

Research Methods in HCI 2nd edition

Section 3: Experimental design

2017年7月18日

高田 雄太

Section 3: Experimental design (pp.45--69)

- 実験デザインは **true-experiments, quasi-experiments, and non-experiments** の3種類に分けられる (Fig.3.1)
 - True-exp.: 複数のグループ or 条件 + 被験者の無作為抽出
 - Quasi-exp.: 複数のグループ or 条件 + 被験者の固定抽出
 - Non-exp.: 一つのグループ or 条件のみ
- **True experiments の特徴** (本章ではこれを主に解説)
 - 最低でも1つの仮説の妥当性確認が目的
 - 最低でも2つのグループ or 条件が必要
 - 通常 dependent variables は定量的に観測可能
 - 結果は統計テストにより評価可能
 - potential biases は除去すること
 - 異なる被験者、時間、場所、実験でも代替可能

3.1 What needs to be considered when designing experiments?

- **仮説:** 良い実験には過去の実態調査の結果に基づく良い明確な仮説
 - #independent variables により #control conditions が決まる
 - Hypothesis 例: [マウス | ジョイスティック | トラックボール] で、[大 | 中 | 小] のアイコンを選択する際のスピードは変わらない。
 - [マウス|ジョイスティック|トラックボール] X [大|中|小] = 9 conditions
 - 目的に応じて最適な評価 metrics を選択, i.e., 速度
 - その他: タイプ速度=入力数/分、タイプ精度=正当数/タイプ数
- **Independent variables をどのように制御するか?**
 - 多くの場合制御困難, e.g., 音声認識エラーはユーザに影響を与えるか
 - 音声認識はエラーなく(あり)、完全認識(一定確率誤認識)という条件の比較
 - 現実的には音声認識にはエラーが発生するため、制御(実験)困難
 - 代替案 Wizard-of-Oz approach (理想形に対するテスト) が存在するが、人間もまた不完全であるため、問題の本質的な解決にはならない

3.2 Determining the basic design structure

- **実験の基本構造** (Fig.3.2)
 - 実験で調査したい independent variables はいくつあるか？
 - 1つ = Basic design, 2つ以上 = Factorial design
 - 各 independent variable はいくつ異なる values を持つか？
 - i.e., いくつ conditions が必要か？
- **実験デザインの選択**
 - independent variables の数に応じて、実験デザインを between-group, within-group, and split-plot design から選択

3.3 Investigating a single independent variable

- **Independent variable が 1 つの仮説例, i.e., basic design**
 - H1: キーボード配列 [QWERTY|DVORAK|アルファベット順] の間でタイプスピードの違いはない
 - H2: [初心者|熟練者]間でオンラインストアへの出品時間に違いはない
 - H3: [米国|ロシア|中国|ナイジェリア]顧客間で業者への信頼に違いはない
 - condition の数 : H1=3, H2=2, H3=4
- **Between-group design** (Fig.3.3)
 - 各被験者はある一つの条件のみに参加, i.e.,被験者数は条件数に依存
- **Within-group design** (Fig.3.4)
 - 各被験者は複数の条件に参加
 - 一つのグループのみすべての条件に参加

※異なる design methods で収集データに対して、それぞれ異なる統計アプローチが必要

3.3.1 Advantages and disadvantages of [between-group | within-group] design

- **Between-group design のメリット・デメリット**

- 1グループ対1実験のため、統計的に cleaner design, learning effectなし
- Within-group と比べて実験時間が短いため、confounding factors も制御容易
- 被験者個々の能力差の影響により、観測したい significant differences が a high-level of “noise” に溶け込み、結果として Type-II エラーを招く恐れ
- Type-II 抑制には被験者数(m)の増加が有効だが、within-group と比べて、各条件(n)に割り当てる必要があり、被験者合計数(m x n) は大きくなる

- **Within-group design のメリット・デメリット**

- 被験者少なくてOK (被験者やコストは限られておりHCIでは重要な要素)
- 被験者者個々の能力差も独立
- 被験者は似たタイプのタスクに取り組むため、後半の結果にlearning effectあり
- Fatigue and Time: 被験者は人間
- Learning effect や fatigue の影響を制御するために、**A Latin Square Design** (次ページ) が一般的に用いられる

[参考] A Latin Square Design: ラテン方格法

- $n \times n$ の行列に n 個の異なる記号を各行、各列に一度だけ出現するよう並べたものをラテン方格 (方陣) と呼び、これに基づく実験計画法をラテン方格法と呼ぶ

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix} \quad \begin{bmatrix} a & b & d & c \\ b & c & a & d \\ c & d & b & a \\ d & a & c & b \end{bmatrix}$$

- 例えば、行に被験者、列にタスクを配置し、すべての被験者がすべてのタスクを異なる順序で取り組む (タスク間の condition を遠いものにする?: counterbalancing; the learning is effectively canceled out, i.e., balanced) ことで Within-group のデメリットを軽減

3.3.2 Choosing the appropriate design approach

- **Between-group と Within-group どちらを選ぶべきか？**
 - それぞれ一長一短、相反する特徴を持つため、実験のコンテキストを踏まえ、ケース・バイ・ケースで選択
- **Between-group を選択するケース**
 - シンプルな/個人能力差異が限定的、LE を受けやすいタスクの場合
 - 個人能力の推定は困難だが、一般的にシンプルで集中する必要のないタスクであればあるほど、その差異は小さくなる
 - Within-group で調査できない問題の場合
 - 前述の仮説 H2 の場合、被験者は初心者と熟練者に同時になれない、仮説 H3 も同様の理由で within-group では調査不可能
 - i.e., conditions ではなく、participants を分ける必要がある
- **Between-group 選択時の注意点**
 - Potential confounding factors, e.g., gender, age, computing/internet experience, に配慮すること
 - 調査したい変数以外の要素は小さい(似た)被験者を選ぶこと

3.3.2 Choosing the appropriate design approach

- **Within-group を選択するケース**

- 複雑な/個人能力差異の大きいタスク、LEを受けにくいタスクの場合
 - 個人の認知も依存するような複雑なタスクは個人能力差異が大きくなる傾向
 - e.g., 2つの異なるフォント間で読む速度に変化はない
- 被験者ストックが少ない場合
 - e.g., universal usability: 専門的な/特殊な条件 + そのエリア在住の被験者となると途端にその数は少なくなる

- **Within-group 選択時の注意点**

- Learning effect への配慮
 - Initial interaction の測定が目的でなければ、十分に時間を与えることは、learning effect の抑制に有効
 - Learning curve は、時間が経つにつれ flat に (Fig.3.5; 少しでも有効!)
 - **HCI において counterbalancing + training/trial は、learning effect の制御によく用いられる**
- Fatigue への配慮
 - タスク数の削減、実験時間の短縮(60-90min)、強制的な休み時間の導入

3.4 Investigating more than one independent variable

- **Factorial design (Fig.3.2)**

- independent variable が複数の実験に対する設計法で、変数をサブセットに分割し、各変数ならびに変数間の interaction effects を評価
 - i.e., 適用対象：一つ以上の independent variable or factor の実験
- 条件の数は各変数のレベルの組み合わせで決定
 - e.g., キーボードの違いによるタイプスピード評価(Table3.2)
 - Independent variables: キーボードとタスク
 - Levels: キーボード=[QWERTY|DVORAK|Alphabetic], タスク=[作文|転写]
 - $3 \times 2 = 6$ 条件で、キーボードの影響とタスクの効果を同時に評価可能
- 各条件を counterbalance に被験者へ割り当てれば、between-group と within-group のいずれも適用可能

- **Split-plot design**

- between-group design と within-group design のハイブリッド
 - e.g., GPSを使った運転(Table3.3)
 - Independent variables: GPS使用有無と年齢帯
 - Levels: GPS=[w/|w/o], 年齢帯=[20-40|41-60|61-]
 - 年齢の影響を between-group, GPS有無の影響を within-group で評価

3.4.3 Interaction effects

- **“The differing effect of one independent variable on the dependent variable, depending on the particular level of another independent variable.”**
 - Independent variables X and Y の間に顕著な interaction が存在する場合、X and Y が dependent variable Z を決定づける
 - HCI の場合、パフォーマンスは複数の要素が影響するため、interaction effects を調査することは極めて重要、結果(design)にも影響
- 例題： [初心者|熟練者] が [マウス|タッチスクリーン]を用いて、選択できたオブジェクト数/時間を計測
 - 調査の結果(Fig.3.6)、いずれも熟練者の方が選択数/分は多いが、マウスの方がよりその選択数向上率 (学習効果) は高い。
 - メモ: ユーザの“熟練度”に“時間”の要素が含まれているような？
 - i.e., タッチスクリーンは学習期間の限られたATMパネルのようなデバイスに適しており、マウスは学習期間の長い desktop のようなデバイスに適している（上手な仮説設定がeffect発見のコツ？）

3.5.2 Reliability of experimental results

- **Reliable experiment**

- 異なる他の研究グループが実験を replicate したとしても、同じ結果が consistent, dependable and stable に得られるということ
- 一方で、物理, 化学, 生物学等のような “hard sciences” は実験結果に変動が大きく replica 困難。このような変動は error として扱われる。
 - cf. Webサーバの応答解析

- **Random errors, aka. “chance errors” or “noise”**

- 現実値から+/-いずれにも振れるオフセット値
- 除去することは不可能だが、サンプルサイズを増やすことで軽減可能

- **Systematic errors, aka. “biases”**

- Random error とは異なり、現実値から一定方向にズれていく値
- サンプルサイズを増やしても、現実値を上回るor下回る (Fig.3.7)
- 主に5つ種類の systematic errors が存在
 - measurement instruments, experimental procedures, participants, experimenter behavior, and experimental environment

3.5.2.1—3.5.2.3 Bias:

- **Bias caused by measurement instruments**
 - 実験器具に起因する bias で、例えばストップウォッチで時間を測る等
 - 追加実験や SW の使用により軽減可能
- **Bias caused by experimental procedures**
 - 実験手順に起因する bias で、タスクのランダム化のない within-group だと learning effect や fatigue による影響で bias が発生
 - 実験の instruction も重要
 - 「早くタスクに取り組んで下さい」 vs. 「自分のペースでタスクに取り組んで下さい」
 - 被験者の認識違いで bias が発生することも; この手の bias 抑制に pilot study は有効
 - 余裕がある場合のみ、予備実験を本番さながらに行い、ミス潰しに活用
- **Bias caused by participants**
 - 特定の年齢/履歴/経験/知見/を持つ被験者は systematic error の原因に
 - 被験者を注意深く選択し、ストレスの無い実験環境/方法、休息を与えること、場合によっては日を変えること等十分に配慮し bias を抑制

3.5.2.4—3.5.2.5 Bias:

- **Bias due to experimenter behavior**

- 実験監督者の言動や身振り、表情等が影響し発生する bias; P.63下参照
 - 被験者の行動を暗に誘導させることはNG
 - 実験する側の対応に差があってもNG
- Experimenter behavior による bias の抑制には、
 - 中立/落ち着き/10分前行動を心掛けること
 - 複数監督者がいる場合は、1人がリードし、それ以外はサポートすること
 - 練られた実験ガイドラインや重要箇所の録画/録音が有効, e.g., TOEIC

- **Bias due to experimental factors**

- Physical environmental factors: 物音、室温、照明、湿度等
- Social environmental factors: 周りの被験者数、被験者同士の関係等
- Experimental factors による bias の抑制には、
 - 部屋/環境は、静かに適度な室温/照明/湿度にし、清潔/均一にする
 - 可能なら被験者を一人ずつ呼び、別室で観察, e.g., 取り調べ室
 - 監督者は、部屋/環境の事前チェックを行う

これら bias を完璧に除去することは不可能だが、抑制には制御や報告が重要

3.6 Experimental procedures

- **HCI の実験におけるライフサイクル (具体例はP.66)**
 - Identify a research hypothesis
 - Specify the design of the study
 - Run a pilot study to test the design, the system, and the study instruments
 - Recruit participants
 - Run the actual data collection sessions
 - Analyze the data
 - Report the results
- Research design を支援する OSS: Touchtone
 - DNS error でリンク切れ？

3.7 Summary

- **Design process**

- Dependent variables はいくつあり、どのように計測するか？
- Independent variables はいくつあり、どのように制御するか？
- Conditions はいくつあるか？
- Experimental design は、Between-group, Within-group, or Split-plot のどれを採用するか？
- Potential bias は発生し得るか？回避/抑制できるか？

- **Experimental design まとめ**

- Between: cleaner で learning effect や fatigue に強い一方で、被験者数が必要であり、個人能力差異にも影響される
- Within: stronger で被験者数は少なく、個人能力差異にも強いが、learning effect や fatigue の抑制が必要
- Split-plot: Between と Within のハイブリッドで independent variables は複数必要