
Vers un environnement informatisé d'évaluation de la qualité ergonomique d'interfaces multimédia

Étude exploratoire

Stéphane Caro

*Université de Bourgogne, IUT de Dijon
Laboratoire sur l'Image les Médiations et le Sensible
en Information-Communication (LIMSIC)
Bd du Docteur Petitjean
B.P. 17867
21078 Dijon Cedex
stephane.caro@u-bourgogne.fr*

RÉSUMÉ. Notre objectif est d'encourager les concepteurs à s'intéresser aux usages de leurs produits en prenant davantage en compte les facteurs humains dans la phase de test qui est souvent focalisée sur les aspects techniques. L'idée dans cette étude est donc de faciliter le recours aux utilisateurs comme source d'évaluation lors des tests de conception en apportant aux concepteurs un environnement logiciel de test pratique à mettre en œuvre. Nous avons mené une expérience de nature à explorer la faisabilité de ce projet.

ABSTRACT. Our objective is to encourage the human-computer interface designers to be more interested in the uses made of their product by taking into account to a higher degree the human-factor in the testing phase which often is too focused on technical aspects. We want to facilitate the integration of the user as a source of evaluation during the testing phase. Our aim is to offer the designer a software environment of practical tests that allow them to have their own products evaluated by their users. We undertook an experiment to explore the feasibility of this project.

MOTS-CLÉS : interface homme-machine, hypermédia, ergonomie, méthode d'évaluation assistée par ordinateur, critères de qualité ergonomique, évaluation par l'utilisateur.

KEY WORDS : man-machine interfaces, hypermedia, ergonomics, computerized evaluation method, ergonomic quality standards, evaluation by the user.

1. Positionnement de l'étude

Les interfaces multimédia (pages web, Cédéroms) et plus généralement les interfaces logicielles souffrent souvent de la non application des connaissances du moment en ergonomie. Les applications informatiques professionnelles bénéficient de moyens de développement importants. Ce n'est pas le cas des interfaces multimédia et en particulier de bon nombre de sites web, ces derniers pouvant être développés à peu de frais. Il n'existe pas de méthode d'évaluation « grand public » des aspects ergonomiques d'une interface multimédia. Certaines tentatives, présentées plus loin, sont dans l'ensemble destinées seulement aux interfaces web et n'adressent pas toujours les dimensions de l'analyse ergonomique.

L'objet de cet article est la présentation d'une méthode de test ergonomique assistée par ordinateur et utilisable par des novices en ergonomie. Après une courte description du projet et un exposé des méthodes existantes (première partie) nous proposons dans la deuxième partie de valider la méthode de test par une expérience exploratoire. Les problèmes soulevés par l'utilisation de cette méthode sont abordés dans la discussion (troisième partie).

Notre objectif ici est d'accroître l'intérêt porté par les concepteurs aux usages de leur produit. Pour cela une voie intéressante est de favoriser la prise en compte des facteurs humains dans la phase de test qui est souvent centrée sur les aspects techniques. L'évaluation ergonomique d'interfaces logicielles rencontre en général deux obstacles. L'un est culturel (en témoigne la fréquente omission de cette étape dans les *process* de conception) et l'autre économique (coût de l'intervention d'experts). La tendance actuelle des solutions économiques est d'aller vers des recueils de recommandations ou de critères de qualités ergonomiques utilisables par les concepteurs eux même, sans exiger d'eux qu'ils soient experts en ergonomie. Un autre seuil de vulgarisation consiste à inciter les concepteurs à recourir aux utilisateurs en leur demandant d'évaluer le produit après une phase d'utilisation des fonctions représentatives de celui-ci — *tests d'utilisations*. Parmi l'ensemble des méthodes d'évaluation, une voie également prometteuse est celle de *l'évaluation automatique* par un logiciel de certaines dimensions ergonomiques prévisibles (lisibilité, densité informationnelle par ex.). Les *méthodes d'inspection ergonomique* sont aussi une source de diagnostic qui complète souvent valablement les tests utilisateurs, car pour l'heure il n'existe aucun programme informatique qui puisse évaluer automatiquement le respect de tous les critères de qualité ergonomiques connus. L'évaluation par des tests utilisateurs à l'aide d'une méthode basée sur l'inspection ergonomique nous semble une voie intéressante si elle peut être assistée par des outils méthodologiques (questionnaire en ligne par ex.).

1.1. Dispositifs analogues existants

Des dispositifs de test de ce type existent déjà pour l'évaluation de sites web comme le questionnaire WAMMI. Ce questionnaire d'évaluation de site web comprend 20 questions (voir fig. 1 pour un aperçu de quelques questions) qui donnent lieu à un diagnostic en 6 dimensions d'analyse du produit. Un exposé de la méthode peut être consulté sur le site de la société qui la commercialise¹.

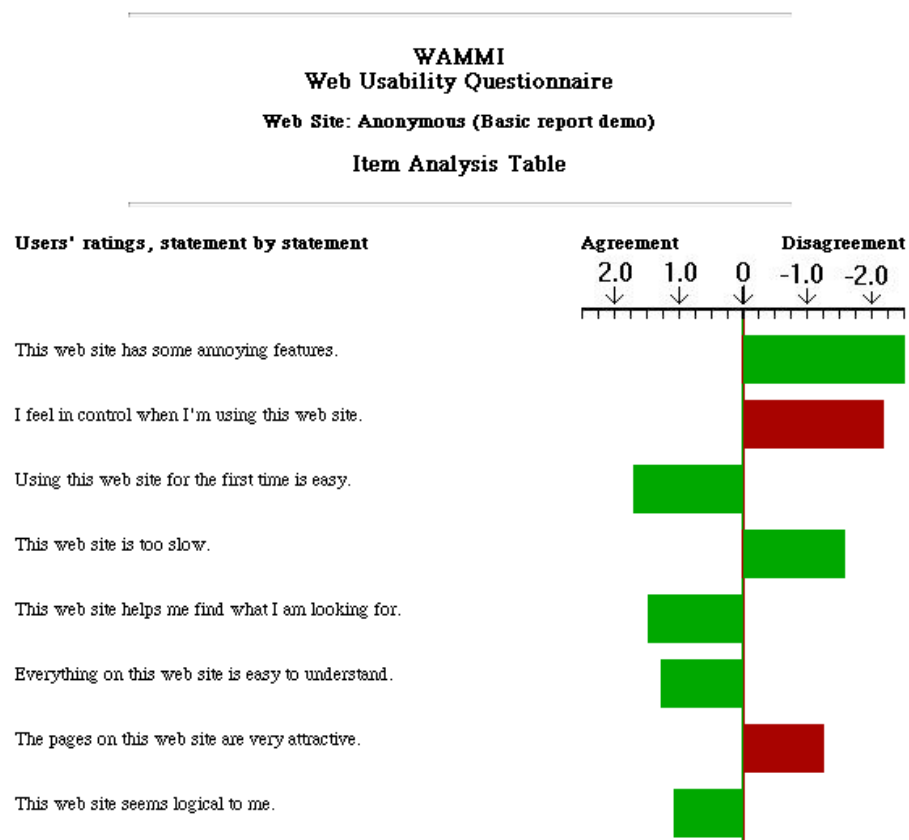


Figure 1. Exemples de questions posées lors d'un test WAMMI.

La figure 2 présente un tableau récapitulatif des résultats d'une évaluation de site selon les 6 dimensions d'analyse choisies :

1. www.nomos.se, Human Factors Research Group & Nomos Management AB, 1998.

- L'attractivité du site.
- Le contrôle de l'utilisateur sur l'interface.
- L'efficacité du site.
- L'aide disponible.
- La facilité d'apprentissage.
- La facilité d'utilisation globale.

WAMMI
Web Usability Questionnaire
Web Site: Anonymous (Basic report demo)
Numeric Summary

| Scale | Median | Mean | Standard Dev. |
|------------------|--------|-------|---------------|
| Attractiveness | 44 ? | 47.87 | 24.08 |
| Controllability | 48 | 48.24 | 23.27 |
| Efficiency | 60 | 57.20 | 23.58 |
| Helpfulness | 66 | 60.70 | 21.46 |
| Learnability | 58 ? | 53.02 | 27.91 |
| Global Usability | 58 ? | 54.73 | 21.71 |

Figure 2. Résultats des 6 dimensions d'analyse.

Le premier inconvénient de la méthode est que le commanditaire ne dispose pas de l'outil d'évaluation. Pour cause, chaque évaluation de site est facturée entre 600 ₣ (évaluation standard) et 1400 ₣ pour un questionnaire avec des questions rajoutées par le commanditaire (www.wammi.com). Par ailleurs ce service est assez spécifique aux sites web et ne distingue pas les dimensions d'analyse liées à l'ergonomie (assistance, facilité d'utilisation, facilité d'apprentissage) de l'utilité du produit (fonctionnalités offertes, performances du système, fiabilité technique) et enfin les considérations esthétiques (design, typographie, couleurs etc.). De plus, le rapprochement qui est fait entre les 20 questions posées aux utilisateurs et le calcul des 6 indices généraux présentés dans le bilan (voir fig. 2) n'est pas transparent pour le commanditaire.

Nous proposons de focaliser notre attention sur l'ergonomie uniquement. En effet cette dimension n'est pas toujours évaluée faute d'outils adaptés. L'utilité est plus facilement évaluable même par des non-spécialistes (fonctionnalités offertes, performances et stabilité du système etc.). L'esthétique du produit quant à elle pose des problèmes de subjectivité difficiles à maîtriser. Il existe déjà des outils d'analyse automatisés qui traitent aussi des sites web pour certaines dimensions ergonomiques.

L'outil *Bobby* par exemple analyse le code HTML des pages et vérifie qu'il permet l'accès aux déficients visuels, indique les versions de navigateur compatibles avec la page et donne une estimation du temps de chargement de la page avec un modem 28,8 kbps (<http://www.cast.org/bobby/>). Deux outils proposés par le *WebMetrics Tool Suite* permettent d'évaluer des dimensions complémentaires (<http://zing.ncsl.nist.gov/webmet/>).

- Le WebSAT analyse les sites du point de vue de l'accessibilité aux différents publics (utilisateurs d'anciennes versions de navigateurs, vitesses de connections dégradées, matériel informatique ancien), l'utilisation des formulaires (présence de boutons d'annulation, de réinitialisation de la saisie, rétroaction après l'envoi...), les performances (vitesse de chargement des images, optimisation de leur type et de leur taille, présence de balises HTML indiquant la taille des images pour en accélérer l'affichage...), maintenance des pages (contrôle des liens, des informations de maintenance, dates de mise à jour et auteurs de la page), la navigation (usage limité du bouton précédent ou *back*, coloration des liens, ouverture de multiples fenêtres...), lisibilité (nombre de liens ou mise en relief typographique excessifs ...).
- Le WebCAT permet de contrôler par une technique de tri de carte (*card sorting*) si l'arborescence du site est bien appariée avec la représentation que les utilisateurs en attendent. Les sujets doivent faire glisser les items de l'arborescence dans les catégories prévues par les concepteurs. On analyse l'accord entre les sujets pour savoir si les termes et regroupements sont bien choisis (voir fig. 3).

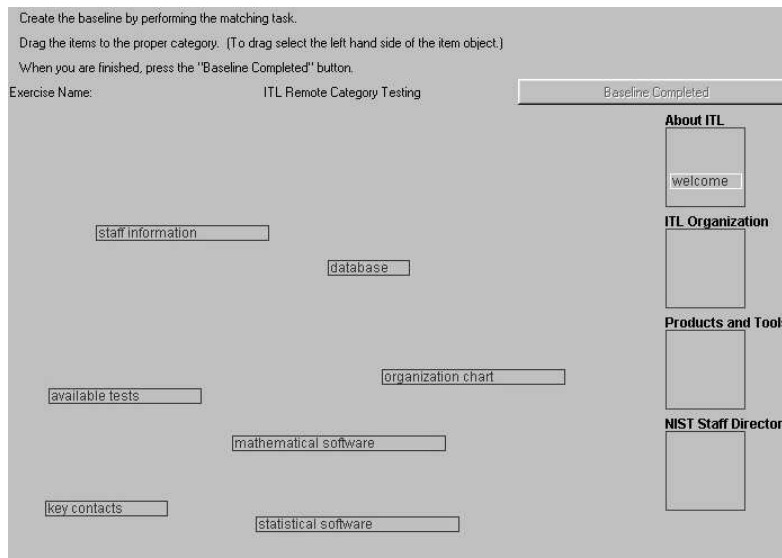


Figure 3. Logiciel d'évaluation des arborescences WebCAT (catégories à droite, items à classer dans les catégories à gauche).

Il manque aux outils existant certaines dimensions d'analyse. L'homogénéité des choix de conception au cours des écrans, la prise en compte de l'expérience de l'utilisateur, la compatibilité avec ses habitudes, la brièveté des séquences d'actions destinées à atteindre un but, autant de dimensions qu'un outils entièrement automatisé ne peut évaluer. L'évaluation humaine reste encore plus complète pour ces dimensions, ce qui nous a conduit à l'assister par des outils appropriés.

1.2. Présentation du projet

L'objectif à moyen terme serait de fournir aux équipes de conception un guide méthodologique succinct accompagné d'un CD ROM ou d'un site web comprenant un logiciel d'évaluation ergonomique par questionnaire. Ce questionnaire en ligne est basé sur les critères de qualité ergonomique recensés à l'INRIA par Bastien, Leulier et Scapin [BAS 98]. Ces critères présentent l'avantage d'avoir déjà fait l'objet de tests d'utilisation avec des sujets non experts en ergonomie [BAS 92]. La liste ci-dessous présente ces critères.

| Liste des critères de qualité ergonomique. | |
|---|--|
| Les 18 critères élémentaires apparaissent en caractères gras. | |
| 1. Guidage | |
| 1.1 Prompting | |
| 1.2 Groupement/Distinction entre items | |
| 1.2.1 Groupement/Distinction par la localisation | |
| 1.2.2 Groupement/Distinction par le format | |
| 1.3 Feed-back immédiat | |
| 1.4 Lisibilité | |
| 2. Charge de travail | |
| 2.1 Brièveté | |
| 2.1.1 Concision | |
| 2.1.2 Actions minimales | |
| 2.2 Densité informationnelle | |
| 3. Contrôle explicite | |
| 3.1 Actions explicites | |
| 3.2 Contrôle utilisateur | |
| 4. Adaptabilité | |
| 4.1 Flexibilité | |
| 4.2 Prise en compte de l'expérience de l'utilisateur | |
| 5. Gestion des erreurs | |
| 5.1 Protection contre les erreurs | |
| 5.2 Qualité des messages | |
| 5.3 Correction des erreurs | |

6. Homogénéité / Cohérence
 7. Signifiante des codes et dénominations
 8. Compatibilité

Dans ce projet de questionnaire en ligne, pour chaque critère (élémentaire), les utilisateurs doivent déplacer un curseur sur une échelle bi-polaire non graduée (selon la technique de l'échelle bi-polaire [BIS 99]). Celle-ci quantifie le respect de chaque critère au sein de l'interface tel que perçu par l'utilisateur (voir figure 4) et donne aux concepteurs un diagnostic ergonomique rapide. Les réponses des sujets permettent également de recueillir des critiques et suggestions d'amélioration du produit à l'aide d'un champ textuel réservé à cet usage pour chaque critère. Nous pensons que même si la qualité scientifique du test n'est pas parfaite, favoriser un contact entre concepteurs et utilisateurs constituerait déjà une avancée positive.

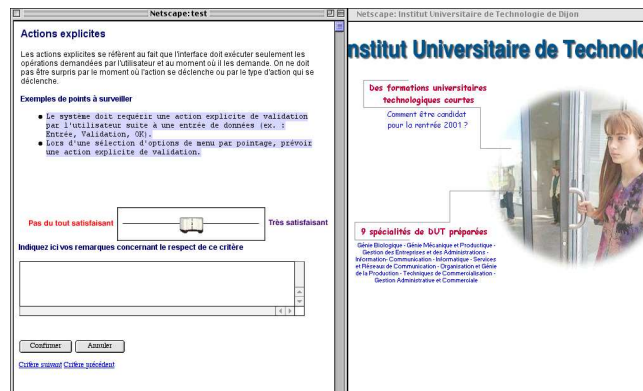


Figure 4. Logiciel d'évaluation à gauche et interface évaluée à droite.

L'hypothèse directrice est qu'un outil d'assistance à l'évaluation ergonomique pourrait être utilisé, tant par des sujets experts que novices au domaine du produit testé. Ainsi des tests d'ergonomie pourraient être conduits rapidement au sein des *process* de production. Nous avons mené une expérience exploratoire sur « support » papier préfigurant l'utilisation de la technique afin d'en mesurer l'efficacité. Pour réaliser ce test nous avons choisi deux logiciels d'administration de systèmes informatiques très différents en phase de prototypage : une interface de type « graphique » (menus déroulants, multifenêtrage, utilisation de la souris) et une interface « texte » (style « minitel »). L'expérience a été menée au sein de la société

8 Revue d'intelligence artificielle. Volume 14 – n° 1-2/2000

Bull au cours d'évaluations ergonomiques d'outils logiciels d'administration de parcs informatiques².

2. Étude réalisée dans le cadre d'un partenariat entre *Bull* et l'action Airelle de l'INRIA Rhône-Alpes entre septembre 96 et août 98 à Echirolles (Isère) sous la direction d'André Bisseret.

2. Expérience exploratoire

2.1. Présentation de l'expérience

On demande aux sujets d'accomplir un certain nombre de tâches représentatives de l'utilisation de chaque logiciel (environ une heure de manipulation). Après ce temps d'utilisation du logiciel, on présente la définition de chaque critère (sur papier) ainsi que quelques exemples choisis parmi ceux qui sont donnés par Bastien et Scapin pour illustrer les définitions. On demande au sujet d'évaluer le respect du critère selon la méthode de l'échelle bi-polaire. Le sujet coche sur un segment de droite non gradué, entre deux pôles (très satisfaisant, pas du tout satisfaisant). En mesurant l'endroit coché on obtient une variable continue. On recueille pour chaque critère les commentaires inscrits par les sujets dans une zone prévue à cet effet. En définitive on obtient une évaluation quantitative et qualitative du respect du critère.

2.2. Objectifs

L'objectif initial était de valider la possibilité d'accélérer la procédure de test des maquettes de logiciels à l'aide d'un outil d'assistance *ad hoc*. Un sous objectif était d'étudier la possibilité d'employer une méthode de test en partie automatisée qui soit utilisable par des novices en ergonomie.

2.3. Méthode

2.3.1. Sujets

Les participants se répartissent en 2 groupes de 4 sujets volontaires non rémunérés. Sur les 8 sujets, 4 sont « experts » dans l'activité d'administration ou l'utilisation des logiciels d'administration et 4 sont « novices » mais possèdent la formation nécessaire à la fonction d'administrateur. Deux sujets parmi les experts ne sont pas ou sont peu familiers avec le matériel *Bull* (un administrateur extérieur et un personnel *Bull* en contrat de courte durée). Les sujets « novices » sont des stagiaires *Bull* en école d'ingénieur ou en niveau équivalent (DESS) et ne sont pas familiers du matériel *Bull*. Le groupe compte 3 femmes et 5 hommes de niveau d'étude bac +3 minimum.

2.3.2. Matériel

Concernant les logiciels testés, les messages d'aide sont disponibles, même si leur contenu est provisoire. Les applications simulent un fonctionnement normal en général, même si les commandes réelles ne sont pas implémentées. Dans le cas où les commandes ne fonctionnent pas complètement, le moniteur indique au sujet le

comportement futur du logiciel. Plusieurs questionnaires sont remis ainsi qu'une consigne écrite.

2.3.3. *Dispositif technique et déroulement de l'expérience*

Les sujets sont placés en situation d'utilisation des logiciels. Les tâches demandées aux sujets sont ajustées en fonction des résultats de pré-tests. Les sujets passent l'expérience en individuel, en présence du moniteur. Ils testent les 2 logiciels (interface « graphique » et interface « texte »). L'ordre de passation des logiciels est contrebalancé (la moitié des sujets commence par l'interface « texte », l'autre moitié par l'interface « graphique »). Après la réalisation des tâches demandées par le moniteur, le sujet remplit les questionnaires d'évaluation. La passation est suivie d'un débriefing (explication des difficultés, réponses aux questions des sujets). Lors du dépouillement des questionnaires, les marques portées sur les échelles bi-polaires sont converties en variable continue à l'aide d'une règle graduée.

2.3.4. *Consignes*

La première consigne donnée aux sujets était la suivante :

L'objectif de la séance est de passer en revue le *design* de deux nouveaux logiciels d'aide à l'administration de systèmes informatiques et d'obtenir votre opinion concernant ces logiciels. Les logiciels sont destinés à l'administration d'une machine nommée *Polykid*. La machine *Polykid* est composée de plusieurs modules interconnectés. Elle est équipée d'une console d'administration. Elle fonctionne comme une machine unique, avec un système d'exploitation unique. Je vais vous demander d'effectuer certaines tâches. Au fur et à mesure que vous utiliserez les logiciels pour exécuter les tâches indiquées ci-dessous, je pourrais vous poser des questions sur ce que vous voyez ou sur ce que vous attendez des fonctions que vous utilisez. Sentez-vous libre de donner des observations pendant la session. Nous vous demandons de « penser tout haut » en indiquant les actions que vous voulez faire et les informations que vous recherchez. Il n'y a pas de question stupide ou de mauvaise réponse. Ce produit est un prototype, ne soyez donc pas étonnés qu'il puisse réagir d'une façon inattendue.

Vous allez tester deux logiciels : l'application *Service Master*, et l'application *Redundancy Manager*.

Pour chaque logiciel une consigne spécifique était présentée à la suite de cette consigne générale. La consigne spécifique explique rapidement à quoi servent les logiciels et présente les tâches à réaliser avec ceux-ci.

2.3.5. Plan d'expérience

Le plan d'expérience est à un facteur inter-sujet, le niveau d'expertise, et un facteur intra-sujet, le type de logiciel.

2.3.6. Variables recueillies

Variable quantitative L'échelle bi-polaire est traduite en une variable « note » (bornée à -5 et +5).

Variable qualitative On recueille les éventuels commentaires des sujets sur le respect du critère ou/et ses suggestions d'amélioration (à la discrétion du sujet).

2.4. Résultats

2.4.1. Analyse quantitative

On peut construire plusieurs graphiques à partir des résultats obtenus³. Les graphiques qui suivent présentent pour chaque type d'interface les moyennes des « notes » obtenues aux 18 critères pour chaque interface (interface « graphique » fig. 5 et interface « texte » fig. 6).

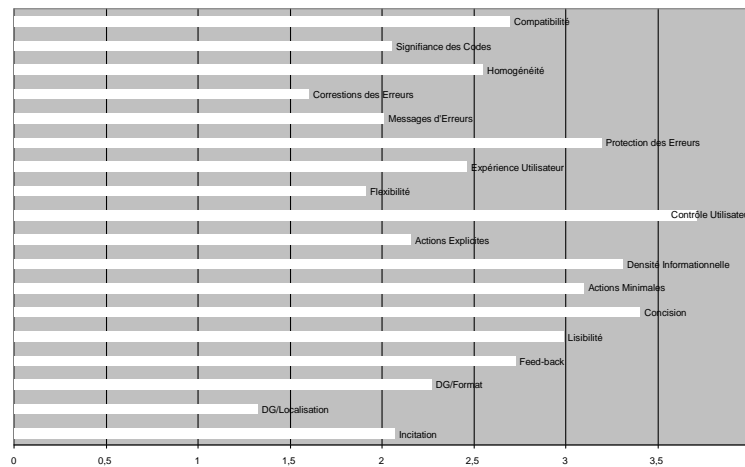


Figure 5. Interface « graphique » : moyenne des notes pour chaque critère, sujets experts et novices.

3. Ces graphiques seraient construits de façon automatique par le logiciel de test en ligne.

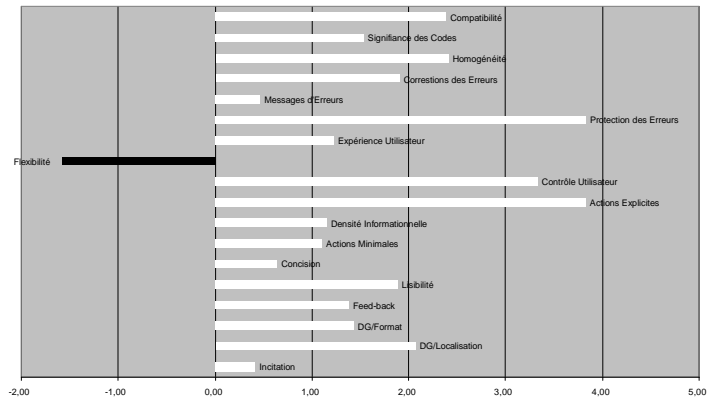


Figure 6. Interface « texte » : moyenne des notes pour chaque critère, sujets experts et novices.

On peut constater que dans l'ensemble les sujets attribuent rarement des notes négatives. Le non respect d'un critère n'est sanctionné par la note, que quand ce dernier semble manifestement enfreint (« flexibilité » dans l'interface « texte »).

Les moyennes obtenues par critère peuvent être analysées en fonction du niveau d'expertise. La figure 7 présente ce type de résultat pour l'interface « texte ».

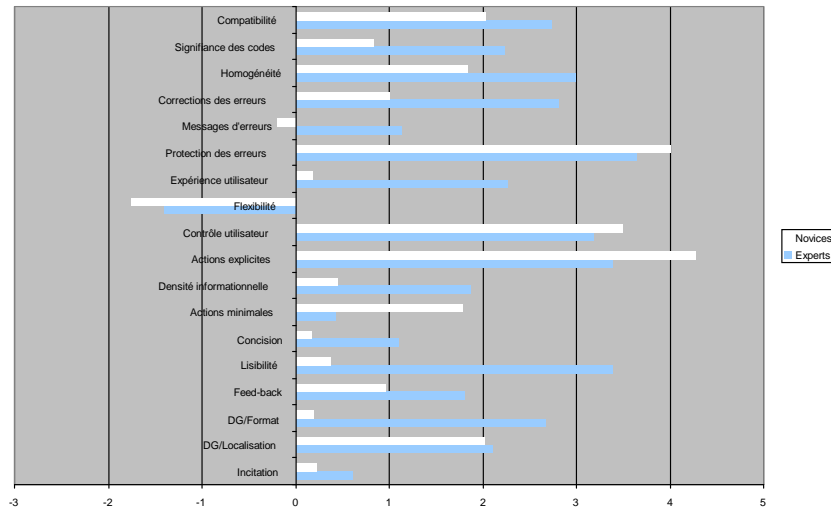


Figure 7. Interface « texte » : moyenne des notes pour chaque critère en fonction de l'expertise.

Il n'y a pas de tendance régulière à noter différemment les critères selon que les sujets sont novices ou experts du domaine d'activité. Pour chaque critère, ce sont tantôt les novices qui mettent une note plus élevée, tantôt les experts.

Il convient également de s'intéresser aux écarts de notation entre les sujets. On peut admettre que, plus les sujets vont considérer le respect d'un critère de la même manière et plus la mesure aura des chances d'être fiable. Le tableau 1 présente les moyennes et écart types obtenus pour les notes de l'interface « texte »⁴.

4. Ce tableau serait construit de façon automatique par le logiciel de test en ligne.

| | Moyenne experts et novices | Écart type experts et novices |
|---------------------------------|----------------------------|-------------------------------|
| Incitation | 0,42 | 1,94 |
| DG/Localisation | 2,07 | 2,00 |
| DG/Format | 1,44 | 2,29 |
| Feed-back | 1,38 | 2,54 |
| Lisibilité | 1,88 | 2,92 |
| Concision | 0,64 | 2,84 |
| Actions minimales | 1,11 | 1,83 |
| Densité informationnelle | 1,16 | 1,96 |
| Actions explicites | 3,83 | 1,12 |
| Contrôle utilisateur | 3,34 | 1,18 |
| Flexibilité | -1,58 | 3,03 |
| Expérience utilisateur | 1,23 | 2,97 |
| Protection/Erreurs | 3,83 | 0,77 |
| Messages d'erreurs | 0,47 | 2,26 |
| Correction des erreurs | 1,91 | 2,71 |
| Homogénéité | 2,41 | 1,28 |
| Signifiante des codes | 1,54 | 2,89 |
| Compatibilité | 2,39 | 1,61 |

Tableau 1. Interface « texte » : moyenne et écart type des notes pour chaque critère.

Les écarts types montrent des différences de notation entre les sujets pour certains critères et cela est vrai dans les deux groupes (experts et novices). La non-flexibilité de l'interface « texte » par exemple est moins sanctionnée par les sujets experts de l'activité d'administration, ceux-ci étant plus coutumiers de ce type d'interface, sous *Unix* notamment. Ceci peut expliquer l'écart type important du critère *Flexibilité*. On peut penser que cette différence d'expérience a pu également influencer la façon de noter certains autres critères (Expérience utilisateur, Signifiante des codes).

Nous avons procédé à une ANOVA avec comme variable indépendante la « note » attribuée par les sujets à chaque interface pour chaque critère (soit 18 notes par sujet pour une interface). Globalement on constate logiquement (les deux interfaces étant très différentes), que les sujets ne notent pas les deux interfaces de la même manière. L'interaction « type d'interface » et « note » est significative [$F(18 ; 108) = 2,75 ; p < .01$]. Les sujets semblent bien différencier les critères entre

eux, ils leurs attribuent des notes significativement différentes [(F(18 ;108) = 3,5 ; $p < .01$], quelle que soit l'interface évaluée et quelque soit leur niveau d'expertise.

Il n'y a pas d'interaction entre l'expérience des sujets et la façon de noter les interfaces [(F(18;108) = 1 ; $p > .43$, NS] encore que pour l'interface « graphique » on remarque une tendance proche des seuils de significativité [(F(18 ;108) = 1,65 ; $p > .06$, NS]. Pour cette interface, les sujets notent quatre critères de façon très différente selon leur expertise (incitation, distinction et groupement par la localisation, actions explicites et homogénéité). Les deux premiers critères sont notés plus favorablement par les sujets experts et les deux derniers par les sujets novices.

2.4.2. Analyse qualitative

Le tableau ci-dessous, présente, pour les 2 premiers critères évalués, les commentaires que les sujets ont mentionné dans la zone prévue à cet effet. Il s'agit ici de l'interface graphique et des critères « Incitation » et « Groupement/Distinction par la localisation ». Les remarques des sujets sont données quelque soit leur pertinence par rapport à l'interface et sont rapportées sans reformatage.

| | | |
|---|-----------|--|
| 1. Incitation (<i>Prompting</i>) | S1 Expert | Help contextuel. Il faut savoir qu'il faut bouger la souris pour faire apparaître les zones. |
| 1. Incitation (<i>Prompting</i>) | S6 Novice | Je n'ai pas pu éteindre le <i>Service Master</i> . Sens du <i>Apply</i> . (Différence avec <i>OK</i> ?) |
| 1. Incitation (<i>Prompting</i>) | S7 Novice | Le format des numéros de téléphone n'est pas normalisé. Le label du nouveau numéro de téléphone est trop éloigné du champ de saisie, on ne le remarque pas |
| 1. Incitation (<i>Prompting</i>) | S7 Novice | On est obligé d'aller dans l' <i>Operator Panel</i> pour démarrer ou stopper une machine |
| 2. Groupement/Distinction par la localisation | S1 Expert | Information liée au numéro de téléphone (<i>Remote maintenance</i>) non trouvée : passage en rappel automatique. |
| 2. Groupement/Distinction par la localisation | S3 Expert | Localisation spatiale de commandes à regrouper. Un onglet à supprimer dans <i>Remote maintenance</i> . |
| 2. Groupement/Distinction par la localisation | S4 Novice | Problème des n° de téléphone à rappeler automatiquement. Quel est la sélection courante ? |

| | | |
|---|-----------|--|
| 2. Groupement/Distinction par la localisation | S6 Novice | Le bouton Stop devrait être en haut à droite, placé près du <i>Reload</i> . Mettre les icônes dans le même ordre que les items de menu. |
| 2. Groupement/Distinction par la localisation | S5 Expert | Accès par <i>Operator Panel</i> plutôt que <i>Polykid</i> . |
| 2. Groupement/Distinction par la localisation | S7 Novice | Je pensais que les fonctionnalités de l' <i>Operator Panel</i> concernaient les modules du <i>Polykid</i> à cause de la hiérarchie. Je pensais aussi que les autres icônes sous <i>Operator Panel</i> étaient des modules. |

Tableau 2. Interface « graphique » : extrait des commentaires des sujets pour les 2 premiers critères évalués

3. Discussion

Nous ne discuterons pas ici les résultats que les logiciels testés ont générés. Ce qui nous intéresse, ce sont les problèmes et difficultés soulevés par l'utilisation de cette méthode, tant dans son application que pour l'exploitation des résultats. Par ailleurs les évaluations obtenues par chaque logiciel ne sont pas comparables, les logiciels étant très différents (interface et objet de chaque logiciel).

L'hypothèse d'une utilisation possible de la méthode d'évaluation proposée, tant par des sujets experts que par des sujets novices dans l'activité du logiciel testé semble être confirmée. On ne constate en effet pas de grosses différences d'évaluation entre les groupes de sujets. Par contre les différences d'évaluations inter-sujets sont importantes.

La partie la plus fertile du dispositif n'est pas forcément le volet quantitatif, bien qu'il puisse mettre au jour des problèmes importants. Ce qui semble plus intéressant est la collecte d'informations lors de l'observation du sujet et ensuite les commentaires et suggestions des sujets sur les questionnaires ou plus informellement lors des débriefings post-expérimentaux. L'analyse qualitative des commentaires des sujets par critère apporte des informations riches, un grand nombre de problèmes sont signalés par les sujets. Toutefois, pour bien les traiter, il convient de posséder un minimum de connaissances dans le domaine de l'ergonomie afin de savoir les corriger avec les dispositifs appropriés. Par ailleurs, il convient de connaître l'interface, les sujets ayant tendance à utiliser les abréviations ou le jargon de chaque interface, sans formuler toujours des phrases complètes. Parfois les remarques ou critiques sont à pondérer, et en tout cas à apprécier à l'aune des standards existants. Ce ne sont pas forcément des remarques ou suggestions à prendre systématiquement en compte. Un pendant informatique de la méthode aurait un inconvénient : les

sujets dessinent souvent sur le papier une nouvelle icône, une nouvelle arborescence ou une proposition de réorganisation d'un l'écran. Le support papier se prête bien à ce type de tâche. Il faut donc prévoir ces outils à proximité du sujet en complément du matériel expérimental.

Les résultats de l'expérience montrent qu'avec une telle méthode, on peut diagnostiquer les critères non respectés par un logiciel sans toutefois en déterminer forcément la cause précise ni la façon de résoudre le problème (travail de l'ergonome professionnel). L'avantage d'une telle technique de test est son côté reproductible et partiellement automatisé. Quoique limités, les résultats peuvent être fournis instantanément dans une version en ligne du questionnaire et le fait de soulever des problèmes majeurs est déjà précieux pour les équipes de conception (surtout si ce diagnostic peut se faire indépendamment de l'intervention d'un ergonome). Peut-être qu'une courte formation pré-expérimentale pourrait améliorer le consensus entre les sujets. Une expérience complémentaire nous semble souhaitable pour valider ce dispositif dans le cas d'une utilisation en ligne. En outre il conviendrait d'utiliser une interface *ad hoc* comprenant des défauts enfrenant tous les critères de qualité ergonomiques et dont on connaîtrait le nombre. Dans ce cadre expérimental il serait possible d'évaluer l'efficacité de la méthode au prorata des défaut signalés par les sujets sur l'ensemble possible. Une expérience conduite par des sujets concepteurs permettrait également de questionner les problèmes déontologiques, liés à la pratique de techniques expérimentale par des sujets non rompus aux méthodes de tests d'utilisation (aidés seulement d'un court manuel papier d'accompagnement du logiciel d'évaluation).

4. Bibliographie

[BAS 92] BASTIEN J.-M.-C., SCAPIN D.-L., « A Validation of Ergonomic Criteria for the Evaluation of Human-Computer Interfaces », *International Journal of Human-Computer Interaction* vol. 4, n° 2, 1992, p. 183-196.

[BAS 98] BASTIEN J.-M.-C., LEULIER C., SCAPIN D.-L., « L'ergonomie des sites web », *Créer et maintenir un service web*, Cours INRIA, Pau, 28 sept. - 2 oct. 1998, Paris, Editions ADBS, p. 111-173.

[BIS 99] BISSERET A., SEBILLOTTE S., FALZON P., *Techniques pratiques pour l'étude des activités expertes*, Toulouse, Editions Octarès, 1999.

Stéphane Caro est docteur en sciences de l'information et de la communication de l'université Stendhal Grenoble III. Il est actuellement maître de conférences à l'université de Bourgogne. Ses activités de recherche concernent la conception et l'évaluation des hypermédias.