**HAMMLAB DATA ANALYSIS APPROACH**

In this supplement, we describe the typical statistical data analysis approach taken for full scope simulator studies in HAMMLAB. Controlled experimentation in realistic operating environments have to balance the rigor and flexibility of the statistical analysis methods. The following sections describe the solutions adopted in HAMMLAB with respect to test logic, statistical and practical significance testing, the use of confidence intervals, comparison of means, familywise error rate, and violations of statistical assumptions.

**Test logic**

In his influential article, “The earth is round (p < .05)”, Jacob Cohen (1994) explains the logical shortcomings of null hypothesis testing, which was the prevailing methodological paradigm in experimental psychology for decades. According to Cohen, null hypothesis testing gives the false impression that causal inferences can be drawn from experiments with a purely deductive logic.

The American Psychological Association responded to this eye-opener by establishing a committee called the Task Force on Statistical Inference (TFSI). TFSI’s mandate was to clarify controversial methodological issues and provide guidelines for the application of statistics in psychology (Wilkinson, 1999). The committee concluded that rigid methodological orthodoxies should be avoided and encouraged researches to give up “practices that institutionalize thoughtless application of statistical methods” (p. 604).

Given Cohen’s convincing arguments and TFSI’s general recommendations, we accept that null hypothesis testing is a meaningless mechanistic ritual. However, controlled experimentation may still be a valuable research tool. As argued by John Stuart Mill (1843), causal inferences are trustworthy if; (a) an effect is present only when the cause is present, (b) an effect is absent when the cause is absent, and (c) both of these relationships are observed. Alternative interpretations of the covariation between cause and effect can then be ruled out. In other words, threats to valid causal inference can be eliminated by comparing matching situations where variables of interest operate, or do not operate. Experimental manipulation may therefore be a powerful technique, despite the logical flaws of null hypothesis testing.

Building strong experiments without the comforting formalism of null hypothesis testing can be a challenge. In HAMMLAB, we have adopted the following approach:

* Experimental manipulations involve comparisons of matching situations where treatments are present or absent. Classical experimental designs that were originally developed for null hypothesis testing may be used to characterize the manipulations and identify appropriate statistical data analysis procedures.
* When the experiment explores new territory, as in comparisons of novel design concepts or technologies, it is perfectly acceptable to rely on data-driven discovery without hypotheses. In situations where theory and/or practical experience informs the test, it may be helpful to establish hypotheses to support the development of powerful test situations, e.g., in an evaluation of competing design principles, or to assess whether the experimental findings are expected or surprising. Hypotheses are written in natural language and anticipate the effects of the experimental manipulations.
* The quality of the dependent variables is given special attention, since poor measurement has been a major weakness in experimental research (Pedhazur & Pedhazur Schmelkin, 1991; Cohen, 1994).
* Statistical analysis is used to evaluate whether robust patterns of results can be extracted from the data. The analysis is not a formal test, but a goal-driven statistical exploration to develop meaningful interpretations of the data-set. TFSI emphasized that statistical techniques should never be more complicated than necessary and promotes the use of graphics to display results and assess assumptions (Wilkinson, 1999).
* Statistical techniques are used to search for unanticipated effects that can contribute to the development of new theories and the formation of hypotheses for later experiments.

**Statistical significance testing**

According to TFSI, it is always better to report actual p-values than to make dichotomous accept-reject decisions with a predefined significance level (Wilkinson, 1999). P-values have a straightforward interpretation and express the likelihood that statistical test results have occurred by chance alone. By providing the p-value associated with a statistical test, we are disclosing the statistical significance of that test in a reliable and understandable way. For example, if the p-value of a test is p=0.65, there is a 65% chance that the observed result is due to random sampling variation and consequently a 35% chance that we have observed a true effect. This would typically be regarded as an unconvincing outcome and a finding that is hard to trust. The outcome of the statistical test would be far more convincing if the p-value had been, e.g., p=0.07 with a 7% chance that the observed result is due to random sampling variation and a 93% chance that we have observed a true effect. Thus, the p-value itself can serve as an accurate and continuous indicator of statistical significance (ranging from p=0.00 to p=1.00) without the introduction of unjustified filtering based on discrete cut-off criteria to separate significant from insignificant results.

**Practical significance testing**

TFSI encouraged researchers to present effect sizes for primary outcomes. Effect size indicators reveal the practical significance of the findings and inform future power analyses and meta-analyses (Wilkinson, 1999). However, the interpretation of effect sizes is not straightforward. Fern and Monroe (1996) identify the following challenges:

* The impact of an effect is a function of its size, but also the perceived value of a unit change on the dependent variable. For example, a 1 percent reduction in the death rate for cancer is a small effect of a new drug but would still be noteworthy due to the magnitude of the cancer disease and the importance of saving lives. Thus, practical significance is more than the effect size.
* The levels of a fixed treatment are arbitrarily determined by the experimenter, and do not represent the population of all possible treatment levels. It is therefore impossible to estimate the population component of variance in the dependent variable attributable to fixed treatments. Thus, effect sizes for fixed effects can only be generalized across studies that have chosen identical levels of experimental manipulation on the fixed treatments.
* The total variance estimate in the denominator of the effect size is strongly affected by subjective judgment. That is, researchers aggregate the data according to their needs and calculate partial effect size indicators (see Cohen, 1973).
* Comparison of effect sizes across experimental designs with different error terms, sample sizes, and/or number of levels on the manipulated variables can be misleading.

Fern and Monroe (1996) still conclude that it is informative to report effect sizes along with statistical significance (i.e. p-values given the approach to statistical testing discussed above). However, the effect size must be interpreted within specific research contexts and in light of the weaknesses listed above.

**Confidence intervals**

According to TFSI, interval estimates should be given for principal effects. It is also recommended to include interval estimates in graphical representations whenever possible (Wilkinson, 1999). The rationale for this is that interval estimates give information about uncertainties The most popular interval estimate is the 95% confidence interval and no other alternatives are considered here.

For within-subject designs, standard confidence intervals around the cell means can produce invalid inferences about mean differences (Loftus & Masson, 1994; Cumming & Finch, 2005). This is because the variance among subjects affects the size of the confidence interval even though subject differences are removed from the error term in the repeated measures analysis of variance. Belia et al. (2005) demonstrated that the majority of researchers misinterpret graphical representations of confidence intervals for within-subject effects by making incorrect inferences about mean differences.

It is possible to calculate and graph customized within-subject confidence intervals that facilitate correct inferences about mean differences (Loftus and Masson, 1994; Cousineau, 2007). However, the customized confidence intervals will no longer provide accurate estimates of the cell means. We do not understand the purpose of presenting inaccurate interval estimates for within-subject effects in order to enable precise graphical illustrations of information that is expressed by statistical significance testing (p-values). The idea of calculating customized within-subject confidence intervals is therefore abandoned.

It is concluded that interval estimates provide potentially valuable information regarding the precision of findings and should be presented for principal effects. Since most researchers misinterpret figures that include confidence intervals for within-subject effects, such graphical representations are not advisable. Customized confidence intervals for within-subject effects can provide information about patterns among means, but this information is redundant and produces inaccurate estimates of the cell means. Interval estimates for the cell means of within-subject effects should therefore include variance among subjects and be reported separately in a tabular format.

### Comparison of means and Familywise error rate

When we conduct many statistical tests to analyse an experiment, it becomes increasingly likely that effects occurring by chance are wrongly judged to be true effects. This risk is often referred to as the *familywise error rate* (see Howell, 1997; p. 349). To solve the problem, statisticians have developed correction procedures that control the overall error rate, such as the Bonferroni test (ibid.). Such corrections make it increasingly difficult to reveal systematic effects as the investigator performs more tests. With our p-value approach to statistical significance testing (see above), the interpretation of statistical significance is always based on the p-value associated with each test, and there is no way of systematically adjusting our interpretations across multiple tests. The best we can do is probably to be skeptical and aware that low p-values can occur by chance when we conduct many statistical tests to analyze an experiment.

In the analysis of rich experimental data from complex work settings, we also have principal doubts regarding the relevance of the correction procedures mentioned above. It is in the nature of such analysis to dig deep and conduct tests from every possible angle, and it seems overly conservative to gradually weaken the exploration tools as a response to increasing efforts. On the other hand, “fishing” in the data and over-interpretation of spurious effect is unacceptable. Our protection against these pitfalls is to search for clear and consistent patterns of results that are corroborated through triangulation of evidence and supported by theory, previous empirical results, and/or operator experience. Isolated findings without basis in theory or real life should be ignored. We can thereby have reasonable confidence in the results without sacrificing the efficiency of the analysis tools.

According to TFSI, pairwise multiple comparison methods, such as the Tukey Honestly Significance Difference (HSD) test, are too conservative when preceded by an omnibus F test in a stagewise testing procedure (Wilkinson, 1999). The committee further points out that comparisons of all possible means may restrict researchers to uninteresting hypotheses and are unnecessary for the understanding of experimental results (ibid.). Thus, comparison of means should be restricted to a few meaningful contrasts, trend analysis, or other investigations of anticipated structure. It can often be appropriate to limit the comparison of means to graphical inspection of the effects. This analysis strategy is described by Howell (2006) as the “minimalist school”, which is a legitimate position, since there are no absolute rules prescribing when the statistical data analysis should stop, and the interpretation of results begins.

**Statistical assumptions**

Parametric statistical tests presume an interval or ratio level of measurement, and normal distribution of the population from which the sample data are drawn. However, many statisticians argue that parametric tests are sufficiently robust to make non-parametric tests unnecessary. This view is implicitly reflected by the strong emphasis on parametric methods in statistical textbooks and software packages. Furthermore, a number of studies have documented the robustness of ANOVA (Maxwell & Delaney, 2000), and non-parametric tests do not even exist for complex factorial experimental designs. However, some statisticians still dispute the general suitability of parametric tests.

The following pragmatic approach is taken in HAMMLAB:

* Parametric tests are used unless distributions are asymmetrical or skewed in opposite directions across treatment populations. ANOVA is not robust to these particular violations of the normality assumption (Kirk, 1995), and whenever these deviations occur, the data may have to be transformed or treated non-parametrically (Howell, 1997). Thus, we are not testing the normality of the distributions as such but search actively for pre-defined distribution patterns that are known to destruct ANOVA. Distributions are evaluated through histograms, normal probability plots, and the Shapiro-Wilk W test (Dell Inc., 2015).
* Parametric tests are used unless the dependent variables are clearly on the nominal or ordinal level of measurement. Performance indicators are otherwise assumed to be in the grey area between the ordinal and interval level of measurement and are therefore treated as if they conform to interval scales (Pedhazur & Pedhazur Schmelkin, 1991).
* When a treatment has more than two levels of experimental manipulation in a within-subject design, statistical assumptions about sphericity and compound symmetry enter the picture (Kirk, 1995; Howell, 1997). These assumptions rarely hold for complex experiments in realistic laboratory environments. The optimal analysis strategy is to employ the multivariate approach to repeated measures first (Dell Inc., 2015). If this test reveals systematic effects, we know that the sphericity and compound symmetry assumption is no longer an issue, and that the power of the multivariate approach was sufficient. Further univariate testing is therefore unnecessary. If, however, the multivariate test suggests that an effect likely occurred by chance, the potentially more powerful univariate test is carried out along with a test of sphericity (Dell Inc., 2015). If the sphericity assumption is violated, the Greenhouse-Geisser and Huynh-Feldt univariate adjustments are calculated (Howell, 1997).

 **Univariate statistical tests**

Statistical analyses of HAMMLAB experiments typically rely on multiple univariate statistical tests (Huberty & Morris, 1989) because, (i) many of the human performance indicators are intended to be conceptually independent, (ii) measures are often inconsistently related across complex simulator experiments, making it hard to defend fixed clusters of measures for multivariate analysis, and/or (iii) HAMMLAB studies usually have an exploratory element where the impact of manipulated variables on a wide selection of human performance measures are investigated for the first time, without the comfort of established theoretical models or strong hypotheses (ibid.).

**References**

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10 (4)

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and psychological measurement*, 33, 107-112

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, 49 (12), 997-1003.

Cousineau, D. (2007). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson’s method. *Tutorial in Quantitative Methods for Psychology*, 1(1), 42-45.

Cumming, G. & Finch, S. (2005). Inference by eye. *American psychologist*, 60 (2), 170-180

Dell Inc. (2015). *Dell Statistica (data analysis software system)*, version 13. software.dell.com.

Fern, E. F., & Monroe K. B. (1996). Effect size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, 23, 89-105

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.), CA: Wadsworth.

Howell, D. C. (2006). Multiple comparisons with repeated measures. Material prepared for the next version of ‘Statistical methods for psychology’. <http://www.uvm.edu/~dhowell/StatPages/>

Huberty, C. J. & Morris, J. D. (1989). Multivariate Analysis Versus Multiple Univariate Analyses. *Psychological Bulletin*, Vol. 105, No. 2, 302-308.

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, 54 (4).

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*, 3rd ed. Pacific Grove, CA: Brooks Cole.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, 1, 476-490

Maxwell, S. E., & Delaney, H. D. (2000). *Designing experiments and analyzing data: A model comparison perspective.* Mahwah, NJ: Lawrence Erlbaum Associates

Mill, J. S. (1843). *A System of Logic.* London: Longmans

Pedhazur, E. J. & Pedhazur Schmelkin, L. (1991*). Measurement, design, and analysis. An integrated approach.* NJ: Hillsdale.

Skraaning, G. (2016). *A Reanalysis of the Work Practice Experiments in HAMMLAB (2009-2014).* OECD Halden Reactor Project, HWR-1194, Halden, Norway

Skraaning, G., Eitrheim, M. H., Lau, N., Nihlwing, C., Hurlen, L. & Karlsson, T. (2010). *Coping with Automation in Future Plants: Results from the 2009 HAMMLAB Experiment.* OECD Halden Reactor Project, HWR-937, Halden, Norway

Skraaning, G., Lau, N., Welch, R., Nihlwing, C., Andresen, G., Brevig, L. H. Veland, Ø., Jamieson, G., Burns, C. M. & Kwok, J. (2007). *The Ecological Interface Design experiment (2005).* OECD Halden Reactor Project, HWR-833, Halden, Norway

Wilkinson, L. (1999). Statistical methods in psychology journals. Guidelines and explanations. *American Psychologist*, 54 (8), 594-604.