Requirements for the Empirical Assessment of Human-AI Work Systems: A Contribution to AI Measurement Science

Gary Klein
Mohammad Jalaeian
Macrocognition, LLC
Robert R. Hoffman
Institute for Human and Machine Cognition
Shane T .Mueller
Michigan Technological University
William J. Clancey
Institute for Human and Machine Cognition

With Special Contributions
by
Margaret Burnett
Oregon State University
Nancy Cooke
Arizona State University
Florian Jentsch
University of Central Florida

This material is approved for public release. Distribution is unlimited. This material is based on research sponsored by the Air Force Research Lab (AFRL) under agreement number FA8650-17-2-7711. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

Cite as:
Hoffman, R.R., Klein, G., Mueller, S.T., and Clancey, W.J. (2021). "Requirements for the Empirical Assessment of Human-AI Work Systems: A Contribution to AI Measurement Science." Technical Report, DARPA Explainable AI Program.

Abstract

The development of AI systems represents a significant investment. But to realize the promise of that investment, performance assessment is necessary. Empirical evaluation of Human-AI work systems must adduce convincing empirical evidence that the work method and its AI technology are learnable, usable, and useful. The theme to this Report is the notion that AI assessment must be effective but must also be efficient. Bench testing of a prototype of an AI system cannot require extensive series of experiments with complex designs. Thus, the empirical requirements that are presented in this Report involve escaping some of the constraints that are imposed in traditional laboratory research. Also, there is a recognition of new constraints that are unique to AI evaluation contexts. Empirical requirements are presented covering study design, research methods, statistical analyses, and online experimentation. The 15 requirements presented in this Report should be applicable to all research intended to evaluate the effectivity of AI systems.

Outline

Introduction

The development of AI systems represents a significant investment. But empirical testing is necessary in order to realize the promise of that investment. This is one of two Reports that express guidance for the empirical evaluation of human-AI work systems.

Who is This Report For?

This Report is primarily for system developers who are preparing to conduct evaluations of the performance Human-AI work systems.

The goal is to promote meaningful research that meets the needs of applied clients, and encourages sponsors to support rather than avoid research.

This Report may be of interest also to managers of AI-related programs and to individuals who are concerned with AI policy and evaluation.

The Focus of This Report

In practice, some form of "evaluation" occurs at every step in the system development process, spanning requirements development, system design, implementation, validation and verification testing, and refinement phases. The word "evaluation" in the present Report is not used in this comprehensive sense. This Report focuses on the empirical evaluation of the Human-AI work system, using human subjects to demonstrate the value of the AI system, or a prototype of an AI system. Furthermore, the word "requirements" in this Report refers to requirements for the methodology of AI assessment, and not to requirements for, say, engineering or design.

This Report focuses on the evaluation of existing, already constructed computer programs. That said, a number of the requirements conveyed in this Report, and a number of the recommendations conveyed in the companion Report, refer to activities that are conducted *prior* to the construction of an AI system. Conformance to the requirements and recommendations about "what to do before the study" may improve the research and facilitate the overall evaluation process.

Organization of This Report

This Report is organized as follows. The two introductory sections ("What Must Experimentation Accomplish?" and "What is Rigor in AI Measurement Science?") are followed by the main section that presents the empirical requirements.

What Must AI Assessment Accomplish?

The Human-AI work system depends critically on the cognitive capacities of both partners, in a context that is complex and dynamic. An empirical investigation that is intended to assess the quality of the work is, in essence, a psychological experiment, one in which the "Equipment" with

which the subjects work is a large-scale computational system. Empirical evaluation has two aspects.

1. *AI performance evaluation must demonstrate that the work method that is shaped by the AI is understandable, learnable, usable, and useful.*

When considered from this perspective, a host of questions about the evaluation method confront the developer:

How do we test for the usefulness and usability of the AI?
How do we evaluate its performance?
What are our measurement scales and metrics?
Is the AI trustworthy?
When is it reliable?
Is the work process that is imposed by the AI one that can be readily learned?
Is the AI valuable to operators in their actual work context?

2. *Empirical evaluation must be a path to discoveries.*

Researchers should be open to surprises, and be prepared to exploit what is learned. Empirical activities that lead to discoveries are as important, if not more important than experiments designed just to prove that a technological intervention is good. When considered from this perspective, a host of questions challenge the developer, such as:

Does the AI enable the user to diagnose AI limitations, edge cases, and difficult situations?
Does the AI empower the user to create kluges and work-arounds?
Does the AI enable the user to learn about what can go wrong?
Does the AI empower the user to recover from mistakes?
Do the task and the AI enable the participant to increase their domain expertise?
Does use of the AI leave the users with a feeling of satisfaction?

What is Rigor in AI Measurement Science?

In the pragmatic context of the empirical evaluation of AI systems, there needs to be an unbiased examination of the understandability, usability and usefulness of the technology, and demonstration of the value the technology adds to performance in the actual work context (see Clancey, 2020). At the same time, the evaluation studies need to be efficient. One way to approach this is to consider what makes for inefficiency in experimentation.

Many variables play a powerful role in determining the AI-enabled work and its results. This means that multiple variables need to be manipulated or controlled. This leads to a disconnect: The time frame for effective experimentation is outpaced by the change in technology and work. It is desirable to avoid the problem of multi-year year evaluation, because the technology has been substantially modified even while the evaluation is taking place. From the perspective of technology development, developers do not want to wait while numerous complex experiments are conducted.

In its word origins, 'rigor' means inflexibility. The word "experiment" carries with it the assumption that the empirical activity must be tightly controlled, as in a laboratory. But more rigor, as defined in laboratory experimentation, is not necessarily better when transposed to context of evaluating AI systems. It can even be worse. There is a tendency to over-control variables and make the tasks more artificial and context-free. Researchers can more easily add rigor to toy problems and laboratory-like tasks than to realistic tasks. The trap here is stripping the context away until one gets findings that don't apply to the sponsor's needs. Thus, unnecessary rigor can create barriers to completing evaluation experiments. We call this 'rigor mortis'.

Case Study in Rigor Mortis

In one famous example, a government agency funded a very large-scale study to compare "glass cockpits" (which were new at the time) with conventional cockpits, to see what the new technology contributed and what were its limitations. A large team of contractors and government researchers were involved in all this work — it was going to be a landmark project, a career-defining set of experiments to serve as a standard for doing good science on an applied question about a human-machine work system. The study involved carefully controlled conditions, carefully selected scenarios, and large numbers of pilots to be participants. It took a year just to design this single, complex experiment. Data were collected on a large number of variables, to make sure little got missed. It took another year to run all the participants. Then came the challenge of how to analyze all the data, and it took another year to develop the evaluation plan. These years of delay made the results less relevant than they would have been years earlier, plus producing such high levels of complexity for the data analysis that no one was willing to step in when the government project monitor transitioned to another program. The project was terminated. The data were never analyzed.

This concern is not fanciful. Rigor mortis events may actually be discouraging government sponsors from conducting evaluations of new technologies, and that is unfortunate because fielding untested systems creates all kinds of risks. One of the authors of this Report (GK) was in a recent meeting reviewing a new military mission, and someone stated that the program would need to include a performance evaluation. Another, more senior person responded that the military no longer seemed very enthusiastic about research and experiments anymore. This statement came as a surprise. During a break GK asked why, and was told it was because of many experiences where the research was too expensive, took too long, and provided answers that were obsolete by the time they arrived.

Case Study in Minimum Necessary Rigor: The "Klinger-Klein Test"

A study by Klinger, Klein, et al. (1993) illustrates the practical constraints that can be involved in the evaluation context, and how it is possible to satisfy the "lightweight yet necessary" requirement despite those constraints. The project involved the design of a workstation and its interfaces for operators on the AWACS air defense platform. A cognitive task analysis revealed 40 problems with the existing interface that made the cognitive work inefficient (e.g., poorly designed displays, unnecessary memory demands, loss of situational awareness). The

results suggested a redesign, which was implemented and then evaluated. But the opportunity for the operators to learn and then perform with the new workstation was very limited, to only four and a half hours. The participants had had hundreds of hours of practice with the existing interface. Yet, their performance with the new interface showed a notable improvement relative to baseline performance. This was a very simple experimental design: One experimental condition (the new interface) compared to a control condition (archived baseline performance data), and a relatively small sample size (18 operators).

This case study illustrates what it means for an evaluation to be **necessary:** It is necessary to demonstrate that the AI results in an improvement in the performance of the work. This case study also illustrates what it means for an evaluation to be **sufficient**: it was a simple experimental design that demonstrated the value-added. In this Report, we offer some requirements for "minimum necessary rigor." Our objectives are to reduce or eliminate excessive expense and excessive time.

What to Do Before the Study

Requirement 1

Do not do a literature review.

There is a significant difference between bootstrapping a research team and responding to a programmatic requirement to produce a literature review as a deliverable. Literature reviews always seem obligatory, but are rarely created soon enough to deeply impact the technology development and evaluation processes, which proceed proceeds apace at risk. Best practice is to identify the traps and challenges discovered in previous work on the topic at hand. The direct path to such a listing would be interviews with five to seven selected leaders or experts in each of the pertinent fields. Those individuals would provide the most succinct and important historical scholarship. That should be obtained prior to the beginning of an AI research and development project.

Study Design

Requirement 2

Design small-scale studies that are targeted to particular hypotheses.

All too often, researchers desire single, complex factorial experiments on the assumption that single, large-scale, large-n experiments can adequately evaluate multiple hypotheses.

Requirement 3

Conduct pilot studies to test and refine the methods, materials, and procedures.

All too often evaluation activity dives into the large-scale experiments, and once started, adjustment of the method andprocedure causes complications. Best practice is to conduct one or

more pilot studies. These are *not* designed to evaluate the primary hypothesis (e.g., the technology intervention is good), but instead are intended to garner assurance that the methodology is sound and the procedure runs smoothly. Almost invariably, pilot studies lead to improvements in the study design and methods or the practicalities of running the evaluation. Pilot studies can be conducted with as few as ten participants.

Requirement 4

Run a two-Condition, between-participants study.

EVALUATION CONDITION	CONTROL CONDITION
Participants 1 through n	*Participants n+1 - 2n*

The simplest required study involves two conditions, which we call Evaluation and Control. Different participants would participate in the two Conditions. The Evaluation Condition would involve the AI, the Control Condition would not. Participants in both Conditions would conduct the same task. This assumes, of course, that the task as completed in the Evaluation Condition is the same as the task that is used in the Control Condition. The purpose of this study is to demonstrate that the technology insertion is good.

An alternative design is to have a Control Condition in which the new technology is inserted, but some crucial element or capability of the new technology is disabled. A number of the key elements of the technology might be hobbled all together. If the results do not clearly distinguish the Control and Evaluation Conditions, something is very wrong. If the results do clearly distinguish the Control and Evaluation Conditions, subsequent studies can engage in more target empirical probes.

Note that in this design it may be not be necessary to actually "run" a Control Condition if there are usable baseline data on performance using the legacy work system.

Requirement 5

Run a two-Condition, within-participants study.

EXPERIMENTAL CONDITION	CONTROL CONDITION
Participants	Participants
1	1
2	2
3...	3...
n	n

The second required study also involves two conditions, which we again call Control and Evaluation, but the in the within-participants study, the participants experience in both Conditions. In other words, this is a repeat-measures design. The early trials in the Evaluation Condition are, effectively, training. While this study has the benefit of affirming or disconfirming the goodness

hypothesis, the design has the distinct benefit of permitting an investigation of the learning curve for using the AI technology. For one thing, if there is no learning curve, something is very wrong. When there is a learning curve, it can be invaluable in projecting the scope of a training regimen.

Requirement 6

The number of participants in the study conditions need not be large.

If there is no clear effect of a technological intervention on a sample of ten participants, then something is very wrong. If the sample size in any one condition is less than about 10, one has the risk of confusing individual differences with main effects or interaction effects, especially if there is some selection bias that influences the variability of the data (e.g., all the participants in one condition were students in evening classes, meaning generally older and more mature, so they give different results compared to the typical college freshman).

All too often, researchers advocate for experiments with large n. The tacit reason is that with increasing sample size the chances of achieving statistical significance are increased. It is further argued that large data sets are readily achievable via online platforms. So why not have a large n if it is easy to get? This too is a mythical belief. Large data sets from complex factorial experiments mandate significant amounts of data analysis, and the explanation of the results gets convoluted. The trade-off is that the low effort to get the data is balanced by the great effort to make sense of the data.

Requirement 7

Training should be minimal.

In the field setting, operators may have to use an AI system with minimum training, entailing a demand to that AI systems be highly learnable, if not intuitive. In the above discussion of the concept of rigor, an evaluation method called the Klinger-Klein Test was used as a case study. Klinger et al. found greatly improved performance on a new interface after only 4.5 hours of training whereas the participants had had hundreds of hours of practice with the existing interface. Researchers may want to evaluate a few groups, receiving different types or amounts of training, but this is not necessary, as long as satisfactory performance is achieved following minimal training.

Statistical Analysis

Requirement 8

Do not use statistical analyses that are too opaque or complicated.

If the AI doesn't yield a dramatic improvement in performance, why go to the trouble of developing it and training people to use it in the field? Especially unnecessary is the concern over achieving statistical significance at the $p < 0.01$ level versus the $p < 0.05$ level, or obtaining results that are described as "nearly" significant. On some interpretations of statistical significance, the decision is binary and therefore "marginal" significance is not a legitimate conclusion (see Hoffman, 2020).

Requirement 9

Be prepared to set a "high bar" for determining whether or not the AI is good.

Interviews with operators and stakeholders have revealed the "high bar" in the field setting. One interviewee said that if he could not achieve an understanding of how an AI system works within ten trials or attempts, that he simply would not use it (Hoffman, et al., 2021).

Requirement 10

Consider practical significance.

The notion of practical (or "material") significance has a considerable history. In 1956, Roger Kirk introduced the phrase "practical significance," saying:

> *"The appeal of null hypothesis significance testing is that it is considered to be an objective, scientific procedure for advancing knowledge. In fact, focusing on p values and rejecting null hypotheses actually distance us from our real goals: deciding whether data support our scientific hypotheses and are practically significant or useful" (pp. 755).*

Appendix A presents a method for calculating practical significance, a method that involves naive learners as participants, followed by an evaluation of their performance by domain experts.

Online Experimentation

Online platforms (both asynchronous and synchronous) can permit the gathering of a great deal of data from many participants. The cost or effort of acquiring the data can be less than when conducting the research in person, as in a laboratory, workplace, or other field setting.

Requirement 11

Identify non-naive participants and prohibit their participation.

Individuals may have participated in online studies that are similar in nature to the researcher's study. Pre-screening questions can be utilized to mitigate this, as can use of a qualification system.

Honesty is a consideration for online experiments, as it is for laboratory experiments. But in the online case, dishonesty may be motivated if a participant feels that they might be disqualified from participating for one or another reason related to the inclusion criteria. Rates of dishonesty seem to vary greatly across experiments in which the online modality has been evaluated. Some researchers have reported that simply encouraging participants to be honest can reduce the problem.

Requirement 12

Limit the scope of recruitment to participants who can connect with low levels of latency.

There can be show-stopping latencies when a participant is running multiple applications, has an unstable internet connection, and gets interrupted by unanticipated software updates. Slow wifi can discourage certain types of interactions.

Requirement 13

Design the online task so that it can be performed by an average person without expertise in a specific domain or field, and with little to no training required.

Forewarning participants that the online task might take time and be complex can reduces the set of individuals who are willing to participate. "This may be one of the larger hurdles we face when designing AMT studies for the evaluation of human-AI work systems" (Karneeb, 2017).

Requirement 14

Throughout recruitment and onboarding, participants have to be given explicit direction that they limit distractions such as cellphones, additional browser tabs, pets, and roommates or family members as much as possible.

In-person studies (in either the laboratory of the field setting) offer strong controls, with everyone in the same environment using the same systems, under researcher oversight. Online experimentation introduces uncontrolled factors that might affect the results. There will be variation in the participants' home environments. There will be interruptions from family members; there will be distractions when participants monitor their emails and phones, or when they multitask during the study; there can be influences when co-located participants compare notes.

Requirement 15

For studies that are conducted online, run a small group of participants in-person and one-on-one with a researcher and an observer so that you can observe and ask questions.

This face-to-face will let you probe deeper about the learners' reasoning and decision making when such investigation is made difficult or impossible in the online context. In addition, a face-to-face pilot study will lend assurance that the planned online methodology is sound.

References

Clancey, W.J. (2020). *Designing agents for people: Case studies of the Brahms work practice simulation framework*. Amazon Kindle Print Replica e-book. [https://www.amazon.com/Designing-Agents-People-Simulation-Framework-ebook/dp/B08D7XK8ZY].

Hoffman, R.R. (2020, September). Concept Blog Episode No. 5: "0.01 and 0.05." [https://www.ihmc.us/hoffmans-concept-blog/]

Hoffman, R.R., Klein, G., Jalaein, M., Mueller, S.T., and Tate, C. (2021). "The Stakeholder Playbook." Technical Report from Task Area 2, DARPA Explainable AI Program.

Karneeb, J. (2017). "Amazon Mechanical Turk: An XAI Overview. Report from the Evaluation Team, DARPA Explainable AI Program.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746- 759). Thousand Oaks, CA: Sage.

Appendix

Calculating Practical Significance: The Practical Significance Ratio

What does it mean for a result to be practically significant? Practical significance is either a difference between groups or conditions that obtains for a plurality or majority of the participants or a difference between groups showing shows a positive impact on human performance and goal achievement.

It is noted that while the positive impact might apply for research results that achieved statistical significance (using traditional parametric tests), but also might apply for research results that did not achieve statistical significance.

It has been suggested that practical significance can be assessed by applying tests that measure effect size, i.e., eta squared (η2) or Omega squared (ω2), or by conducting a power analysis in support of a null hypothesis significance test. All these methods are, in essence, calculated as a difference between group averages divided by a measure of variability. The problem here is the assumption that practical significance is expressed by the raw data. *The raw numbers "do not know where they came from." Practical significance* ***must*** *be evaluated by bringing in information from outside the data.*

Basic and applied research can be distinguished on many factors. Table A.1. lists the qualities of research that contribute to practical significance (adapted from Hoffman and Deffenbacher, 1993).

Table A1. Qualities of research that contribute to practical significance (adapted from Hoffman and Deffenbacher, 1993).

Pre-Conditions	
Ecological character of the research method, materials, tasks	Is the research method ecologically valid, ecologically relevant, ecologically salient, ecologically representative?
Ecological character of the theoretical foundations	Are the foundational hypotheses, theories ecologically valid, ecologically relevant, ecologically salient, ecologically representative?
Post-Conditions	
Ecological character of the results	Ecological Utility: The results help you do things.
	Ecological Novelty: The results help you do new things
	Ecological Generality: the results help you do things in diverse contexts
Effectivity character of the results	Actionability: Do results entail a procedure for creating change?
	Effectivity: Assuming actionability, what is the ease of translating results into applications or practice?

Utilizing the factors listed in Table A.1, practical significance can be calculated by the Practical Significance Ratio Method (PSR). The PSR is patterned after the Content Validity Ratio method (Lawshe, 1975).[1] The CVR is useful for quantitatively assessing the validity of each scale item with a small group of raters. In order to get reasonably stable psychometric estimates for evaluating the items' communality, a rule of thumb (that is fairly well supported by the psychometric literature) is that one wants five to seven respondents (the "5+2" rule; see Crispen and Hoffman, 2016).

Who are the people best positioned to make judgments on the qualities listed in Table B.1?: Highly experienced domain practitioners. The PSR method involves asking a group of experts whether each of the Table A.1 factors is manifested in the results.

Care must be taken to rigorously define expertise for the given domain so that genuine experts can be identified and their participation solicited. Methods for proficiency scaling are well-documented and well understood (see Hoffman, 1998, 2019; Hoffman, et al., 2014). The selected experts must be individuals who have worked actively in the domain recently and (most importantly) are in a position to make judgments or decisions about the allocation of personnel, effort, or resources.

The PSR analysis might involve data on more than one performance or outcome measure, for convergence on the determination. The expert raters are told the nature of the criterion measure (dependent variable). It might be time-to-task completion, it might be response correctness or accuracy. Additionally, the raters are shown the frequency distributions, which informs them of the span, range and skew of the dependent measure.

The PSR can be applied in two analyses, one being an analysis of performance and the other being the analysis of impact.

Analysis of Performance or Outcome Data

Each rater determines whether the obtained finding is either:

(1) Practically Significant —a game changer
or
(2) Valuable—a worthwhile improvement
or
(3) No practical significance

[1] Content Validity has been defined in different ways. A good definition is the degree to which an assessment instrument is relevant to and representative of the theoretical concept that it is designed to measure. In other words, does the instrument (consisting of some number of individual ratings scale items) measure what it is intended to it measure? In traditional psychometrics, this is evaluated by looking for a correlation between the scale items and some other, previously-established psychometric instrument that is known to evaluate the target concept. While this works in traditional psychometrics (e.g., personality scales, intelligence scales, etc.) it does not apply well to some other situations and contexts.

These alternatives are intended to give the rater latitude to make context-sensitive judgments, such as when even a few seconds can make a difference in performance (success or lives saved, etc.), even if only for a few of the participants.

The PSR is a transformation of the proportion of raters who rated an item as either "Practically Significant" or "Valuable" to the total number of participants. The formula is:

$$PSR = \frac{n_{p,sv} - (N/2)}{N/2}$$

where $n_{ps,v}$ is the number of participants who rated the result as either Practically Significant or Valuable and N is the total number of participants. The proportionalization results in a scale that ranges between −1 (perfect disagreement) and +1 (perfect agreement). When fewer than half of the participants rate the result as either Practically Significant or Valuable, the ratio will be negative. Values above zero indicate that over half of panel members agree that the result is either Practically Significant or Valuable. For the CVR, the decision heuristic is that the value should be 0.50 or greater (Ayre and Scally, 2914; Lawshe, 1975). It seems reasonable to apply this threshold to the PSR.

An alternative form for the PSR involves asking slightly different questions.

Impact Analysis: Performance, Productivity, Effectiveness

The raters can be asked to consider these four focus questions, derived from the Hoffman-Deffenbacher (1993) ecological analysis (see Table A.1, above):

- Would this finding change how resources are allocated?
- Would this finding make the work more efficient and productive?
- Would this finding enable the work system to avoid undesirable outcomes?
- Would this finding mean that there is a need to create a new role with new responsibilities?

These questions about the gains (productivity, performance, effectiveness) and avoidances (wasted time, risks, responsibility gaps, coordination costs) can each be answered by one of three responses:

(1) Definitely, (2) Perhaps, (3). Likely Not

A logic similar to that for Performance Analysis now applies: If more than half of the raters rate each of the obtained differences as (1) or (2) on each of the four focus questions, then the obtained difference is of Practical Significance according to Impact Analysis.

Subsequent deliberations can consider whether an investment based on the finding of Practical Significance is worth the risk. This is a matter of policy, which might, for example, rely on some form of cost-benefit analysis.

Steps of the PSR Method

1. CVR analysis of the pre-conditions by 5 peers.

2. If CVRs are $\geq$ 0.5, conduct a Likert evaluation of results by 5 peers in terms of the Pre-conditions.

IF aggregated Likert < 0.5, Practical significance = No

IF aggregated Likert is $\geq$ 0.5, proceed to step 3

3. CVR analysis of the post-conditions by 5 peers.

4. If CVRs are $\geq$ 0.5, conduct a Likert evaluation of results in terms of the Post-conditions by 5 peers on the Post-conditions.

IF aggregated Likert $\leq$ 0.5, Practical significance = No
IF aggregated Likert $\geq$ 0.5, Practical significance = Yes

Control Considerations

For the above analyses, it may be deemed prudent to obtain judgments from a group of experts who are *not* informed of any results from parametric null hypothesis statistical tests. Comparison of results from Informed and Not-Informed groups would be permit adjustment of the conclusion if bias is revealed (i.e., bias induced by belief in arbitration by statistical significance tests).