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

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

With Special Contribution
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). "Recommendations for the Empirical Assessment of Human-AI Work Systems: A Contribution to AI Measurement Science." Technical Report, DARPA Explainable AI Program.

Abstract

This Report is a companion to the Report titled "Requirements for the Evaluation of Human-AI Work Systems." Whereas that Report focused on the minimum necessary empirical requirements for the assessment of AI systems, this Report provides additional recommendations and technical details to assist the developers of AI systems. Recommendations are presented covering study design, research methods, measurement, statistical analyses, and online experimentation. This guidance should be applicable to all research intended to evaluate the effectivity of AI systems.

Outline

Introduction

The Purpose of This Report

The primary purpose of this Report is to express recommendations for AI Measurement Science for the methodology of empirical evaluation of the performance of the Human-AI work system.

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.

Organization of This Report

All research involving human subjects must be based on clear experimental designs, specified research methods, and statistical data analysis. Measures of performance must have clear and unambiguous operational definitions. However, the evaluation of AI systems is its own context, which brings unique factors and considerations.

While the companion Report presented "Requirements for the Empirical Evaluation of Human-AI Work Systems", this Report provides "Recommendations for the Empirical Evaluation of Human-AI Work Systems." The recommendations are optional; the requirements are necessary.

The work in which users operate with the assistance of AI systems is complex cognitive activity. Thus, the empirical assessment of an AI system is technically a psychological experiment. The 29 recommendations expressed in this Report cover the following aspects of research methodology: What to do Before the Study, Study Design, Measurement, Research Methods, Statistical Analysis, Form and Format for Data Graphs, and a special consideration of the challenges of Online Experimentation.

What to Do Before the Study

In the case of the development of AI systems, technology design and development typically commence even during the program proposal phase, without an initial evaluation of the cognitive requirements. Notional computational architectures are envisioned, ones that are supposed to be solutions of great promise. But because the developers "jump in," they may be basing their envisioned technologists on limited notions of the cognitive work. Considerable time and effort have to be expended in the early months of the project in the attempt to explain to explain concepts and methods that might be unfamiliar.

Recommendation 1

Prior to project start, engage cognitive scientists in an evaluation of the cognitive requirements for the envisioned work system.

The work in which users operate with the assistance of AI systems is complex cognitive activity. Thus, the empirical assessment of an AI system is technically a psychological experiment. Given this, it is valuable and arguably necessary for cognitive and experimental psychologists to be involved in the phases of requirements development, system design, and not just in the evaluation of existing, already constructed AI systems.

The Human-AI work system depends critically on the cognitive capacities of both partners, in a context that is complex and dynamic.

It is important to keep in mind the fact that human-machine work systems are cognitive work systems. Even before a Program is announced, an advisory group would forge recommendations regarding the specification of the cognitive requirements for Proposers. Cognitive work analysis reveals what the user really needs, which may not be what the researchers think the user needs.

Recommendation 2

The team of researchers who are designing and conducting the evaluation studies should be interdisciplinary.

The team should include: (1) subject matter experts who are accomplished in the target domain, (2) engineering design specialists (in Human-Computer Interaction, Computer-Supported Cooperative Work, and Human Factors), (3) Cognitive/Social scientists, and (4) specialists in psychological research methodology.

One way to ensure usability and usefulness of the AI is to engage domain experts in the AI system design and development processes. Having guided the technology development process from the start, at the end of the evaluation (if successful) one has a first cohort of practitioners who are facile at using the new technology. They would be the best people to begin training others, including other trainers.

Recommendation 3

Stop calling the evaluation studies "experiments."

"Experiment" is a term that entails increased (unnecessary) rigor. The empirical evaluation of AI systems can be referred to in any number of ways: "AI performance tests," "evaluation studies," "empirical evaluations," "exploratory investigations," or "AI assessments." (These are the terms used throughout this Report and the companion Report on Requirements.) The purpose is to avoid the slippery slope that leads to complex designs with multiple factors and controls. The research team can claim that additional studies or experiments can be conducted later on, if warranted. But in the meantime, a streamlined evaluation can get underway.

Recommendation 4

Conduct a Pre-mortem.

A Pre-mortem is a process in which all of the members of the development team think about reasons why the project might not work (Klein, 2007, 2021). Reasons can span any aspect of the project, from system architecture, to interfaces, and to the evaluation process. The reasons are collectively assessed in an all-team meeting, to generate concepts of how to anticipate and, avoid or mitigate traps or shortcomings. The Pre-mortem activity can take as little as 30 minutes.

Recommendation 5

Stop calling the research participants "subjects."

The major psychology professional societies have long recommended against the use of the word. It is dehumanizing and disrespectful. If anything, the participants in AI evaluation studies should be treated as valued colleagues, without whom the research would not be possible.

Recommendation 6

Write generic research protocols and protocols that assert that the empirical evaluation is exempt from IRB review.

Evaluation studies rely upon human participants, and therefore qualify as scientific research, and therefore must be reviewed and approved by an Institutional Review Board (IRB), and subsequently reviewed and approved by a Human Research Protections Office of the contracting agency. IRB review and final approval can be quite time-consuming.

Generic Proposals

Researchers often engage in systematic or long-term research, consisting of a number of studies, each of which uses similar or related methods. Researchers can submit for IRB review a proposal for such research in toto, rather than a proposal for each individual study. Once the generic proposal is approved by the Institutional IRB and by the HRPO of the contracting agency, the research can proceed without the necessity for each individual study to undergo IRB review.

Generic proposals lay out the possibilities for a series of experimental designs (measures, controls, etc.), including examples of possible materials and descriptions of methods. The submission to the institutional IRB can assert that specific protocols will be formed on the basis of findings as they develop. The submission declares that the researchers will inform the IRB Chair of any modifications to the protocol that may impact the IRB's determination.

Exempt Proposals

Protocols for the evaluation of human-machine work systems, constituted as psychological investigations, will typically involve standard and accepted psychological and psychometric measures and methods — standard measures of performance, demographic questionnaires, Likert scales for judgment tasks, commonly-used materials or standardized instruments (e.g., mental tests, surveys, etc.). In addition, the research participants' tasks will likely be similar or identical

to the tasks that they ordinarily conduct in their work. Thus, they would not entail potential risk or stress beyond that experienced in their daily lives.

These things considered, the protocol can be deemed exempt from IRB review (see the Code of Federal Regulations). But ironically, it takes an IRB (or an IRB Chair) to make that determination. However, once such a determination has been made and the HRPO of the contracting agency has accepted that determination, the empirical evaluation of the AI can commence.

However, if any individual study, even if standard in its overall methods, involves any manipulation that entails potential risk or stress, then that individual study must undergo IRB and HRPO rreview. If there is any doubt that a particular study might not need to undergo IRB review, the question should be addressed to the researchers' IRB.

Study Design

Recommendation 7

The participant's task should be ecologically valid.

Experience with large-scale projects (e.g., aircraft, highly automated naval vessels, spacecraft, self-driving vehicles) and their ensuing failures indicates that evaluation of automation is crucial early in design, *before* construction, and not just during iterative prototyping involving empirical assessment in authentic work settings (Clancey 2020; Clancey, et al., 2011).

In the empirical assessment of AI systems, the task presented to participants should be one that is representative of the tasks that are conducted in the work domain for which the AI has been created. In the experimental laboratory, the tasks presented to participants are often simplified versions of real-life tasks or experiences. They are abstracted away from their real-world context in service of control (and rigor). But they often end up artificial, so-called "toy" problems. In AI evaluation, there is usually pressure to resort to artificial tasks because they are easiest to design and present. In many AI evaluation projects the researcher defines the task, independently of the end users. The task is shaped by the researchers' model of the task. Researchers put themselves in the shoes of the user, and assume what the user needs. Thus, the task which is presented to users are low on ecological relevance. The task is completely removed from its "real world" task context (see Clancey, 2020).

At the extreme, the tasks for the participants preserve the context and complexities of the operational environment. One wants the "nasty variability of the world" to be manifest in the tasks, because that variability is certain to characterize the work context into which the technology is to be inserted. If the evaluation is sterilized, the results may not carry over. Ideally, evaluators can try out their AI system during actual work, or perhaps in a training exercise, or failing that, in a simulation. Also, it is well known that the cognitive requirements of the task matter more than the surface features, so researchers need not worry much about matching the look-and-feel of the actual task (so-called fidelity) — but should worry more about capturing the things that make the actual task difficult.

Recommendation 8

At the start of the procedure, each participant needs to be presented clear instructions on the nature of the task, the rationale for the task (i.e., its ecological validity), the materials, and the AI technology that is to be used. The instructional phase should also include practice examples.

In some AI evaluations, the instructions given to the participants in the evaluation study were about the "buttonology" of how to use the technology. We have seen researchers who are not familiar with experimentation and methodology not appreciate that they have to prepare clearly articulated material that explains the domain, the task, and the materials. Participants learn during the initial instructions; they begin to form mental models of the task and the technology that they will be using. It is important for the initial instructions to include some practice trials. And data from those trials are worth analyzing, as they provide a window into the earliest moments of the learning curve.

Recommendation 9

Expand the two-Condition within-participants study into a counterbalanced design.

The primary purpose of the between-groups and within-participants (repeat measures) studies is to simply show that at least in some circumstances the technology insertion is good. One of the requirements in the companion Report ("Requirements for the Evaluation of Human-AI Work Systems") is to conduct a simple within-participants design having Control and Evaluation Conditions as a repeat measure. We refer here to a second two-condition within-participants study as a recommendation. This is because if the between-groups study does not show that the technology intervention results in performance gain, there is no point in conducting the counterbalanced within-participants study, even though there might be a within-participants effect of an independent variable.

Rigor in laboratory research always involves the assertion that a within-participants design has to include counterbalancing to control for order effects.[1] If for example, a participants does well in the Evaluation Condition after first experiencing the Control Condition, their good performance may be due to practice and might not have much to do with the technology per se. If a participant does poorly in the Control Condition after experiencing the Evaluation Condition, it might be that the degradation in performance was due to the changes in the work method, and might not reflect a loss of the value added by the AI. This and other possible outcomes can be the subject for subsequent targeted studies.

[1] In the context of epistemological validity, a within-participants design needs not only CE and EC conditions, but also EE and CC. These latter two conditions give you a better picture of the learning curve and also control for the fact that the participants have had more time to learn the task. In EC and CE conditions, they have to relearn and not just keep learning. This complication is not manifest in the Recommendations in this Report as it runs counter to Requirement 1, that the evaluation be "lightweight" and avoid rigor mortis.

Recommendation 10

An option is to control for effort.

In the evaluation of human-AI work systems, the Control Condition involves the withholding of the AI that is being evaluated. Apart from this difference, one wants the differences between the Evaluation and Control Conditions to be "fair." The participants in a Control Condition would not have to be trained on the use of the AI. So perhaps their poorer performance is due to the lack of training rather than the technology insertion per se. The participants in the Evaluation Condition would have more new things to learn, implying a higher level of cognitive effort. This may work against any advantage accrued by the new technology, leading to an underestimation of impact.

In one design for a control study, some participants would be given specialized training, and a Control group spends the same amount of time reading irrelevant training material. In the field setting people given readings are somewhat unlikely to actually study them, whereas the training for a new AI tool would probably be mandated. After the two groups get their training, there would be a performance test. It could be argued that if the Control group (given the training readings) performed as well as the Evaluation group, why bother with the specialized training?

Recommendation 11

An option is to control for "task demands."

If an empirical evaluation is to really nail things down, certain minimal comparisons and certain appropriate control conditions and procedures ae advisable. The most salient of these is in the two-Condition between-groups study. It is well-known in laboratory psychology that research participants try to figure out the experimenter's hypotheses. It is a near certainty that the participants in an AI evaluation study will realize that they are being asked to evaluate some new technology, and will likely entertain the notion that researchers want to adduce evidence that the AI technology is good.

The easiest way to control for this is to say to participants, in the initial instructions, something like the following: "*We are conducting this study in order to evaluate the XYZ AI system. We are not sure whether it is good or not, and need you to tell us.*" A post-experimental question can ask directly about possible bias induced by task-demands.

Another way to test for task demands is deliberately bias a control group, by saying in the initial instructions that the study involves evaluating an AI system developed elsewhere in order to see whether it is useful and reliable, implying that the researchers themselves question the goodness of the AI system. If performance of that biased control group approximates that of an unbiased control group, one can infer that any measured effects of the other independent variables are not due to bias.

Recommendation 12

The group of participants in the evaluation studies should include some participants who are practitioners in a domain having job tasks that are similar to or the same as the tasks in the AI evaluation studies.

Not all of the participants should be "college students" or some slice of the general population. That said, evaluation does involve a need to see what happens when novices first learn the tasks. The issue of participant selection can get complicated if the AI system is designed for a variety of types of operators, e.g., newbies as well as seasoned vets. Researchers may want to consider studying the Group x Treatment interaction. Do some types of participants perform much better (or worse) than others? If so, researchers can become detectives and sort out what is going on.

Recommendation 13

Debrief the participants.

A great deal can be learned about what is going on in the reasoning of participants from post-study cognitive interviews. But not all of the participants in the evaluation studies need to be debriefed. More can be learned from cognitive interviews with 10 people than can be learned from a morass of numerical Likert scale data collected from hundreds of participants.

Measurement

Measures of performance can take many forms: time to task completion, proportion of critical sub-tasks completed within some time interval, and so forth. Historically, performance measures are measures of hits (correct performance), errors (mistakes made), accuracy, and time (HEAT measurement). The demonstration of progressively improving performance is an indicator of learnability and usability, but it not an unambiguous measure of understandability or usefulness. A participant might come to perform well but not really have a good understanding of the AI. An AI-enabled work system might enable good performance in the research setting, but that is no assurance that the AI would be valuable to operators in their actual work context. Thus, it is common for measures other than HEAT measures to be helpful. Researchers must imagine what measures are going to be sensitive to the effects that are expected.

Recommendation 14

Wherever possible, researchers should utilize archived data, past training data, or other data sets that reflect baseline performance that is, performance at the key tasks using the legacy work system.

Archived performance data can serve as a Control condition in evaluation studies.

Recommendation 15

Since the human-AI work system is intrinsically a cognitive work system, it is necessary to evaluate participants' mental models of the domain, the task, and the AI tools they are using.

Generally, the methods are some form of "Think Aloud" task. Details about methods for eliciting user mental models are presented in the May 2018 DARPA XAI Task Area-2 Technical Report titled "A Guide to the Measurement and Evaluation of User Mental Models"

Recommendation 16

Despite the desirability of utilizing a rich palette of measures, researchers are cautioned to avoid trying to measure everything.

Loading up on measures always complicates the analysis and interpretation of the results (e.g., the need to assess high-order interactions, look for cross-correlations, etc.).

Loading up the measures has another bad consequence; it punishes the participants. First there is the demographic form participants have to complete. It asks about such things as the participant's background and experience at tasks in the domain under study. Next, multiple measures are made during task performance. There can be forms to fill out at the completion of each trial or block of trials. (various judgments). Then, after the trials are completed, there are yet more measures taken, often focused on the participant's reactions to and comments about the task. Other measures have been used as well, such as psychometric tests to evaluate the participant's cognitive style, trust in the technology, etc. One study we learned of involved having the participants fill out multiple Likert scale instruments on top of everything else. While it is valuable for evaluation studies to use multiple measures, this can be overdone, and it places a burden on the participants.

Recommendation 17

Do not give a measure more than one interpretation, use distinct measures.

For instance, participants could perform well using an AI tool even if they have a reductive mental model of how the AI works. So a measure of performance should not be interpreted as a measure of mental model quality. Better to include an independent test of mental model quality just to be sure. Here is another example: A researcher might assume that a high rating of trust in an AI tool AI is a measure of reliability. In fact, people often rely on technologies when they don't fully trust them, because they appreciate the technology's foibles and know how to work around them.

Recommendation 18

In tasks that require participants to make judgments, try to rely on scales that have been psychometrically validated.

It is reasonable and often desirable to develop new measurement scales to elicit participant's judgments. Creating good scales and scale items is a skill that benefits from guidance by a psychometrician. The wording of scale items and the anchors in ratings scales can impact the overall levels of agreement, which can result in a ceiling effect that reduces the reliability of the measure.

It is possible to rely on selected items taken from scales that have previously been psychometrically validated. This is a perfectly legitimate thing to do (see Hoffman, Mueller, Klein and Litman, 2018). For example, many evaluation studies use the NASA-TLX measure of workload, and regard that instrument as a means of measuring mental workload. In fact, it is an instrument for measuring task load. Only a few of the items refer to what is called mental effort or mental workload. There is ample precedent in the psychometrics literature for using only selected items from a validated scale. The general guidance is: "Ask participants the questions to which you seek answers."

Recommendation 19

If you are using judgment scales, consider the alternative ways of obtaining responses.

For example, two-alternative forced-choice tasks are relatively powerful. Research shows that participants can get frustrated by having too many choices. Likert scales often have odd numbers of options (ranging from "agree" to "disagree"), but if you use an even number of options (i.e., no mid-point) you can force participants to take one side, and later you can recode these into binary responses.

Recommendation 20

Wherever possible, include a "free response" option for any judgment scales or questionnaires that you use. This option may turn up information you hadn't thought about before.

We have all experienced pre-formulated feedback forms that have questions with a limited set of response options, options that do not adequately capture our sentiments or experiences. The empirical evaluation of AI systems should not so constrain the participants.

Statistical Analysis

Recommendation 21

Think of statistical tests as exploratory tools.

Misuses and limitations of traditional significance testing (see Gigerenzer, 2004; Kirk, 1996; Schmidt, 1996; Yates, 1951) motivate an approach that considers parametric null hypothesis significance testing as an exploratory tool, not a means of automating scientific judgment by calculating significance levels or effect sizes. For example, a statistical "effect" is sometimes interpreted as a causal effect of the independent variable. As another example, significance levels are sometimes reported well beyond the second decimal place, which is unnecessary. Graphs of results are sometimes used to *imply* that the results show meaningful effects or differences when there has been a failure to achieve statistical significance.

The reliance on significance testing can be tempered by a consideration of the practical significance of the results. For experiments having domain practitioners as participants, those participants could have had many hours of previous experience and practice at the tasks. That being the case, if their performance on the AI-enabled task (Experimental Condition) shows marked improvement, that would be a clear indicator of a practically significant result. See the Appendix in the companion Report ("Requirements for the Empirical Assessment of Human-AI Work Systems") for another method of evaluating practical significance.

Recommendation 22

Explore the performance curve.

A frequency diagram is a fundamental way of depicting results on performance, and has stood the test of time (Gigerenzer, 2004, 2015). Frequency x performance histograms have scores on the performance measure as the values on the x-axis and frequency (of participants receiving each of

the scores) on the y-axis. Frequency diagrams have been surprisingly absent from presentations on AI evaluations. For performance functions, the data are often messy and non-normal; distributions are typically skewed. Null hypothesis significance tests may yield misleading results if the assumptions of the tests are not met.

Recommendation 23

Explore the learning curves.

Graphs showing learning curves are often absent from presentations on AI evaluations; the results are only described by the findings of statistical significance tests along with group averages. Learning curves can be revealed by creating histograms having scores on the (performance) measure as the values on the y-axis, and time or trials on the y-axis. A learning curve should manifest for individual participants, as well for an aggregation (average, median, mode).

Recommendation 24

Explore the overlap of the distributions.

What a researcher seeks is minimal overlap of the distributions for the Experimental and Control conditions. On the other hand, complete lack of overlap is generally not to be expected. It is well understood in statistics that a statistically significant difference can be found and yet the better performers in a Control Condition out-perform the poorer performers in some other condition (e.g., the Evaluation Condition). The amount of overlap can be striking, even nearly total, and even when a statistically significant difference is obtained. Figure 1 illustrates this using data from an actual AI evaluation study. Nearly all of the participants in the Control condition (blue histobars) outperform the participants in the Evaluation Condition whose performance falls in the first two quartiles for the Evaluation Condition (orange histobars). A noteworthy proportion of the participants in the Control Condition outperform the median performance of participants in the Evaluation Condition.

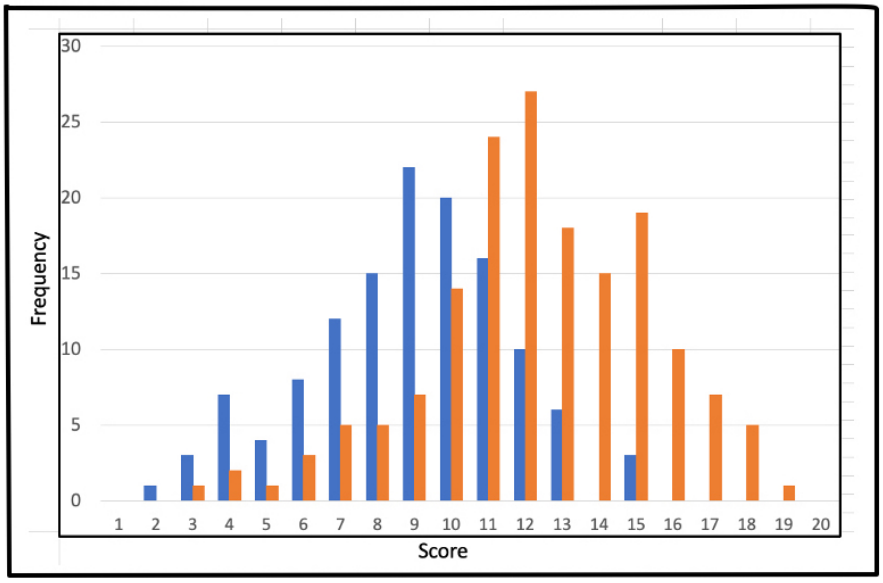

Figure 1. An example of distribution overlap.

The examination of distributions also involves skew. Sometimes it is the skew that contributes to the achievement of practical significance, as the following Case Study illustrates.

Case Study in the Exploratory Analysis of Distributions

In a study conducted by Anderson and his colleagues at Oregon State University, (Anderson, et al., 2021), groups of 25 men and 25 women were compared in terms of elements of cognitive and learning styles. Data were represented in a frequency diagram having the number of style elements on the y-axis and semi-decadal age groups on the x-axis. The male-female distributions showed considerable overlap, the group averages seemed close. But the statistical tests showed a significant difference for men and women for most of the age groups. Upon inspection, it was clear that the male-female distributions were skewed, and skewed in opposite directions. It was this skew that was the interesting discovery, not the fact that statistical tests showed significant differences.

Recommendation 25

"Outlier" data points are not things to be ignored, discarded, or kept out of the statistical analyses.

Some graphs showing data distributions use dots to indicate "outliers." The removal of a participant's data point(s) is often done solely for the sake of forcing the data set to confirm to the assumptions of the parametric statistical tests. But especially in performance evaluation, there is no such thing as an outlier. The highest achievers represent what is humanly possible. Poorest performance can be a pointer to a training problem or a selection problem, and should not be ignored even if the participant was not performing with due diligence. Post hoc interviews are recommended. Probing the meaning of "outliers" is a significant path toward deeper understanding.

To justify the removal of any data point, outlier or not, there must be *independent* evidence that the participant was not performing with due diligence. It should be kept in mind that participants who are not "outliers" might also have failed to perform with due diligence.

Form and Format for Data Graphs

Recommendation 26

It is recommended that AI developers reference the graphing standards provided by the American Psychological Society and the Human Factors and Ergonomics Society.

Many of the problems seen in data graphs used in presentations and technical reports derive from the reliance on readily-available software. These invariably have defaults for graph style and graphical elements that are not in accord with best practice in human factors and experimental psychology. Graphing Improprieties take a number of forms: the use of microfonts, the use of thin lines for the x-and y-axes, the use of under-specific labels for the x- and y-axes, too much clutter, and so forth.

Following are some of the specific recommendations from the American Psychological Association and the Human factors and Ergonomics Society:

- *Graphs should have clear, readable labels for the x- and y-axes.*
- *Graphs should always have a title, placed below the figure, that says clearly and succinctly what is being graphed* (e.g., "Performance as a function of experimental condition").
- *Graphs should always include a legend that labels the conditions (groups, trials, etc.) that compose the independent variable.*
- *Graphs should be accompanied by text that calls out he major or salient findings.*

American Psychological Association graphing guidelines [https://apastyle.apa.org/style-grammar-guidelines/tables-figures/figures]

Human Factors guidelines
[https://www.researchgate.net/publication/220457627_Guidelines_for_Presenting_Quantitative_Data_in_HFES_Publications]

Recommendation 27

It is recommended that the data graphs presented by all Performer Teams on a given AI development project be consistent in their graphing design, its style and format.

Experience has shown that the Performer Teams on projects utilize different graphing styles. Conformance to accepted standards on the part of all the Performer Teams on a project is highly recommended as it would ensure consistency and enhance the clarity of communication for the project as a whole.

Recommendation 28

Avoid the creation of graphs with too much clutter, such as multiple colors, layers, 3-D perspectives, etc.

While complex graphs have their uses, in presentations they are usually overwhelming. There is insufficient time for viewers to both understand the meaning of the graphical elements (shapes, colors, etc.) and then unpack the meanings of the results.

Recommendation 29

A graph should always be accompanied by text that calls out the main finding of interest that is the focus of the graph.

This is often lacking, especially in presentations. In reports, the text should be placed after the Figure.

Experimentation Using Online Modalities

In recent years, researchers in a variety of disciplines have been making more use of distance modalities to conduct research that involves collecting data from persons. This is particularly true for social sciences including psychology. Online research has been called a "technological revolution" is social-psychological research (Buhrmester, Talaifar, and Gosling, 2018, p.152). Modalities such as that provided by Amazon's Mechanical Turk have been utilized in research that has focused on the assessment of new technologies, including AI systems. Other online platforms are being released often, including ones that are specialized to one or another research domain (for demographics, surveys, and other empirical activities).

Advantages of Online Experimentation

Online research has some distinct advantages. For example, it allows researchers to obtain data from a broad demographic, typically more diverse than is available in the traditional population of research participants (college students) (Casler, Bickel, and Hackett, 2013). On the other hand, data quality can be in question for participants who are not native English speakers, and it can remain difficult to find samples of hard-to-reach populations (Buhrmester, Talaifar, and Gosling, 2018, p. 152).

Most recently, reliance on online platforms has come to be regarded as a necessity due to the impact of the COVID virus. The online modality result in the gathering of a great deal of data from a great many persons. The "cost" or effort of acquiring the data can be less than when conducting the research in person, as in a laboratory or field setting.

The Downside

Because large amounts of data can be collected via distance modalities, researchers may find themselves loading up the burden on the participants, with various questionnaires and judgment tasks.

Because large amounts of data can be collected via distance modalities, the advocation of Mechanical Turk contributes to the drive to achieve statistical significance, which encourages the collection and analysis of large samples. One researcher commented, "If we can use Mechanical Turk and get lots of data the easy way, why not?

A particular concern in discussions of the pros and cons of online experimentation is with the issue of whether the data collected in a distance modality are of high quality (Buhrmester, Kwang and Gosling, 2011). The person-to-person conduct of a psychological experiment is a skill that requires practice and interpersonal finesse, especially in the comportment of the researcher as she or he interacts with the participants. In the in-person modality, participants are always able to ask questions of the researcher directly, and this can help ensure that the methods and tasks are properly understood and properly conducted. In the in-person modality the experienced researcher can always adapt on the fly, and at the same time maintain the integrity of the work.

Nancy Cooke at Arizona State University (Cooke, 2021) has provided detailed case studies in how the methods and procedures designed for in-person research on Human-AI systems had to be adapted for the online situation. Margaret Burnett and her colleagues of Oregon State University (Dikkala, Burnett, et al., 2021; Ko, Latoza and Burnett, 2021) tell the following tale about the difficulties of online experimentation.

Case Study in Online Difficulties

Privacy/security risks faced both researchers and participants. For example, the IRB protocol prepared at the outset of the project did not cover videorecording. So that was not allowed. Screen-sharing proved problematic, since the participant might share information not covered by the approved protocol (e.g., email) or material that some might regard as offensive. On the part of the researchers, their need to have remote access to their computers also posed risks. Installing our technology on participants' computers was not covered by our IRB either, and could create participation barriers for people who are uncomfortable installing software—countervailing our goal of gathering diverse participants. We mitigated privacy/security risks by sharing only the relevant window, and collecting only audio and by running our programs on the researchers' computers and displaying it to participants, except for the shared Google Drawings document. Our Zoom Data Collection mechanisms offered some alluring features, such as automatic audio transcription—but our optimism was misplaced. Some transcripts did not arrive, requiring tech support; others were too low quality to use. Finally, synchronizing distinct repositories of information was painstaking, because we used a second platform for markup (Google Drawings), without a unifying video.

The Oregon case study continues with a tale about ways in which some problems were mitigated.

Case Study in Mitigation

We mitigated control risks using three strategies. First, even though it was in the online modality, our study was one-on-one, offering natural conversation and presence/attention checks. Examples include asking participants to think aloud if they fell silent, and interrupting participants if Wi-Fi became problematic or onscreen items were not visible. Second, we mitigated distractions by minimizing context switches (e.g., moving from screenshare to browser—particularly disruptive on small screens). Third, we relied on replaying games and browser-based shared state to mark up the data because participants interacting with the "live" system risked them going "off the rails" of our study sequence. We set up contingency plans for choppy/lost internet, inaudible communications, computers losing power, and broken links—most of which arose. Payment for in-person studies usually involves exchanging cash, with a duplicated signed receipt producing an audit trail for all parties. We chose not to e-transfer cash to participants because that would introduce a sampling bias by requiring participants to bank with particular services, and some lacked audit trails. Ultimately, we elected to transfer money between researchers using Zelle, and from researcher to participants using Amazon Gift eCards. For gift card payment, the researcher confirmed that the participant acknowledged receipt, both verbally and via email. Amazon notified the

researchers upon redemption, increasing workload via associating gift card numbers to participants.

As this case study shows, distance modalities involve challenges that must be navigated when conducting evaluation studies. Some of the main challenges are expressed below.

Challenge 1

Not being able to determine participants' hardware/software platforms introduces the possibility of considerable experimental noise.

Researchers have no control over participants' technical environment, e.g., mouse, keyboard, and monitor, internet quality/dependability, OS, other simultaneously running applications. The devices participants are to use in a remote scenario have to depend on technology that is accessible to them. For example, display luminance may be a manipulation in a virtual environment (e.g., teamwork in a simulated dark environment), but a participant's monitor and brightness settings may invalidate the intended effects. Online experiments can require technology that would enable participants to freely pan, zoom, and mark-up the material that was presented to participants in the tasks. Making the simulation work across platforms can require downsampling high-resolution images that the tasks necessitate, or (worse) substantially delaying feedback on annotations.

Challenge 2

Researchers are advised to design remote studies to include all necessary workspace elements on the participant's screen.

Challenge 3

It is important for the online tasks and task instructions to be clear, if not intuitive.

Despite what one might initially suppose, online participants have been found to be quite attentive to the tasks in online experiments. But the participants in a study are likely to manifest varying levels of attention, lapses of attention, and inattention (Buhrmester, Talaifar, and Gosling (2018).

Challenge 4

Some researchers have found that participants are not willing to spend long periods of time on complex tasks.

This is naturally a concern for evaluations on human-machine work systems in contrast with the many tasks posted on Mechanical Turk that take only minutes to perform.

Challenge 5

It is a challenge to determine whether to use attention checks or approval rates.

Occasional "attention check" questions are sometimes embedded in the task, but they do not guarantee attention and may contribute to attrition. As general feature of online studies, precisely because large samples and lots of data are easy to collect, the experimental designs load up on measures, questionnaires, and other things for the participants to do. Buhrmester, Talaifar, and

Gosling (2018) advise researchers to engage with participants who have a 95% or greater "approval rate," and avoid the use of attention checks.

Challenge 6

Ensuring a participant's well-being is more difficult during a remote study than during an in-person study, even if extra precautions are taken.

"Due to high level of unemployment and financial hardship during the COVID-19 crisis, people who would have never considered participating in a study may be incentivized to sign up. This suggests that researchers may need to be more vigilant than normal of potential ethical issues such as coercion amidst remote experimentation." (Cooke, 2021).

Challenge 7

It is a challenge to ensure that no personally-identifying information is collected.

Third-party applications can automatically record log-ins. Applications can require a participant to set up an account to use them, and automatically identify the respondent's location, which can potentially be used to triangulate a participant's identity.

Challenge 8

The payment mechanism can introduce sampling bias, since they can require participants to have certain kinds of bank accounts, interests in shopping with particular companies, etc.

Challenge 9

It is a challenge to manage communications among the researchers.

Nancy Cooke and her colleagues at Arizona State University (Cooke, 2021; Cooke, Demir and Huang, 2020) have provided detailed case studies in how the methods and procedures designed for in-person research on Human-AI systems had to be adapted for the online situation. Cooke (2021) tells the following story about the challenges of communication among the members of the research team.

Case Study in Communication

Experiments often involve numerous experimental artifacts including consent forms, step-by-step procedures, stimuli, data sensors, and lab journals. Whereas shared physical spaces are normally useful for organizing these artifacts, remote studies tend to rely on physically-distributed space (i.e., study personnel workspaces) or virtual platforms to organize materials. This may change (or exacerbate) the coordination demands of getting the right artifacts to the right people. Multiple monitors, overall increased screen real-estate and dual-tasking may be required in order for researchers to manage the digital artifacts and data streams... recording a participant in the lab requires video cameras to be placed strategically through the space and recording software to be configured. However, recording a participant during a remote study typically requires personnel to press the 'record' button and ask the participant to turn on their webcam.

Challenge 10

Evaluation studies involving tasks in which participants have to work as team face additional logistical hurdles.

Research on AI-enabled human teaming typically relies on virtual task environments, sometimes using well-known game platforms, and sometimes modeled after the actual work contexts, such as the operation of drones by teams that include an AI team member. Although many testbeds for evaluation research were designed on the assumption that the participants would be co-located, the testbeds can be adapted for online experimentation. (see Mercado et al., 2016; Novitzky, et al., 2016).

The training of participants for Human Teams-AI evaluation requires more time and effort, as they must be trained on the individual work and the team work. Furthermore, feedback during training must be tailored to each participant to ensure that all the participants have achieved a baseline competence. This even more complicated when teams involve heterogeneous roles. For the actual experiments, there can be a need for complicated "onboarding" procedures. The researcher's interface must that allow them to monitor one or more participants' behaviors (e.g., multiple monitors or devices).

Cooke (2021) provides a review of this work, including a discussion of how remote-hosting applications and technologies can be adapted to support interactive task environments in the distance modality for human-AI teaming research.

Challenge 11

Researchers need to be prepared for the considerable effort needed to engage in the recruiting process, and in scheduling and confirming the participants' sessions with enough time to cancel or reschedule them as needed.

Challenge 12

Researchers need to be prepared for an extraordinary number of no shows or cancellations.

A major logistical issue for research that uses virtual testbeds for Team-AI evaluations involves the fact that only a handful of participants can be engaged in the team task at a given time. Thus, much of the advantage of online research (lots of data that are easy to get) is lost.

Challenge 13

In online Human-AI team evaluation it is extraordinarily difficult to attempt to make physiological or psychophysiological measurements.

Conclusion About Online Modalities

Cooke (2021) concludes with a positive take on the research using the online modality:

> "Technical and logistical issues are common in Human-AI Team experiments, and they appear to occur more frequently in remote studies, particularly those that are assembled from multiple applications not intended for research.... it may take more

> time to conduct a remote experiment than an in-person experiment... But experiments have shown that experimental control is possible even when human-subjects and complex team tasks are involved."

References

American Psychological Association graphing guidelines [https://apastyle.apa.org/style-grammar-guidelines/tables-figures/figures]

Buhrmester, M.D., Kwang, T., and Gosling, S.D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6*, 3-5.

Buhrmester, M.D., Talaifar, S., and Gosling, S.D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science, 13,* 1490154.

Casler, K., Bickel, L., and Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's Mechanical Turk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*, 2156-2160.

Chunpir, H.I., and Ludwig, T. (2017). A software to capture mental models. In *Universal Access in Human-Computer Interaction*, Part III (pp. 393-409). New Work: Springer International Publishing. [doi: 10.1007/978-3-319-58700-4_32].

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].

Clancey, W.J. (2019). "Critical thinking about AI and explanation." Presentation for The National Academies Board on Human System Integration (BOHSI) Panel: Explainable AI, System Transparency, and Human Machine Teaming. *Proceedings of the Human Factors and Ergonomics Society 2019 Annual Meeting* (p.631-635). Santa Monica, CA: Human Factors and Ergonomics Society. [DOI: https://doi.org/10.1177/1071181319631100].

Clancey, W.J., Lowry, M., Nado, R., and Sierhuis, M. (2011, August) Software productivity of field experiments using the Mobile Agents Open Architecture with workflow interoperability. In *Proceedings of the IEEE Fourth International Conference on Space Mission Challenges for Information Technology* (SMC-IT) (pp. 85-92). IEEE Computer Society: Palo Alto, CA.

Cooke, N. (2021). "Remote research methods for Human–AI–Robot teaming." Manuscript, Center for Human, AI, and Robot Teaming, Program in Human-Systems Engineering, Polytechnic School, Arizona State University, Mesa, AZ.

Cooke, N., Demir, M., & Huang, L. (2020, July). A Framework for Human-Autonomy Team Research. In *Proceedings of the International Conference on Human-Computer Interaction* (pp. 134-146). Cham, Switzerland: Springer.

Dikkala, R., Burnett, M.M., et al. (2021). Doing remote controlled studies with humans: Tales from the COVID trenches. In *Proceedings of the ACM/IEEE 14th International Conference on Cooperative and Human Aspects of Software Engineering* (CHASE 2021). [https://conf.researchr.org/home/chase-2021]

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33,* 587–606.

Gigerenzer, G. (2015). *Simply rational: Decision making in the real world.* Oxford: Oxford University Press.

Hoffman, R.R., Mueller, S.T., Klein, G., and Litman, J. (2018). "Metrics for Explainable AI." Technical Report form Task Area 2, DARPA Explainable AI Program.

Human Factors guidelines [https://www.researchgate.net/publication/220457627_Guidelines_for_Presenting_Quantitative_Data_in_HFES_Publications]

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

Klein, G. (2007). Performing a project premortem. *Harvard Business Review, 85*(9), 18-19.

Klein, G. (2021, 14 January). The Pre-Mortem Method. *Psychology Today* Blog. [https://www.psychologytoday.com/us/node/1156332/preview]

Ko, A.J., LaToza, T.D., Burnett, M.M. (2015). A practical guide to controlled experiments of software engineering tools with human participants. *Empirical Software Engineering, 20* (1), 110-141. [DOI 10.1007/s10664-013-9279-3]

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., and Procci, K. (2016). Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors*, *58*, 401-415.

Novitzky, M., Dougherty, H. R., & Benjamin, M. R. (2016, November). A human-robot speech interface for an autonomous marine teammate. In *Proceedings of the International Conference on Social Robotics* (pp. 513-520). Cham, Switzerland: Springer.

Schmidt, F. (1996). APA Board of Scientific affairs to study issue of significance testing, make recommendations. *Score, 19*, 416-428.

Yates, F. (1951). The influence of "statistical methods for research workers" on the development of the science of statistics. *Journal of the American Statistical Association, 46*, 19-34.

.