WISS に適切な評価実験デザインとは？

宮下芳明*

概要．本稿では，WISS 査読方針における「投稿論文の評価実験に問題がある場合」の扱いについて，その変遷と弊害について述べ，評価実験はどうあるべきかを考察した．完璧でないプロトタイプシステムが議論材料としては有益であるように，完璧でなくても議論材料として有益な評価実験手法が必要であると主張し，群間比較に頼らない「シングルケース実験」の可能性について主張した．

WISS はワークショップでありながら査読がある．それにあたり，「評価実験をどのように扱うか」について以下のような変遷がある．

WISS2010-11: 仮に荒削りであっても未来を切り拓くような研究，議論を呼ぶような研究であれば，評価実験の記載の有無を問わず高く評価します．

WISS2012-15： 実験については，加点のみで減点はしないものとします．有意義な評価がなされている場合には，同じ内容で評価がない場合よりも高く評価します．逆に，評価実験に問題がある場合，その理由のみで不採択にすることはせず，最終原稿から評価を取り除くという条件で，評価以外の部分の価値で採否を判断します．ただし，これは，研究途中の成果を議論するワークショップとしての特別措置であり，きちんと成果を論文誌や国際学会で発表する場合には，適切な評価を行うことが重要な要件になります．後者の扱いで採録された場合には，査読コメントや発表での議論を参考に，適切な評価を行った上で，論文誌や国際学会で発表を行うことを推奨します．

WISS2016: 評価実験については，加点のみで減点はしないものとします．有意義な評価がなされている場合には，同じ内容で評価がない場合よりも高得点となります．逆に，評価実験に問題がある場合，その部分を採録判定には含めず，それ以外の部分で採否を判断します（「採録判定時のコメント」にもその旨の表示を行います）．

WISS2017-19: 「評価が行われていない」ことを理由として減点とはしませんが，有意義な評価がなされている場合には，同じ内容で評価が無い場合よりも高得点となります．但し，評価手法に明らかな問題があり，読者に誤った情報を与えると判断された場合，減点対象となります．

* 明治大学 総合数理学部 先端メディアサイエンス学科

このように 2010-11 年は評価実験の記載の有無を問わない方針だったのに対し，2012 年以降は「しっかり評価実験を行っている論文がより高得点になる」方針に変わっている．本稿で筆者が注目したいのはそれではなく，「評価実験に問題がある場合」の扱いの変化（下線部）であり，それは 2017 年を境に厳しくなっている．「減点」と記されているが，「採録の判定は，査読結果として得られる評価値を前提とした上で，プログラム委員会における議論をもって決定します」と査読方針に記されていることから，評価実験に問題があった場合，軽微な減点を超えて論文を不採択に転落させるほどの影響を及ぼしかねないことがわかる．論文投稿目的が「採択されること」だとすると，評価実験を載せることによって不採択になるリスクは高くなり，「評価が行われていないことを理由に減点されない」ことも明記されていることから，論文を実装報告にとどめたほうが有利…つまり，評価実験をつけるぐらいなら削除し，紙面を埋めるために，本来の目的と無関係な機能を増やしたり，その説明を冗長に記したり，作例・写真等を活用したりすれば良いという発想になりかねない．だが，主観的な設計思想と実装報告にとどまっていては，「思想に共感するか否か」ぐらいしか議論できず，益は少ない．投稿者からすれば，提案システムに効果があることは自明に思えるほどの確信があるかもしれないが，効果が本当にあるかどうかは，何らかの客観的な評価がないと信用されないし，それを示すことが研究者の責務でもある．

さて，WISS で発表される「提案システム」はプロトタイプであり，完璧なものでないことが多いが，議論対象として有益とみなされている．ならば同じように，完璧ではない評価実験でも，議論可能な対象にならないものだろうか？

HCI 分野においては，心理学の実験手法として正しさが保証されているやり方で「提案システム」「従来システム」の実験群（処置群）・統制群（対照群）

の比較が行われることが一般的である．しかし「対象となる標本が無作為に抽出されたもの」「それらの測定値が正規分布をしていること」といった前提条件を満たすことが困難なことも多い．また，ワークショップで議論するプロトタイプ段階なのに，最初からそのような厳密かつ大規模な実験を行うこと自体にもリスクがあるかもしれない．たとえば大数と統計に頼るあまり，ますます人間ひとりひとりを観察しなくなり，本質を見誤る恐れもある．

そこで筆者が現時点で注目しているのは，シングルケース実験（単一事例実験，N=1 実験，少数 n 計画）である[1]．これは，一つの事例あるいは一人の被験者を対象にして，途中から提案システムを使わせる介入を行い，その前後の変化をみて介入が有効だったかどうかを調べる方法，及びその応用のことである．書籍[2]には「単一か少数である被験者・被験体がたまたま特殊という可能性もあるため，そこで得られた知見をただちに一般的なものとはとらえられないという問題がある．この実験の方法論は成熟しているが，心理学全体の中では，やや限定された分野で用いられているように思われる」と記されているが，別の書籍[3]では「心理学の歴史においては，ただひとりだけの被験者（ときには研究者自身）を対象にした実験的研究のほうが長い歴史をもっており，重要な成果を上げてきた．」「臨床や教育の分野の研究，とくに行動療法とよばれる治療アプローチの適用と効果の評価において用いられている」と記されている．国内では，書籍 [4]でこうした方法論が 1990 年にまとめられ，その後も盛んに関連書籍が発刊されている[5][6][7]．実験計画法についても[4]で網羅的に整理されており，反復型実験計画，非反復型実験計画，基準移動型実験計画，交替操作型実験計画，その他の実験計画と分類がなされている．

シングルケース実験では，従属変数の継続的な測定を行い，介入の効果が反映されるかどうかを検討する．介入の効果がはっきりと検出できるように，介入前の測定値の推移（ベースライン）を安定させる実験統制を積極的に行い，安定してきたらそこで介入を行う．もちろんそれだけでは「介入前後にそれ以外の出来事があり，それが変化の原因になっているのではないか（履歴の脅威）」，「時間の経過による自然変化ではないか（成熟の脅威）」といった可能性への対処が十分ではない．このため，脅威に対する防衛力を高め，研究の内的妥当性を高めるデザイン（ABA デザイン，ABAB デザイン，多重ベースラインデザイン）がある．ABA デザインは，介入後に，もう一度ベースラインに戻してみる手法，ABAB デザインは，2 度目のベースライン期の後にさらにもう一度介入を行い，介入結果を再確認するとともに，好ましい行動の定着をはかる手法である．「いったん行動が変容したら要件を除いても元に戻らない」場合は，多重ベースラインデザインを用いる．たとえば自転車に乗れるようにする何らかの訓練法で，いったん介入を経て自転車乗りに成功してしまうと，そこで訓練を中止したところで，再び自転車に乗れなくなるわけではない．こうしたときには，複数のベースラインを対象にし，被験者ごとに介入時期を変えて検証する．

筆者は，音楽演奏支援システムを対象に，シングルケース実験について考察を行い，まとめている[1]．そこでは，シングルケース実験ならではの恩恵を享受でき，方法論に則ることでよりよい知見が得られることも結論した．ターゲットユーザに該当する被験者を多く集めにくいとき，被験者に大きな個人差があるときは多標本実験を実施しにくいので，シングルケース実験の恩恵がある．欠点を挙げるならば，[4][5][6]のように体系立てて解説した書籍が，製薬や教育など様々な分野にわたるものの，HCI 分野の評価実験を念頭に置いていないことが挙げられる．また，心理学史においてシングルケース実験はいったん全否定されたのちに復権を果たした経緯[4]があるため，[2]のようにシングルケース実験の意義を認めない書籍や思想も少なくない．もし，シングルケース実験手法をより HCI 分野に特化させ，実験手法として効率的かつ有効な手法に改良できれば，異なる名前をつけることによって切り分けていくことも重要かもしれない．WISS では常に新しく多様なインタラクション技術が発明されているが，同じように，WISS だからこその，新しい評価実験手法も発明されてほしい．そのための議論が必要である．

参考文献

[1] 宮下芳明. インタラクション研究でのシングルケース実験についての考察, エンタテインメントコンピューティングシンポジウム 2019 論文集, Vol.2019, 2019.

[2] 高野陽太郎, 岡隆. 心理学研究法, 有斐閣, 2017.

[3] 南風原朝和, 下山晴彦, 市川伸一. 心理学研究法入門 ―調査・実験から実践まで, 東京大学出版会, 2001.

[4] 岩本隆茂, 川俣甲子夫.シングル・ケース研究法 ― 新しい実験計画法とその応用, 勁草書房, 1990.

[5] マイケル・ハーセン, デーヴィド・H.バーロー. 一事例の実験デザイン ― ケーススタディの基本と応用, 1997.

[6] 平山尚, 藤井美和, 武田丈.ソーシャルワーク実践の評価方法 ― シングル・システム・デザインによる理論と技術,中央法規出版, 2002.

[7] Ronald D. Franklin, David B. Allison, Bernard S. Gorman. Design and Analysis of Single-Case Research. Psychology Press, 2014.