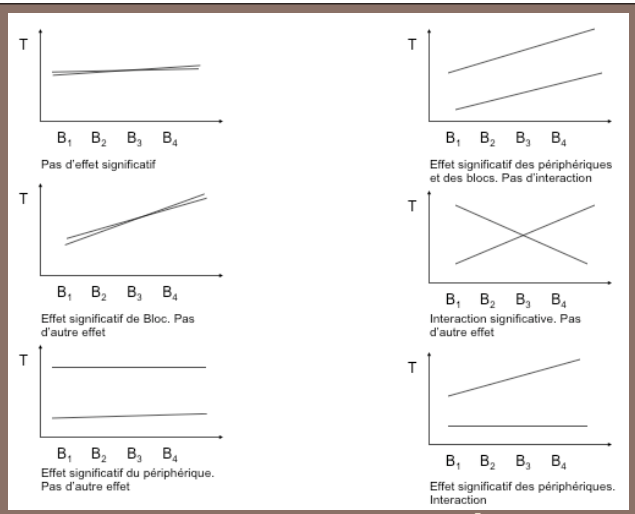



# INTERFACES 3D: ÉVALUATION

Géry Casiez http://www.lifl.fr/~casiez
RVI Master 2 spécialité IVI – Université de Lille 1

## Importance de l'évaluation



- Retour qualitatif
  - Au plus tôt dans le processus de développement (dans le cadre de la CCU)
  - Evaluations informelles (petit nombre de participants)
  - Evaluations formelles (grand nombre de participants)
- Résultats quantitatifs
  - Evaluation finale
  - Démontrer qu'une interface est meilleure qu'une autre par exemple

## Méthodes d'évaluation



- Evaluations ponctuelles
  - Participation d'utilisateurs représentatifs des utilisateurs finaux
  - Différentes étapes de conception
  - Différents niveaux d'évaluation
  - Qualitatif/quantitatif
- Questionnaires
- Interviews et démos

## Méthodes d'évaluation



- Evaluation heuristique
  - Evaluation séparée par plusieurs experts qui évaluent en appliquant un ensemble d'heuristiques et de guides de conception
  - Manque de guide de conception et d'heuristiques en 3D
- Cognitive walktrough: évaluation d'une interface en considérant plusieurs tâches de base qu'un utilisateur exécuterait
  - Evaluer la capacité de l'interface à réaliser chaque tâche de base
  - Utile pour les utilisateurs débutants

## Méthodes d'évaluation

5

- Expériences contrôlées
  - Choix d'un petit nombre de facteurs dont on veut mesurer les effets
  - Choix de mesures
  - Construction d'un plan expérimental
  - Choix d'un ou plusieurs groupes d'utilisateurs

## Participants

6

- Les gens qui participent à une expérience sont appelés "participants"
- Eviter d'utiliser le terme "sujet"
- Utiliser le terme participant pour toute référence explicite à l'expérience (ex:, "tous les participants ont obtenu un taux d'erreur important…")
- Les commentaires généraux ou les conclusions peuvent utiliser d'autres termes: " ces résultats suggèrent que les utilisateurs ont moins de chances de …"

## Variable indépendante

7

- Une variable indépendante est une variable qui est manipulée à travers la conception de l'expérience
- Ex: périphérique, type de retour, apparence d'un bouton, mise en page, sexe, âge, expertise, etc.
- Les termes variable indépendante et facteur sont synonymes
- "Indépendant" signifie "indépendant des participants"

## Condition de test

8

- Les valeurs prises par une variable indépendante sont les conditions de test (niveaux)
- Donner des noms à la fois aux variables indépendantes (facteurs) et aux conditions de test (niveaux)
- Ex:

| Facteurs | Niveaux |
|---|---|
| Périphérique | Souris, trackball, joystick |
| Type de retour | audio, tactile, retour de force |
| Tâche | Pointage, dragging |
| Visualisation | 2D, 3D, animée |

# Variable dépendante

9

- Une variable dépendante est une variable représentant les mesures ou observations d'une variable indépendante
- Ex: temps de réalisation, vitesse, précision, taux d'erreur, nombre de touches appuyées, vitesse d'apprentissage, etc.
- Donner un nom pour la variable dépendante avec les unités
- "Dépendant" signifie "dépendant des participants"
- Exemples:
  - Temps de réalisation (ms), vitesse (mots par minute, nb de sélections par minute, etc), taux d'erreur (%) ...

# Variable de contrôle

10

- Conditions ou facteurs qui (a) peuvent influencer une variable dépendante, mais (b) qui ne sont pas étudiés et dont on peut s'accommoder d'une certaine façon
- Une façon de les contrôler – est de les traiter comme des variables de contrôle
- Une variable de contrôle est gardée constante d'un test à l'autre
- Ex:, éclairage d'une pièce, bruit de fond, température
- L'inconvénient est d'avoir trop de variables de contrôle qui rendent l'expérience moins généralisable (cad., applicable à d'autres situations)

# Variable de contrôle

11

- Nombre moyen d'images par seconde
- Latence moyenne
- Retard du réseau
- Distorsion optique
- ...

# Variable aléatoire

12

- Au lieu de contrôler tous les facteurs, certains peuvent varier de manière aléatoire
- De tels facteurs sont des variables aléatoires
- Plus de variabilité est introduite dans les mesures (---), mais les résultats sont plus généralisables (+++)

## Variable de confusion

13

- Une variable qui varie systématiquement avec une variable indépendante est une variable de confusion
- Ex: Si trois périphériques sont toujours testés dans le même ordre, la performance des participants peut s'améliorer avec l'entraînement; ex., de la 1ère à la 2nde condition, et de la 2nde à la 3e condition; par conséquent "l'apprentissage" est une variable de confusion (parce qu'elle varie systématiquement avec le "périphérique")

## Intra-sujets, Inter-sujets

14

- L'administration des niveaux d'un facteur est soit intra ou inter-sujets
- Si chaque participant est testé sur chacun des niveaux, le facteur est dit intra-sujets
- Si chaque participant est testé sur seulement un niveau, le facteur est dit inter-sujets. Dans ce cas, des groupes séparés de participants sont utilisés dans chaque conditions.
- Les termes "mesures répétées" et "intra-sujets" sont synonymes.

## Intra vs. inter Sujets

15

- Question: Lors de la conception d'une expérience, vaut-il mieux utiliser des facteurs intra-sujets ou inter-sujets?
- Réponse: Ca dépend!
- Discussion:
  - Parfois un facteur doit être inter-sujets (e.g.,sexe, age)
  - Parfois un facteur doit être intra-sujets (e.g., session, bloc)
  - Parfois on a le choix. Dans ce cas, il faut faire un compromis
  - Avantage intra-sujets: la variance due aux pré-dispositions des participants est normalement la même dans toutes les conditions (cf. inter-sujets)
  - Avantage inter-sujets: évite les phénomènes d'interférences (ex: utiliser deux claviers avec une organisation différente des touches)

## Plan d'expérience

16

- Le plan d'expérience fait référence à l'organisation des facteurs, niveaux, procédures … dans une expérience
- Exemple:
  - "Plan 3 x 2 intra-sujets" correspond à une expérience avec deux facteurs, ayant 3 niveaux dans le premier, et 2 niveaux dans le second. Il y 6 conditions de test au total. Chacun des facteurs est intra-sujets signifiant que tous les participants testent toutes les conditions
- Note: Une conception mixte est aussi possible
  - Dans ce cas, les niveaux d'un facteur sont administrés à tous les participants (intra-sujets) alors que les niveaux d'un autre facteur sont administrés à des groupes différents (inter-sujets).

# Contre balancement

17

- Pour une conception intra-sujets, la performance des participants peut s'améliorer avec l'entraînement d'une condition de test à une autre.
- Pour compenser, l'ordre de présentation des conditions est contre-balancé.
- Les participants sont divisés en groupes, et un ordre différent est administré à chacun des groupes
- L'ordre suit un carré latin

# Carré latin

18

- La caractéristique définissant un carré latin est que chaque condition apparaît seulement une fois dans chaque ligne et colonne.
- Ex:

Carré latin 3 X 3

| | | |
|---|---|---|
| A | B | C |
| B | C | A |
| C | A | B |

Carré latin 4 X 4

| | | | |
|---|---|---|---|
| A | B | C | D |
| B | C | D | A |
| C | D | A | B |
| D | A | B | C |

Carré latin 4 X 4 balancé

| | | | |
|---|---|---|---|
| A | B | C | D |
| B | D | A | C |
| D | C | B | A |
| C | A | D | B |

Note: Dans un carré latin balancé chaque condition précède et suit chaque autre condition un nombre égal de fois

# Analyse statistique

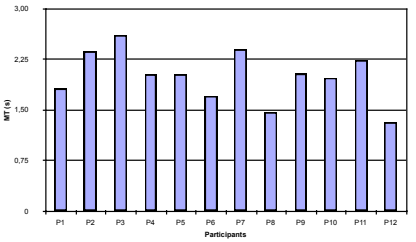
19

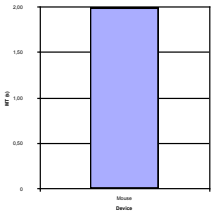
- Statistiques descriptives
- Exemple
  - La performance de 12 participants a été mesurée pour 2 périphériques (souris et tablette)
  - Plusieurs mesures de performance ont été réalisées
  - Chaque mesure de performance est une variable dépendante
  - Une des variables est le temps de réalisation (T)
  - La variable indépendante est le périphérique

# Statistiques Descriptives

20

| Participant | Souris |
|---|---|
| P1 | 1.81 |
| P2 | 2.36 |
| P3 | 2.6 |
| P4 | 2.02 |
| P5 | 2.02 |
| P6 | 1.7 |
| P7 | 2.39 |
| P8 | 1.46 |
| P9 | 2.03 |
| P10 | 1.97 |
| P11 | 2.23 |
| P12 | 1.31 |
| Moyenne | 1.99 |
| Ecart type | 0.38 |



# Evaluations comparatives

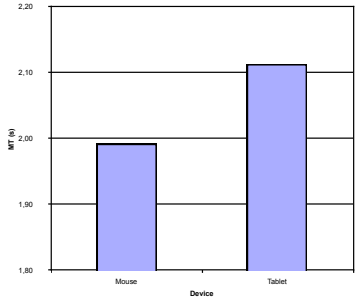
21

- Le résultat précédent, seul, n'est pas très intéressant
- L'objectif est souvent de comparer une ou plusieurs conditions
- Les conditions sont les niveaux de la variable indépendante
- Dans l'exemple, la variable indépendante est "Périphérique" et les niveaux sont "Souris" vs. "Tablette"

# Evaluations comparatives

22

| | T (s) | |
|---|---|---|
| Participant | Souris | Tablette |
| P1 | 1.81 | 1.9 |
| P2 | 2.36 | 2.17 |
| P3 | 2.6 | 2.47 |
| P4 | 2.02 | 2.03 |
| P5 | 2.02 | 2.03 |
| P6 | 1.7 | 2.12 |
| P7 | 2.39 | 2.5 |
| P8 | 1.46 | 1.89 |
| P9 | 2.03 | 2.13 |
| P10 | 1.97 | 2.02 |
| P11 | 2.23 | 2.55 |
| P12 | 1.31 | 1.54 |
| Moyenne | 1.99 | 2.11 |
| Ecart-type | 0.38 | 0.29 |




Les études comparatives sont plus intéressantes mais est-ce que ce résultat est plus intéressant?

# Hypothèse nulle

23

- Déclarer comme "hypothèse statistiquement nulle" quelque chose qui est logiquement l'opposé de ce que l'on croit.
- Appeler cette hypothèse *H0*
- Montrer à partir des données que *H0* est fausse, et doit être rejetée
- En rejetant *H0*, on confirme ce en quoi on croit

# Hypothèse nulle

24

- L'hypothèse nulle est rejetée ou non

| | | Etat du monde | |
|---|---|---|---|
| | | H0 | H1 |
| Décision | H0 | Acceptation correcte | Erreur de type II $\beta$ |
| | H1 | Erreur de type I $\alpha$ | Rejet correct |

# ANOVA

25

- C'est le principal outil statistique utilisé dans le domaine de l'interaction homme-machine pour évaluer des expériences
- Utilisé pour répondre à des questions du type "Est-ce que le temps pour accomplir telle tâche varie différemment suivant le type de technique d'interaction utilisé? "

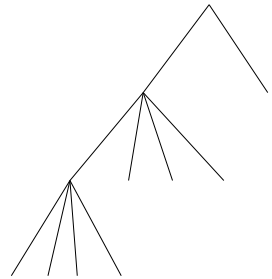
# Plan expérimental

26

Facteur 1 (ex: Périphérique): 2 niveaux (e.g. souris and tablette)

Facteur 2 (ex: Bloc): 4 niveaux

Facteur 3 (ex: ID): 4 niveaux



# Conditions d'utilisation

27

- Indépendance des données
- Distributions normales
- Homogénéité des variances

# ANOVA avec un facteur intra-sujets

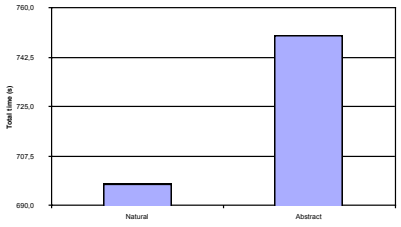
28

- Exemple
  - Apparence d'une icône avec 2 niveaux: naturel et abstrait
  - Mesure du temps de sélection (secondes)
  - 10 participants

# ANOVA avec un facteur intra-sujet

29

| Participant | Naturel | Abstrait |
|---|---|---|
| P1 | 656 | 702 |
| P2 | 259 | 339 |
| P3 | 612 | 658 |
| P4 | 609 | 645 |
| P5 | 1049 | 1129 |
| P6 | 1135 | 1179 |
| P7 | 542 | 604 |
| P8 | 495 | 551 |
| P9 | 905 | 893 |
| P10 | 715 | 803 |
| **Moyenne** | **697.7** | **750.3** |
| **Ecart-type** | **265.13** | **258.75** |




**Mauchly's Test of Sphericity[b]**

Measure: MT

| Within Subjects Effect | Mauchly's W | Approx. Chi-Square | df | Sig. | Epsilon[a] Greenhouse-Geisser | Huynh-Feldt | Lower-bound |
|---|---|---|---|---|---|---|---|
| T | 1.000 | .000 | 0 | . | 1.000 | 1.000 | 1.000 |

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept
Within Subjects Design: T

# ANOVA avec un facteur intra-sujet

30

**Tests of Within-Subjects Effects**

Measure: MT

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| T | Sphericity Assumed | 13833.800 | 1 | 13833.800 | 33.359 | .000 |
| | Greenhouse-Geisser | 13833.800 | 1.000 | 13833.800 | 33.359 | .000 |
| | Huynh-Feldt | 13833.800 | 1.000 | 13833.800 | 33.359 | .000 |
| | Lower-bound | 13833.800 | 1.000 | 13833.800 | 33.359 | .000 |
| Error(T) | Sphericity Assumed | 3732.200 | 9 | 414.689 | | |
| | Greenhouse-Geisser | 3732.200 | 9.000 | 414.689 | | |
| | Huynh-Feldt | 3732.200 | 9.000 | 414.689 | | |
| | Lower-bound | 3732.200 | 9.000 | 414.689 | | |

Effet significatif $F_{1,9} = 33.36$ $p < 0.0001$ de l'apparence de l'icône sur le temps d'acquisition.

Nombre de degrés de liberté (nb de niveaux – 1),
(nb de niveaux – 1) x (nb de participants – 1)

**Pairwise Comparisons**

Measure: MT

| (I) T | (J) T | Mean Difference (I-J) | Std. Error | Sig.[a] | 95% Confidence Interval for Difference[a] Lower Bound | Upper Bound |
|---|---|---|---|---|---|---|
| 1 | 2 | -52.600* | 9.107 | .000 | -73.202 | -31.998 |
| 2 | 1 | 52.600* | 9.107 | .000 | 31.998 | 73.202 |

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

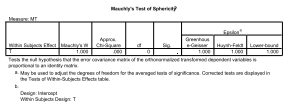
a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Si $p < 0.05$, il y a 95% de chances que la différence observée n'est pas due au hasard

# ANOVA avec un facteur intra-sujet

31

- Retour à l'exemple précédent




**Tests of Within-Subjects Effects**

Measure: MT

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| T | Sphericity Assumed | .088 | 1 | .088 | 4.510 | .057 |
| | Greenhouse-Geisser | .088 | 1.000 | .088 | 4.510 | .057 |
| | Huynh-Feldt | .088 | 1.000 | .088 | 4.510 | .057 |
| | Lower-bound | .088 | 1.000 | .088 | 4.510 | .057 |
| Error(T) | Sphericity Assumed | .214 | 11 | .019 | | |
| | Greenhouse-Geisser | .214 | 11.000 | .019 | | |
| | Huynh-Feldt | .214 | 11.000 | .019 | | |
| | Lower-bound | .214 | 11.000 | .019 | | |

$F_{1,11} = 4.51$ $p=0.057$

p-value > 0.05 pas d'effet significatif sur le temps. Peut-on conclure qu'il n'y a pas de différence entre les deux périphériques?

# Analyse de puissance

32

- *La puissance*, qui varie entre 0 et 1, est la capacité à détecter un effet, s'il existe
- Plus elle est proche de 1, plus l'expérience a de chances de détecter un effet
- *Puissance* > .80 est généralement considéré comme acceptable; i.e., si *p* est significatif et *Puissance* > .80, il y a de fortes chances que l'effet existe vraiment.

Exp. curseurs

**Tableau ANOVA pour Technique**

| | ddl | Somme des carrés | Carré moyen | Valeur de F | Valeur de p | Lambda | Puissance |
|---|---|---|---|---|---|---|---|
| Sujet | 9 | 1231492.000 | 136832.444 | | | | |
| Catégorie pour Technique | 1 | 13833.800 | 13833.800 | 33.359 | .0003 | 33.359 | 1.000 |
| Catégorie pour Technique * Sujet | 9 | 3732.200 | 414.689 | | | | |

Exp. Périph.

**Tableau ANOVA pour Technique**

| | ddl | Somme des carrés | Carré moyen | Valeur de F | Valeur de p | Lambda | Puissance |
|---|---|---|---|---|---|---|---|
| Sujet | 11 | 2.294 | .209 | | | | |
| Catégorie pour Technique | 1 | .088 | .088 | 4.510 | .0572 | 4.510 | .482 |
| Catégorie pour Technique * Sujet | 11 | .214 | .019 | | | | |

# Analyse de puissance

33

- Augmenter la puissance de l'analyse en augmentant le nombre de participants

# 95% CI

34

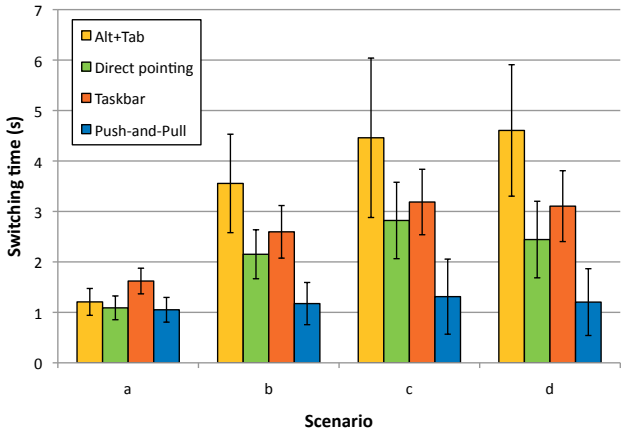



**Figure 4. Mean switching time for SWITCHING TECHNIQUE and SCENARIO. Error bars represent 95% confidence interval.**

# ANOVA avec un facteur inter-sujets

35

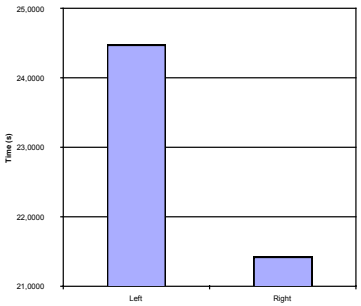
- Exemple:
  - Tester si une interface ou une technique d'interaction fonctionne mieux avec les gaucher ou les droitiers (ou hommes vs femmes)
  - 2 groupes de participants sont nécessaires: 5 gauchers (G) et 5 droitiers (D)
  - La variable indépendante est la latéralité avec 2 niveaux, Gauche et Droite
  - La variable dépendante est le temps (secondes) pour accomplir la tâche.

# ANOVA avec un facteur inter-sujets

36

| gauche | droite | |
|---|---|---|
| 25.6 | 14.3 | |
| 23.4 | 22 | |
| 19.4 | 30.4 | |
| 28.1 | 21.1 | |
| 25.9 | 19.3 | |
| *24.48* | *21.42* | *Moyenne* |
| *3.29* | *5.84* | *Ecart-type* |

Gros écart entre les moyennes. Différence statistiquement différente?




ANOVA Table for Time(s)

| | DF | Sum of Squares | Mean Square | F-Value | P-Value | Lambda | Power |
|---|---|---|---|---|---|---|---|
| Handedness | 1 | 23.409 | 23.409 | 1.043 | .3371 | 1.043 | .142 |
| Residual | 8 | 179.616 | 22.452 | | | | |

→ Augmenter le nombre de participants

# Conception intra vs. inter sujets

37

- Intra-sujets: 2 fois plus puissant avec 2 fois moins de participants (si 2 niveaux)…
- … mais demande 2 fois plus de temps
- Quand c'est possible, la conception intra-sujets est préférée pour les groupes de petite taille

## ANOVA avec un facteur intra-sujets et un facteur inter-sujets

38

- Exemple: Contrôler si le contre-balancement annule les effets d'apprentissage
- Retour sur l'exemple des icônes abstraites et concrètes
- L'ordre a été contrebalancé entre les sujets

## ANOVA avec un facteur intra-sujets et un facteur inter-sujets

39

**ANOVA Table for Completion Time (s)**

| | DF | Sum of Squares | Mean Square | F-Value | P-Value | Lambda | Power |
|---|---|---|---|---|---|---|---|
| Group | 1 | 67744.800 | 67744.800 | .466 | .5142 | .466 | .091 |
| Subject(Group) | 8 | 1163747.200 | 145468.400 | | | | |
| Icon Type | 1 | 13833.800 | 13833.800 | 30.680 | .0005 | 30.680 | .999 |
| Icon Type * Group | 1 | 125.000 | 125.000 | .277 | .6128 | .277 | .074 |
| Icon Type * Subject(Group) | 8 | 3607.200 | 450.900 | | | | |

- Pas d'effet de groupe significatif ($F_{1,8}$ = 0.466, ns)
- Pas d'interaction significative Type d'icône x Groupe ($F1,8$ = 0.277, ns)
  - → pas de transfert d'apprentissage asymétrique

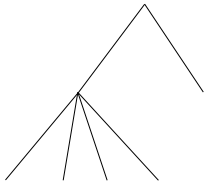
## ANOVA avec deux facteurs intra-sujets

40

- Exemple: 2 facteurs
  - Périphérique P1, P2
  - Bloc B1, B2, B3, B4

Facteur 1 (Périphérique): 2 niveaux

Facteur 2 (Bloc): 4 niveaux



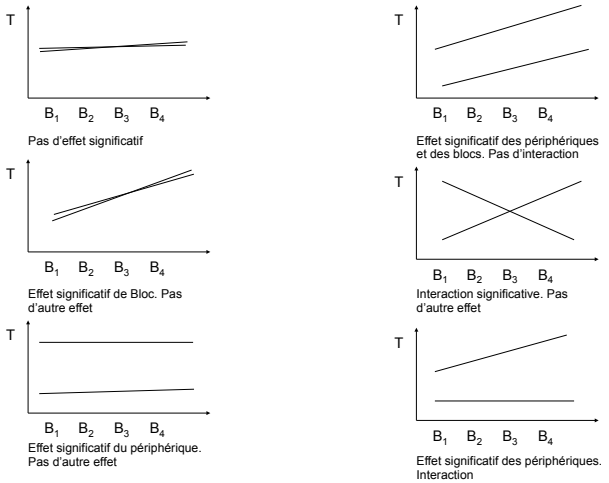
# ANOVA avec deux facteurs intra-sujets

41

- Effet significatif principal: Périphérique ou/et Bloc
  - e.g. Effet significatif entre P1 et P2
  - On peut trouver une différence significative entre les blocs mais sans pouvoir conclure sur l'effet d'apprentissage
  - Pour connaître l'histoire complète: Etudier l'interaction Périphérique x Bloc

# Interaction

42




# ANOVA avec deux facteurs intra-sujets et un facteur inter-sujets

43

- Même conception que précédemment
- Facteur inter-sujets: contre-balancement des périphériques

# Aussi…

44

- 4 facteurs…
- … n facteurs

## Coder l'expérience

45

- Toujours enregistrer les données brutes!
  - Evite les erreurs
  - Evite les oublis

## Questionnaires

46

- Retour qualitatif
- Utiles à tous les stades de conception
- Connaître la facilité d'utilisation d'un produit, le degré de fatigue…
- Critères 3D:
  - présence
  - confort utilisateur

## Présence

47

- «feeling of being there»
- Comment la mesurer?
  - échelle de 1 à 100
  - questionnaire
  - réactions des utilisateurs suite à l'apparition d'événements
  - tests de mémoire

## Confort

48

- simulator sickness
- similaire au mal des transports
- défaut de correspondance entre différentes informations qui arrivent au cerveau
- effets secondaires dans le monde réel
- Fatigue des bras, mains et yeux
- Utilisation de questionnaires

## Construction de questionnaires

49

- Les questions doivent s'enchaîner logiquement
- Les réponses à une question ne doivent pas être influencées par les précédentes
- L'enchaînement doit aller du général au particulier et rester logique

## Réponses

50

- Sur une échelle: Echelle de Likert
- Questions fermées : oui/non, QCM
- Questions ouvertes

## Echelle de Likert

51

- La personne indique son degré de satisfaction pour une question sur une échelle
- L'échelle comporte 5 ou 7 niveaux
- Echelle bipolaire
- Possibilité de comparer des techniques

## Types de questions

52

- Evaluer la facilité d'utilisation de la souris:
  1. Très difficile
  2. Difficile
  3. Normal
  4. Facile
  5. Très facile

## Echelle de Likert: analyse des réponses



- Analyse séparée des réponses ou regroupement par sommation
- Analyse séparée: pas possible d'utiliser une ANOVA
  - Tests non-paramétriques
- Regroupement: possibilité d'utiliser une ANOVA

## Exercice



- Proposez un plan expérimental pour comparer les techniques de ray-casting et de main virtuelle.