EVALUATION DES INTERFACES UTILISATEUR : TAXONOMIE ET RECOMMANDATIONS

Joëlle Coutaz, Sandrine Balbo

Laboratoire de Génie Informatique-Institut IMAG
BP 53, 38041 Grenoble Cedex 9
Tel. +33 76 51 48 54, +61 75 95 33 31
E-mail : Joelle.Coutaz@imag.fr, Sandrine_Balbo@Bond.edu.au

Résumé : Si l'on reconnaît la nécessité de produire des logiciels utiles et utilisables, l'évaluation ergonomique est le plus souvent jugée trop coûteuse, réductrice et peu fiable. A cela, nous répondons qu'à chaque situation correspond un ensemble de méthodes idoines. Dans cet article, nous proposons un espace taxonomique qui inventorie les dimensions qu'il convient de considérer pour orienter le choix d'une méthode d'évaluation : moyens humains, ressources matérielles, connaissances requises, facteurs situationnels et nature des résultats fournis. Chacun de ces points est analysé en détail.

Mots-clés : Evaluation des interfaces utilisateur, taxonomie des méthodes d'évaluation, choix de méthode d'évaluation.

1. Introduction

La facilité d'utilisation des logiciels est devenue un argument publicitaire. Si l'on reconnaît la nécessité de produire des logiciels utiles et utilisables, l'évaluation ergonomique est le plus souvent jugée trop coûteuse, réductrice et peu fiable. Les concepteurs et les développeurs de logiciels ne sont pas toujours convaincus que le rapport "coût/bénéfice" de l'évaluation ergonomique soit à leur avantage. A cela, nous répondons qu'à chaque situation correspond un ensemble de méthodes idoines. En génie logiciel, l'Assurance Qualité se pratique à façon, en fonction des ressources et des objectifs. A chaque cas, sa méthode. Nous reprenons ce principe à notre compte en l'appliquant à l'évaluation ergonomique.

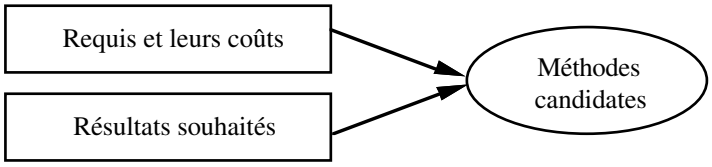

Figure 1 : Choix d'une méthode d'évaluation en fonction des coûts et des résultats recherchés.

La figure 1 traduit notre point de vue. Toute technique d'évaluation implique des requis (par exemple, les ressources matérielles et les moyens humains mis en jeu) et produit des résultats (par exemple des données quantitatives de performance). Connaissant le coût des requis de chaque méthode, la nature des résultats souhaités et les moyens disponibles dans l'environnement, chaque évaluateur doit pouvoir identifier une ou plusieurs méthode(s) qui réponde(nt) à son analyse de coût et de risque.

Il convient maintenant d'identifier les dimensions qui permettent de caractériser puis d'évaluer les requis et les résultats d'une méthode. C'est ce que nous nous proposons de faire dans cet article. Nous offrons ainsi un espace de classification qui, utilisable comme aide au choix d'une méthode d'évaluation, va bien au-delà des taxonomies usuelles. Au paragraphe suivant, nous rappelons brièvement les classifications les plus courantes puis nous développons notre propre cadre de réflexion. On trouvera dans [Balbo 94] une description détaillée de ce travail.

2. Classifications usuelles

La classification la plus fréquente est fondée sur la distinction entre approches prédictives et approches expérimentales [NIELSEN 90, JEFFRIES 91, BASTIEN 91, SALBER 93]. Senach parle respectivement de méthodes analytiques et de méthodes empiriques [SENACH 90]. Les méthodes prédictives ne nécessitent pas la présence d'utilisateur ni l'implémentation, même partielle, du système. De fait, elles conviennent dès les premières étapes du cycle de vie d'un logiciel : analyse des besoins et conception. Ces techniques permettent, à partir d'une description du système, voire d'un modèle de l'utilisateur, d'identifier des problèmes potentiels d'utilisabilité. Selon les modèles, théoriques (PUMS [YOUNG 90], ICS [BARNARD 87]) ou heuristiques (appel à des experts pratiquant par exemple la technique structurée des "cognitive walkthrough" [LEWIS 90]), les descriptions utilisées sont formelles, semi-formelles ou informelles.

Les méthodes expérimentales reposent sur le recueil de données comportementales de sujets mis en situation. Le recueil s'effectue sur support audiovisuel, et/ou par capture automatique d'actions ("monitoring") ou encore par interviews et questionnaires. Les données observées font ensuite l'objet d'une analyse. L'analyse porte généralement sur les chemins suivis et les procédures adoptées par les utilisateurs, les temps d'exécution, la nature, la cause et la fréquence des incidents [VALENTIN 93]. Elle conduit à une synthèse qui indique entre autre le degré de gravité des incidents et si possible, des recommandations.

A l'inverse des méthodes prédictives, les approches expérimentales :

- exigent la présence d'utilisateurs,
- fournissent des données observées du monde réel,
- se pratiquent généralement une fois franchie la phase de conception.

Pourtant, comme les méthodes prédictives, les approches expérimentales peuvent être appliquées dès les premières étapes du cycle de vie sous forme par exemple de scénarimage (en anglais : "story-board") ou produit par un générateur d'interfaces de type "présentation" [NIGAY 94]. On n'insistera jamais assez que dès l'analyse des besoins, une maquette même rudimentaire peut utilement servir à comprendre le problème, à impliquer et convaincre le client, à construire les modèles de tâche et de l'utilisateur. Selon les outils de développement disponibles, la nature du projet et la compétence des architectes du logiciel, la maquette sera jetée ou bien se transformera progressivement en prototype puis en produit.

De manière générale, les méthodes prédictives et expérimentales se complètent et peuvent se pratiquer tout au long du processus de développement. La pratique itérative "tests-corrections" définit les fondements de l'évaluation formative [HIX 93] : chaque évaluation fournit de nouveaux enseignements dont la prise en compte conduit à une nouvelle version. La conception itérative convient naturellement au processus de développement en spirale mais elle s'inscrit également dans un cycle de vie en V avec retours arrière. Poussée à l'extrême, la conception itérative conduit à un processus de développement en étoile centrée sur l'évaluation [Hix 93]. A l'évaluation formative, il faut opposer l'évaluation globale (en anglais : "summative") pratiquée une seule fois en fin de mise au point ou pour comparer des logiciels existants [HIX 93].

Whitefield et ses co-auteurs organisent leur classification en fonction de la présence de l'utilisateur et du système [WHITEFIELD 91]. Ici, le critère directeur est la ressource matérielle (le système informatique) et les moyens humains (les sujets). La présence de l'utilisateur (ou du système) peut être virtuelle ou réelle. Une "présence virtuelle" de l'utilisateur (ou du système) signifie que la méthode ne nécessite pas l'intervention d'un utilisateur réel (ou l'utilisation d'une version implémentée du système) mais suppose une représentation (formelle ou non) de l'utilisateur représentatif (ou du système).

Le découpage entre méthodes prédictives et expérimentales met en lumière les deux grands courants de recherches et de pratiques dans le domaine de l'évaluation ergonomique. Si cette distinction est utile comme élément de structuration du domaine, elle n'offre aucun support au problème du choix d'une méthode face aux contraintes contextuelles du processus de développement. La classification de Whitefield et de ses co-auteurs va dans le sens de nos préoccupations mais elle mérite d'être approfondie. Nous présentons au paragraphe suivant une taxonomie plus précise motivée par le problème du choix de méthodes d'évaluation.

3. Une taxonomie pour le choix de méthodes d'évaluation

La figure 2 présente de manière synthétique les dimensions de notre espace de classification. Ces dimensions représentent cinq catégories de préoccupations : les moyens humains, les ressources matérielles, les connaissances requises, les facteurs situationnels, et la nature des résultats fournis. Nous analysons chacun de ces points dans les paragraphes qui suivent.

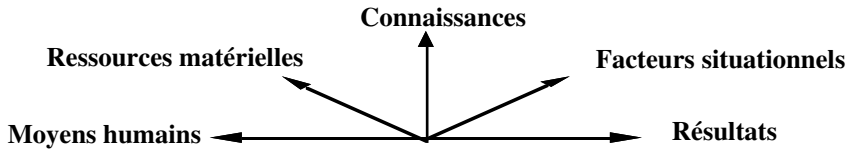

Figure 2 : Taxonomie centrée sur le problème du choix de méthodes d'évaluation ergonomique. Se reporter aux figures 3 à 7 pour un affinement des cinq axes.

3.1. Moyens humains

Les moyens humains désignent l'ensemble des acteurs impliqués directement ou indirectement dans l'activité d'évaluation. Nous distinguons trois classes d'intervenants : les sujets, les évaluateurs spécialistes et une catégorie qui regroupe tous les autres acteurs (les dévelopeurs, les concepteurs d'interface, les clients, les agents commerciaux, etc.).

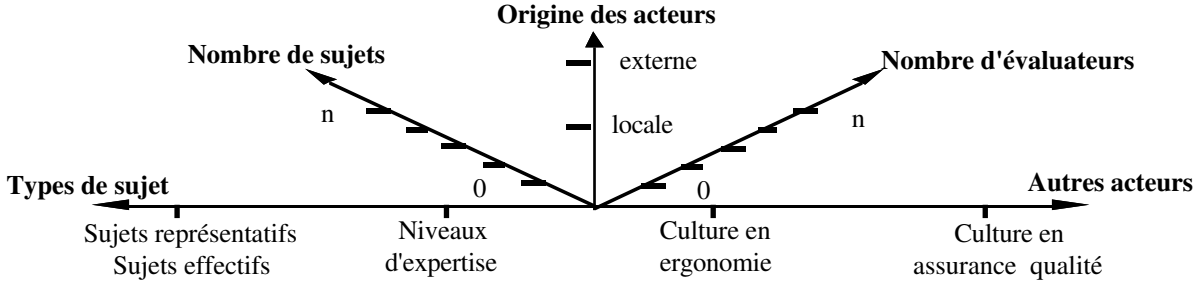

Figure 3 : Le sous-espace des ressources humaines.

Comme le montre la figure 3, les acteurs peuvent appartenir à l'organisation chargée du développement (il s'agit alors d'acteurs *locaux*), ou bien sont recrutés à *l'extérieur*. Cette distinction a de l'importance dans l'évaluation des coûts du recrutement. Les facteurs situationnels (voir paragraphe 3.4) et notamment le type de projet, l'enveloppe budgétaire et le planning conditionneront l'origine et le nombre de sujets et de spécialistes qu'il sera possible de retenir.

En l'absence de sujet, il faudra avoir recours aux approches prédictives. Si l'on choisit une méthode heuristique, il faudra prévoir entre *trois et cinq spécialistes* pour une couverture raisonnable des problèmes potentiels d'utilisabilité [POLLIER, NIELSEN 90]. Sans sujet ni spécialiste, seuls les modèles théoriques pourront être appliqués. Il conviendra alors de s'interroger sur l'accessibilité de ces modèles aux évaluateurs qui

n'ont pas nécessairement une formation en ergonomie ou en psychologie(voir paragraphe 3.3).

Si l'on adopte une méthode expérimentale, le réalisme des données recueillies dépendra de la représentativité des sujets et des scénarios. Bien que la règle ne soit pas absolue, Valentin et ses co-auteurs recommandent de commencer par définir la population [VALENTIN 93]. Cette démarche est bien adaptée lorsque l'on dispose de l'*utilisateur effectif (réel)* du futur logiciel. S'il faut avoir recours à des *utilisateurs représentatifs*, la définition préalable des scénarios et d'un modèle de l'utilisateur peuvent utilement diriger le recrutement des sujets. Typiquement, les sujets sont organisés par groupes en fonction de leur niveau d'expérience dans le domaine, voire dans l'utilisation du produit, et leur familiarité avec l'informatique.

Pour la dernière catégorie d'acteurs (c'est-à-dire les intervenants autres que les sujets et les spécialistes), il est important de s'interroger sur leur niveau de culture en matière d'assurance qualité (au sens large du génie logiciel) mais aussi sur leur attitude vis-à-vis des principes de l'ergonomie. A fortes réticences doit correspondre, nous semble-t-il, une technique d'évaluation légère de type heuristique.

3.2. Ressources matérielles

Les ressources matérielles recouvrent tous les moyens physiques impliqués dans l'activité d'évaluation. La figure 4 montre le sous-espace que nous proposons comme élément de structuration. Nous convenons de faire la distinction entre les ressources servant d'instruments de capture et les ressources matérielles faisant l'objet d'une évaluation.

Ressources matérielles pour la capture. Ces ressource incluent les questionnaires et interviews, les régies vidéo (éventuellement portatives) ou laboratoires spécialisés, les ordinateurs. Toutes ces ressources peuvent être caractérisées par un coût mais aussi par la *portée*, la *nature* et le *niveau d'abstraction* des données qu'elles permettent de recueillir. Pour illustrer l'intérêt de ces dimensions, considérons la capture automatique par ordinateur.

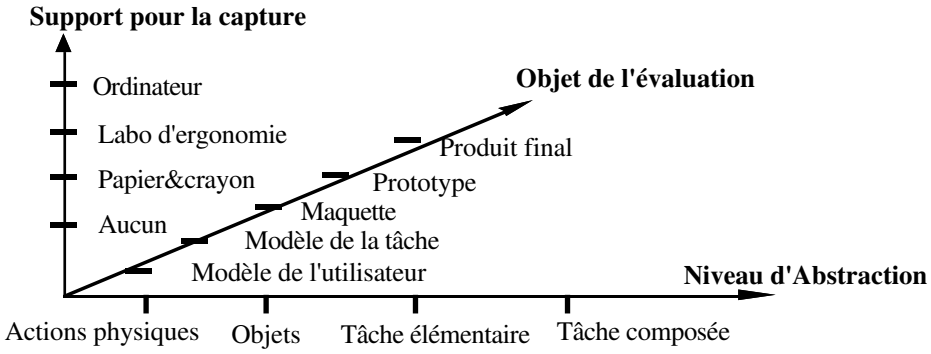

Figure 4 : Le sous-espace des ressources matérielles.

Actuellement la capture par ordinateur permet d'enregistrer des actions physiques de type clavier-souris et au mieux la manipulation d'objets simples dans le cadre de tâches élémentaires [MACLEOD 93, HAMMONTREE 92]. Lorsque la manipulation d'objet est prise en compte, seuls sont repérables les objets prédéfinis des boîtes à outils (par exemple, les menus et les boutons de Motif). En conséquence, la capture automatique ne couvre que les phénomènes de bas niveau d'abstraction et de plus, pour un niveau d'abstraction donné, présentent des ruptures de portée : les actions physiques ne sont pas toutes repérées (et notamment les métacomportements verbaux et gestuels) et les objets manipulables ne sont pas tous observés (et notamment les objets construits à façon au-

dessus des bibliothèques graphiques). Si l'objectif est une évaluation fiable, ces limitations exigent de pratiquer des tests complémentaires. Le projet NEIMO vise à combler ces lacunes [SALBER 93].

Ressources faisant l'objet de l'évaluation. Comme le montre la figure 4, tous les produits du processus de développement peuvent faire l'objet d'une évaluation (se reporter au modèle en étoile de Hix&Hartson [Hix 93]. Ce peut être : le modèle de tâche, le modèle de l'utilisateur, les spécifications externes du système, une maquette, un prototype, un produit, etc. L'objet de l'évaluation peut être décrit de manière formelle, semi-formelle ou informelle ; il peut admettre plusieurs formes de supports (papier, maquette carton, ordinateur). A chacune de ces formes de présentation correspondent un coût de développement et des compétences.

3.3. Connaissances requises

Comme le montre la figure 5, les connaissances requises pour mettre en application une méthode sont de deux sortes: les premières ont trait au *coût cognitif d'accès* à la méthode. Par exemple, toutes les approches heuristiques nécessitent une formation en ergonomie cognitive. Si l'évaluateur dont on dispose n'a pas cette compétence, il est sans doute préférable pour la fiabilité des résultats de s'orienter vers un modèle prédictif théorique par exemple le modèle Keystroke [CARD 83] (sous réserve que ce niveau d'abstraction soit suffisant pour l'étude) ; ou encore appliquer les recommandations de Nielsen qui permettent de pratiquer une "ingénierie de l'utilisabilité à faible coût" [NIELSEN 89].

La seconde classe de connaissances requises concerne les *descriptions nécessaires en entrée* au bon fonctionnement de la méthode : modèle de l'utilisateur, modèle de la tâche, spécifications externes de l'interface, définition de scénarios. Plus le spectre des informations d'entrée est large, plus la méthode est coûteuse à mettre en œuvre. En revanche, on peut espérer obtenir en sortie des résultats plus robustes et concis.

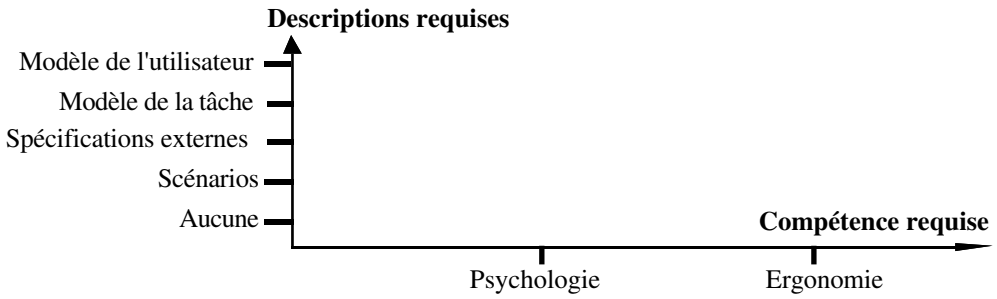

Figure 5 : Le sous-espace des connaissances requises.

3.4. Facteurs situationnels

Les facteurs situationnels définissent le contexte de l'évaluation. Comme le montre la figure 6, nous proposons de distinguer : l'étape du processus de développement, le lieu dans lequel l'évaluation peut être pratiquée, la typologie des applications et des techniques d'interaction, l'enveloppe budgétaire et les contraintes de planning.

Etape du processus de développement : il convient de noter que l'étape du cycle de vie est orthogonale à l'objet de l'évaluation. En particulier, une maquette (objet de l'évaluation) peut servir à l'analyse des besoins tout comme elle peut servir à valider des spécifications externes. Toutefois, le niveau de réalisme de la maquette variera en fonction de l'étape et de la nature du projet, du type donc de cycle de développement : durée de vie prévue du logiciel (version unique d'un logiciel expérimental publicitaire pour valider des concepts ou versions multiples de logiciel pérenne à forte pénétration industrielle, etc.).

Lieu de l'évaluation : l'évaluation peut être pratiquée dans l'entreprise cliente ou bien en laboratoire. Les expériences de laboratoire sont plus faciles à contrôler mais on devra s'interroger sur la validité "écologique" des données recueillies : le comportement en laboratoire n'est pas nécessairement celui observé in situ.

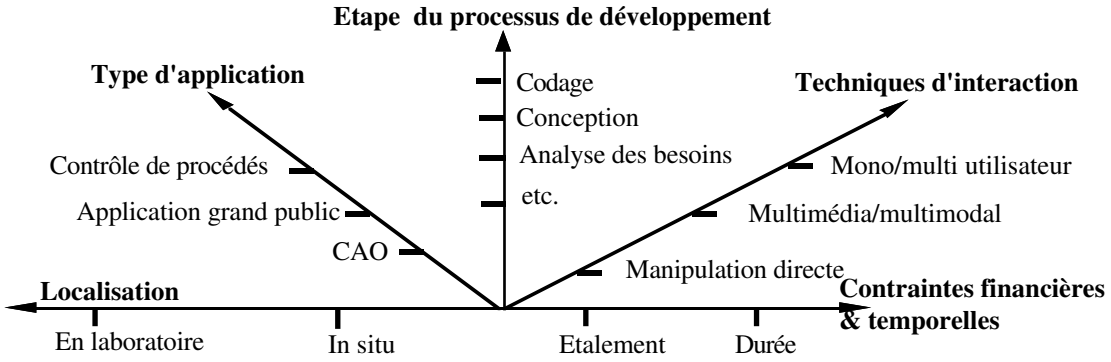

Figure 6 : Le sous-espace des facteurs situationnels.

Typologie des applications et des tâches : on convient généralement de distinguer les applications de type CAO caractérisées par des tâches à structure variable, les applications de type contrôle de procédés où l'on relève au contraire des tâches opératives de routine, les applications grand public qui militent en faveur d'un dialogue guidé par le système, etc. A l'évidence, les exigences sur la nature du recueil varient en fonction du domaine applicatif. En particulier, le réalisme des données recueillies est un critère prioritaire pour les applications où la sécurité est un facteur dominant. Dans ce cas, l'évaluation expérimentale est incontournable.

Typologie des interfaces : interfaces graphiques à manipulation directe, interfaces multimédia et multimodales (on trouvera dans [NIGAY 94] une définition précise de ces termes), interfaces multi-utilisateur. Aucun modèle théorique ne peut prétendre aujourd'hui assurer l'évaluation des interfaces multimédia et multimodales ni permettre l'évaluation des collecticiels. Pour ces nouvelles interfaces, les heuristiques émergent à peine [SUTCLIFFE 94]. A l'heure actuelle, les méthodes d'observation constituent, à notre sens, les seules approches raisonnables.

Enveloppe budgétaire et planning : l'enveloppe budgétaire et le planning conditionnent le choix d'une technique d'évaluation. Le planning définit deux formes de contraintes temporelles : la durée que l'on peut dédier à l'évaluation (par exemple, trois jours au total) mais aussi son étalement dans le temps (trois jours répartis, par exemple, au maximum sur deux semaines).

3.5. Résultats fournis

La figure 7 synthétise les dimensions qui permettent de caractériser les résultats d'une méthode d'évaluation. Certaines dimensions ont trait au contenu, d'autres à la qualité, d'autres encore au support de restitution.

Le contenu s'étudie selon deux volets orthogonaux : le niveau d'abstraction des informations (niveaux tâche, syntaxique, lexical et physique) et le type d'information. A son tour, une information d'un type donné se distingue par :
a) sa valeur sémantique : nombre d'erreurs et degré de gravité, nombre de recommandations transgressées, temps d'exécution et d'apprentissage, degré de satisfaction de l'utilisateur, traces, etc.
b) la tranche d'exécution du scénario à laquelle elle correspond : avant l'exécution du scénario, pendant et après ;

c) sa nature : quantitative (il s'agit d'un résultat numérique comme un temps d'apprentissage) ou qualitative (résultat non numérique comme une liste de problèmes d'utilisabilité) ; subjective (qui représente une opinion) ou objective (c.-à-d. une mesure observée) ; explicative si elle permet de comprendre la raison du problème et corrective si elle fournit le remède au problème d'utilisabilité.

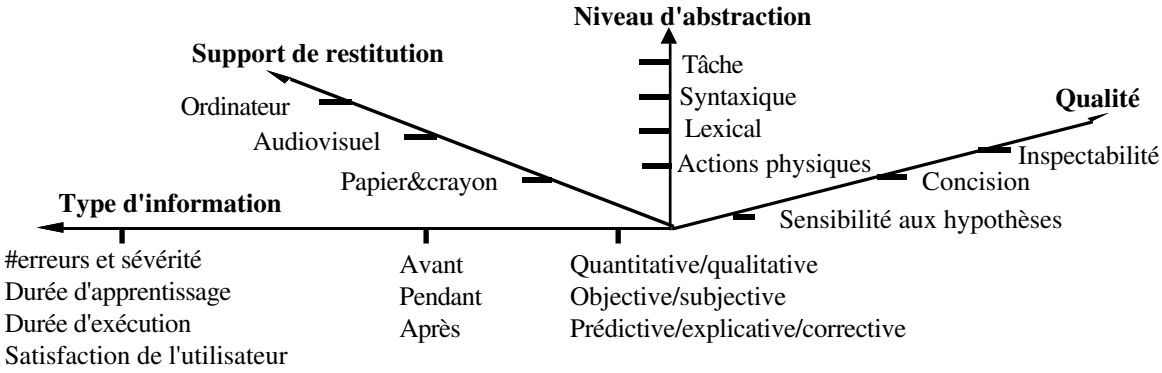

Figure 7 : Le sous-espace des résultats fournis par une méthode.

Sous l'en-tête "qualité" nous regroupons des critères généraux qui dénotent la robustesse des résultats de la méthode d'évaluation : sensibilité de l'information aux hypothèses (en particulier GOMS est sensible à l'hypothèse du comportement expert), concision de l'information qui joue sur les temps d'analyse (comme contre-exemple, citons le manque de concision des enregistrements audiovisuels), inspectabilité de l'information qui permet au spécialiste de parcourir les résultats et de comprendre les phénomènes.

Le support de restitution est aussi un ingrédient important dans la présentation des résultats. Nous distinguerons entre la présence et l'absence de support automatisé. Aujourd'hui, la plupart des outils informatisés assurent l'inspectabilité du recueil des données comportementales mais vont rarement au-delà de la détection automatique des problèmes. De plus, les problèmes détectés relèvent des niveaux d'abstraction les plus bas et s'appuient uniquement sur des informations quantitatives objectives.

4. Conclusion

Nous venons de présenter les dimensions qu'il convient de considérer dans le choix d'une méthode d'évaluation ergonomique. Nous ne prétendons pas à l'exhaustivité mais nous estimons avoir fourni une bonne couverture de l'espace actuel et à venir des méthodes d'évaluation.

5. Remerciements

Nous tenons à remercier nos collègues de l'atelier "Model for Developing High Impact Formative Evaluation Methods", CHI'94, qui ont contribué à la mise en place des idées ici présentées.

6. Références

[BALBO 94]
S. Balbo, "Un pas vers l'évaluation automatique des interfaces utilisateurs", Thèse de l'Université Joseph Fourier, Grenoble, Septembre 1994.

[BARNARD 87]
P.J. Barnard, "Cognitive Resources and the Learning of Human-Computer Dialogs", in Interfacing Thought, edited by J.M. Carroll, The MIT Press, pp. 112-158, 1987

[BASTIEN 91]
C. Bastien, "Validation de critères ergonomiques pour l'évaluation d'interfaces utilisateurs", INRIA, Unité de recherche INRIA_Rocquencourt, rapport de recherche n° 1427, Programme 3, Mai 1991

[BASTIEN 93]
J.M. Bastien & D.L. Scapin, “Preliminary findings on the effectiveness of ergonomic criteria for the evaluation of human-computer interfaces”, in Proceedings of the InterCHI'93 conference, Amsterdam, April 1993

[DOWELL 89]
J. Dowell & J. Long, “Toward a conception for an engineering discipline of human factors”, Ergonomics, vol. 32, No 11, pp. 1513-1535, November 1989

[HAMMONTREE 92]
M.L. Hammontree, J.J. Hendrickson & B.W. Hensley, “Integrated data capture and analysis tools for research and testing on graphical user interfaces”, in Proceedings of the CHI'92 Conference, pp. 431-432, ACM Press, Monterey, 3-7 May 1992

[HIX 93]
D. Hix & R. Hartson, “Developing User Interfaces, Ensuring Usability through Product & Process”, Wiley, 381 pages, 1993

[JEFFRIES 91]
R. Jeffries, J.R. Miller, C. Wharton & K.M. Uyeda, “User interface evaluation in the real word : a comparison of four techniques”, Proceedings of the CHI’91 Conference on Computer Human Interaction, New Orleans, ACM New York, pp. 119-124, 1991 May

[LEWIS 90]
C. Lewis, P. Polson, C. Wharton & J. Rieman, "Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces", in Proceedings CHI'90, ACM, 1990, pp. 235-242.

[MACLEOD 93]
M. Macleod, “DRUM, Diagnostic Recorder for Usability Measurement”, NPL, DITC HCI group, Teddington, Middlesex, TW11 OLW, UK, 1993

[NIELSEN 89]
J. Nielsen, “Usability Engineering at a Discount”, in Designing and Using Human Computer Interfaces and Knowledge-Based Systems, Salvendy & Smith Eds., Elsevier North Holland, pp. 394-401, 1989

[NIELSEN 90]
J. Nielsen & R. Molich, “Heuristic evaluation of user interfaces”, Proceedings of the CHI’90 Conference on Computer Human Interaction, Seatle, ACM New York, pp. 349-256, 1990

[NIGAY 94]
L. Nigay , “Conception et modélisation logicielles des systèmes interactifs : application aux interfaces multimodales”, thèse de l’Université Joseph Fourier, Grenoble, Janvier 1994

[POLLIER 91]
A. Pollier, “Evaluation d’une interface par des ergonomes : diagnostics et stratégies”, Rapport de recherche INRIA no 1391, Février 1991

[SALBER 93]
D. Salber & J. Coutaz, “Applying the Wizard of Oz Technique to the Study of Multimodal Systems”, East-West HCI conference, Moscow, August 1993

[SCAPIN 90]
D.L. Scapin, “Des critères ergonomiques pour l'évaluation et la conception d'interfaces utilisateurs”, Acte du XXVI Congrès de la SELF, 3-5 Octobre, Montréal, Canada, 1990

[SENACH 90]
B. Senach, “Evaluation ergonomique des interfaces homme-machine : une revue de la littérature”, Rapport de recherche INRIA n° 1180, Programme 8, Communication Homme-Machine, Mars 1990

[SUTCLIFFE 94]
A. Sutcliffe & P. Faraday, “Designing Presentation in Multimedia Interfaces”, CHI’94 proceedings, Boston, pp. 92-98, 1994

[VALENTIN 93]
A. Valentin, G. Vallery & R. Lucongsang, “L’évaluation ergonomique des logiciels, une démarche itérative de conception”, ANACT, Colection Outils et Méthode, 151 pages, 1993

[WHITEFIELD 91]
A. Whitefield, F. Wilson & J. Dowell, “A framework for human factors evaluation”, Behaviour and information technology, vol. 10, no. 1, pp. 65-79, 1991

[YOUNG 90]
R.M. Young & J. Whittington, “A knowledge Analysis of interactivity”, Proceedings of INTERACT'90, edited by D. Diaper, G. Cockton & B. Shackel, Elsevier Scientific Publishers B.V, pp. 207-212, Cambridge (United Kingdom), 27-31st August 1990