Evaluation quantitative des interfaces

Modèles prédictifs de performance

Etudes utilisateurs

1

Introduction

- Objectifs
  - Evaluer l 'utilisabilité des interfaces
  - Evaluation **quantitative** ou qualitative des performances des utilisateurs
- Diverses situations
  - dans un laboratoire dédié
  - en situation réelle
  - avec ou sans la collaboration des utilisateurs
- Types différents d 'évaluation
  - Evaluation quantitative des performances des utilisateurs
    - repose sur des expérimentations contrôlées
    - test d 'hypothèse
  - Evaluation prédictive des performances

Mountaz Hascoët, Univ. Montpellier II, LIRMM

2

I. Evaluation empirique des performances

Evaluation qui nécessite la présence d'utilisateurs

- Méthodes quantitatives
  - repose sur des expérimentations contrôlées
  - recueil de données quantitatives
  - analyse statistique des résultats

- Méthodes qualitatives
  - Questionnaires
  - Evaluation de la conformité aux recommandations
    - Heuristiques (Checklist)
      - 1. Le comportement du système est prédictible et consistent
      - 2. Feed-back satisfaisant
      - 3. La mémoire de l'utilisateur n'est pas surchargée
      - 4. Le dialogue est orienté vers la tâche
      - ...
    - Grille d 'évaluation
  - Incidents critiques
  - Walkthrough

Tests utilisateurs

- Objectifs
  - Explorer une nouvelle technique d 'interaction ou de présentation
    - ex: interaction bi-manuelle, pie-menus, toolglass
    - ex: algorithme de placement, 3D/2D, etc.
  - Choisir entre des alternatives de conception
    - ex: menu fugitifs / menus déroulants
  - Evaluer (comparer) un (des) système(s)
    - mesure des performances pour une tâche donnée
- Moyens
  - Conception d 'expériences contrôlées
    - Choix des variables et des résultats à prouver
    - Plan d'expérience
    - Réalisation de l'expérience et recueil des données
    - Analyse des résultats

Expériences contrôlées

Expérience contrôlée

=

choix de variables

+

plan d ’expérience

+

déroulement du test et recueil des données

+

analyse des résultats

Choix des variables

- Variables indépendantes
  - Ce sont les **variables manipulées** pendant l'expérience
  - à une variable sont associées des modalités
    - modalité = valeur possible pour la variable
  - par exemple :
    - variable indépendante : type de procédé de pointage
    - modalités : souris/joystick/clavier
- Variables dépendantes
  - Ce sont les **variables mesurées** pendant l'expérience
  - Par exemple:
    - temps nécessaire pour sélectionner une cible
- Résultat à prouver:

  variables indépendantes => des variables dépendantes

Variables dépendantes

- Exemples
  - Temps nécessaire pour accomplir une tâche
  - Nombre de tâches accomplies en un temps donné
  - Nombre d'erreurs
  - Nombre d'erreurs impossible à corriger
  - Temps passé à corriger des erreurs
  - Nombre de commandes utilisées (ou pourcentage)
  - Nombre de caractéristiques dont l'utilisateur peut se souvenir après l'expérience
  - Nombre de fois où le système d'aide est utilisé
  - Temps passé à consulter l'aide
  - Pourcentage de commentaires positifs/négatifs

Plan d'expérience

Le plan d ’expérience est composé de

- Choix des sujets et répartition des sujets sur les modalités
  - conception mono-sujet ou pluri-sujets
  - répartition équilibrée
- Banc d'essai:
  - ensemble de tâches à accomplir que l'on teste sur plusieurs interfaces pour les comparer
- Scénarios
  - description détaillée des bancs d ’essais, situation opérationnelle, de la durée du test.
- Test témoin
  - permet de déceler les problèmes du test
    - scénarios mal définis, effets des extrêmes (tests trop facile trop difficile => masquent les différences) , test trop long, etc.

Choix des sujets : conception inter-sujets

- Choisir une population cible
- Diviser la population en groupe de même taille
  - autant de groupes que de modalités de la variable indépendante testée
- Répartition des sujets sur les modalités de la(les) variable(s) indépendante(s)
  - chaque groupe est soumis à une modalité
  - Ex: Groupe1 -> souris, Groupe 2 ->clavier, Groupe 3 -> joystick
- Conséquences:
  - bruit lié aux différences entre les sujets
    - minimiser les différences entre sujets
    - compenser par nombre de sujets
  - la seule différence systématique entre les sujets est la modalité à laquelle ils sont soumis

Choix des sujets : conception intra-sujets

- Choisir une population cible
- Chaque sujet est soumis à toutes les modalités de la variable indépendante
- Contrôler
  - effets de bords liés à l 'interaction entre les tests des différentes modalités
    - apprentissage
    - biais
- Avantages
  - moins de bruit => nécéssite moins de sujets
- Inconvénient
  - contrôle des effets de bord
    =>complique la répartition équilibrée

Répartition équilibrée

- Cas de la conception pluri-sujets
  - groupe 1 -> modalité 1
  - groupe 2 -> modalité 2
  - ...
- Cas de la conception mono-sujet
  - Carré Latin
    - Nb de sujets = nb total de modalités

	Mod. 1	Mod. 2	Mod. 3	...
Sujet 1	1	2	3	
Sujet 2	2	3	1	
Sujet 3	3	1	2	
...				

- Conception répliquée
  - Réplication du carré latin

  => Nb de sujets = k * nb total de modalités

Réalisation du test et recueil des données

- Contrôle de l'environnement opérationnel
  - attention aux facteurs "nuisibles"
    - charge des machines, lumière, bruit, etc.
- Recueil de données pour les variables dépendantes
  - - évaluation subjective par l'utilisateur
  - - temps d'exécution des tâches du scénario et performance globale
  - - taux d'erreur/temps passé à corriger des erreurs
  - - mode opératoire et enchaînement des actions
  - - utilisation effective des commandes et des informations présentes à l'écran
    - occulométrie
  - - paramètres physiologiques comme l'état nerveux

Les collecteurs de données (ou mouchards)

- 70 % des professionnels de tests ergonomiques utilisent des mouchards.
  - 80 % d'entre eux ont créé leurs propres mouchards
  - 20 % l'ont acheté.
- Fonctionnalités minimales du parfait Mouchard:
  - temps d'éxécution de tâches
  - nombre d'erreurs, temps de correction, etc.
  - commentaires/annotations
  - questionnaires
- Fonctionnalités plus évoluées
  - résumé des informations rassemblées
  - analyses statistiques
  - contrôleur d'enregistrement video
- Qualités requises
  - Ne pas dégrader les performances du système espionné
  - Etre "invisible" de l'utilisateur

Mouchards (suite)

- Chez Sun: un ensemble d'outils qui
  - - captent les évènements utilisateurs -> fichier de log.
  - - filtrent, traduisent et regroupent les évènements reçus pour en extraire des évènements caractéristiques de l'interaction entre utilisateur et un système donné.
  - - contrôleur video, synchronisation avec les évènements, contrôle des vues, etc.
- Chez Microsoft
  - Observer
    - annotation de video en temps réel
  - Tracker
    - intercepte les événements tels que:
      - sélection dans un menu, frappe sur une touche, mouvement de fenêtre, etc.
  - Reviewer
    - facilite l'analyse des fichiers de log et des videos issues de Tracker et Observer

Mouchards (suite)

- Chez Xerox: EVA (Experimental Video Annotator)
  - lien direct avec le magnétoscope
  - interface à boutons personalisables pour permettre annotation en temps réel de la vidéo
- Chez Apple
  - Event logger
    - fichier de log avec évènements utilisateurs estampillés.
  - Observation logger
    - Fichier de log avec annotations manuelles de l'observateur au cours de l'expérience.
    - Le Fichier est directement associé à la video
      - facilite son analyse

Analyse des résultats

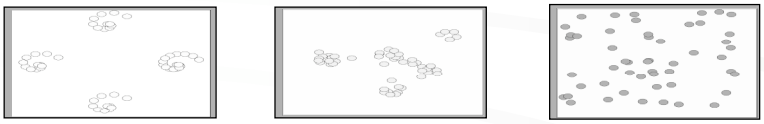
- Test statistique
  - Vérifier que les différences entre les résultats ne sont pas le seul effet du hasard
- Choix du test statistique
  - Test d'hypothèses
    - Tester l'hypothèse H0 (Pas de différence entre les différents traitements) contre H1 (différence significative)
    - Choisir le seuil de 1ère espèce: P(rejeter Ho | Ho est vraie)
      - entre 1% et 5%
- Exemples de tests usuels
  - Test de Student ou T-Test
    - Si la variable indépendante a deux modalités
  - Analyse de Variance et Test de Fisher
    - Si la variable indépendante a plus de deux modalités
  - Ces tests pré-supposent que la distribution des données est gaussienne (suit une loi normale)

Bémols

- Certains aspects plus difficiles à couvrir
  - Facilité d 'apprentissage
  - Stratégies, planification, décomposition des tâches.
- Coûteux
  - Ex: un expérimentateur met une semaine pour évaluer un nouvel éditeur par tests utilisateurs!
- Nombreuses sources d 'erreur
  - Ex: les tests utilisateurs
    - Sujets/ Expérimentateur
      - biais lié à l 'attente d 'un résultat particulier ou de connaissances préalables
    - Environnement opérationnel
    - Analyse statistique
      - domaine très vaste
        - » test très simple + légère modification => test très compliqué
        - » => erreur de test ou erreur d 'interprétation des résultats

Exemple d 'étude utilisateur (suite)

- Objectif
  - Comparer trois techniques différentes pour afficher des informations



  - Tâche considérée
    - simplification du problème initial
    - retrouver le nombre le plus proche d 'un nombre donné
- Choix des variables et des résultats à prouver
  - Variable indépendante: type de placement
    - modalités: algo1, algo2, algo3
  - Variables dépendantes
    - temps nécessaire pour réaliser la tâche
    - nb clics nécessaires pour réaliser la tâche

Exemple d ’étude utilisateur (suite)

- Plan d'expérience
  - Conception mono-utilisateur équilibrée

	Algo1	Algo2	Algo3	...
Sujet 1	1	2	3	
Sujet 2	2	3	1	
Sujet 3	3	1	2	
...				

  - Banc d'essai:
    - trouver les 5 nombres les plus proches de 21 dans un ensemble de 50 nombres

Exemple d ’étude utilisateur (suite)

- Scénarios
  - `Afficher page d ’accueil - explication du test`
  - `Ecran transition (1)`
  - `Afficher écran d ’essai`
    - `Trouver les 5 nombres avec un algo, tests pour rien`
  - `Ecran transition (2)`
  - `Afficher premier test`
    - `Trouver les 5 nombres avec un algo`
    - `(répéter 3 fois avec le même)`
    - `puis idem avec un autre algo`
  - `Durée prévue 15 mn par utilisateur`
- Test témoin
  - vérifier sur quelques cobayes la compréhensibilité du test, sa durée, les résultats

Exemple d ’étude utilisateur (suite)

- Réalisation de l'expérience
  - Application pour
    - dérouler le scénario (enchaînement automatique des écrans et gérer dialogue utilisateur) et
    - contrôler les compteurs (enregistrement et écriture)
- Analyse des résultats
  - Analyse de Variance et Test de Fisher
    - algo1 significativement meilleur que algo2 pour les 2 variables dépendantes (idem avec algo3)
    - algo2 et algo3 pas de différence significative
      - => on ne peut rien dire
      - nécessite de poursuivre le test avec population plus importante

Evaluation quantitative*

- Conception de questionnaires(types de questions)
  - Réponses guidées

    Le 12 novembre est une date idéale pour placer un examen?

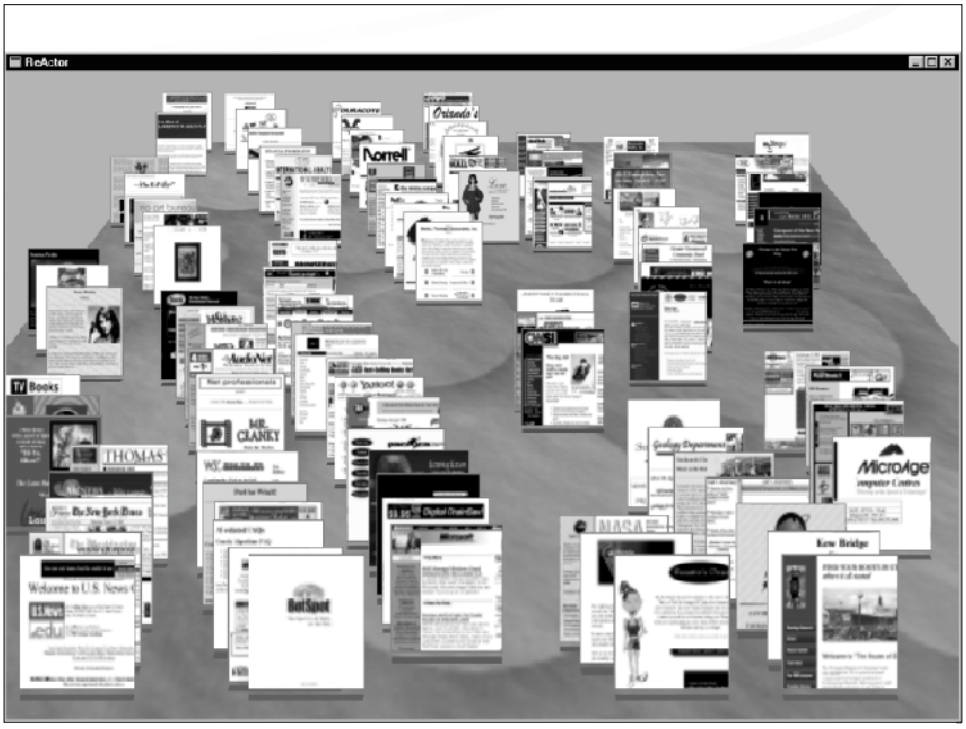
    ❑ Tout à fait d'accord ❑ à peu près d'accord ❑ pas d'accord ❑ pas du tout d'accord
    - *NB: Forcer le choix nb pair de réponses (4-6)*
  - Choix multiple

    Quel est votre éditeur de texte préféré?

    ❑ Emacs ❑ Word ❑ FrameMaker ❑ Vi
  - Ordonnancement

    Placer ces éléments par ordre d'importance (1-3: important,très important..., 0 inutile)

    ❑ Répertoire ❑ Gestion de plusieurs communications ❑ Messagerie
  - Différentiations sémantiques
    - Entourer le nombre qui représente le mieux votre avis sur les outils proposés
    - `Simple  3 2 1 0 1 2 3 Complexe`
    - `Intuitif  3 2 1 0 1 2 3 Non-Intuitif`
- **: cf aussi cours et td précédents pour les autres méthodes*



DataMountain

- Objectifs:
  - Améliorer la gestion de collections de pages www
- Principe
  - S'appuyer sur la mémoire visuelle et spatiale
- Conception
  - Visualisation
    - Plan incliné
  - Interactions
    - Stratégies particulières de déplacement pour éviter les recouvrements
    - Clic => Centrage de la page sélectionnée
    - Info-bulles

Expérimentation contrôlée (Variables)

- Variables
  - indépendantes
    - Type d'interface
      - IE
      - DataMountain1
      - DataMountain2
    - Type de questions
      - Imagette
      - Titre
      - Résumé
      - Titre+résumé+imagette
  - Dépendantes
    - T: temps mis pour effectuer la tâche
    - E: nb de pages visitées avant de trouver la page demandées
    - N: nb échecs avant 1 minute
    - A: appréciation subjective de l'utilisateur

Expérimentation contrôlée (Choix des sujets et scénarii)

- Scénario de base:
  - Classer 100 pages web
  - Retrouver des pages décrites par l'un des 4 types de description de pages envisagés.
- Conception inter-sujets
  - 1 groupe avec IE4
  - 1 groupe avec DM1
  - 1 groupe avec DM2
    - En tout 30 sujets, utilisateurs expérimentés de IE4
  - L'ordre de présentation des types de questions est aléatoire

Expérimentation contrôlée (Résultats)

- T : temps mis pour effectuer la tâche
  - Globalement
    - DM2 > DM1 > IE4 mais
  - Avec description Titre
    - IE4 > DM1
- E: sélection d'une mauvaise page
  - Globalement
    - DM2 > DM1
    - DM2 > IE4
- A: appréciation subjective de l'utilisateur

  … questionnaire « un peu » partial …

Data Mountain: Using Spatial Memory for Document Management

Questionnaire Item	IE4	First Data Mountain	Second Data Mountain
I like the software	3,4 (1,0)	3,3 (1,2)	3,7 (0,7)
The software is efficient	3,6 (1,1)	2,9 (1,2)	3,3 (0,8)
The software is easy to use	3,6 (1,1)	4,0 (0,9)	4,0 (1,1)
The software feels familiar	4,0 (0,7)	3,3 (1,2)	3,4 (1,2)
It is easy to find the page I am looking for with the software	3,4 (1,0)	3,3 (1,0)	3,4 (1,0)
Organizing web pages is easy with the software	3,4 (1,1)	3,7 (1,1)	3,8 (0,8)
If I came back a month now I would be still be able to find many of these web pages	3,2 (0,8)	4,1 (0,6)	3,6 (1,3)
I was satisfied with my organization scheme	3,2 (1,2)	3,4 (0,8)	3,1 (1,2)
My organizing scheme was very similar to the organization in my home Favorites folder	3,9 (1,5)	3,6 (1,4)	3,4 (1,6)

User satisfaction averages for on a five point scale where 1=disagree, 5=agree
(standard deviations in parentheses)

Conclusion

- Des réponses quantitatives qui permettent d'apprécier les gains d'utilisabilité
- Des réponses qualitatives qui permettent de réviser la conception d'une interface

II. Evaluation prédictive des performances

- Méthodes d'évaluation utilisables au moment de la conception
- Méthodes analytiques
  - reposent sur l'analyse de l 'exécution des tâches et sur des modèles cognitifs et conceptuels
- Modèles cognitifs d 'exécution des tâches

  => modèles théoriques et généraux
  - Modèle du processeur humain
  - Modèle Keystroke
  - Modèle GOMS
  - ...

Le Processeur Humain: modèle prédictif

Individu = système de traitement de l 'information

- Système composé de 3 sous-systèmes
  - sous-système sensoriel
  - sous- système moteur
  - sous- système cognitif
- Chaque sous-système dispose de
  - mémoire locale
    - m = capacité de la mémoire
    - d = persistance de la mémoire
  - processeur
    - t = cycle de base du processeur
- Référence Bibliographique
  - *Card, Moran, Newell, The psychology of HCI, Ed. LEA*

Système sensoriel

- Système sensoriel visuel
  - m = 17 lettres
  - d = 200 ms
  - t = 100 ms
  - 2 stimulis espacés de moins de 100 ms ont tendance à fusionner
    - vitesse minimale de rafraîchissement pou r percevoir une animation
    - 10 images/sec. (film 18 mm -> 18 images/sec.)
- Système sensoriel auditif
  - m = 5 lettres (ou équivalent)
  - d = 1500ms
  - t = 100 ms
- Remarques générales
  - l 'intensité du stimulus agit sur la durée du cycle
    - plus le stimulus est intense plus la durée du cycle diminue

Système moteur

- Un mouvement = suite de micro-mouvements
  - t = 70 ms
- Loi de Fitts
  - Le temps nécéssaire pour sélectionner une cible est proportionnel à la distance à parcourir pour atteindre la cible et inversement proportionnel à la taille de la cible
  - TempsDpt = 0.1 log (2D/L)
    - D = distance à la cible
    - L = largeur de la cible
    - log2 (2 D /L) peut être vue comme l'indice de difficulté d'une sélection

D (cm)	L (cm)	TempsDpt (s)
20	1	0.4
10	0.1	0.8
30	0.5	0.7

Variantes de la loi de Fitts

- La loi de Fitts revue par Card, Moran, Newell
  - Le temps T pour sélectionner une cible avec la souris est
  - $T = K + L \log_2 (2 D / S)$
    - K : constante évaluée pour la souris à 1.03 s par Card, Moran, Newell
      - délai de prise en main de la souris et de clic sur le bouton
      - (délai minimal : 1 s).
    - L constante évaluée pour la main à 0.1 s par Card, Moran, Newell
      - (délai minimal : 0.08 s).
- Autres variantes
  - Version Welford
    - $T = L \log_2 (D / L + 0{,}5)$
  - Version MacKenzie
    - $T = K + L \log_2 (D / L + 1)$

Système cognitif

- 1 cycle permet la reconnaissance->action
  - m = 7 mnèmes (± 2) mémoire à court terme
  - d = 10-100 ms
  - t= 70ms
- Remarques
  - Si la mémoire de travail est saturée le système se dégrade
    - Ex: dans une liste trop longues (plus de 7 items) proposée à un sujet dans le temps d 'un cycle, certains items sont perdus. Les seuls items qui restent sont ceux placés au
      - début/ fin/ milieu de liste
  - La notion de mnème correspond à une *unité cognitive*
    - => RER = 1 Mnème ou 3 Mnèmes selon contexte.

Processeur Humain: application

- Mesure du temps de réponse minimal d 'individus à un stimulus issu d 'un système informatique
  - Un individu assis devant son écran doit appuyer sur la barre d 'espace lorsqu 'un symbôle apparaît. Quel est son temps de réponse (Trép)?

```
0. Le symbôle apparaît
1. Le système sensoriel est activé
     Un cycle est nécessaire pour que l 'image soit représentée
3. Le système cognitif est activé
     Un cycle est nécéssaire pour que l 'image soit connectée à une
       réponse
4. Le système moteur est activé
     Un cycle pour appuyer sur la barre d 'espace
```

  - Bilan :

    Trép = $t_{sensoriel} + t_{cognitif} + t_{moteur}$

    Trép = 100 + 70 + 70 = 240 ms (<= temps moyen)

    Valeurs pour utilisateur très rapide ou très lent:

    105 ms à 470 ms

Processeur humain: application

- Evaluation comparative : Joystick/ flèche (touche) / souris
  - Souris:
    - T = 1.03 + 0.096 log2 (D /S +0,5)
  - Joystick
    - T = 0.99 + 0.220 log2 (D /S +0,5)
  - Flèches:
    - T = 0.98 + 0.074 (Dx /Sx + Dy /Sy)
    - où Dx est le nombre de fois où il faut appuyer sur une flèche verticale
- Conclusion
  - la souris > autres procédés de pointage.
  - Taux d'erreurs: les flèches > autres procédés de pointage.

Le modèle Keystroke

- Card Moran Newell 1983
- Objectif:
  - Décomposition en tâches élémentaires et génériques pour prédire le temps d'execution
- Opérateurs:
  - K (Keystroking) : frappe
  - P (Pointing) : désignation
  - H (Homing) : rapatriement de la main
  - D (Drawing) : dessin
  - M (mental activity) : activité mentale
  - R (response time) : temps de réponse du système

Exemple

- Tâche: Découper un bout d'image avec un éditeur de dessin bitmap
  - P Pointer vers le menu
  - K Choisir la commande sélection
  - P pointer au coin supérieur gauche
  - K Appuyer sur le bouton de la souris
  - P Pointer au coin inférieur droit
  - K Relacher le bouton de la souris
  - P Pointer vers le menu déroulant
  - K Appuyer sur le bouton de la souris
  - P Choisir la commande couper
  - K Relacher le bouton de la souris

Keystroke

- Evaluation expérimentale des temps d'execution des différents opérateurs:
  - K : 0.2 secondes
  - P : Loi de Fitts modifiée => 0.8 et 1.5 secondes
  - H : 0.4 s
    - ie P + H _ 7 K
  - D : 0.9 n + 0.16 l pour n segments de longueur moyenne l
  - M : 1.35 s
  - R :
    - Max (0, n - t) = temps d'attente
    - n = temps de traitement
    - t = temps exploité par l'utilisateur
- La principale difficulté consiste à placer les opérateurs M

Règles pour le placement des opérateurs M

- Règle 1 : Insérer M devant chaque sous-méthode
  - Par ex devant tous les K qui ne font pas partie d'une chaîne d'arguments
    - ls -a /usr <=> MKK MKK MKKKK
- Règle 2 : Supprimer M s'il peut être anticipé
  - par ex: sélection avec la souris:
    - déplacer la souris + clic ° MPMK car le K est anticipé
    - c'est donc MPK
- Règle 3: si MKMKMKMK constitue un mot alors simplifier par MKKKK
- ...

Keystroke : application

- Evaluation comparative de procédés de déplacement
  - Méthode 1: souris
  - Méthode 2: clavier
- Méthode 1:
  - - Prendre la souris la déplacer au point désiré et sélectionner
    - M1 = H (souris) + P (Pointeur) + K (Clic) + H (Retour)
  - - Insertion des opérateurs M:
    - M1 = H + M+P + M+K + H
  - - Elimination des opérateurs superflus (anticipation)
    - M1 = H + M+P+K + H
      - TM1 = 3.45 s

Keystroke : Exemple

- Méthode 2 :
  - - tant que le curseur n'est pas sur la ligne cible, taper Ctrl-n
  - - tant que le curseur n'est pas sur le mot cible, taper esc-f
- Formule Keystroke:
  - M2 = K (touche ctrl) + a * K (touche n) + b * (K (touche esc) + K (touche f))
  - insertion/simplification des M:

M2 = M + K + a * K + M + b * (K + K)
= (1 + a + 2b) 0.2 + 2.7

- Méthode 1 / méthode 2
  - - M1 (Souris) meilleure que M2 (Clavier)
  - <=> a + 2b > 2.75
  - Mais pour un utilisateur expérimenté:
  - - M1 meilleure que M2 <=> a + 2b < 9,5

Bilan Keystroke

- Avantages
  - Analyse quantitative
    permet de comparer les différents choix lexicaux et syntaxiques possible d'une interface.
  - Simplicité
- Inconvénients
  - Problème du placement de l'opérateur M
  - Imprécision des mesures de base:
    - moyennes ne tenant pas compte de variations importantes:
      touches spéciales / touches usuelles
      type de sélection
  - Concentré sur l'aspect syntaxique et lexical de l'activité de l'utilisateur.
    - gain de performance au niveau lexical peut s'estomper devant l'accomplissement global d'une tâche

Conclusion

- Choisir sa méthode d'évaluation selon...
  - ... la phase dans laquelle elle intervient
    - Au moment de la conception
      - méthodes analytiques d'évaluation
      - méthodes qualitatives
  - ... le style de l'évaluation
    - en laboratoire
      - toutes (tests utilisateurs, méthodes analytiques, méthodes qualitatives, etc.)
    - en situation réelle
      - méthodes qualitatives
      - tests utilisateurs restreints
- Ne pas sous-estimer les coûts de l 'étude
  = se donner les moyens d'une évaluation fiable

Mountaz Hascoët, Univ. Montpellier II, LIRMM