*, 4 (1989), 662–680.

[46] . Yang, D., Zhang, D., Chen, L., and Qu, B. NationTelescope: Monitoring and visualizing large-scale collective behavior in LBSNs. *Journal of Network and Computer Applications 55*, August (2015), 170–180.